UNIVERSIDADE DA CORUÑA
DEPARTAMENTO DE COMPUTAÇÃO

TESE DOUTORAL

PROBLEMÁTICA SOBRE O IMPACTO DA EXPANSÃO AUTOMÁTICA DE CONSULTAS E DESAMBIGUAÇÃO SEMÂNTICA NA AVALIAÇÃO DOS SISTEMAS DE RECUPERAÇÃO DE INFORMAÇÃO

Doutorando: Francisco João Pinto
Directores: Carme F. Pérez-Sanjulián
Antonio Fariña

A Coruña, Julho de 2008

Ph. D. Thesis supervised by
Tese doutoral dirixida por

Carme Fernández Pérez-Sanjulián
Departamento de Galego-Portugués, Francés e Linguística
Facultade de Filoloxía
Universidade da Coruña
Campus da Zapateira, s/n
15071 A Coruña (España)
Tel: +34 981 167000 ext. 1739
carme@udc.es

Antonio Fariña
Departamento de Computación
Facultade de Informática
Universidade da Coruña
Campus da Elviña, s/n
15071 A Coruña (España)
Tel: +34 981 167000 ext. 1383
fari@udc.es

Resumo

Nos últimos anos, se está produzindo um enorme incremento da quantidade de informação disponível em formato digital. As bases de dados textuais não são uma excepção. O aumento de bibliotecas digitais, bases de dados documentários, e sobretudo o grande crescimento da Web, fazem que as colecções de texto cresçam em tamanho e número de forma exponencial. No entanto, dispor de grandes quantidades de documentos não resulta de interesse se não existem meios que permitam obter a informação desejada num contexto determinado. É por isso que surge com força o problema da Recuperação de Informação que deve contemplar não só como tem de ser estruturada a informação para facilitar o acesso à mesma, senão também a forma em que dito acesso se realiza. Entre outros aspectos, uma parte fundamental do processo de recuperação de informação é o modo em que um usuário realiza a formulação de uma consulta que deverá plasmar adequadamente sua necessidade de informação, e cômo dita consulta é interpretada pelo sistema de recuperação para posteriormente obter aqueles documentos que contenham informação relevante.

Para que o processo de recuperação seja efectivo, e por tanto se recuperem o maior número de documentos relevantes possíveis, dada uma consulta, é essencial ser capaz de determinar o conjunto de palavras que expressam semanticamente a necessidade de informação do usuário. Deste modo, em lugar de recuperar simplesmente os documentos que contêm as palavras inclusas por um usuário em sua consulta, é possível recuperar documentos que contenham termos relacionados com os expandidos. Por outra parte, a existência de palavras que podem ter um significado ambíguo por exemplo a polissémia faz interessante o processo de desambiguação semântica; isto é, seleccionar o sentido mais apropriado de uma palavra ambígua num contexto determinado. Geralmente para esta tarefa é necessário contar com uma fonte de informação semântica, isto é, recursos linguísticos como dicionários, tesauros, etc.

Nesta tese doutoral se explora o uso do recurso linguístico WordNet como meio

para realizar expansão automática de consultas, e se estuda cômo o processo de desambiguação permite melhorar a efectividade do processo de recuperação de informação. Baseando-nos no sistema LEMUR, levou-se a cabo a implementação de um sistema de recuperação que inclui a possibilidade de utilizar diversas técnicas de expansão de consultas em nossos experimentos. Os resultados empíricos obtidos sobre grandes colecções de textos de referência (TREC) demonstram que a desambiguação de palavras permite obter a máxima vantagem na expansão de consultas com WordNet.

Resumen

En los últimos años, se está produciendo un enorme incremento de la cantidad de información disponible en formato digital. Las bases de datos textuales no son una excepción. El aumento de bibliotecas digitales, bases de datos documentales, y sobre todo el gran crecimiento de Web, hacen que las colecciones de texto crezcan en tamaño y número de forma exponencial. Sin embargo, disponer de grandes cantidades de documentos no resulta de interés si no existen medios que permitan obtener la información deseada en un contexto determinado. Es por ello que surge con fuerza el problema de la Recuperación de Información que debe contemplar no sólo cómo ha de ser estructurada la información para facilitar el acceso a la misma, sino también la forma en que dicho acceso se realiza. Entre otros aspectos, una parte fundamental del proceso de recuperación de información es el modo en el que un usuario realiza la formulación de una consulta que deberá plasmar adecuadamente su necesidad de información, y cómo dicha consulta es interpretada por el sistema de recuperación para posteriormente obtener aquellos documentos que contengan información relevante.

Para que el proceso de recuperación sea efectivo, y por tanto se recuperen el mayor número de documentos relevantes posibles, dada una consulta, es esencial ser capaz de determinar el conjunto de palabras que expresan semánticamente la necesidad de información del usuario. De este modo, en lugar de recuperar simplemente los documentos que contienen las palabras incluidas por un usuario en su consulta, es posible recuperar documentos que contengan términos relacionados con aquéllos (expandidos). Por otra parte, la existencia de palabras que pueden tener un significado ambiguo (p.ej. polisemia) hacen interesante el proceso de desambiguación semántica; esto es, seleccionar el sentido más apropiado de una palabra ambigua en un contexto determinado. Generalmente para esta tarea es necesario contar con una fuente de información semántica, es decir, recursos lingüísticos como diccionarios, tesauros, etc.

En esta tesis doctoral se explora el uso del recurso lingüístico WordNet como medio para realizar expansión automática de consultas, y se estudia cómo el proceso de desambiguación permite mejorar la efectividad del proceso de recuperación de información. Basándonos en el sistema LEMUR, se ha llevado a cabo la implementación de un sistema de recuperación que incluye la posibilidad de utilizar diversas técnicas de expansión de consultas en nuestros experimentos. Los resultados empíricos obtenidos sobre grandes colecciones de textos de referencia (TREC) demuestran que la desambiguación de palabras permite sacar el máximo partido en la expansión de consultas con WordNet.

Abstract

In the last years, an enormous increase of the amount of information available in digital format has been taking place. Textual databases are not an exception. The wide spread use of digital libraries, documental databases, and mainly the big growth of the Web cause text collections to grow continuously in both size and number. Nevertheless, just storing such large amount of documents is not usually enough. For those large amounts of documents to become useful, it is also necessary to develop tools that permit us to obtain the relevant information in a certain context. Not only does Information Retrieval deal with how documents should be structured to make it easier to recover some information from a large text collection, but also how such retrieval process is performed. Apart from other important issues, the way in which a user formulates a query (showing an information need) and how it is interpreted by a retrieval system are of special interest in order to be able to recover those documents containing relevant information for a given query.

Determining the set of words that semantically represents the necessity of information of a user for a given query is an important issue that should be taken into account in order to obtain an effective retrieval. That is, to retrieve a set of documents as small as possible containing as many relevant documents as possible. By focusing in the meaning of the words included in a query, instead of just searching for all those documents that contain such terms, it is possible to retrieve documents that might not contain those words, but others related to them (expanded terms). Additionaly, the existence of words that can have different meanings depending on the context (for example polisemic words) makes it interesting to perform semantic disambiguation. Semantic disambiguation aims at choosing the most appropiate meaning of a word for a given context. It usually needs to count on a semantic information source. Some typical Linguistic resources like dictionaries, thesauri, etc are usually handled for this sake.

In this PhD thesis we study the use of the linguistic resource WordNet as the way

to perform automatic query expansion, and we show how disambiguation allows us to improve the effectiveness of the information retrieval process. We have implemented an information retrieval system that is based on LEMUR. Our system permit us to use different query expansion approaches in our experiments. The empirical results obtained over some large reference text collections (TREC) show up that word disambiguation is a powerful tool that permit us to obtain improved effectiveness when query expansion based on WordNet is performed.

Á Deus quem me tem dado fortaleza e entendimento.

Á todos aqueles que lutam para um mundo melhor.

Á Carol, minha esposa, pelo seu carinho e inspiração.

Agradecimentos

Á Agência Espanhola de Cooperação Internacional (AECI) pela bolsa de estudos otorgada para a realização do curso de doutoramento.

Ao Ministério da Educação da República de Angola pela bolsa de estudos otorgada para a realização da tese de doutoramento.

Cabe aqui um agradecimento a todos que contribuíram decisivamente para o desenvolvimento desse trabalho.

Ao Professor Álvaro Barreiro García que me deu a possibilidade de introduzir-me na área da recuperação de informação na Web e por dispor-me dos meios suficientes para que a realização deste trabalho fosse efectiva.

De maneira especial aos Professores Nieves Brisaboa, María do Carme Fernández Pérez-Sanjulián e Antonio Fariña que me acolheram carinhosamente e incentivaram-me para chegar ao fim.

Ao meu querido amigo Professor Esteban García Martín, que sempre esteve ao meu lado apoiando-me psicológicamente nos momentos mais difíceis.

A todos que de alguma forma me auxiliaram, motivaram e torceram para que esse trabalho chegasse ao fim, especialmente a minha família e os meus amigos.

Contribuições

Esta tese centra-se no âmbito das Bases de Dados textuais que contém texto proveniente da linguagem natural. Neste entorno é importante que quando um usuário deseja obter uma determinada informação exista um Sistema de Recuperação de Informação que facilite dita tarefa. Dito sistema deve ser capaz de recuperar documentos de uma maneira *eficiente* (por exemplo um usuário Web não deseja esperar vários segundos para obter uma resposta), e *efectiva*, devolvendo aqueles documentos mais relevantes de cada à informação solicitada pelo usuário através de uma consulta. Nesta tese revisa-se o processo de recuperação de informação (RI), começando por estruturas básicas de RI como são os índices invertidos. A continuação centra-se no seu objectivo principal, a qualidade do processo de RI. O principal objectivo desta tese é o estudo de cômo a expansão de consultas e a *desambiguação do sentido das palavras* permitem melhorar a efectividade da recuperação de informação. Com estas duas técnicas, favorece-se a possiblidade de que documentos que não seriam recuperados pelo simple facto de não conter os termos que o usuário inclui na sua consulta possam ser recuperados já que a expansão destes termos incluem na busca outros termos sinónimos. Para conseguir isso, faz-se uso do recurso linguístico WordNet como o meio para levar a cabo *a expansão de consultas* e desambiguação do sentido das palavras. Neste caso a desambiguação joga o papel primordial de tratar de reduzir o ruído que ineludivelmente é incluido no processo de recuperação de informação ao expandir os termos de uma consulta. Os resultados experimentais concluem que a desambiguação é beneficiosa quando se faz expansão de consultas. Em resumo, as principais contribuições deste trabalho são:

- Análise de técnicas tradicionais de indexação como são os índices invertidos e sua aplicação nos sistemas de recuperação de informação. Também se analisa o comportamento dos usuários dos sistemas de recuperação de informação na Web, frente ao caso dos usuários de sistemas tradicionais.

- Descrição dos conceitos fundamentais dos diferentes algoritmos existentes para a desambiguação do sentido das palavras (WSD). Além disto, se apresentam os principais enfoques utilizados em WSD e se descrevem os recursos linguísticos mais utilizados em particular WordNet. Finalmente se revisam os algoritmos de aprendizagem automática que são normalmente utilizados na WSD.

- Desenvolvimento para a formulação de consultas, concretamente a expansão de consultas e WSD mediante o uso de WordNet 2.1. Por um lado, a tese trata um tema da actualidade e útil no caminho para a elaboração de um buscador semântico que resolva os problemas dos actuais buscadores sintácticos como a polissemia. Por outro lado, avalia-se o sistema de forma sistemática, utilizando as medidas mais comuns na avaliação dos sistemas de recuperação de informação.

- Utilizando como base o sitema de recuperação de informação Lemur, se inclue a possiblidade de agregar a expansão de consultas e WSD como meio para melhorar a qualidade de recuperação de informção. Assim levou-se a cabo a avaliação do sistema de recuperação da informação Lemur mediante medidas de precisão e *recall*, considerando diferentes modos de consulta: 1) mediante a consulta original, 2) expandindo a consulta com sinónimos e 3) desambiguando previamente o sentido das palavras mediante o uso de WordNet com uma metodologia da avaliação baseada na simulação sob um modelo do usuário.

Conteúdo

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

A recuperação de informação (IR) estuda a representação, armazenamento, organização e acesso aos elementos de informação. A representação e a organização dos elementos de informação devem proporcionar um fácil acesso ao usuário de um sistema de recuperação de informação (SRI) pela informação do seu interesse [15] (cap: 1).

O usuário, por sua parte, deve traduzir a sua necessidade de informação numa consulta que pode ser processada pelo motor de busca ou sistema de informação. Geralmente, esta transformação produz uma série de palavras chave que resumem a descrição da necessidade de informação do usuário.

Com base a consulta do usuário, o objectivo fundamental de um sistema de IR é recuperar a informação que seja útil ou relevante para o usuário. É importante diferenciar entre recuperar informação e recuperar dados.

A recuperação de dados, no contexto de um sistema de IR, consiste basicamente em determinar quais são os documentos de uma colecção que contém as palavras chave introduzidas pelo usuário, o qual não é suficiente para satisfazer a necessidade de informação do usuário. De facto, um usuário de um sistema de IR está interessado em obter informação sobre um determinado tema.

Num sistema de recuperação de informação se obtém objectos que podem ser pouco precisos ou com poucos erros que não serão detectados pelos usuários. A principal razão para isto é o facto de que a recuperação de informação deve tratar texto em linguagem natural que nem sempre está bem estruturado e pode ser semanticamente ambíguo. Por outra parte, um sistema de recuperação de dados (como

por exemplo, uma base de dados relacional) administra dados que apresentam uma estrutura e uma semântica bem definida.

Um sistema de recuperação de informação deve, de alguma maneira, interpretar o conteúdo dos elementos de informação ou documentos da colecção e ordená-los de acordo com um grau de relevância para a consulta do usuário. A dificuldade não está unicamente em saber como extrair esta informação senão também em saber utilizá-la para determinar a relevância de cada documento. Em consequência, a noção de relevância é um conceito chave na recuperação de informação. De facto, o objectivo primordial de um sistema de IR é recuperar todos os documentos que são relevantes para a consulta.

1.1 Evolução histórica

Durante aproximadamente 4.000 anos o ser humano tinha organizado informação para mais tarde recuperá-la e utilizá-la. Um exemplo típico é o índice dos conteúdos de um livro. A medida em que o volume de informação cresce torna-se necessário construir uma estrutura especializada para assegurar um acesso mais rápido a informação armazenada. Uma antiga e popular estrutura para este propósito é uma colecção de palavras ou conceitos seleccionados com ponteiros associados aos documentos (ou informação) relacionada: o índice. De uma forma ou outra, os índices são os núcleos dos sistemas de recuperação de informação modernos.

O processo de indexação, conhecido originalmente como catalogação, é a técnica mais antiga para a identificação do conteúdo dos elementos para facilitar a sua recuperação. Na época dos Egípcios, em Babilónia, as bibliotecas ordenavam por tema as tabelas *cuneiformes*[99]. Praticamente até ao século XIX os temas de indexação se converteram em hierárquico. Em 1963 la Livraria do Congresso dos Estados Unidos iniciou um estudo para a automatização mediante computadores dos arquivos bibliográficos. Desde 1966 até 1968 a Livraria do Congresso trabalhou com o seu projecto piloto MARC 1. MARC (Machine Readable Cataloging) estandariza a estrutura, conteúdos e codificação dos registos bibliográficos. O sistema se fez operativo em 1969 [14].

O primeiro sistema comercial de catalogação é DIALOG, desenvolvido por Lockheed Corporation em 1965 para a NASA. Se comercializou en 1978 com três arquivos governamentais dos índices de publicações técnicas. En 1988, quando foi vendido a Knight-Ridder, DIALOG continha uns 320 índices usados por mais de 91.000 assi-

nantes de 86 países [81].

A introdução dos computadores para assistir na catalogação não cambiou o modo de operação básica dos catalogadores ou indexadores humanos encarregues para determinar os termos adequados para cada elemento a indexar. A estandarização das estruturas de dados, como por exemplo as empregadas em MARC, permitiu compartilhar índices entre distintas entidades e reduziu a sobrecarga manual para a manutenção do catálogo. No entanto, o processo ainda requeria que o catalogador introduzisse termos para a sua posterior localização. O usuário passou a realizar uma busca física entre os cartões do catálogo a dispor-se de uma busca baseada em computadores e que mostrava o resultado numa tela por meios equivalentes à dos cartões originais.

Nos anos 80, a redução significativa do custo da capacidade do processamento e a memória nos computadores permitiu o acesso ao texto completo de um elemento, com relação aos termos manualmente especificados [120]. Este facto permitiu trocar a forma de indexar e buscar; não é necessário introduzir manualmente para ser indexados, e se mostra ao usuário aqueles elementos que maior relevância apresenta com a busca realizada.

Com respeito ao processo de recuperação de informação, é importante distinguir os dois pontos de vista do problema de recuperação de informação: A perspectiva humana e a perspectiva do computador. Desde o ponto de vista centrado no computador, o problema de recuperação de informação consiste basicamente em construir índices eficientes, processar as consultas dos usuários com um rendimento adequado e desenhar algoritmos de ordenação que melhoram a qualidade do conjunto de respostas. Desde a perspectiva humana, o comportamento dos usuários, compreende as suas principais necessidades e determina os efeitos destas necessidades sobre a organização e operação dos sistemas de recuperação. Numa história mais recente as bibliotecas tinham sido as primeiras instituições que tinham adoptado os sistemas de IR para recuperar informação. Na primeira geração, estes sistemas consistem basicamente numa automatização de tecnologias prévias (por exemplo, catálogos de cartões), permitindo unicamente buscas sobre autores e títulos das obras. Na segunda geração se aumentou as funcionalidades de buscas, permitindo consultas sobre temas, palavras chave e outras características mais complexas. Na actualidade se está desenvolvendo a terceira geração centrando-se em aspectos de interface gráfica, características de hipertexto e em arquitecturas de sistemas abertos.

Pelo contrário, a princípios dos anos 90 um facto cambiou a percepção de recupe-

ração de informação: a introdução da Word Wide Web. A Web se está convertendo num *repositório* universal de conhecimento humano e cultural que tem permitido compartilhar ideias e informação numa escala nunca antes conhecida. O seu êxito está baseado numa concepção de uma interface de usuário padrão (*estándar*) que sempre opera da mesma forma, não importando o entorno computacionalmente utilizado. Como consequência aos usuários se ocultam detalhes como os protocolos de comunicação, a localização das máquinas e os sistemas operativos. E no entanto qualquer usuário pode criar os seus próprios documentos Web e enlaçá-los com outros documentos Web sem nenhum tipo de restrições. Isto constitui um aspecto chave visto que converte a rede num novo meio de publicação acessível para todo o mundo [6].

Com respeito a recuperação de informação, se pode considerar que tem havido três câmbios fundamentais com relação à recuperção devido as melhorias tecnologicas e o *boom* da Web. Em primeiro lugar, se tem reduzido drasticamente o custo por ter acesso a várias fontes de informação, o que permite chegar a um volume de audiência não possível anteriormente. Segundo, os avanços em todos os tipos de comunicação digital tem proporcionado um maior acesso as redes, o que implica que a fonte de informação pode estar disponível incluso se está fisicamente localizada a grandes distâncias, e com tempos de resposta reduzidos. E em terceiro lugar, a liberdade para publicar qualquer tipo de informação que qualquer julgue interessante tem contribuido para a popularidade da Web. Pela primeira vez na história muitas pessoas têm acesso gratuito a um imenso meio de publicação.

Fundamentalmente, o baixo custo, a grande capacidade de acesso e liberdade de publicação tem permitido e animado à gente a utilizar a Web como um meio altamente interactivo para compartilhar informação.

Não obstante as características e o êxito da Web, também a rede tem introduzido novos problemas: encontrar informação útil na WWW é normalmente uma tarefa tediosa e difícil. Por exemplo, para satisfazer uma necessidade de informação o usuário deve navegar através do espaço de hiperenlaces buscando a informação de interesse. Contudo, a navegação é geralmente ineficiente. Realmente, o principal obstáculo parte da ausência de um modelo de dados bem definido na Web, o que implica que a definição e a estrutura da informação são de baixa qualidade.

Estas dificuldades têm atraído e renovado o interesse na recuperação de informação e as suas técnicas como soluções prometedoras, pelo que a IR se tem convertido num aspecto tecnológico chave para o futuro desenvolvimento e evolução da World

Wide Web. Há uma larga história de experimentação em IR. A investigação começou com experimentos para indexar linguagens a princípios dos anos sessenta e tem continuado a volta de mais de quarenta anos de experimentação com os sistemas de IR. A avaliação IR se tem convertido num campo de investigação activo.

A quantidade de informação em que uma pessoa pode ter acesso cresce exponencialmente, graças sobretudo a Internet, e ainda o tipo de esta informação é cada vez mais variada, a informação textual hoje em dia é a predominate. Aproximadamente 90% do total da informação que maneja uma empresa é texto. Podemos encontrar textos em documentos, manuais, informes, correios electrónicos, faxes e também em páginas Web. Somente para este último meio há estimações de que a quantidade de texto disponível é de pelo menos 6 terabytes [6]. Neste escenário se entende o crescente interesse pelos sistemas de acesso a informação e, em geral, pelas tarefas de análise automático do conteúdo textual.

Na segunda metade do século XX se produz o que se tem chamado incremento ou explosão de documentos, um crescimento exponencial da massa de documentos, de todo tipo e em todo suporte. Isto tem posto de relevo o problema de recuperação de informação. Quer dizer, a necessidade de seleccionar documentos concretos que resolvam necessidades informativas concretas. O problema se centra fundamentalmente em seleccionar em função do conteúdo dos documentos; outro tipo de selecção, que consiste em que todos os campos tais como (título, nome do autor, data da edição etc.) possam ser usados os que tenham a informação mais relevante para a realização da busca, ao tratar-se de informação estruturada que pode ser processada mediante tecnologia convencional, facilitando assim ao usuário o acesso aos documentos do seu interesse [192].

A via de abordar dito problema de recuperação de informação é a indexação manual, o conteúdo dos documentos é examinado e analisado pelas pessoas expertas, e descrito por estas pessoas utilizando as chamadas linguagens documentais: um tipo de linguagens artificiais controlados e desenhados especificamente para descrever o conteúdo temático dos documentos. O resultado de estas descrições pode ser armazenado de forma que se facilitem buscas posteriores entre estas descrições, seleccionado assim os documentos que podem responder a determinadas matérias. Em princípio esta forma de armazenamento eram os ficheiros clássicos em papel ou cartolinas, ordenados por diversos critérios e, posteriormente, as bases de dados convencionais dos computadores.

A indexação manual, pelo contrário mesmo quando se armazenam e administram

os seus resultados com computadores, tem sérios inconvenientes. Em primeiro lugar, é um processo caro e custoso, deve ser levado a cabo por pessoas especializadas e se trata de uma tarefa que requer muito tempo; não se trata, pois de uma questão somente de elevados custos económicos, o tempo necessário para indicar os documentos é maior que o que estes demoram em produzir-se. É impossível processar nem sequer uma mínima parte dos documentos que se produzem; o alto grau de absorção de uma boa parte da documentação actual agrava o referido problema.

O segundo grande problema de indexação manual é o da inconsistência. Tem-se comprovado experimentalmente que distintos indexadores descrevem o mesmo documento de maneira distinta (apesar de utilizar a mesma linguagem controlada por eles) [97]. Incluso o mesmo indexador, em diferentes momentos, produz descrições dos mesmos documentos. É difícil produzir depois uma recuperação eficaz, partindo de descrições de conteúdos inconsistentes. Qual ou quais são os materiais que se deveriam buscar para satisfazer uma determinada necessidade de informação?

Isto nos leva a um terceiro problema. Para seleccionar os documentos que resolvam uma necessidade de informação, é preciso descrever dita necessidade, e fazê-la com a mesma linguagem controlada que se utilizou para descrever os documentos; se para isto foi necessário pessoal especializado, para formalizar as necessidades de informação também será preciso, o usuário recorrer a intermediários, a este pessoal especializado, para obter resultados satisfatórios.

Contudo na actualidade, uma boa parte dos documentos estão disponíveis em formato electrónico. Em determinadas ocasiões, documentos em suporte de papel estão também em formato electrónico, pois têm sido elaborados mediante máquinas electrónicas (por exemplo um processador de texto); em outros casos existem somente e directamente em suporte electrónico. Seja o que fosse, este facto introduz um câmbio substancial, pois ao estar o documento completo num suporte legível por computador, pode ser processado por programas informáticos e é possível expor uma indexação totalmente automática.

A indexação automática contudo, não está isenta de problemas. O principal problema está no facto de que um documento contém muita informação, mas debilmente estruturada, ou menos estruturada de uma forma que não é o suficientemente explícita para que os programas informáticos actuais possam entendê-la. Uma solução simple deste problema é o que se tem vindo conhecendo como buscas em texto livre, ou também como busca de subcadeias. Isto é, a selecção por parte de um programa informático de aqueles documentos que contém tal ou qual palavra. Normalmente

se poderia buscar por mais de uma palavra, e neste caso poderão indicar restrições adicionais mediante operadores booleanos, operadores de proximidade, etc.

Esta solução simple tem os seus inconvenientes: Os mais importantes são os derivados da sinonímia e da polissemia. Dado que um mesmo conceito pode se expressar com palavras distintas, sinónimos, nem sempre se pode saber qual delas terá sido utilizada em cada documento, por outro lado posto que uma mesma palavra se pode referir a conceitos diferentes, podemos nos encontrar numa situação em que muitos documentos que contém uma determinada palavra em realidade tratem sobre temas que nada têm a ver com o que nos interessa.

No âmbito dos sistemas de recuperação de informação (RI), a constatação da dificuldade de formular consultas que se mostram efectivas recuperando informação relevante tem suscitado um grande interesse pelas técnicas de modificação de consultas. Em concreto, neste presente trabalho de tese doutoral nos centraremos fundamentalmente na expansão de consultas e desambiguação do sentido das palavras utilizando como recurso linguístico WordNet.

A expansão de consultas tem resultado ser uma técnica eficaz para melhorar a efectividade dos sistemas de RI. Esta melhoria se pode conseguir automaticamente por meio da inclusão de novos termos na consulta original fazendo uso de recursos linguísticos.

A expansão de consultas com WordNet tem mostrado ser potencialmente relevante para aumentar o *recall*, porque permite recuperar documentos relevantes que poderiam não conter os termos da consulta. Propusemos uma expansão de consulta baseada na sinonímia de WordNet, mas fazendo uso da desambiguação.

1.2 A desambiguação do sentido das palavras

Actualmente o recurso mais valioso para o ser humano é a informação. Com a tecnologia informática, tem sido possível manipular uma grande quantidade de informação e armazená-la em forma electrónica gerando documentos de diversos géneros e idiomas. Além disto, a maior parte desta informação se encontra disponível na Web, onde se estima que 80% se representa textualmente [233]. Devido a que esta massa de informação cresce exponencialmente, surge à necessidade de processá-la automaticamente para facilitar muitas tarefas. Por tal motivo se recorre num dos ramos de inteligência artificial: O Processamento de Linguagem Natural

(PLN). O enfoque principal desta área consiste em criar métodos, técnicas e ferramentas computacionais que permitam realizar análises de informação escrita ou oral e que facilitem a busca e a organização da referida informação. Uma das tarefas mais difíceis no processamento automático da linguagem humana é a resolução do problema da ambiguidade das palavras. Esta resolução do problema é fundamentalmente necessária, dado que uma palavra pode ser interpretada de diferentes formas, quer dizer, possue mais que um significado ou sentido. Este fenómeno linguístico se conhece como polissemia.

Determinar o sentido correcto de uma palavra num texto ou numa conversação é uma tarefa constante na comunicação humana que, contudo, raras vezes causa problemas. Pelo contrário, para um computador se converte numa enorme dificuldade, ainda que as palavras sejam óbvias para um ser humano. Isto se deve a que as máquinas interpretam os textos como um conjunto de palavras sem nenhum significado e/ou valor. Para alcançar a compreensão e interpretação adequada da informação é necessário analisar ao fundo cada uma das palavras e assim obter a interpretação ou sentido mais apropriado. Por esta razão, surge a necessidade de investigar métodos que ajudem a determinar o significado adequado das palavras que existem dentro de um texto com a ajuda do conhecimento, para além do mesmo texto. Encontrar estes métodos e alternativas para solucionar este tipo de problemas é precisamente uma das motivações deste trabalho.

Um dos problemas mais importantes na linguagem natural e que afecta de forma directa à recuperação de informação é a ambiguidade léxica. A ambiguidade das expressões linguísticas é inerente a todos os níveis léxicos: morfológico, sintáctico e semântico. Isto leva consigo uma inevitável perda de informação e uma introdução de ruído ao realizar inferências linguísticas. O problema se faz presente de forma especial quando se tratam os problemas de variação terminológica e de multilinguismo ao ter que seleccionar os elementos de expansão ou tradução tanto nas consultas como nos documentos já seja a outro idioma como a uma representação conceptual.

Resulta difícil determinar como afecta a Recuperação de Informação a ambiguidade léxica em cada um dos níveis léxicos. A consideração explícita da ambiguidade léxica requer um processamento linguístico: etiquetado morfossintáctico para determinar a categoria gramatical e desambiguação do sentido das palavras (WSD, *Word Sense Disambiguation*). Este processamento linguístico não está isento de erros e introduz uma variável difícil de quantificar na hora de avaliar os efeitos de um processamento linguístico na Recuperação de Informação. Estes erros podem prejudicar

a recuperação compensando os possíveis benefícios no uso de técnicas linguísticas.

A eleição do sentido mais apropriado para uma palavra polissémica se tem convertido numa das tarefas mais fundamentais dentro da área do PLN que é chamada Desambiguação do Sentido das Palavras ou, em inglês, *Word Sense Disambiguation* (neste trabalho de tese doutoral se utilizarão as siglas WSD para referir-se a esta tarefa). Esta tarefa é necessária para alcançar o desenvolvimento da maioria das aplicações e outras tarefas do PLN. Até agora, os métodos propostos têm alcançado resultados limitados ao número de palavras a desambiguar, dados os recursos linguísticos utilizados. Assim, a falta de informação, sobre o uso da maioria das palavras, isto é uma das principais causas destas limitações. Assim que, a recopilação dos recursos linguísticos suficientes para WSD é extremamente importante.

A língua é um dos aspectos fundamentais do comportamento humano e é um componente crucial de nossas vidas. Estuda-se em diversas disciplinas académicas. Cada disciplina define o seu próprio sistema de problema se tem os seus próprios métodos para tratá-los. A linguística, por exemplo estuda a estrutura da língua, enquanto a psicolinguistica por outra parte, estuda os processos de produção e do entendimento humano da língua. A linguística de computo se encarrega da solução de problemas que têm que ver, por exemplo, com a identificação da estrutura de orações ou com o modelado do conhecimento e o raciocínio bem como com a definição de estratégias que permitam o uso da linguagem em tarefas específicas.

No âmbito das áreas de investigação, pode-se dizer que estamos vivendo um momento de revolução no pensamento linguístico. Dentro do meio académico constantemente têm surgido questões sobre o que é a língua, como esta se organiza e como a mesma deve ser estudada e ensinada actualmente. Grande parte dessa revolução veio como consequência da inserção da tecnologia dentro da área de estudos das ciências humanas, mais especificamente com o advento do computador digital que trouxe novas perspectivas no âmbito da investigação científica.

Segundo [18], o computador, com a sua grande capacidade de armazenamento e processamento de dados, "começa a desempenhar nas ciências humanas, o papel transformador que o telescópio teve na física e nas ciências exactas.

É pertinente afirmar que essa revolução está relacionada ao desenvolvimento da Linguística de Corpus. O termo **Linguística de Corpus** apareceu pela primeira vez em um livro denominado *Corpus Linguistics*, organizado por [1], onde se encontra associado à utilização de computadores como ferramenta de análise na área de investigação linguística.

Tradicionalmente, o termo corpus tem sido usado por linguistas para designar um conjunto de dados de certa língua natural que pode ser utilizado como base para investigações linguísticas. Levando-se em consideração todas as características importantes atribuídas a essa abordagem, a definição abaixo é bem adequada:

Um conjunto de dados linguísticos (pertencentes ao uso oral ou escrito da língua, ou a ambos), sistematizados segundo determinados critérios, suficientemente extensos em amplitude e profundidade, de maneira que sejam representativos da totalidade do uso linguístico ou de algum dos seus âmbitos, dispostos de tal modo que possam ser processados por computador, com a finalidade de propiciar vários resultados e úteis para a descrição e análise [18].

Partindo dessa definição, pode-se deduzir que o uso de corpus numa investigação implica numa observação detalhada de fenómenos linguísticos em amostras autênticas de uso da língua. Apesar do esforço de alguns linguistas do passado, que corajosamente utilizaram corpus que eram colectados, mantidos e analisados manualmente, fica evidente que, antes da tecnologia dos computadores, esse tipo de estudo era praticamente inviável, não só devido à dificuldade de se fazer uma análise detalhada de palavras dentro de um contexto, mas também devido a pouca confiabilidade dos resultados, uma vez que o ser humano não é adequado para trabalhos dessa natureza.

Além disso, factores como o trabalho e a grande quantidade de pessoas que estudam com base nesse tipo de abordagem envolveriam também e contribuiriam para a inviabilidade desse tipo de investigação durante muitos anos. Mesmo assim, investigações de base em corpus em linguística e lexicografia foram produzidas desde o século XVIII, até a década de cinquenta com os estruturalistas norte-americanos (para uma resenha das abordagens de base em corpus antes do aparecimento de recursos computacionais, ver [63]. Felizmente, com o advento do computador e a possibilidade de processar e armazenar um grande número de informações, hoje esse tipo de investigação se tornou mais viável, fazendo parte de uma área de estudo que se encontra em plena expansão.

A partir da metade da década de oitenta, um grande número de estudos estruturados a partir da metodologia com base em corpus passou a ser mais frequente na área de investigações linguísticas contemporâneas. Com a redução dos custos das máquinas, tornou-se possível armazenar e processar um grande número de dados para fins de investigação sem grandes dificuldades.

Para se ter uma breve ideia, o primeiro corpus linguístico electrónico foi lançado

em 1964, o *Brown University Standard Corpus of Present-day American English*. Esse corpus continha 1 milhão de palavras, uma quantidade de dados considerada impressionante para a época, mas pouco significativa se comparada a um corpus actual como o BNC (*British National Corpus*), que conta hoje com aproximadamente 100 milhões de palavras.

Portanto, pode-se perceber que a utilização intensiva das máquinas viabilizou o processamento automático de uma grande quantidade de dados de línguas naturais, fazendo com que a utilização de corpus se tornasse uma rotina entre diversos investigadores da área.

Além disso, é importante destacar que esse tipo de abordagem serve como uma alternativa para aqueles investigadores que resistem ao distanciamento da teoria linguística em relação à língua utilizada no cotidiano. A Linguística de Corpus permite que não somente os linguistas, mas também professores, tradutores, lexicógrafos e outros profissionais da área possam questionar paradigmas da teoria linguística que foram estabelecidos sem uma abordagem empírica.

De acordo com [18]. A Linguística de Corpus ocupa-se da colecta e da exploração de corpus, ou conjuntos de dados linguísticos textuais colectados criteriosamente, com o propósito de servirem para investigação de uma língua ou variedade linguística. Esse tipo de abordagem tem como objectivo a exploração da linguagem por meio de evidências empíricas extraídas com o auxílio do computador.

Um dos aspectos que atrai muito a atenção dos linguistas de corpus é a busca por evidências que mostrem certa padronização do léxico, isto é, "uma regularidade nos tipos de associação a que se submetem as palavras de uma língua [18]. A descrição desse tipo de padrão já foi estudada por vários autores, como afirma [18].

Os padrões de uma palavra podem ser definidos como todas as palavras e estruturas com as quais são regularmente associadas e que contribuam para seu significado. Um padrão pode ser identificado se uma combinação de palavras ocorre com relativa frequência, se é dependente de uma palavra específica, e se há um significado claro associado.

A área que se ocupa sobre a descrição de padrões se preocupa com questões de cômo descubrir quais os padrões lexicais (relações que caracterizam os sentidos das palavras e que contribuem na resolução do problema da ambiguidade das palavras) dos quais a palavra faz parte; investigar se a palavra se associa regularmente com outros sentidos específicos; determinar em quais estruturas ela aparece e estabelecer

se há uma correlação entre o uso/sentido da palavra e as estruturas das quais ela participa e verificar se a mesma está associada com uma certa posição organizacional no texto [18].

Todas as questões citadas acima se referem à padronização como regularidade expressa na recorrência sistemática de unidades co-ocorrentes de várias ordens (lexical, gramatical, sintáctica etc.). Segundo [18], esses padrões podem ser formalizados em três conceitos principais:

1. Colocação: associação entre itens lexicais, ou entre o léxico e campos semânticos. Por exemplo, em termos lexicais, *stark* associa-se a *contrast*; *sheer*, a *scale*, *number* e *force*. Em termos de campos semânticos, *jam* relaciona-se com itens no campo dos alimentos: *tarts*, *butty* e *doughnuts*.

2. Coligação: associação entre itens lexicais e gramaticais. Por exemplo, *start* é mais comum com sintagmas nominais e orações -ing, enquanto *begin* é mais usado com um complemento *to*.

3. Prosódia semântica: associação entre itens lexicais e conotação (negativa, positiva ou neutra) ou instância avaliativa. O nome deve-se ao facto de certas palavras prepararem o ouvinte, ou o leitor, para o conteúdo semântico que está por vir, da mesma maneira que a prosódia na fala indica para o interlocutor que tipos de sons estão por vir a seguir. Por exemplo, *cause* tem uma prosódia semântica negativa, pois se associa a palavras desfavoráveis como *problem(s)*, *damage*, *death(s)*, *disease*, *concern* e *cancer*. Já *provide* possui uma prosódia semântica positiva ou neutra, uma vez que se associa a palavras desse tipo, tais como *assistance*, *care*, *jobs*, *opportunities* e *training*.

Um dos aspectos centrais de estudo tanto em linguística teórica como em linguística computacional é a polissemia, o qual é um problema controvertido para o entendimento da linguagem. Apesar da grande quantidade de bibliografia dedicada ao respeito bem como a existência de várias teorias e orientações, a polissemia segue sendo um problema teórico de difícil solução [186]. Em particular, na área de Processamento de Linguagem Natural (PLN), a polissemia se considera desde meio século como o maior problema por resolver [Weaver55] e as competições Senseval de sistemas de desambiguação léxica (Word Sense Disambiguation, WSD) revelaram a imensa dificuldade da tarefa [114].

Este trabalho focaliza a expansão de consultas e desambiguação do sentido das palavras de forma automática, usando o recurso linguístico WordNet. O problema da

ambiguidade é comum em muitas aplicações, como a Recuperação de Informações, a Tradução Automática (TA), a Extracção de Informações, a Análise de Conteúdo, etc. Na TA e em outras aplicações multilingues, em particular, os "sentidos" de uma palavra ambígua em uma língua-fonte (LF) correspondem à sua tradução na língua alvo (LA).

A ambiguidade lexical é causada, fundamentalmente, pela existência de algumas relações semânticas interlexicais, principalmente a polissemia e a homonímia. De acordo com a classificação adoptada neste relatório [137], na **polissemia** uma mesma palavra tem dois ou mais significados diferentes, mas relacionados entre si, sendo que, normalmente, somente um dos significados se ajusta a um determinado contexto. Na **homonímia** duas ou mais palavras com significados totalmente distintos, sem traços comuns, são idênticas quanto ao som (homofonia) e/ou à grafia (homografia).

O problema da ambiguidade lexical pode, ainda, ser classificado como ambiguidade categorial ou ambiguidade de sentido [220]. A **ambiguidade categorial** ocorre quando as duas ou mais opções de significados de uma dada palavra são de diferentes categorias gramaticais. Na tradução, um exemplo de ambiguidade categorial causada pela relação de homonímia é a palavra do inglês *field*, que pode ser traduzida para as palavras "campo" (substantivo) ou "interceptar" (verbo), em português. Já um exemplo de ambiguidade categorial derivada da relação de polissemia é a palavra do inglês *eats*, que pode ser traduzida em português como "mantimentos, víveres, gêneros alimentícios" (substantivos) ou "come" (verbo "comer" conjugado na terceira pessoa singular, presente do indicativo). A **ambiguidade de sentido**, por sua vez, ocorre quando as duas ou mais opções de sentido (ou tradução) de uma dada palavra têm a mesma categoria gramatical. Alguns exemplos são a palavra *know*, que pode ser traduzida como "saber" ou "conhecer", como um caso de polissemia, e a palavra *light*, que pode ser traduzida como "leve" ou "luz", como um caso de homonímia.

A ambiguidade categorial é, em geral, muito mais simples que a de sentido, uma vez que pode ser resolvida, na maioria das vezes, pela análise das características sintácticas das palavras, realizadas por procedimentos de etiquetação gramatical ou análise sintáctica, por exemplo. Procedimentos dessa natureza alcançam, actualmente, resultados bastante satisfatórios. A resolução da ambiguidade de sentido, por sua vez, exige a análise da semântica das palavras e, eventualmente, a análise do uso de tais palavras (realizadas por procedimentos de análise semântica e pragmática, por exemplo). Portanto, o foco da maioria dos trabalhos voltados para o tratamento da ambiguidade lexical está no problema da ambiguidade de sentido,

considerando tanto a relação de polissemia quanto a de homonímia. A área que se ocupa do tratamento desse problema é denominada **Desambiguação léxica de sentido** (DLS), em inglês *Word Sense Disambiguation* (WSD).

Para realizar a desambiguação de maneira automática, é necessário incorporar um módulo de WSD aos processos de interpretação e/ou geração da língua. Para a construção de um módulo de WSD, várias questões devem ser analisadas, as quais dão origem a muitas decisões de projecto. Essas questões incluem o conjunto de palavras a desambiguar, quais os possíveis sentidos dessas palavras, o método adoptado para a desambiguação, cômo o módulo será avaliado, etc. Os trabalhos de WSD existentes apresentam, de facto, muitas variações, de acordo com as decisões tomadas com relação a essas e a muitas outras questões. Este relatório procura justamente analisar e discutir essas questões, bem como apresentar os principais trabalhos de WSD já propostos desde o surgimento dessa área, os quais foram desenvolvidos de acordo com diferentes decisões com relação às tais questões.

Quando se trabalha sobre a estrutura semântica de um documento é necessário utilizar conhecimento sobre as estruturas da linguagem; dito conhecimento deve ser de tipo morfológico, sintáctico, semântico e pragmático. O morfológico nos proporciona informação de cômo se constroem as palavras. O conhecimento sintáctico por outro lado nos dá informação de cômo combinar as palavras para formar frases, enquanto o semântico está enfocado a saber quê significam as palavras e cômo contribui o significado das mesmas ao significado completo da frase. Por último o conhecimento pragmático, proporciona-nos informação de cômo o contexto afecta à interpretação das frases. Todas estas formas de conhecimento linguístico, têm o problema associado da ambiguidade [157]. O objectivo geral dos sistemas de PLN, é o tratamento da língua a fim de ser interpretada da mesma maneira que o fazemos os seres humanos.

A tarefa da desambiguação do sentido das palavras consiste na associação de uma palavra, dada num contexto, com uma definição ou significado que a distingue de outros significados atribuíveis à dita palavra. Qualquer sistema de PLN precisa um módulo com estas características. WSD não é um fim em si mesma, senão que é uma etapa necessária para realizar como são a análise sintáctica ou a interpretação semântica em tarefas do PLN, bem como para o desenvolvimento de aplicações finais, tais como: recuperação de informação classificação de textos [118], análises de discurso e tradução automática [212] entre outras.

Por exemplo, um sistema de recuperação de informação tradicional responderá

à pergunta: Quais são as plantas que vivem no deserto? com todos os documentos que contenham os termos plantas e deserto independentemente do seu significado. Em alguns destes documentos o termo planta apareceria com o sentido de ser vivo, enquanto em outros significaria indústria. Se o sistema de recuperação de informação fora capaz de distinguir os sentidos dos termos da consulta, devolveria somente os documentos nos que se usa o sentido de ser vivo. Para isso, o sistema deve integrar um módulo de WSD, tanto para desambiguar os termos da consulta como os termos dos documentos indexados.

Ultimamente, ressalta-se o diálogo necessário entre a linguística teórica e a linguística computacional: por um lado não é possível um progresso significativo nos aspectos computacionais da polissemia sem avanços sérios nas questões teóricas; por outro lado, o labor teórico pode beneficiar-se dos resultados da linguística computacional e ao mesmo tempo encontrar sua comprovação nas aplicações do processamento da linguagem natural.

Uma posição de bastante relevância na área de WSD é a de quem sustentam a falta de concordância entre o tipo de conhecimento sobre os sentidos oferecido pelas fontes léxicas e o necessário para desambiguar ocorrências no texto [111] [223]. Uma opinião crescente na comunidade computacional é que o contexto desempenha um papel central na resolução da polissemia e por isso tem que ser parte integrante da sua solução [186].

1.2.1 Descrição do problema

A tarefa da WSD surgiu como um problema na área da Tradução Automática em 1949 (ver [227] e [100]) ao tratar de traduzir automaticamente uma palavra polissémica de um idioma para outro. Para alcançar este propósito era essencial saber a que sentido se referia uma palavra num idioma original para eleger a expressão apropriada no idioma destino. Em 1960 Bar Hillel encontrou-se com a seguinte frase em inglês "*Little John was looking for his toy box. Finally, he found it. The box was in the pen. John was very happy*" [91]. A palavra em inglês "pen" se poderia referir a uma *pluma* ou pena para escrever ou a uma creche para crianças. Qual é o sentido mais adequado para a palavra "*pen*"? Este problema não afecta somente na Tradução Automática, senão também nas áreas onde a precisão da busca e a análise da informação são essenciais.

Outro exemplo para demonstrar a dificuldade desta tarefa é a palavra polissé-

mica gato em, onde dois dos seus sentidos mais populares para estas palavras são: (1) animal mamífero felino e (2) máquina com encaixes que serve para levantar objectos pesados. A ambiguidade apresenta-se quando não se tem nenhuma ideia a que sentido se está referindo a palavra. Para tratar de resolver este problema se recorre na análise do contexto, quer dizer, as palavras que rodeiam a palavra que se pretende desambiguar. Por exemplo, na oração "*O gato necessita alimentos para sobreviver*". se podem encontrar as palavras que há no contexto que são chaves para determinar o sentido da palavra "gato", neste caso, as palavras "*alimento*" e "*sobreviver*" ajudam a determinar o sentido da palavra sobre o qual se refere ao animal mamífero felino. A desambiguação de uma palavra não é um problema para um ser humano dado que graças á interpretação do enunciado ele pode determinar a relação semântica existente entre as palavras do contexto e a palavra que se quer desambiguar. Desafortunadamente, a interpretação automática de um enunciado invólucra a correcta desambiguação da palavra. Assim a tarefa da desambiguação é uma tarefa prévia á interpretação considerando unicamente informação léxica orientada a estabelecer estatisticamente a sua ocorrência com palavras presentes no seu contexto imediato.

Nos anos 1970 e 1980, se apresentaram os primeiros sistemas de WSD. Estes sistemas estavam baseados em regras. Para estes sistemas se usaram fontes de conhecimento (Corpus e regras) que se desenvolveram manualmente centrando-se especificamente na desambiguação de algumas palavras específicas. Mais tarde, nos anos 1990 surgiram métodos baseados em técnicas de Aprendizagem Automática (AA) (refere-se a [155]) para um estudo). O propósito da AA é utilizar métodos e técnicas que permitam a máquina adquirir conhecimentos por meio de exemplos, neste caso, sobre o uso dos sentidos de uma palavra. A informação dos exemplos se encontra representada de maneira que os métodos e técnicas de AA possam interpretá-las. Com ajuda de estes algoritmos se criaram sistemas de WSD baseados em Métodos Supervisionados quer dizer, sistemas que dependem de conhecimentos previamente organizados para realizar a tarefa da desambiguação. Este conhecimento se centrava em forma de texto e/ou documentos de várias fontes (notícias, jornais, publicações, etc.) onde as palavras a estudar têm sido marcadas pontualizando o seu sentido (ver [100] e [145]). Assim estes corpus reúnem exemplos de uso de essas palavras especificando explicitamente o sentido em cada caso. As técnicas de AA tomam esta informação como conjunto de treinamento resultando em sistemas supervisionados de WSD [145].

Por suposto, os métodos supervisionados dependem do uso deste conjunto de

dados fazendo que o conhecimento dado ao problema do computador seja limitado. Quer dizer, a cobertura no nosso sistema somente pode enfocar-se naquelas palavras etiquetadas no corpus. Ainda mais, a criação destes conjuntos de dados, é muito custosa tanto em termos de recursos humanos como em tempo.

Antes da necessidade de evitar esta dependência, surgiram os métodos baseados em corpus que utilizam outros recursos: os métodos não supervisionados e não supervisionados. Os métodos supervisionados requerem um conjunto de dados etiquetados para o treinamento. Em câmbio os métodos não supervisionados, utilizam outros recursos, tais como dicionários, modelos matemáticos e incluso conjunto de dados não etiquetados. Ainda que não são tão eficazes como os métodos supervisionados, contam com a vantagem de poder ter uma cobertura mais ampla nas palavras graças a que já não depende de dados etiquetados.

Um enfoque recente entre os sistemas não supervisionados e minimamente supervisionados, é o uso da Web como um recurso lingüístico [76] para desambiguar, em contraste com outros métodos supervisionados, estes sistemas geralmente não utilizam técnicas de Aprendizagem Automática. Contudo apesar de ter a Web como recurso léxico, os resultados estão longe de ser exactos (ver [4] como um exemplo neste caso). Muitos destes sistemas utilizam a Web como uma fonte para obter mais exemplos e criar um subconjunto de dados não etiquetados automaticamente [143], mas ainda assim não são suficientes.

A diferença dos métodos mencionados, o enfoque que outros trabalhos têm sido propostos enfocam um ponto de vista distinto em quanto ao uso da Web: utilizá-la directamente como conhecimento desambiguador. As hipóteses que se têm pretendido obter ao longo destes documentos são que ao existir muitos documentos (muita informação) na Web se pode dizer que existem evidências sobre o uso adequado de palavras polissémicas. De esta maneira, a informação contida na Web se pode utilizar para apoiar directamente a desambiguação das palavras polissémicas e facilitar o processo de eleição do sentido mais adequado. Por esta razão, é necessário a criação de uma metodologia que faça explícito o conhecimento desambiguador contido implicitamente na Web, mas esta não é a motivação e o objectivo principal deste trabalho de tese doutoral, mas sim a motivação e o objectivo principal de este trabalho consiste fundamentalmente na expansão de consultas baseada em tesauros e desambiguação do sentido das palavras utilizando o recurso linguístico WordNet.

1.3 O sistema de recuperação de informação *Lemur*

Nesta secção vamos ver basicamente cômo funciona um Sistema de Recuperação de Informação. Para tal utilizaremos a ferramenta LEMUR [132], que trabalha de forma eficiente, estendida e utilizada pela comunidade de Recuperação de Informação e do Processamento de linguagem natural (PLN).

Como Sistema de recuperação de informação *Lemur* nos permite realizar todas as etapas desde a indexação até a avaliação. *Lemur* nos fornece um API com que desenvolvemos as nossas extensões em Java, e está desenhado para funcionar em sistemas operativos Unix o GNULinux e também funciona sob o sistema operativo Windows.

As ferramentas básicas são as seguintes:

- **BuildIndex**: Gera um índice a partir de uma colecção de documentos.
- **ParseQuery**: Processa as consultas para submetê-las a avaliação.
- **Reteval**: Realiza o processo de avaliação das queries y nos devolve um ranking de documentos.

Vejamos primeiro os conceitos com que devemos ter em conta:

Lemur é um conjunto de ferramentas (toolkit) para facilitar o modelado de linguagem e recuperação de informação. Inclui tecnologias como recuperação distribuída, com perguntas estruturadas, resumos automáticos, filtrado, e classificação. Permite programar facilmente as suas próprias personalizações.

Se podem construir sistemas básicos de recuperação de textos utilizando métodos de modelado de linguagem, além de métodos tradicionais tais como os baseados no modelo de espaço vectorial e *Okapi*. Também é utilizado numa ampla variedade de tecnologias da informação como o filtrado e a busca de respostas.

É útil tanto para alunos que estão aprendendo as características fundamentais dos sistemas de recuperação de informação como para investigadores em processamento de linguagem natural e recuperação de informação que não querem escrever o seu próprio indexador, e preferem concentrar-se em desenvolvimento de novas técnicas e algoritmos. Para além da indexação, proporciona alguns algoritmos de

recuperação de linha do fundo, como *Okapi* e *Divergencia KL* para uso e comparações.

Pode ser útil para construir os seus sistemas de busca: IR básico ad hoc, IR distribuído, IR utilizando perguntas estruturadas, IR usando indexações distribuidos, agrupamento de documentos e resumos.

1.3.1 Características

Indexação:

- *Stemming (stemmers de Porter e Krovetz)*: A técnica de stemmer de Porter consiste na transformação de uma palavra na sua raiz sintáctica, as técnicas de análise funcionam com os plurais, formas de gerúndio, sufixos do tempo passado etc. Por exemplo, "computer" e "computing" ambas se representam mediante "comput". Pelo contrário a técnica de Krovetz é um stemmer lexical linguístico de validação devido aos processos envolvidos na morfologia linguística. O stemmer utiliza o processo de consulta de dicionário ao fim de verificar todas as eliminações que ocorrem nas seguintes etapas: i) Transformação do plural para o singular; ii) Conversão da forma passada para a forma presente; iii) Eliminação do ing. O stemmer krovetz requere parámetros adicionais ao passo que o stemmer de Porter não requere parámetros adicionais.
- Omite *stopwords*: Portanto os stopwords constituem o ficheiro que contém a lista de palavras comuns, que elimina artigos, preposições, conjunções e algumas outras palavras mais frequentes.
- *Reconhece os acrônimos*: Os acrônimos formam o ficheiro que contém a lista de acrônimos. Os acrônimos são as palavras formadas pelas iniciais, por exemplo E.U.A é o acrônimo de "Estados Unidos de América".
- Reconhece as propriedades a nível de *token*, como a *part of speech* ou o reconhecimento de identidades: A gramática tradicional classifica as palavras baseadas em oito partes do discurso: O verbo, o substantivo, o pronome, o adjectivo, o advérbio, a preposição, a conjunção e a preposição. Cada parte do discurso não explica como a palavra é, mas como a palavra é usada. De facto, a mesma palavra pode ser um substantivo numa frase e um verbo ou um adjectivo noutra frase.

- *Permite a indexação de passagem.* Lemur utiliza PassageIndexer para a construção de um índice de passagem para uma colecção de documentos. Os documentos são segmentados em passagens com um número de termos de passagem, utilizando o parámetro PassageSize. Para a construção do índice segue-se alguns passos para a execusão de uma aplicação Lemur considerando os parámetros que Lemur proporciona [132].

Recuperação:

1, 3, 13, 74, 22 ...,

- *Ad hoc retrieval (TFIDF, Okapi, KL-divergence)*: Lemur é uma ferramenta que suporta distintos modelos de recuperação como por exemplo TFIDF, Okapi e KL-divergence. O parámetro para seleccionar o modelo de recuperação é retModel e toma os seguintes valores: 0 para TFIDF, 1 para Okapi e 2 para KL. RetEval executa vários experimentso de recuperación (com/sem feedback) para evaluar os diferentes modelos de recuperação.
- *Passage retrieval*: Os documentos são segmentados em passagens e um modelo de linguagem é construído para cada passagem. Se realiza um ordenamento (ranking) de passagens, de acordo com a probabilidade que a consulta poderia ter sido gerada por cada uma delas. Os documentos são ordenados baseado-se na contagem da sua melhor passagem.
- *Cross-lingual retrieval*: A ferramenta Lemur permite ao usuário usando determinados parámetros realizar a busca numa língua e recuperar os diferentes documentos em linguagens cruzadas.
- *Relevance feedback*: La ideia principal é a modificação de consulta existente baseada na valorização de relevancia.

Clustering de documentos: O agrupamento de documentos (clustering), é o acto de coleccionar os documentos similares em grupo, onde a similaridade é alguma função em um documento. Lemur fornece dois APIs para o agrupamento de documentos: El clúster (grupo) API, que define os clústeres por si mesmo e o clústerDB API, que define cômo os clústeres são persistentemente armazenados.

Summarization: Consiste numa aplicação que constrói um sumário ou resumo baseado na selecção da sentença (frase). Esta aplicação gera um sumário para um documento com selecção da sentença. Para usá-la, seguem-se as etapas

gerais de funcionamento de uma aplicação de Lemur, utilizando os parámetros correspondentes.

Processamento simples de texto: Na secção seguinte se explica cômo se realiza o processamento de texto.

1.3.2 Tipos de dados/documentos que indexa Lemur

[TRECParser]: Reconhece o texto nas etiquetas *TEXT*, *HL*, *HEAD*, *HEADLINE*, *TTL*, e *LP*. Por exemplo:

```
<DOC>
<DOCNO> document_number </DOCNO>
<TEXT>
Index this document text.
</TEXT>
</DOC>
```

WebParser : Elimina as etiquetas HTML e reconhece o seguinte formato:

```
<DOC>
<DOCNO> document_number </DOCNO>
Document text here could be in HTML.
</DOC>
```

1.3.3 Etapas fundamentais para a indexação e recuperação de informação

Nesta secção falaremos do "texto" referindo-se á informação contida tanto nos documentos da colecção como nas consultas do usuário.

- **Normalização do texto**. Antes de passar incluso ao preprocessado é conveniente normalizar os documentos. Esta normalização inclui eliminar dos documentos caracteres estranhos, lixos que pode aparecer, normalizar valores numéricos e de datas, etc.

- **Preprocessado do texto**. É bem conhecido que o processado dos documentos da colecção e das consultas (o básico é *stopper* y *stemmer*) aumenta em grande medida a bondade do sistema de RI.

- **Preparação do texto**. Uma vez preprocessados os documentos tem que adequar o formato que tinham ao que aceitem a ferramenta de RI que utilizemos. Neste caso tem que parsear o texto tendo em conta que Lemur aceita os formatos vistos anteriormente (*TrecParser* y *WebParser*). Qualquer dos dois é válido para Lemur.

- **Parsing do texto**. Uma vez que estão preparados os ficheiros lançamos o primeiro script de Lemur, que converte o/os ficheiros da colecção ou da consulta a um formato interno de Lemur.

- **Indexação**. Una vez que estão parseados os ficheiros da colecção lançamos outro *script* de Lemur, encarregado de criar o índice correspondente em ficheiros. Uma vez que estes índices estão criados podemos lançar quantas vezes queremos as consultas para fazer a recuperação de informação. Este passo de indexação e os prévios de preparação dos documentos da colecção se fazem *offline*; não nos vai importar quanto tempo demoram, já que se farão uma única vez.

- **Recuperação de informação**. Cada vez que tomamos uma consulta de um ficheiro, preprocessaremos essa consulta, adequaremos o seu formato, lançaremos o *script* de *parsing* e outro *script* de *Lemur* para fazer a recuperação de informação. Isto sim será um processo online, pelo que nos interessará que seja rápido para que o buscador funcione em tempo real e não se demore.

- **Leitura e compreensão dos resultados obtidos**. O último passo consistirá na visualização dos resultados, pelo que previamente terá que comprender quê nos devolve *Lemur*.Toda informação acima citada pode encontrar-se no sítio de Lemur.

1.4 Motivação e objectivos da investigação

O que se pretende com este trabalho é estudar cômo a expansão automática de consultas baseadas em tesauros e a desambiguação do sentido das palavras utilizando o recurso linguístico WordNet e com uma metodologia TREC de avaliação,

pode melhorar a recuperação de informação. Para tal se realiza o desenvolvimento de um sistema de recuperação. Finalmente dito sistema é avaliado para obter resultados que permitam comparar os resultados obtidos em realizar as "consultas originais" com respeito às obtidas ao utilizar a expansão de consultas, desambiguação do sentido das palavras, ou ambas.

1.5 Organização da tese

O conteúdo da tese doutoral está organizado da seguinte maneira:

No capítulo 2 se descreve de forma geral uma introdução relacionada com as tecnologias ás Bibliotecas Digitais, seguidamente se faz uma análise do lugar que elas ocupam na nova sociedade de informação em desenvolvimento, assim como na área da recuperação de informação, descrevendo os serviços que podem ofercer os meios virtuais na informação multimédia e por último se descreve a Biblioteca Virtual Galega como um exemplo prático de uma Biblioteca Digital que maneja distintas estratégias para facilitar a Recuperação de Informação.

No capítulo 3 se realiza uma descrição pormenorizada dos conceitos fundamentais sobre a recuperação de informação e as técnicas de indexação do texto. Assim como se explicam as diferenças entre os sistemas de recuperação de informação na Web e os sistemas de recuperação de informação tradicional.

No capítulo 4 se apresentam os comentários sobre os antecedentes dos outros trabalhos realizados no âmbito da expansão de consultas. Depois se apresentam os modelos de recuperação de informação assim como se descreve o modelo de usuário utilizado no presente trabalho da tese doutoral.

No capítulo 5 se apresentam os conceitos fundamentais dentro da desambiguação dos sentidos das palavras e se analisam os diferentes enfoques existentes na desambiguação dos sentidos das palavras, assim como os conjuntos de dados utilizados. Por outro lado se analisa WordNet, o recurso linguístico usado para a expansão de consultas e desambiguação dos sentidos das palavras. Além disto, se apresenta uma breve explicação de alguns algoritmos de aprendizagem automática mais utilizados nas experimentações.

No capítulo 6 se introduz as metodologias e ferramentas da avaliação de um sistema de recuperação de informação usadas de maneira padrão na recuperação de informação. Se avalia o SRI Lemur mediante medidas de precisão e *recall*, conside-

rando diferentes modos de consulta: i) Mediante a consulta original; ii) Expandindo a consulta com sinónimos mediante o uso do recurso linguístico WordNet, utilizando a metodologia da avaliação TREC baseada em simulação. iii) Desambiguando o sentido das palavras. Por outro lado se explicam os conceitos fundamentais no âmbito da avaliação de um SRI. A continuação se explica as provas realizadas comentando-as e comparando os seus resultados, destacando as melhorias obtidas ao expandir e ao desambiguar as consultas com as distintas categorias sintácticas sob o recurso linguístico WordNet.

No capítulo 7 se explicarão as conclusões obtidas neste presente trabalho de tese doutoral e se explicarão as perspectivas do possível trabalho futuro a ser realizado.

Também se incorporam dois apêndices A e B com a informação detalhada. No apêndice A se apresenta a lista de palavras comuns que nos permite eliminar durante o processo de indexação, os artigos, preposições, conjunções e algumas outras palavras mais frequentes e no apêndice B se descreve o desenvolvimento de uma aplicação que nos permite realizar a formulação da expansão de consultas e WSD, com o uso do recurso linguístico WordNet, no âmbito da RI.

Capítulo 2

As bibliotecas digitais

Neste capítulo para além de se apresentar de forma geral uma introdução relacionada com as tecnologias ás Bibliotecas Digitais (Digital Library), também se faz uma série de análises sobre o lugar que elas ocupam na nova sociedade da informação em geral e em particular, na área da recuperação de informação, descrevendo assim os serviços em que estes meios virtuais tão ricos em informação multimédia podem oferecer. Na última secção deste capítulo se exemplifica e se descreve a Biblioteca Virtual Galega, pois este tipo de serviços web podem em termos das suas capacidades dar coesão as sociedades virtuais com interesses comuns promovendo a interacção das pessoas que as integram.

2.1 Introdução

O potencial educacional e de preservação e disseminação cultural e científica existente nas Bibliotecas Digitais é inegável. Por outro lado, Bibliotecas Digitais estão entre os mais complexos e importantes sistemas de informação na actualidade. Tal complexidade é justificada principalmente pela inerente interdisciplinaridade envolvida na construção desse tipo de sistema. Bibliotecas Digitais integram resultados de investigação de disciplinas tais como Gerenciamento de Bancos de Dados, Recuperação de Informação, Engenharia de Software, Multimédia/Hipermédia, Interacção Humano-computador, entre outras.

A importância das Bibliotecas Digitais, radica no caso de que é um sistema de informação que contém colecções de documentos de diferentes tipos que permite

ao usuário utilizando uma determinada ferramenta de consulta, recuperar os documentos do seu interesse dado um contexto determinado, pois ao existir muitos documentos (muita informação) nas Bibliotecas Digitais existem evidências de que que muitas palavras sejam polissémicas. A recuperação de informação trata antes de tudo as necessidades de informação de um usuário e expressá-las em forma de consultas, para obter um conjunto de documentos que forneçam a informação expressada por essas necessidades. Portanto, isto resulta de interesse para as Bibliotecas Digitais, visto que o problema da recuperação de informação está relacionado com a interactividade do usuário com os sistemas de informação, as interfaces de busca, softwares desenvolvidos para a recuperação de informação de forma não ambígua dado o contexto. Além disto os documentos de consulta podem ser encontrados numa Biblioteca Digital de formas que o usuário, utilizando uma determinada ferramenta de busca obtenha os que ele considera relevantes dado o contexto.

A recuperação e disseminação de informação no ambiente Web são dificuldades que existem actualmente, pois podem estar de forma desestruturada e desorganizada segundo um padrão aceite na área de organização, armazenamento e recuperação de informação. Alguns recursos que podem minimizar essas dificuldades são tanto Bibliotecas Digitais, que possuem acesso simultâneo e remoto às informações de forma eficiente, quanto ao serviço de personalização, que permite ao usuário uma interacção personalizada baseada no seu perfil. O problema de prover esses recursos se encontra na onerosidade e dificuldade do processo de desenvolvimento desse tipo de biblioteca devido à grande quantidade de processos e elementos envolvidos na sua construção.

É flagrante o surgimento da sociedade da informação. Sobre a razão do papel crescente da informação, resulta de fundamental importância para o desenvolvimento económico, social, cultural e político dos países em geral. Muitas tentativas à determinação de um conceito para a sociedade da informação são relativamente recentes, sendo foco de constantes estudos em muitos países e organizações internacionais, objectivando definições mais exactas à extensão e influência na sociedade [159].

Essa preocupação, entretanto, fica mais evidente e intensiva nos anos 90, quando foi liberado o uso da rede Internet, criada em 1969, e utilizada apenas pelo meio académico, e posteriormente estendida aos meios de comunicação de massa.

A adopção dessa rede intensifica cada vez mais o acesso geral e irrestrito às informações devido às facilidades proporcionadas pela infra-estrutura e conjunto de

programas e protocolos que permitem aos usuários utilizá-la sob diferentes maneiras. Ao longo da história da Internet surgem diversos tipos de ferramentas (Gopher, Ftp, Telnet , correio electrónico, entre outros), e de serviços (Archie, Usenet, World Wide Web, entre outros).

As inovações tecnológicas implementadas para uso da interface WWW, proporcionaram a grande expansão para a intermediação de serviços e produtos nas áreas académica e comercial, e esse avanço transformou a Internet, em pouco tempo, num grande repositório universal do conhecimento humano nunca antes imaginado.

A concepção de uma interface de uso padrão (WWW), independente do tipo de ambiente computacional utilizado, possibilitou aos usuários e institucionais, tornarem-se também criadores de textos/informações, transformando a Web um meio de publicação e acessibilidade irrestritas. Esse universo sem fronteiras tem atraído a atenção de milhões de pessoas de todos os lugares do mundo, que actualmente executam tarefas por meio dessa rede. Apesar da disponibilidade desses recursos, o acesso aos repositórios não tem sido fácil; encontrar informações úteis, é frequentemente tedioso e tarefa difícil.

Quando da recuperação de informação necessária para resolver os seus problemas, o usuário deve "navegar" no espaço hipertextual (links Web), que é vasto e quase sempre desconhecido, trazendo um grande desconforto durante a realização da investigação, muitas vezes ineficientes devido à recuperação não satisfatória. Vários serviços indexam essas informações, por meio de motores de busca ou sistemas de recuperação de informação, cujos resultados apresentam problemas, pois a falta de uma infra-estrutura sólida e estável quanto ao tratamento dos recursos ali dispostos tem feito da Web um sistema de informação não muito bem disciplinado.

Essa situação é ocasionada principalmente pela falta de uma padronização, a nível mundial, que oriente a estruturação da informação com base em identificadores de forma e de conteúdos dessas páginas, lembrando que, conforme a ISO (International Standard Organization) [105], padrões são acordos oficializados, contendo especificações técnicas ou outros critérios para serem utilizados, consistentemente, como regras, normas ou definições de características.

Para encontrar soluções à essa questão, desde o início dos anos 90, várias iniciativas têm sido desenvolvidas pelas organizações responsáveis pelo gerenciamento como também por instituições de investigações e provedores de informação. Dentre os estudos existentes, encontra-se a questão da recuperação de informação, que tem como foco a forma de interactividade do usuário com os sistemas de informação

e as interfaces de busca, sem esquecer dos instrumentos (software) desenvolvidos para o seu acesso, a criação de códigos para a localização de objectos na rede, de forma inequívoca e, finalmente, a questão da descrição da informação contida nesse ambiente.

Essas questões envolvem o objectivo de identificar, localizar e recuperar a informação de forma única e não ambígua dado o contexto. Quanto a descrição dos vários tipos de conteúdos, verifica-se o desenvolvimento de conjuntos de metadados, também conhecidos por esquemas ou formatos de metadados, e esses mecanismos técnicos estão sendo utilizados já há algum tempo pelas organizações que gerenciam conteúdos na Internet para conferir o suporte necessário à organização dessas informações, sendo que a construção conhecida por sintaxe (de estrutura e elementos de dados) de cada tipo de metadados varia dependendo das necessidades especificas de seus usuários.

No momento, além da questão de gestão dos recursos na Web, verifica-se também a influência da Internet no ambiente da biblioteca e nas funções e serviços por ela desempenhados. De acordo com vários autores e especialmente por [126], a Internet é o recurso electrónico que tem provocado maior impacto nos serviços e operações de bibliotecas e nas actividades dos bibliotecários.

Dessa maneira os processos tradicionais de organização e prestação de serviços se modificam dia a dia. Um dos pontos críticos que a comunidade bibliotecária enfrenta actualmente é como efectuar o controle e a descrição de milhões de itens disponíveis na rede, e incorporá-los nas coleções.

Inúmeros estudos e experiências para compatibilizar procedimentos já existentes para o controlo e acesso aos dados bibliográficos com o ambiente de novos suportes de informação estão em andamento a fim de adaptar as práticas tradicionais para esses recursos. Entretanto, iniciativas por investigadores de outras áreas avançam, gerando novos conhecimentos para a organização e tratamento dos estoques de informação. Os resultados dessas investigações na área da Ciência da Informação, apontam que a influência dos vários tipos de padrões de metadados, empregados para a descrição de recursos Web e para as bases de dados, influenciam as metodologias utilizadas pelas bibliotecas no tratamento da informação [126]. Para isso, faz-se necessário conhecê-los para subsidiar a sua utilização ou adaptação, e a adopção de metadados, desde o momento da geração do recurso de informação irá oferecer condições necessárias para a sua recuperação e subsídios para o tratamento dos dados por parte das organizações responsáveis pelo controle bibliográfico nacional e

universal.

Em concreto as Bibliotecas Digitais modernas, e especialmente os sistemas de busca na Web estão especialmente orientados para melhorar as capacidades de recuperação.

2.2 As bibliotecas digitais

Um número considerável de informações localizadas em Websites apontam actualmente para publicações electrónicas ou para textos que estão armazenados em Bibliotecas Virtuais e Digitais, e as tecnologias que propiciaram essas condições encontram-se agrupadas na computação (computadores e tecnologias de armazenamento), redes (terminais e aplicativos para distribuição), e nos conteúdos (texto, imagem, vídeos, sons, música). Esse último aspecto apresenta também mudanças significativas quanto ao aspecto das linguagens de marcação SGML, HTML, XML [61], assim como nos formatos de apresentação para o usuário: ASCII (American Standard Code for Information Exchange); Bitmaps , formato de imagem que incluem o GIF (Graphic Interchange Format), JPEG (Joint Photographic Experts Group), PNG (Portable Network Graphics), BMP (bitmap OS/2 ou Windows), TIFF (Tagged Image File Format), entre outros; formato PDF (Portable Document Format); formato PostScript, formato Latex/Tex [198].

Todas essas implementações deram as condições necessárias para a consolidação das publicações electrónicas, cujos projectos propiciaram as condições para a edição nesse suporte e podem ser entendidas, segundo [198], como "qualquer tecnologia de distribuição de informação de forma que possa ser visualizada pelo computador e que utiliza recursos digitais para adquirir, armazenar e transmitir informação de um computador para outro", estando disponíveis em formato físico como disquetes, CD-ROM, ou por acesso on-line, principalmente pela Internet, utilizando-se de recursos como correio electrónico, File Transfer Protocol (FTP) e WWW.

Essas experiências consolidaram conhecimentos e proporcionaram subsídios para o desenvolvimento de Bibliotecas Digitais que, conforme [59], é uma das formas mais avançadas e complexas de sistemas de informação, pois frequentemente envolve “suporte de forma colaborativa, preservação de documento digital, gerenciamento de base de dados distribuída, hipertexto, filtros de informação, recuperação de informação, módulos de instrução, gerenciamento de direitos autorais, serviços de informação multimédia, serviços de referência e respostas às questões enviadas, busca de

recursos, e disseminação selectiva.

Muitos autores têm desenvolvido estudos objectivando a conceituação de uma Biblioteca Digital. Para [13] o conceito de biblioteca digital está na analogia com um lugar onde se encontra um repositório contendo uma coleção organizada de publicações (que possam ser impressos) e outros artefactos físicos, combinados com sistemas e serviços que facilitem o acesso físico, intelectual, e disponível por longo tempo. Entretanto [61] considera que há diferentes conotações sobre Biblioteca Digital para diferentes grupos de profissionais. Para os de tecnologia da informação é um mecanismo poderoso para gerenciar bases de dados distribuídas; já para a comunidade de negócios ela representa um mercado novo, e quanto à comunidade da ciência da informação é uma nova forma para ampliar, distribuída e remotamente, o acesso a recursos de informação.

A definição de Biblioteca Digital, baseando-se nos estudos realizados até o presente momento, demonstram que ainda está em construção, e em muitas situações confundem-se como Bibliotecas Virtuais ou electrónicas. Sobre esse aspecto [27] proporciona, por meio de uma revisão bibliográfica sobre o assunto, um quadro com a evolução histórica e as características das bibliotecas em ambiente digital, e contribui para uma melhor compreensão sobre a actual conjuntura que passam as bibliotecas de modo geral. De forma geral uma biblioteca é simplesmente uma base dados de documentos, onde um usuário em particular pode encontrar os documentos do seu interesse.

Apesar dessa situação quanto a redefinição das bibliotecas na sociedade da informação, a Biblioteca Digital tem, cada vez mais, um papel fundamental no planeamento estratégico dos novos serviços de informação com a finalidade de facilitar o acesso universal ao património científico e cultural [142]. Segundo a autora, "se uma biblioteca é um conjunto organizado de documentos, uma biblioteca digital é um conjunto organizado de documentos eletrónicos, ou seja, um conjunto organizado de objectos digitais, onde os metadados são a chave para essa organização e formas de acesso à informação contidas nesse conjunto de dados". Dessa forma, a criação de uma Biblioteca Digital implica conhecer todos os processos de tecnologia de informação (hardware, software, armazenamento, protocolos, etc.), e da biblioteca (definição do modelo de metadados, padrões a serem adoptados, nível de detalhamento da descrição, metodologias para recuperar a informação organizada, entre outros requisitos), destacando-se os metadados (dados sobre os dados) que serão a chave fundamental para proporcionar uma recuperação eficiente, eficaz e fácil de informações/documentos úteis para o usuário.

2.2.1 Metadados

A representação e mediação do conhecimento é o elemento-chave para acesso rápido e adequado à informação, e essa necessidade originou várias formas para organizá-la. O controle se manifesta em duas premissas: a descrição, que ao mesmo tempo é uma operação (catalogação) e um produto, e a classificação, que representa as relações existentes entre tópicos referentes a diversas áreas. O objetivo da descrição é fornecer uma representação da fonte, permitindo identificá-la, localizá-la e representá-la nos catálogos correspondentes.

A adopção do catálogo propiciou o controle da informação, incluindo os vários instrumentos normativos de apoio à organização documental, extensivamente utilizados pela comunidade internacional das áreas de biblioteconomia e ciência da informação, e pelas disciplinas relacionadas, tais como museus, arquivos, gerenciadores de registros, documentos, resumos e indexação, tesauros, e sistemas de recuperação da informação [71].

Recuperar a informação contida nas publicações sempre foi o principal objectivo dos instrumentos elaborados pelas bibliotecas e sistemas de informação. Segundo [197] "a questão crucial é que o processo de recuperação depende muito das etapas de indexação e armazenamento, os quais determinam, em grande medida, a estratégia de melhor possível para as buscas feitas num sistema de recuperação da informação". Com o desenvolvimento de tecnologias de informação e comunicação, novas formas para estruturar e disponibilizar a informação são desenvolvidas para acesso por meios electrónicos, como por exemplo os catálogos on-line, mais conhecidos pela sigla OPAC (On-line Public Access Catalog) e com acesso pela Internet.

Assim, a estruturação e organização de conteúdos (textos, gráficos, entre outros) devem estar categorizados e auxiliados por um sistema de navegação intuitivo e confiável [177]. Além disso, o tratamento da informação, de acordo com padrões internacionais para o armazenamento em bases de dados e disponíveis em ambientes digitais, continua sendo imprescindível e importante para a recuperação de informação adequada.

Optimizar a recuperação de informações sempre foi um desafio para os organizadores de repositórios de informação. Actualmente, segundo [12] "bibliotecários e especialistas da informação se esforçam para desenvolver métodos para a descrição, organização e recuperação de objectos digitais remotamente e eles não estão só nesse desafio, porque criadores, provedores e usuários de fontes electrónicas nos

meios académicos, públicos e comerciais também se acham dedicados à investigação sobre esse vasto campo de informações. Cada grupo tem abordado o problema de organização e acesso, a partir dos seus próprios esquemas de referência.

Dessa forma estudos realizados resultam no desenvolvimento de métodos e padrões para a organização de recursos de informação on-line, designados como esquemas de metadados, ou formatos de metadados. Metadados, genericamente, vem sendo definidos como dados sobre dados, ou informações sobre a descrição e a localização de informações existentes na Internet, com o objectivo de permitir a sua recuperação de forma mais adequada por meio dos Websites.

Estudos e experiências disponíveis na literatura desde a década de 90 demonstram a importância do uso de elementos metadados para a estruturação de sistemas de informação, e relatam sobre propostas para o tratamento dos recursos de informação em meio digital, e nesses últimos anos tem sido foco de análise tanto pela comunidade de biblioteconomia e documentação como também por outras áreas responsáveis pela organização e gerenciamento de recursos de informação em geral.

A origem do termo "metadados" se inicia nos anos 60, mas aparece com maior frequência na literatura sobre sistemas de gerenciamento de bases de dados a partir dos anos 80, sendo empregado para identificar as informações auto-descritivas e de auto-controle dos dados contidos nas bases [222]. Até à poucos anos somente alguns teóricos da área tinham conhecimento da palavra metadados, entretanto, o entendimento do conceito de metadados tem se tornado fundamental para os autores, produtores e usuários de serviços de informação [153], e tem sido estudado por muitos especialistas, incluindo bibliotecários, e as conclusões levam à mesma essência que é constante e que tentam solucionar questões sobre o controle e descrição de recursos disponíveis na Internet.

Tendo como característica a observância a padrões internacionais para a descrição do conjunto de dados, os metadados devem estar alinhados com as regras previstas por padrões desenvolvidos por organizações de renome na área, e que são designados como formatos de metadados. Esses formatos, que servem a distintas necessidades e audiências, podem ser utilizados para descrever os mesmos recursos para múltiplos propósitos, tendo como função fornecer as definições que estabelecem a organização padronizada de conteúdos e condições para o intercâmbio por meio magnético [95].

Muitos formatos de metadados são especificações estabelecidas por consenso de determinadas comunidades que gerenciam recursos de informação em suporte digital, com o fim de atender necessidades de informação específicas. Nesse sentido, vários

projectos estão sendo realizados para permitir o estabelecimento de um padrão geral de formatação de metadados com parámetros em âmbito mundial.

De um modo geral um formato deve ter sempre uma especificação formal da estrutura e da semântica; deve definir um conjunto coerente de atributos e das terminologias adoptadas na descrição. Conforme [228], o requisito formal significa que o formato é mantido por uma agência autorizada e responsável pelas especificações conferindo credibilidade ao esquema dos dados; o requisito estrutural possui características codificadas dos valores contidos; e o requisito semântico determina atributos cujo significado é compreendido pelo homem e pela máquina dentro de uma coerência indiscutível.

Complementando, [10] identifica que um formato de metadados deve permitir atingir o objetivo dos metadados, facilitando a identificação, localização, recuperação e uso da informação pelo usuário. Cada formato é construído sob um conjunto de especificações e necessidades, e elaborados por especialistas nas áreas em que foram implementados.

No desenvolvimento de bibliotecas digitais, a formatação de acordo com uma semântica da estrutura com regras bem definidas para a descrição dos componentes, e de uma sintaxe estabelecida para codificação e transferência de dados são condições para o estabelecimento de sistemas robustos de acesso a informação digital. Como exemplo pode-se citar a Biblioteca Digital de Teses da USP, que estabelece padrões para a descrição e estruturação de dados designados pelo Dublin Core, e complementados com metadados locais.

A partir das experiências já existentes, a "nova concepção de biblioteca", substanciada em todos os tipos de bibliotecas emergentes, requererá dos profissionais da informação um alto grau de conhecimento das tecnologias e dos padrões necessários para sua estruturação e gestão. O desafio que as tecnologias da informação e comunicação estabelecem nesse novo século, e a forma como estarão sendo utilizadas é que irão proporcionar o diferencial das organizações responsáveis pelo controle bibliográfico nacional e universal.

2.3 Serviços

Apesar de que as Bibliotecas Digitais surgiram com base aos documentos digitais que contém, torna cada vez mais claro as mudanças que se estão produzindo

nas mesmas ou seja a medida que se vai prestando maior atendimento aos serviços solicitados pelos usuários ao mesmo tempo se vai abandonando os enfoques centrados ao desenho dos conteúdos. As necessidades dos usuários e interesses são os elementos centrais das Bibliotecas Digitais a volta dos quais se organiza o espaço virtual que, por conseguinte é a própria Biblioteca. As políticas implementadas são as que determinam quem pode aceder a quê conteúdos e a quê serviços, bem como a quantidade, qualidade e classe de serviços implementados.

Os serviços mais comuns nas Bibliotecas Digitais são: i) Serviços de busca: Nos permite realizar buscas por dados estruturados e buscas por contéudos nas Bibliotecas Digitais. ii) Serviços de interconexão: Estes são seviços implementados nas Bibliotecas Digitais, entre estes serviços se encontram os emails, chats e grupos de notícias. de usuários; iii) Serviços de difusão de informação: Estes serviços são de tipo publicitário e consistem aos serviços de distribuição proporcionada por outros usuários, administrada por bibliotecas ou por empresas associadas; iv) Serviços associados a perfis do usuário: A ideia básica é que diferentes usuários têm diferentes interesses, gostos, etc. e que estas preferencias são até certo ponto constantes, trata-se de facto, de a biblioteca seja capaz de identificar à pessoa que se está conectando e que actue em consequência oferencendo-lhe por exemplo o interfaz de usuário, os enlaces nas páginas Web etc.; v) Serviços de valor acrescentado: Este tipo de serviço englobam serviços de tradução automática, de preparação de bibliotecas, etc. Neste contexto na seguinte subsecção descreveremos os serviços de busca dada a sua importância no marco da presente tese doutoral, pois são aqueles que facilitam ao usuário o acesso aos documentos que pretende encontrar.

2.3.1 Serviços de busca

O facto de existirem tanto dados estruturados como dados não estruturados faz com se desenvolvam dois tipos de buscas: Buscas por dados estruturados e busca por contéudos:

1. Busca por dados estruturados: Este tipo de busca consiste em que todos os campos (título, nome do autor, data da edição etc.) possam ser usados os que tenham a informação mais relevante para a realização da busca.

2. Busca por conteúdo: É um dos campos mais activo na investigação desde o surgimento da Internet e chama-se recuperação de textos ou recuperação de informação e abarca uma série de métodos e técnicas para a recuperação de

informação nas bases de dados, obtendo assim os resultados de busca. As técnicas de recuperação de textos estão em diferentes níveis de desenvolvimento. O inglês é a língua que conta com mais ferramentas básicas tais como: dicionários, tesauros, sinónimos e antónimos, redes de associação semântica de palavras, lematizações etc.

Actualmente várias Bibliotecas Digitais têm implementado serviços de busca por conteúdo que em geral, não vão além da busca de documentos que contenham verdadeira (s) palavra (s) ou cadeia (s) de caracteres (raiz de uma palavra). Segundo a língua que se trate as técnicas de recuperação de texto estão em diferentes níveis de desenvolvimento. Apesar que o inglês é a língua com mais ferramentas, em espanhol assim como em outras línguas europeias ainda têm um longo caminho a percorrer. Numerosos grupos de investigação em Linguistica Computacional em Espanha e Ibero-América, estão desenvolvendo todo tipo de ferramentas para o processamento automático da língua de Cervantes. Na actualidade se conta com dicionários electrónicos e ferramentas de lematização.

Existe uma Federação de Bibliotecas Digitais que portanto, é um aspecto importantíssimo a se ter em conta, pois é uma área de investigaçao muito activa e em constante desenvolvimento, pois nesta área apresenta-se novos problemas que a Federação de Bases de dados não fomentava anteriormente. Em [30] e [31] é necessário que se possam explorar as capacidades de recuperação de textos das Bibliotecas Digitais Federadas e, além disto, em [32] é necessário desenvolver tecnologia que permita federar Bibliotecas Digitais cujos fundos estejam em línguas diferentes. A proliferação das Bibliotecas Digitais por todo o mundo originou o surgimento da Federação de Bibliotecas Digitais com a necessidade de aglutinar esforços para a não repetição de trabalhos já realizados de catalogação e digitalização das mesmas obras nas diferentes bibliotecas assim como facilitar aos possíveis usuários a obtenção de documentos de busca que considerem pertinentes.

Temos como exemplo de uma Biblioteca Virtual, a Biblioteca Virtual Galega (Figura 2.1), surgiu com o próposito de suprir a carência de conteúdos e literatura em idioma galego existente na Web. Foi desenvolvida completamente no Laboratório de Bases de Dados da Universidade da Coruña, em colaboração com a Área de conhecimento de Filologia Galega e Portuguesa (incluso no Departamento de Galego-Português, Francês e Linguística) de dita universidade. Cabe destacar que, apesar de seu título, a Biblioteca Virtual Galega da Universidade da Coruña não só é uma biblioteca senão também uma editorial virtual denominada eDixital e que atende aos dois objectivos básicos: i) Facilitar o conhecimento da literatura galega, e oferecer

uma plataforma editorial, eDixital, que permitisse aos escritores em activo, tanto consagrados como novatos, publicar as suas produções; ii) Oferecer um ponto de acesso centralizado a qualquer tipo de informação sobre Galiza.

Neste âmbito encontra-se a questão da recuperação de informação, que tem como foco a forma de interactividade do usuário com os sistemas de informação e as interfaces de busca, sem esquecer dos instrumentos (softwares) desenvolvidos para o seu acesso, a criação de códigos para a localização de objectos na rede, de forma inequívoca e, finalmente, a questão da descrição da informação contida nesse ambiente. Essas questões envolvem o objectivo de identificar, localizar e recuperar a informação de forma única e não ambígua. Neste contexto, para além, da questão de gestão dos recursos na Web, verifica-se também a influência da Internet no ambiente da biblioteca e nas funções e serviços por ela desempenhadas. De acordo com vários autores e especialmente por [203], a Internet é o recurso electrónico que tem provocado maior impacto nos serviços e operações de bibliotecas e nas actividades dos bibliotecários.

Figura 2.1: Biblioteca Virtual como metáfora de uma biblioteca real.

Resulta importante salientar que durante o processo de realização das buscas utilizam-se várias técnicas tais como:

1. Uso da metáfora cognitiva: É uma técnica amplamente estudada na área da Interacção Pessoa-Computador [191]. Um exemplo clásssico são os processa-

dores de textos que utilizam à metáfora de máquina de escrever, além disto, utiliza-se a metáfora do arquivo para facilitar a busca de autores e obras.

2. Aproximaçao navegacional: Esta técnica é estudada dentro do campo da Interacção Homem-Máquina e mais concretamente na área de recuperaçao de informaçao [15] (cap: 2), pois permite aos usuários realizar as consultas utilizando distintas técnicas com o objectivo de melhorar os resultados de busca (o que implica a conhecer a linguagem de consultas). A Recuperação de informação está relacionada com a interactividade do usuário com os sistemas de recuperação, interfaces de busca, etc. que permita a busca da informação de forma não ambígua dado o contexto.

A recuperação de informação do tipo texto é o ramo da computação que se dedica ao estudo do armazenamento e recuperação de documentos texto que podem estar contidos numa Biblioteca Digital. O seu principal objectivo é recuperar esses documentos de maneira a satisfazer as necessidades do usuário, as quais são expressas por meio de uma consulta (query). A consulta é uma representação do desejo do usuário diante do sistema, ela é composta por termos-chave que são os elementos básicos do sistema. Um sistema de RI é composto por várias fases tais como a fase de indexação, quando os índices são criados para futuras consultas, a fase de consulta, quando o usuário é capaz de informar os seus desejos (consulta) ao sistema, etc. Durante a fase de indexação, criamos índices. Um exemplo tradicional de índice é o dos índices invertidos que permitem recuperar rapidamente um documento apartir dos termos que contém.

Na fase de consulta, o usuário fornece ao sistema uma consulta que representa o documento que ele deseja recuperar. Essa consulta é comparada aos índices dos documentos previamente armazenados. Os documentos cujos índices sejam considerados semelhantes à consulta são recuperados e ordenados segundo sua semelhança. Durante principalmente a fase de consulta, a representação conceitual de um documento desempenha um papel fundamental. Pois essa representação possibilita que definamos uma função que meça a semelhança entre documento e a consulta. Os documentos são geralmente textos ou partes do texto de documentos e o principal objectivo de um sistema de RI é recuperar informação (contida nos documentos) que possa ser útil ou relevante para o usuário. Tal informação (de interesse do usuário) é normalmente chamada de necessidade de informação do usuário.

Capítulo 3

Recuperação de informação

Neste capítulo se abordam os conceitos relacionados com a recuperação de informação (RI, por suas siglas). Nas primeiras secções se descreve o processo de recuperação de informação, assim como as várias técnicas de indexação utilizadas neste processo, seguidamente se apresentam as principais diferenças de um sistema de recuperação de informação na Web com respeito a um sistema de recuperação de informação tradicional. Posteriormente se analisa brevemente os tipos de motores de busca. Finalmente na última parte deste capítulo se descreve a indexação com aproximação linguística.

3.1 Introdução

A adequada recuperação de informação está directamente associada com a tarefa do usuário e a vista lógica dos documentos do sistema de recuperação de informação.

Por um lado o usuário de um sistema de recuperação de informação deve traduzir a sua necessidade de informação numa consulta em linguagem adecuada para o sistema de recuperação. Geralmente, esta transformação produz uma série de palavras chave que resumem a descrição da necessidade de informação do usuário. Neste caso se considera que o usuário está realizando uma tarefa de recuperação. Em câmbio, se consideramos o exemplo de um usuário que tem interesse num tema pouco definido (por exemplo, basquete) sobre o que realiza uma busca para a seguir simplesmente examinar os documentos obtidos, pode consultar um documento sobre "Los Ángeles Lakers" e aí seguir a outros documentos sobre "Los Ángeles" e de aí seguir a docu-

mentos sobre "Turismo em California". Nesta situação se considera que o usuário está navegando sobre os documentos da colecção e não está buscando. O processo de navegação também é considerado um processo de recuperação de informação, no qual os objectivos não estão claramente definidos no início e cujo propósito final pode sofrer câmbios durante a interacção com o sistema [15] (cap: 1).

Os sistemas de recuperação de informação clássicos permitem unicamente realizar a tarefa de recuperação, enquanto os sistemas baseados em hipertexto estão especialmente preparados para uma rápida e ágil navegação sobre os documentos. De facto, as bibliotecas digitais modernas, e especialmente os sistemas de busca na Web estão especialmente orientados para a combinação de ambas as tarefas com objectivo de melhorar as capacidades de recuperação.

Tanto as tarefas de recuperação como de navegação se consideram, na gíria do World Wide Web, acções "*pull*" já que o usuário solicita essa informação de maneira interactiva. Em contraste, encontram-se as acções "*push*", que realizam a recuperação de informação de maneira automática utilizando agentes softwares que enviam a informação directamente para o usuário.

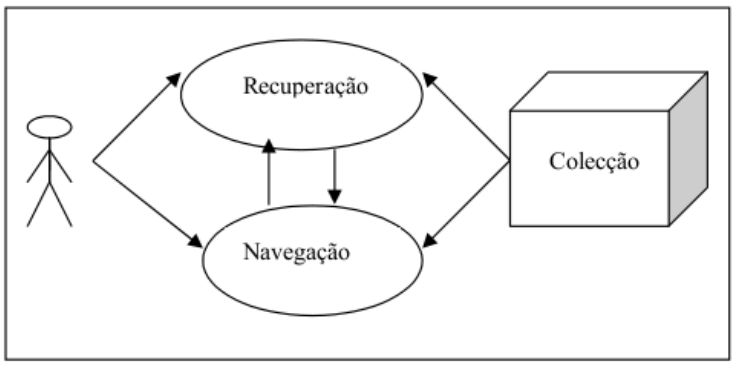

Figura 3.1: Interacção do usuário com o sistema de recuperação

Com respeito à vista lógica dos documentos de uma colecção, frequentemente é representada por meio do conjunto de termos indexados, também denominadas palavras chave. Estas palavras chave podem ser extraídas directamente do texto dos documentos ou bem, podem ter sido especificadas por expertos na matéria. Em ambos casos, sem importar a sua origem, estas palavras chave representam a vista lógica dos documentos.

A tecnologia actual faz possível que os documentos possam ser representados com todo o conjunto de palavras que os conformam. Neste caso, considera-se que

o sistema de recuperação adopta uma visão lógica de texto completo. No entanto, sobre esta base pode ser necessário reduzir o número de palavras chave consideradas, principalmente devido ao elevado número de documentos disponíveis (como por exemplo, na World Wide Web). Em geral se utilizam técnicas de eliminação de palavras comuns, de obtenção de raízes gramáticais ou de identificação de grupo de nomes. Estas operações recebem o nome de operações de texto, provocando a redução de complexidade da representação dos documentos e convertendo a vista lógica de texto completo num conjunto de termos indexados.

3.2 O processo de recuperação de informação

Nesta secção se descreve o processo de recuperação de informação que se realiza num sistema de busca genérico. Na seguinte se descreve a arquitectura global de um sistema deste tipo com base ao exposto em [15](cap: 1). A nível geral, o processo de recuperação se inicia com a definição da base de dados de texto. Para isso é necessário especificar:

a) Os documentos que farão parte da colecção.

b) As operações que se realizarão sobre o texto dos documentos. Estas operações serão as que transformarão os documentos originais e gerarão a vista lógica dos mesmos.

c) O modelo do texto, isto é, a estrutura do texto e os elementos a recuperar. Uma vez que se definiu a vista lógica dos documentos, passa ao processo de indexação que se encarregará de construir o índice associado ao texto. O índice é uma estrutura de dados crítica já que permitirá o acesso eficiente a grandes volumes de dados. É possível a utilização de múltiplas estruturas de dados, no entanto, a mais frequentemente empregada hoje em dia é o ficheiro invertido ou índice invertido. Os recursos empregados pelo índice (tipicamente medidos em tempo e espaço de armazenamento) durante a geração e construção do índice são recuperados através do processo de busca.

A partir da indexação da base de dados de documentos, o processo de recuperação se inicia. Para isso, o usuário especificará uma necessidade de informação que será analisada e transformada seguindo as mesmas operações de consulta (operações lógicas ou de conjuntos) podem ser aplicadas antes de obter a consulta real,

que representa a necessidade de informação original do usuário. Seguidamente, a consulta é processada para obter os documentos recuperados da colecção. Este processamento deve ser realizado com uns tempos de resposta mínimos, ao utilizar como base a estrutura do índice construída previamente. Antes de ser enviados os documentos recuperados ao usuário, são ordenados de acordo a um critério de relevância baseado na consulta original do usuário.

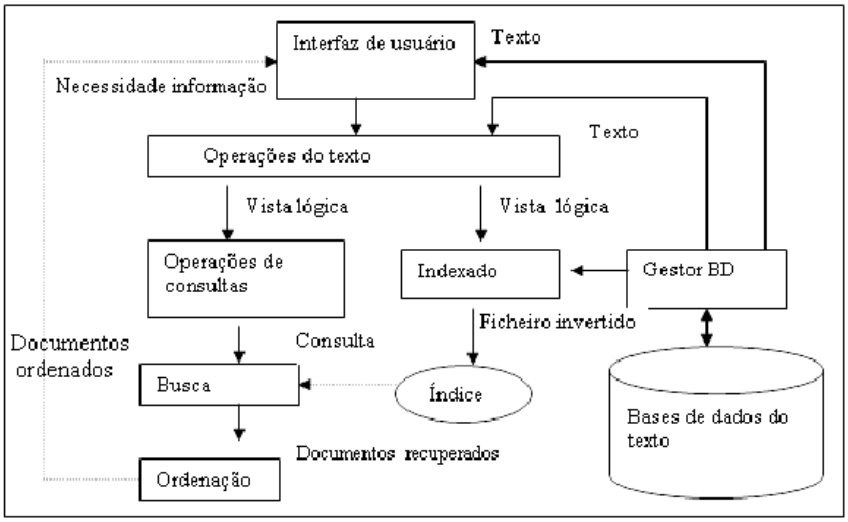

Figura 3.2: Arquitectura de um sistema de recuperação de informação

Por si só, cada um dos componentes de um sistema de recuperação, constitui uma parte fundamental no conjunto do sistema cujo desenho e implementação podem afectar drasticamente ao rendimento oferecido pelo sistema, tanto desde o ponto de vista da qualidade das buscas como desde a perspectiva dos tempos de resposta. No entanto, o núcleo de todos os sistemas de recuperação de informação está no índice, já que é esta estrutura de dados a que representa o conjunto de documentos contidos na colecção e permite a realização de busca sobre os mesmos de maneira eficiente.

Em consequência, na seguinte secção se examinam com detalhe algumas técnicas de indexação bem conhecidas, já que constituem a base fundamental na recuperação de informação.

3.3 Técnicas de indexação

A seguir se mostram as várias técnicas de indexação tradicionais. Esta secção é especialmente importante já que o conhecimento destas técnicas é chave para a realização e compressão de novos desenvolvimentos, já que a maior parte se baseia em variações e/ou extensões de técnicas aqui apresentadas.

Em concreto, se descrevem principalmente as várias técnicas dos ficheiros invertidos. Com tudo, considera-se relevante descrever como primeiro método a inspecção de texto completo, que ainda que não se trate propriamente de um método de indexação, é utilizado para a recuperação de informação directamente a partir de documentos, pelo que pode se considerar como o método mais básico e fácil de indexação.

3.3.1 Inspecção do texto completo

A forma mais directa e simples de localizar os documentos que satisfazem uma determinada consulta, materializada numa série de termos de busca, é a busca directa em todos os documentos destes termos. Em caso que a consulta engloba vários termos ou cadeias de caracteres unidos mediante operadores de lógica booleana, precisa-se um passo adicional para determinar se os documentos encontrados satisfazen a expressão booleana [54].

O estado de arte com respeito a este método de localização de informação se centra basicamente na localização de cadeias de caracteres em documentos. Um algoritmo elementar para tal efeito é o seguinte:

- Comparar os caracteres da cadeia de busca com os correspondentes caracteres do documento.
- Se ocorrer um erro, deslocar a cadeia de busca uma posição à direita dentro do documento e continuar até que se localiza a cadeia ou se atinge a parte final do documento.

A simplicidade deste algoritmo oferece um pobre rendimento. Considerando m o número de caracteres da cadeia de busca e n o número de caracteres do documento, a complexidade do algoritmo tende a O(m*n). Knuth et al. Expõem em [116] um algoritmo que requer O(m+n) comparações. A ideia básica deste algoritmo consiste

em deslocar a cadeia de busca em mais de um caracter quando se prediz um erro. Baseia-se num preprocessamento de cadeia de busca, cujo tempo de execução tende a O(m). No entanto, o algoritmo que proporciona um melhor rendimento é o proposto por Boyer e Moore em [24]. O conceito sobre o que se assenta este algoritmo consiste em realizar as comparações da direita à esquerda, de tal forma que se ocorre um erro, a cadeia de busca pode ser deslocada até m posiciones à direita. Esta técnica faz as buscas, em determinados casos alcançam tempos sublineais (sendo $O(\frac{n}{m})$ no melhor dos casos).

Por outra parte, existem outras soluções que se afastam dos enfoques mais tradicionais. Por exemplo, em [9] se propõe um método baseado na teoria de autómatas finitos que permitem a localização de múltiplas cadeias simultaneamente. E em [237] se propõe um algoritmo tolerante aos erros de soletro.

A nível geral, as principais vantagens do método de inspecção do texto completo são as ausências de sobrecarga do espaço de armazenamento e o mínimo esforço necessário para inserções de novos documentos e/ou actualizações dos existentes. Obviamente, a principal desvantagem são os tempos de resposta que proporciona, que são especialmente elevados se a colecção de documentos é de grandes dimensões.

Como se comentou previamente, este método não se pode considerar estritamente um método de indexação, basicamente porque não utiliza um índice para agilizar o processo de localização de informação. No entanto, é relevante já que pode ter aplicação em cooperação com algum outro método de indexação que realize um filtrado de documentos ou inclusive através de hardware de propósito específico a tais efeitos [96].

3.3.2 Ficheiros invertidos

A técnica de ficheiro invertido, ou de índice invertido, é um mecanismo orientado a palavras para a indexação de uma colecção de documentos de texto para agilizar as tarefas de busca. A estrutura de um ficheiro investido se compõe de dois elementos: o vocabulário e as ocorrências [161]. O vocabulário é o conjunto formado por todas as ocorrências de diferentes palavras que aparecem no texto. Associado a cada palavra se encontra uma lista de documentos onde se pode encontrar essa palavra. O conjunto de todas essas listas é o que constitui as ocorrências (ver Figura 3.2).

Alguns autores diferenciam o conceito de ficheiro invertido e lista invertida. Um ficheiro é o modelo descrito, enquanto que numa lista invertida cada elemento da

lista aponta a uma posição de um documento. Isto simplesmente é um problema de granularidade do direccionamento que abrange desde as posições do texto até blocos lógicos [161].

A utilização do ficheiro invertido aumenta a eficiência nas buscas várias ordens de magnitude, pelo que é uma estrutura básica para grandes volumes de informação. Por outra parte, a principal desvantagem é a quantidade de espaço necessário para armazenar a estrutura do índice que pode variar entre 10% e 100% do texto original, ou inclusive mais. Além disto, é necessário ter em conta as penalizações para as inserções de novos documentos e as actualizações nos documentos já existentes, que requerem acessos ao índice para sua modificação.

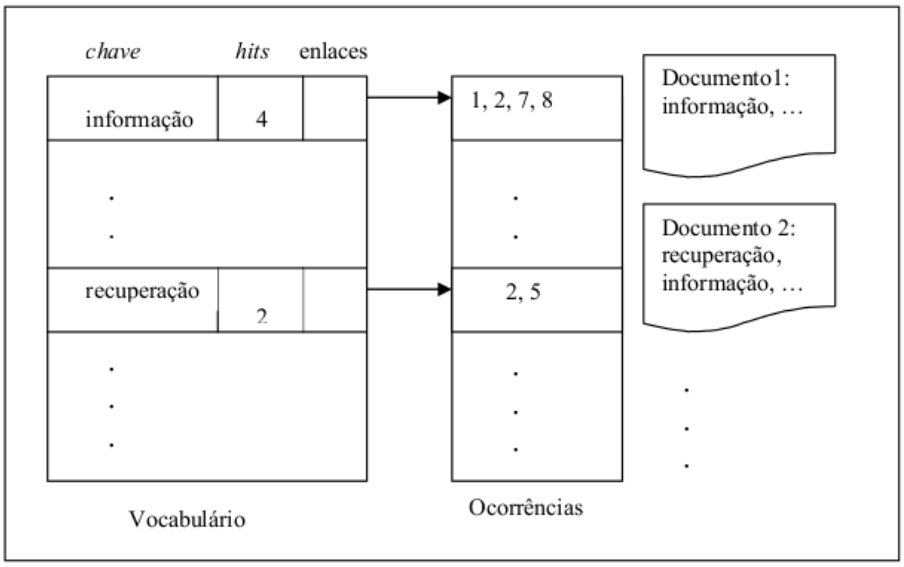

Figura 3.3: Indexação mediante ficheiro invertido

Normalmente, impõem-se uma série de restricções na hora de construir os índices, que consequentemente afectarão directamente às buscas realizadas a posteriori [80]. Alguns exemplos de restricções são os seguintes:

- Em alguns casos, pode ser útil utilizar um vocabulário controlado, que constituirá o vocabulário que será indexado. Aquelas palavras do texto não pertencentes ao vocabulário não serão indexadas, e portanto, posteriormente não poderão ser buscadas.
- Normalmente se dispõe de uma lista de palavras comuns ou stpowords (artigos, preposições, etc.) que por razões de volume de informação não se incluem no

índice, e portanto não serão localizadas na busca.

- Existe um conjunto de regras que se encarregam de definir o início e o fim de uma palavra para ser indexada. Estas regras se encarregam do tratamento de espaços em branco, signos de pontuação, etc. e podem ter um impacto significativo nos termos indexados.

É importante destacar que estas restricções encarregadas de determinar o que será indexado são críticas para a efectividade das buscas, e portanto são um parámetro chave no momento da construção do índice.

O espaço utilizado pelo vocabulário se pode considerar reduzido. Segundo a lei de Heaps o vocabulário cresce segundo $O(kn^{\beta})$, onde β é uma constante entre 0 e 1 que depende do texto e varia entre 0.4 e 0.6 e k é um factor que também depende do texto [84]. Por exemplo, considerando 1 Gigabytes de texto o vocabulário aproximadamente será de só 5 megabytes. Por outro lado, este espaço pode ser reduzido considerando a técnica de Stemming ou outras técnicas de normalização de texto.

Em câmbio, as ocorrências requerem um espaço muito maior. Em geral, o espaço extra é de O(n), sendo n igual ao tamaño do texto já que cada palavra que aparece nos documentos deve ser referenciada uma vez. É importante ter em conta que as palavras comuns, que são as palavras mais frequentes no texto ou as stopwords não serão armazenadas no índice, e ainda nesses casos o tamanho do índice pode ser importante. Na literatura se oferecem diversas cifras ao respeito e ainda que a nível geral seja muito difícil concretar em único valor, a tendência actual é à redução do tamanho de índices abaixo de 100% do texto original [34].

O processo de busca num ficheiro invertido requer três passos:

- Busca no vocabulário: cada uma das palavras ou padrões presente na consulta é isolada e buscada no vocabulário. É importante destacar que as buscas de frases ou por proximidade se dividem em palavras individuais.

- Recuperação das ocorrências: para cada palavra do vocabulário se recupera sua lista de ocorrências associada.

- Manipulação das ocorrências: as ocorrências são processadas para resolver operações de lógica booleana, frases ou proximidade.

As consultas formadas por uma única palavra podem ser buscadas utilizando diversas estruturas de dados com o objectivo de melhorar o rendimento: árvore B, arrays ordenados. Nas seguintes secções se descreverão com detalhes o modo de operação e o rendimento oferecido por algumas técnicas. No entanto, a nível geral, simplesmente armazenado por ordem alfabético as palavras do vocabulário se conseguem um rendimento muito competitivo, já que uma palavra pode ser localizada no vocabulário realizando uma busca binária com um custo de $O(\log n)$, sendo n o número de palavras do vocabulário.

Se a consulta está formada por uma única palavra, o processo finaliza com o envio da lista de ocorrências. Pontualizar que no caso de que o padrão coincida com múltiplas palavras se faz necessária à união de múltiplas listas de ocorrências.

As consultas que envolvam a várias palavras são ligeiramente mais complexas de resolver em caso dos ficheiros invertidos. Cada elemento ou palavra deve ser buscado de maneira separada, devem-se recuperar cada uma das listas e realizar as operações correspondentes em caso que tenham utilizado operadores lógicos na busca. Isto implica a realização de operações de união, intersecção ou resta de conjuntos.

As consultas de proximidade ou de frases incorporam um maior nível de complexidade, já que para cada palavra se deve obter uma lista ordenada de documentos. As listas devem ser combinadas, para o qual se percorrem tentando localizar documentos onde as palavras apareçam em sequência (no caso de uma frase) ou o suficientemente cerca (para o caso das buscas de proximidade). No caso de que alguma das listas seja o suficientemente reduzida se pode realizar uma busca binária no resto apartir dos valores da menor das listas. Além disto, para que o índice possa suportar este tipo de busca é necessário armazenar, para cada documento, as posições e onde aparece cada uma das ocorrências do termo, o que aumenta sensivelmente o tamanho do índice.

Ao nível geral as principais vantagens da técnica de ficheiro invertido se centram na velocidade de localização de palavras. Por outra parte as principais desvantagens se centram na sobrecarga de armazenamento necessário para o índice, que ainda que anteriormente pudesse sofrer penalização grave (entre 50% e 300% segundo Haskin em [82]), na actualidade graças às técnicas de compressão a sobrecarga se reduziu a valores menores de 100%. Por outra parte, dependendo da técnica utilizada o custo de actualização e reorganização do índice pode ser demasiado elevado, pelo que a sua utilização não é aconselhável em meios altamente dinámicos. E finalmente, um

dos problemas mais relevantes se apresenta pela necessidade de realizar operações sobre as listas de ocorrências obtidas, que podem ser especialmente prejudiciais se o número de elementos e/ou de listas é muito elevado. A continuação se descreve várias técnicas de organização do vocabulário de índice invertido.

Arrays ordenados

Um ficheiro invertido implementado usando uma estrutura de um array ordenado armazena a lista de palavras chave (o vocabulário) num array ordenado, incluindo o número de documentos associados a essa palavra chave e um enlace a lista de documentos (ver Figura 3.3). Para localizar um termo no array se realiza uma busca binária, ainda que no caso de dispositivos de armazenamento secundários devem ser adaptados a estrutura de armazenamento e o algoritmo de busca [80].

A utilização desta estrutura para o armazenamento do índice apresenta problemas na hora de gerir as actualizações do índice (principalmente, inserções, de novos termos) já que o custo é elevado. Por outra parte, as principais vantagens se centram na facilidade de implementação e o alto rendimento oferecido.

A construção de um ficheiro invertido utilizando uma estrutura de um array ordenado se divide em vários passos. Inicialmente, o texto é analisado lexicamente para a obtenção das diferentes palavras junto com os documentos onde podem ser localizadas. Esta primeira parte do processo constitui a principal carga de tempo e armazenamento durante a elaboração do índice. Em segundo lugar a lista de termos deve ser ordenada alfabeticamente, criando a lista de documentos onde se localiza cada termo [80]. E opcionalmente, pode-se realizar um processamento do ficheiro invertido para calcular os pesos dos termos, ou realizar reorganizações, compressões do arquivo.

A primeira parte do processo, consiste na obtenção da lista inicial de palavras, descompõe-se em diferentes operações que vão desde a eliminação das palavras comuns até a obtenção da raiz da palavra (processo denominado comummente “Stemming”).

A segunda parte do processo consiste na inversão da lista de termos. Esta operação se costuma realizar com uma ordenação dos termos mantendo os elementos duplicados. O principal problema que apresenta esta técnica é que impulsionou o desenvolvimento de outras variantes descritas nas seguintes secções.

Normalmente, na técnica de ficheiro invertido se armazenam os dois componentes

(vocabulário e ocorrências) de maneira separada. Como se comentou anteriormente o vocabulário possui umas dimensões mediamente reduzidas frente ao grande tamanho que pode chegar a ocupar a lista de ocorrências. O processo de busca, em consequência, centra-se na localização dos termos utilizando uma busca binária no vocabulário para depois aceder directamente à lista de ocorrências associada.

Sobre a técnica básica de indexação usando arrays ordenados surgiram diversas melhorias. Entre elas cabe destacar a proposta por Harman et.al. em [80], onde se apresenta uma técnica que melhora a construção dos ficheiros invertidos para grandes conjuntos de dados. Basicamente, nesta técnica se elimina o passo intermédio de ordenação da lista utilizando um tipo essencial de árvores binárias. Por outra parte, em [58] desenharam FAST-INV, uma técnica para geração de um ficheiro invertido baseada na grande quantidade de memória disponível hoje em dia nos computadores e a ordem inerente dos dados de entrada [58].

Árvores B

Outra estrutura de implementação utilizada para a técnica de ficheiros invertidos se baseia nas árvores B. As árvores B são um caso específico das árvores de busca, cujo componente mais conhecido são as árvores binárias. Numa árvore binária cada nó interno contém uma chave de tal forma que a subárvore esquerda contém chaves menores do que a do pai e a subárvore direita contém chaves maiores. Neste tipo de árvores se adapta de maneira adequada para ser utilizadas directamente em memória principal.

Em câmbio, quando se precisa aceder a memória secundária, as árvores de busca m-arianos apresentam um melhor rendimento já que os nós internos são maiores. Em concreto, as árvores B são tipo concreto destas árvores. Uma árvore B de ordem m se define como [15] (cap: 2):

- A raiz e todos os nós internos da árvore têm entre m e 2m chaves.
- Se k_i é a chave da posição e i–ésima , então todas as chaves do filho (i -1)-ésimo são menores, enquanto todas as chaves do i-ésimo filhos são maiores.
- Todos os nós folhas têm a mesma profundidade.

Em [44] se apresenta as árvores B como uma estrutura eficiente para os ficheiros invertidos em situações de dados dinámicos, onde o número de actualizações é

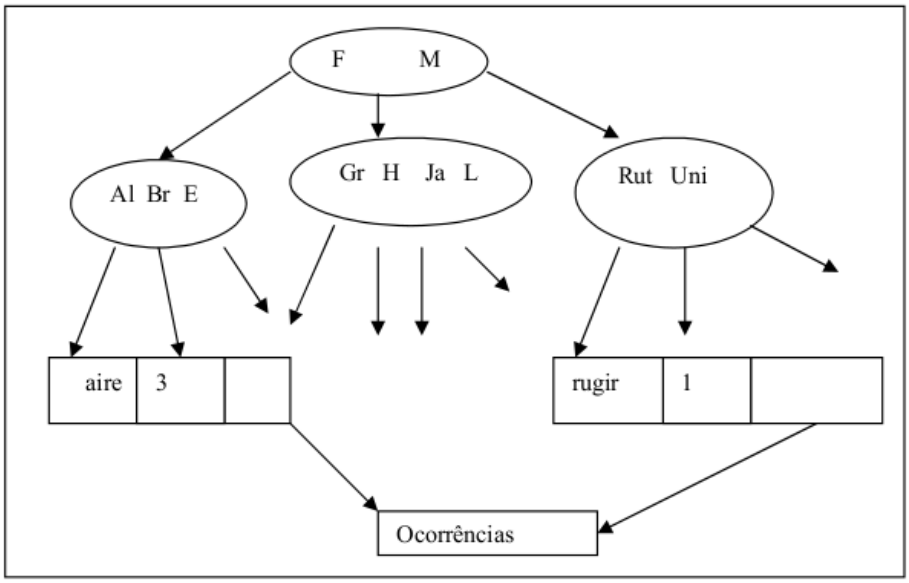

Figura 3.4: Exemplo de árvore B prefixo

preponderante sobre as buscas.

No entanto, são Bayer e Unterauer ([16]) quem propõem a actualização de um tipo especial de árvores B, as árvores B prefixos, que utilizam prefixos das palavras como chaves primárias na árvore. Este tipo de árvore demonstra ser especialmente útil no caso de armazenamento de índices de textos, como é o caso dos ficheiros invertidos.

Cada chave é a palavra mais curta que permite distinguir as chaves armazenadas no nível seguinte, e a chave não tem porque ser, necessariamente, um prefixo de um termo real do índice. Os nós folhas armazenam as palavras chave juntos com os dados associados aos termos de vocabulário, tipicamente a lista de ocorrências, Figura 3.4.

A principal vantagem da utilização da árvore B se centra na facilidade para realizar inserções no índice frente ao caso dos arrays ordenados para inserir um novo registo simplesmente se localiza o ponto de inserção, se não há suficiente espaço nesse nó, divide-se e se promociona uma chave ao nível anterior. Este processo se repete recursivamente até chegar ao nó raiz, que em caso de não dispor mais espaço aumenta a profundidade da árvore [15] (cap: 2).

Assim mesmo a busca é normalmente mais rápida do que o caso de um array ordenado, já que o número de comparações necessárias tanto faz à profundidade da

árvore. Por outra parte, as principais desvantagens se centram na quantidade de armazenamento empregado (devido à presença de nós intermédios, que podem ser de grande tamanho) e a complexidade de implementação.

3.4 Compressão dos ficheiros invertidos

As técnicas de compressão aplicadas a ficheiros invertidos operam principalmente sobre a lista de ocorrências ou lista invertida, se bem, também se podem aplicar técnicas de compressão sobre o conjunto de palavras que conformam o vocabulário. A compressão das palavras chave incluídas no vocabulário não permite obter umas reduções significativas no espaço de armazenamento requerido, ao estar optimizadas para a compressão de documentos completos. Simplesmente mencionar a técnica dos códigos de Huffman canónicos [92], onde cada palavra é substituída por um código deste tipo, cuja longitude depende da frequência da palavra.

A compressão nas listas invertidas se consegue mediante a numeração dos documentos sequêncialmente desde 1, ordenando as entradas da lista invertida, representando a sequência de identificadores como uma sequência de saltos e empregando técnicas de representação compacta de inteiros pequenos [156]. Por exemplo, a seguinte lista invertida:

$$1, 4, 17, 91, 113 \ldots,$$

Poderia ser representada como a seguinte sequência de diferenças (que pode ser comprimida):

$$1, 3, 13, 74, 22 \ldots,$$

As técnicas de compressão aplicadas às sequências deste tipo se dividem em duas classes [156]. Os métodos globais que usam o mesmo tipo de codificação para todas as entradas, e portanto apresentam a vantagem de ser gerais, mas insensíveis à frequência de cada termo. Os métodos locais se caracterizam por calcular cada código tendo em conta um ou mais parámetros da distribuição dos valores (normalmente, a frequência de cada termo).

Um dos métodos globais mais simples é o código *gamma* proposto em [53], onde

cada inteiro x é representado mediante a codificação de $1 + \lfloor \log_2 x \rfloor$ em unário, seguido por un código $\lfloor \log_2 x \rfloor$ bits que representa o valor de $x - 2^{\lfloor \log_2 x \rfloor}$ em binário. Esta codificação gera códigos de longitude variável, mas onde cada conjunto de bits pode ser univocamente decodificado num inteiro positivo. Na Tabela 3.1 se mostram alguns valores de códigos gamma.

A utilização de códigos de longitude variável permite representar de maneira mais sucinta os valores menores (e mais frequentes) do que os valores maiores (e menos frequentes), o que supõe uma melhoria sobre a codificação binária padrão.

Por outra parte, os códigos locais obtém uma vantagem adicional ao considerar a variabilidade das frequências de cada palavra. Por exemplo, a palavra "el" terá associada uma sequência de diferenças formada por valores pequenos, comummente 1. Em câmbio, palavras mais raras terão listas invertidas formadas por diferenças muito maiores. Portanto, os métodos locais, que adaptam sua codificação em função da frequência de cada palavra, podem obter umas taxas de compressão maiores, ainda sem ser gerais.

As técnicas de compressão de listas invertidas aplicadas a grandes volumes de documentos (como é o caso dos sistemas de busca) permitem realizar compressões de índice por um factor de seis [17]. Esta aproximação apresenta a desvantagem de que as listas invertidas devem ser decodificadas ao ser recuperadas, mas esta decompressão é rápida. Por outro lado, existem técnicas de compressão que inserindo uma pequena quantidade de informação adicional de indexação em cada lista permitem evitar uma grande parte do processo de decodificação [156].

Uma característica importante das listas invertidas comprimidas é que as melhores percentagens de compressão se conseguem nas listas de maior tamanho, isto é, os termos mais frequentes. Portanto, não é necessário eliminar as palavras comuns do índice no momento da criação do ficheiro invertido, senão que a decisão da utilização ou não das palavras comuns fica relegada ao momento da consulta.

Em consequência, a compressão das listas invertidas permite obter uma melhora no espaço de armazenamento utilizado pelo índice, e algo que possivelmente possa ser considerado de maior relevância, é a possibilidade de inserir as palavras comuns no índice sem aumentar drasticamente o tamanho final da estrutura.

Além disto outra forma de reduzir o tamanho total da colecção é comprimir o texto com técnicas estatísticas semi-estáticas tales como tagged Huffman [162] ou Dense Codes [33] que reduz o tamanho do texto aproximadamente 35% do tamanho

x	Gamma(x)
1	0
2	100
3	101
4	11000
7	11011
15	1110111
63	11111011111

Tabela 3.1: Codificação gamma para listas invertidas

original, à vez que permitem manter a colecção comprimida todo o tempo. Claro que sim, é preciso descomprimir à hora de mostrar os documentos aos usuários finais. Este facto junto com a grande capacidade de memória dos sistemas actuais tem favorecido a proliferação de índices invertidos que se mantém completamente em memoria [216].

3.5 Recuperação de Informação na Web

World Wide Web se originou ao fim dos anos 80 [19] e nenhum dos seus criadores pude imaginar o impacto que teria. A quantidade de informação textual disponível estava estimada na ordem de 6 terabytes em meados de 1999, através de mais de 800 milhões de páginas distribuídas em mais de 3 milhões de servidores Web [6]. Além disto, há de ter em conta de que outros meios, como imagens, áudio, vídeo, estão também disponíveis, o que converte a World Wide Web numa grande e ubíqua bases de dados sem estruturas, sobre o que os usuários desejam localizar informação, que são necessárias ferramentas que permitam gerir, recuperar e filtrarão a documentação disponível nessa grande base de dados.

De facto, ninguém séria capaz de localizar determinada informação simplesmente explorando a amplitude da Web, num tempo razoável. Pelo contrário são necessárias ferramentas que ajudem aos usuários a localizar aquelas páginas Web mais relevantes para as suas necessidades de informação. Em concreto, poucos anos depois do início e extensão da World Wide Web surgiram, os primeiros sistemas de recuperação de informação na WWW.

Basicamente se define em três formas diferentes de sistemas de busca na Web. As duas primeiras se caracterizam por dispor das suas próprias estruturas de dados sobre as que realizam o processo de busca, enquanto o terceiro tipo carece de

ditas estruturas. O primeiro tipo são os denominados robôs (*robots*) ou motores de busca que tentam indexar a totalidade da Web como uma base de dados de texto. O segundo se corresponde com os directórios Web, que classificam os documentos Web numa ontologia. O terceiro tipo está representado pelos metabuscadores ou multibuscadores, que acedem a outros buscadores para a obtenção dos resultados de busca, realizando uma combinação do total dos resultados obtidos.

Na seguinte secção se descrevem com detalhe as características que convertem em únicos aos sistemas de recuperação de informação na Web, frente aos sistemas de recuperação de informação tradicional.

3.5.1 Características

Os sistemas de recuperação de informação na Web, apresentam umas características completamente diferentes aos sistemas de recuperação de informação tradicionais, principalmente devido às próprias características do seu meio de trabalho. A World Wide Web, assim mesmo, os próprios usuários destes sistemas de recuperação apresentam umas características totalmente diferentes aos seus homólogos de sistemas tradicionais.

Em concreto, as principais características que convertem aos sistemas de recuperação da informação na Web em únicos são as seguintes [98] [15] (cap: 13):

- **Volume de dados**. Por uma parte, o volume da informação disponível hoje em dia em Internet faz com que estes sistemas de informação devem tratar quantidades enormes de dados que se escapam do conteúdo tratado pelos sistemas tradicionais. Por outra parte, o crescimento exponencial de World Wide Web apresenta para os sistemas de recuperação uns aspectos complexos de resolver. O número de servidores Web está estimado em 2,4 milhões segundo NetSizer [163], o número de páginas Web disponíveis estava estimado em 350 milhões de páginas em Julho de 1998 [25], enquanto estimações mais recentes cifram o número de páginas em 800 milhões de páginas, com 6 terabytes de texto associados em meados de 1999 [6], e num valor próximo aos 1700 milhões hoje em dia.

- **Dinamismo**. O conteúdo da World Wide Web está mudando dia à dia, com aparições de novas páginas, modificações de páginas existentes e eliminações de páginas obsoletas. De facto, estima-se que 40% das páginas Web mudam

cada mês [110]. Os sistemas tradicionais estão desenhados para o tratamento de bases de dados textuais com uma grande componente estática.

- **Heterogeneidade**. Na Web se encontra disponível grande variedade de tipos de documentos: páginas Web, documentos textuais puros, imagens, ficheiros de áudio, vídeos, etc. Por outra parte, existem documentos escritos em multidões de linguagens, de facto, estima-se que em Internet se empregam mais de 100 linguagens diferentes.

- **Distribuição**. Pelas características intrínsecas da World Wide Web, os documentos se encontram dispersos sobre milhões de computadores e plataformas. Estes computadores estão interconectados sem nenhuma topologia predefinida, no largo de banda e a fiabilidade de cada rede de interconexão sofre grandes variações. A Web está baseada nos hiperenlaces e isto é o que constitui e conforma a rede de documentos. De facto, em média, cada página Web apresenta mais de 8 enlaces a outras páginas Web.

- **Redundância e falta de estrutura**. A maioria dos sistemas tradicionais de recuperação de informação parte de uma base de dados ou ao menos documentos estruturados. Pelo contrário, na Web cada documento HTML foi elaborado de maneira independente, pelo que não existe uma estrutura de definição de documentos. Além disto, é necessário ter em conta aos dados duplicados, já que se estima que aproximadamente 30% das paginas Web são duplicadas [43], e isto sem ter em conta aos detalhes de redundância semântica.

- **Qualidade**. A Web se considera como um novo meio de publicação, no entanto, não existe nenhum processo editorial prévio, o que provoca que o conteúdo informativo nem sempre seja de uma qualidade adequada. Em Consequência, os dados disponíveis na Web, podem ser falsos, inválidos, obsoletos, com erros de escritura (léxicos, gramaticais, etc.). Estima-se que os erros tipográficos para palavras comuns se produzem em 1 de cada 200, e no caso de palavras mais complexas (por exemplo, sobrenomes estrangeiros) num de cada 3 [160].

- **Usuários**. Os usuários de um sistema de recuperação de informação na Web apresentam um comportamento totalmente diferente aos usuários tradicionais em muitos aspectos. A nível geral, os usuários procuram uma grande variedade de conceitos sem nenhum tipo de nexo comum, as consultas se realizam de forma vaga e difusa, especificando muito poucas palavras chave e comprovando muito poucos resultados.

Estes aspectos constituem as principais diferenças entre os sistemas de recuperação na Web e os tradicionais, o que também marcará certas diferenças nas estruturas de dados empregadas para a busca. De facto, inicialmente os sistemas de busca na Web, basearam-se na tecnologia desenvolvida para o meio de recuperação de informação em meios estáticos e de pouco volume de dados. No entanto, cada dia se faz mais patente a necessidade de aplicar novas técnicas de indexação e representação dos dados para melhorar o rendimento deste tipo de sistemas.

3.5.2 Tipos de motores de busca

No meio dos sistemas de recuperação de informação para a Web se consideram tipicamente três tipos básicos: robôs ou motores de busca, directórios e *metabuscadores* [15] (cap: 13).

Cada um destes sistemas apresenta umas características particulares e permite resolver um tipo determinado de consultas. A nível geral, os robôs geram um grande volume de informação, apesar que a qualidade do seu conteúdo nem sempre está garantida, enquanto os directórios Web realizam uma classificação de páginas Web pelo que a qualidade do seu conteúdo é muito elevada, enquanto a quantidade é muito reduzida. O caso dos metabuscadores é atípico e especial, já que não dispõem de nenhuma estrutura de dados própria, senão que acedem a outros sistemas de busca para realizar as consultas dos usuários.

O principal problema que apresenta este tipo de sistemas em conjunto é o ocultismo que os rodeia sobre as técnicas e tecnologias empregadas para o seu adequado funcionamento. Ainda que existam casos pontuais de buscadores que publicam trabalhos relacionados com sistemas comerciais a fim de evitar a cópia ou imitação, por parte de competidores, dos mecanismos empregados. Apesar disto, nas seguintes secções, se descrevem as estruturas de dados empregados por estes tipos de sistemas, baseando-se em diferentes investigações publicadas ao respecto.

Robôs

Estes sistemas se caracterizam por modelar a Web como uma base de dados de texto. A principal diferença com os sistemas de recuperação de informação tradicionais consiste em que as consultas se devem realizar sem acesso ao texto, utilizando unicamente os índices disponíveis, já que, de outra maneira, seria demasiado custoso

arquivar uma cópia dos documentos da Web ou demasiado lento o acesso às páginas Web.

Este tipo de sistemas de busca prima à quantidade de informação disponível para os seus usuários. O objectivo ideal consiste na indexação da totalidade da Web, devendo para isso dispor de estruturas de dados adequadas para suportar este volume de informação. Estes sistemas são apropriados para buscas específicas ou concretas, já que em outro caso a quantidade de resultados pode extravasar-se ao usuário e a qualidade das respostas pode não ser sempre a óptima.

A maior parte dos robôs disponíveis na Web utiliza uma arquitectura centralizada, onde os robôs percorrem a Web enviando páginas a um servidor central onde são indexadas. Estes robôs, também denominados crawlers, spiders, wanderes ou walkers, não são mais do que agentes softwares que enviam petições aos servidores Web. A maioria utiliza uma estrutura de ficheiro invertido para a indexação do texto recebido dos robôs, índice que é armazenado de maneira centralizada para responder às consultas de usuários.

Também existem robôs que apresentam uma solução distribuída para a resolução do problema, como o caso de Haverst [23]. Este sistema se baseia numa arquitectura distribuída para reunir e distribuir os dados, ainda que apresente o inconveniente de requer a coordenação de múltiplos servidores Web para sua prática, algo que hoje em dia é impensável.

A principal potência deste tipo de sistemas de busca se centra no volume de informação que examina, o que converte em ideais para buscas específicas. Em câmbio, o seu principal inconveniente está baseado na pouca qualidade de certa informação indexada, o que pode provocar a aparição de documentos de baixa qualidade entre os resultados.

O primeiro passo para qualquer robô de busca consiste na obtenção do primeiro conjunto de URLs e percorrê-las de maneira adequada, aspecto que pode melhorar as páginas indexadas [41]. Além disto, é necessário ter em conta o padrão (*estándar*) de exclusão de robôs [119] que permite a um administrador restringir o acesso a seu lugar Web a este tipo de agentes.

No entanto, o aspecto principal de qualquer sistema de busca se encontra na estrutura de dados empregada, e é o aspecto central desta dissertação. Em concreto, os robôs se caracterizam por empregar variantes da técnica de ficheiro invertido [15] (cap:13). Actualmente, o tamanho de um ficheiro invertido se situa em aproximada-

mente 30% do tamanho de texto, o que implica que para um conjunto de 100 milhões de páginas Web se gera um índice de aproximadamente 150 gigabytes, ainda que este tamanho pudesse ser reduzido utilizando técnicas de compressão [236].

O processo de busca se realiza com base a uma busca binária no vocabulário da estrutura de ficheiro invertido, e em caso que a consulta esteja constituída por múltiplos vocábulos, os resultados individuais devem ser combinados para obter a resposta final. Esta etapa de combinação de resultados pode ser pouco eficiente se os resultados individuais são muito numerosos [15] (cap: 13). Outros aspectos mais específicos das buscas, como as buscas por proximidade, não são implementados por alguns robôs já que a sua inclusão no índice é demasiado custoso a nível de espaço de armazenamento, ainda que sim existam motores de busca que proporcionam esta utilidade, no entanto os detalhes de implementação não foram publicados. Por último, antes de mostrar os resultados se realiza uma ordenação com base ao ajuste de cada documento à consulta, com o objectivo de mostrar ao usuário unicamente aqueles documentos que são mais relevantes.

Estes passos constituem, a nível geral, as tarefas de um motor de busca genérico para a Web. A seguir, e de maneira mais detalhada, com base à informação publicada em [29] e [168] se descreve um dos paradigmas destes sistemas de busca na Web hoje em dia.

Na (Figura 3.7) se mostra a arquitectura de um motor de busca. A descarga das páginas Web se realiza por meio de vários robôs distribuídos, coordenados por meio de um servidor de URLs. As páginas obtidas se enviam ao "*storeserver*" que as comprimirá e armazenará num repositório. Cada página se identifica por meio de um identificador de documento (denominado docID). O indexador se encarrega de ler do repositório um documento, descomprimí-lo e analisá-lo sintácticamente. Cada documento se converte num conjunto de ocorrências de palavras denominado "*hits*". Um hit representa à palavra, sua posição no documento e uma aproximação do tamanho da fonte e outras características. O indexador distribui estes hits em conjunto de barris, criando um primeiro índice parcialmente ordenado. Além disto, o indexador realiza outra função importante, durante a análise sintáctica obtém os enlaces entre páginas e armazena essa informação (enlace mais texto do enlace) num ficheiro de enlaces.

O "*URLresolver*" a partir do ficheiro de enlaces converte as URLs relativas em URLs absolutas e a sua vez em identificadores de documentos, insere o texto do enlace no índice e gera uma base de dados de pares de identificadores de documentos

que representam aos enlaces, o que será utilizado para o cálculo de *PageRank*. O "*sorter*" acede ao conteúdo dos barris e os reordena por identificador de palavra para gerar o índice invertido, a partir de qual se criará o novo vocabulário.

O buscador utiliza directamente o vocabulário, o índice invertido e o *PageRank* para responder às consultas dos usuários.

O sistema de busca empregado por Google se baseia em diversas estruturas de dados, algumas das quais têm um papel mais preponderante nas tarefas da indexação off-line, enquanto outras se empregam activamente na busca. As estruturas de dados definidas na arquitetura de alto nível são as seguintes:

- **O repositório**: contém o texto HTML completo de todas as páginas Web. Cada página se armazena comprimida, arquivando para além disto, o identificador de documento, a longitude e sua URL.
- **O índice de documentos**: armazena informação sobre os documentos, incluindo o estado do documento, um ponteiro ao repositório, um checksum e várias estatísticas.
- **Vocabulário**: contém a totalidade das palavras indexadas. Mantém-se em memória principal para agilizar as buscas.
- **Listas de hits**: corresponde-se com uma lista de ocorrências de uma palavra concreta num documento concreto, incluindo informação sobre posição, fonte e características do texto (maiúsculas, minúsculas, etc.). Esta lista se armazena comprimida utilizando um algoritmo de compressão especialmente desenhado para este caso.
- **Índice invertido**: é a estrutura de dados empregada directamente pelo buscador.

A estrutura mais importante para uma busca eficiente é, obviamente, o índice invertido, e um aspecto fundamental é a ordem no que se dispõem os identificadores de documentos na lista. Basicamente se definem duas alternativas: realizar um armazenamento ordenado por identificador, com o qual as operações de combinação prontas se vêem aceleradas. Ou bem, armazenar os identificadores ordenados segundo um critério de importância dos documentos, o qual converte em trivial as buscas por uma única palavra (e no caso de múltiplas palavras provavelmente os

resultados se encontrem entre os primeiros documentos), e penaliza sensivelmente as operações com vários vocábulos.

Por último simplesmente temos a destacar o algoritmo de ordenação dos resultados, denominado PageRank, e exposto em [168] e [41]. Sem entrar em detalhes, comentar que é um algoritmo baseado em votações entre páginas Web, um enlace desde a página A à B se interpreta como um voto de A à B . Aquelas páginas com mais votos se consideram melhores, o qual a sua vez lhe dá maior importância aos seus próprios votos. A ordenação dos resultados depois da busca se realiza com base aos termos buscados e à importância de cada página.

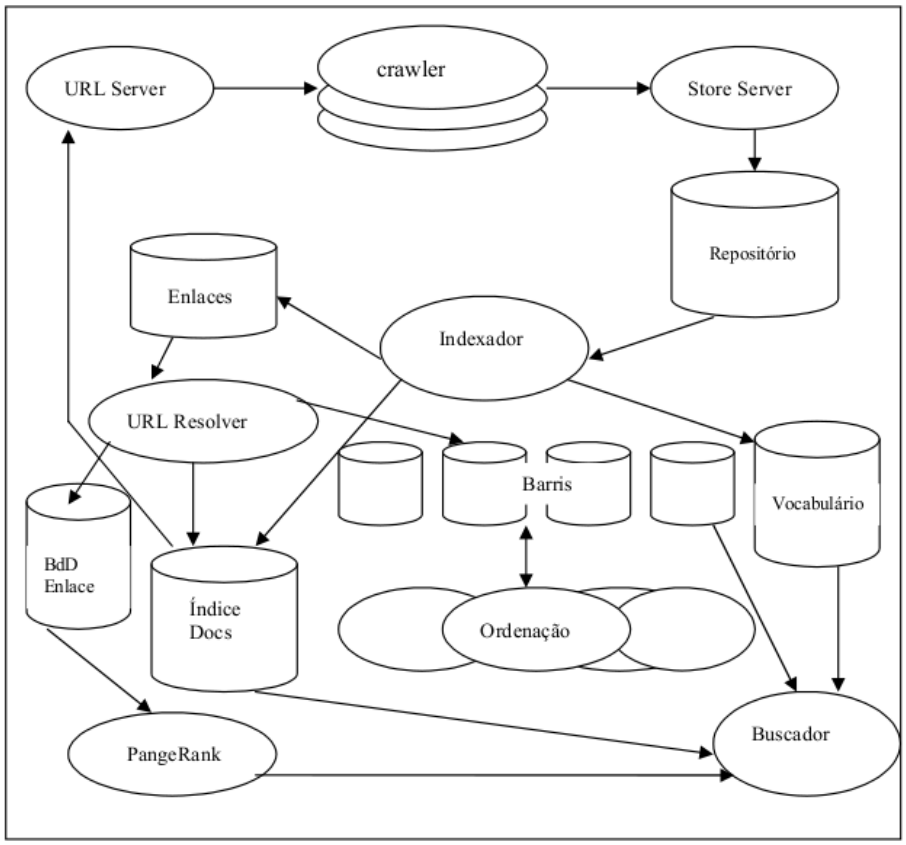

Figura 3.5: Arquitectura de alto nível de um motor de busca

Directórios Web

Este tipo de sistemas de recuperação na Web se caracteriza por combinar a busca com a navegação. Os directórios constituem as ontologias da Web [98], ao propor-

cionar uma classificação de páginas Web baseada numa hierarquia de categorias. A principal característica destes sistemas é a pequena percentagem da totalidade de documentos disponíveis na rede gerido por estes sistemas (se consideram que indexam menos de um 1% de todas as páginas), ainda que se garanta a qualidade dos documentos obtidos como proposta a uma consulta.

Se pode definir um directório como uma taxonomia hierárquica que classifica o conhecimento humano [15](cap:13), portanto, um directório Web é uma taxonomía hierárquica que classifica a informação disponível na World Wide Web. Normalmente um directório Web está constituído por um gráfico dirigido cíclico de categorias às que se associam documentos Web. A construção do gráfico costuma ser bastante flexível, de tal forma que um nó possuirá um número variável de nós filhos e nós pais. Assim mesmo, os documentos podem estar associados com qualquer categoria, sem estar restringidos unicamente aquelas categorias folhas.

O exemplo por excelência de directório se considera Yahoo! [239], o qual ainda que não é o primeiro directório em aparecer na Internet (este posto fica reservado para Galaxy [67], é considerado o directório mais completo ao constar já desde 1.999, arredor de 150.000 categorias e aproximadamente um milhão de páginas Web classificadas.

A principal vantagem destes sistemas se centra na qualidade dos resultados, já que ante uma consulta, na maioria dos casos, os resultados serão muito relevantes. Conjuntamente, uma potente utilidade destes sistemas de recuperação de informação se centra na possibilidade de realizar buscas limitadas a uma zona do gráfico de categorias, o que limita amplamente a temática dos resultados obtidos. De facto, isto constitui uma das características diferenciadoras destes sistemas onde se combina a navegação com a busca.

Pelo contrário, o seu principal problema se centra na capacidade de categorização junto com o crescimento acelerado da World Wide Web. Os esforços com base à classificação automática por meio da utilização de técnicas de clustering e outras, foram investigados desde à vários anos e continua sendo investigados. No entanto, pelo momento o processamento da linguagem natural não é 100% efectivo na extracção dos termos relevantes de um documento [15] (cap:13). Portanto, na maioria dos casos a classificação se realiza manualmente por um número limitado de pessoas. Por outro lado, ainda que não seja perfeita, comparada com a catalogação automática, a catalogação manual é a mais precisa já que os expertos se encarregam de organizar os directórios e índices de tal maneira que se facilite o processo de busca

[117].

Este tipo de sistemas se costuma dividir naqueles baseados numa categorização automática (ou ao menos semi-automática) ou manual. Dentro do primeiro grupo, merece especial menção o projeto Taper [7] [8] [103] onde se pretende que a resposta a uma consulta seja uma lista de temas, frente à tradicional lista de documentos. Para isso, baseiam-se numa análise estatística dos documentos, para,com base ao mesmo, classificar os documentos em algum lugar da hierarquia contribuindo também informação dos enlaces existentes. No entanto, as taxas de erro que devem suportar este tipo de sistema são bastante elevadas, em concreto, em [103] a melhor taxa obtida é de 21%.

Frente aos sistemas anteriores, os sistemas com classificação manual garantem umas taxas de erros mínimas. O máximo expoente (como se comentou anteriormente) continua sendo Yahoo!, que conta [67] com um conjunto de pessoas encarregadas de verificar e analisar documentos Web agrupados por áreas temáticas representativas de uma porção da hierarquia. No entanto, existem outros projectos mais interessantes a este respecto como o ODP ou OpenGrid. O projecto ODP (Open Directory Proyect, [167] se baseia na premissa de que nenhuma companhia será capaz de dispor do número de empregados suficientes como para categorizar a totalidade da Web. Em consequência, neste projecto se permite os milhares de voluntários familiarizados com um determinado tema classificarem e gerir uma parte do directório. Por outra parte, o projecto OpenGrid [133] pretende empregar a opinião de milhares de navegantes (não voluntários como no caso anterior) para a ordenação dos documentos na Web. Para isso se baseia numa ligeira modificação ao padrão HTML que permite incorporar informação adicional aos enlaces. No entanto, este sistema continua sendo uma proposta sem nenhum modelo implementado, frente ao caso de ODP com um sistema funcionando e utilizado por grandes companhias.

No entanto, ambos os tipos de sistemas (automáticos e manuais) requerem uma verdadeira estrutura de dados adequada para a informação que devem gerir, especialmente na hora de realizar buscas. Não obstante, a informação disponível na literatura sobre estes aspectos se pode considerar bastante reduzida, se bem, como o resto de sistemas de recuperação de informação actuais, a base sobre a que se assentam estes sistemas é uma estrutura de ficheiro invertido.

Um directório Web está constituído por três componentes básicos que representam à informação armazenada no mesmo. Por uma parte o vocabulário representa os vocábulos indexados (tanto nos documentos como nas categorias do directório),

existe uma estrutura que representa a hierarquia de categorias existente no directório e se mantém uma base de dados de documentos com a informação básica sobre cada um (URL, título e descrição).

E outra parte também é necessário definir as estruturas que relacionam estes três componentes. Em primeiro lugar se requer um índice para a relação entre as palavras e os documentos, igualmente para a relação entre palavras e categorias, e uma terceira estrutura que associa cada documento com as categorias às que pertence (ou o que é o mesmo, cada categoria com os documentos que contém). A seguir se listam as principais estruturas destes sistemas:

- **Vocabulário**: consiste na estrutura que armazenará todas as diferentes palavras às que fazem referência aos documentos e/ou as categorias, constituindo portanto, um bloco comum para várias estruturas de listas invertidas.
- **Base de dados de documentos**: é a estrutura que contém a informação básica sobre cada documento. Normalmente está constituída pela URL da página Web, o título, uma breve descrição e opcionalmente, alguma informação estatística adicional.
- **Índice invertido documentos-palavras**: é a estrutura que armazena os documentos que se encontram associados com cada palavra. Será o núcleo para a implementação da busca e a implementação dependerá de cada directório Web, ainda que sempre sobre a base de uma estrutura de lista invertida.
- **Índice invertido categorias-palavras**: de forma comum as próprias categorias dispõem de certas palavras chave associadas que constitui uma representação da temática associada à dita categoria. Normalmente este índice se armazenará de maneira separada, ainda que fosse possível a sua inclusão conjunta com os documentos, ainda que o rendimento oferecido pudesse ver-se drasticamente afectado.
- **Hierarquia de categorias**: é fundamental para um eficiente funcionamento do sistema uma estrutura de dados que represente o gráfico dirigido acíclico existente entre as diferentes categorias. Esta estrutura de dados será acedida continuamente durante a navegação do usuário através da ontologia, pelo que primarão os acessos aos nós filho de uma categoria, e em geral a todos os seus descendentes directos ou indirectos. Tipicamente, a estrutura adequada para tais efeitos se costuma ajustar a alguma variante de uma árvore modificada para dar suporte aos múltiplos pais que podem possuir as categorias.

- **Índice invertido categorias-documentos**: nesta estrutura se armazenará os documentos associados a cada uma das categorias. Desta maneira, se constituirá uma lista de documentos para cada uma das categorias da hierarquia. Considera-se mais eficiente o armazenamento numa estrutura separada para a sua posterior utilização tanto na operação da navegação como busca do usuário. Considerando unicamente a navegação através das categorias, seria mais adequado a incorporação à estrutura hierárquica a informação dos documentos associados a uma categoria. Não obstante, como as buscas constituem outra operação básica e predominante nestes sistemas, considera-se mais conveniente o seu tratamento numa estrutura separada.

Sobre estas estruturas de dados vários processos realizam operações tanto de consulta como de actualização. Por uma parte, o processo de categorização (com a indexação que implica estar associada inerentemente) centra as suas tarefas na inserção de novos documentos dentro da ontologia do directório. Este processo, comummente num diretório Web se realiza de maneira manual, ainda que em caso de realizar-se automaticamente não supõe nenhuma modificação sobre as estruturas de dados empregadas. Além disto, este processo não é visível em nenhum momento pelo usuário.

O processo de categorização realiza as incorporações de novos documentos ao directório, para o qual procede a sua incorporação à base de dados de documentos, ao mesmo tempo em que lhe associa uma ou várias categorias da hierarquia na que se situará dito documento. Uma vez que o documento foi inserido, inicia-se um processo de indexação que processa as palavras chave associadas ao documento (por meio do título, descrição, etc.). A modificação de um documento simplesmente consiste na mudança de algum dos parámetros de sua inserção, enquanto o apagado consiste na eliminação do documento da base de dados e todas as relações associadas; ainda que de forma comum estes processos sejam relativamente pouco frequentes.

No entanto, os processos directamente acessíveis para os usuários são os de navegação e busca, e portanto, aqueles onde o tempo de resposta é fundamental. O primeiro deles, a navegação, é característico destes sistemas e se centra num percurso da hierarquia através das suas diferentes categorias, mostrando em cada caso as diferentes categorias inferiores e os documentos associados. Basicamente, este processo requer o acesso à estrutura de dados associada com a hierarquia para a obtenção dos seus nós inferiores, e também a utilização da relação que associa documentos e categorias para a localização dos documentos associados com a categoria actual (em caso de existir documentos associados).

O processo de busca é o mais complexo de todos e o que requer um rendimento mais optimizado. A busca comum que unicamente recupera os documentos relevantes para uma série de palavras chave, requer o acesso ao índice invertido de documentos e palavras chave através do vocabulário, do mesmo modo como se realiza no caso dos robôs, ainda que com um índice sensivelmente menor. Além disto, estes sistemas costumam comprovar se existe alguma categoria relacionada com a consulta por meio de uma busca no índice de categorias e palavras, processo análogo ao anterior.

No entanto, um dos valores adicionados dos directórios Web o constituem as consultas restringidas a uma zona da hierarquia de categorias, o que a sua vez constitui um modelo de busca mais complexo. Neste caso é necessário, uma vez realizada a busca comum, contrastar os documentos obtidos com os documentos pertencentes a uma zona da ontologia e realizar a intersecção para a obtenção do resultado final. Isto implica, por uma parte um percurso completo através de uma parte da estrutura hierárquica e um acesso à estrutura que associa categorias com os seus documentos.

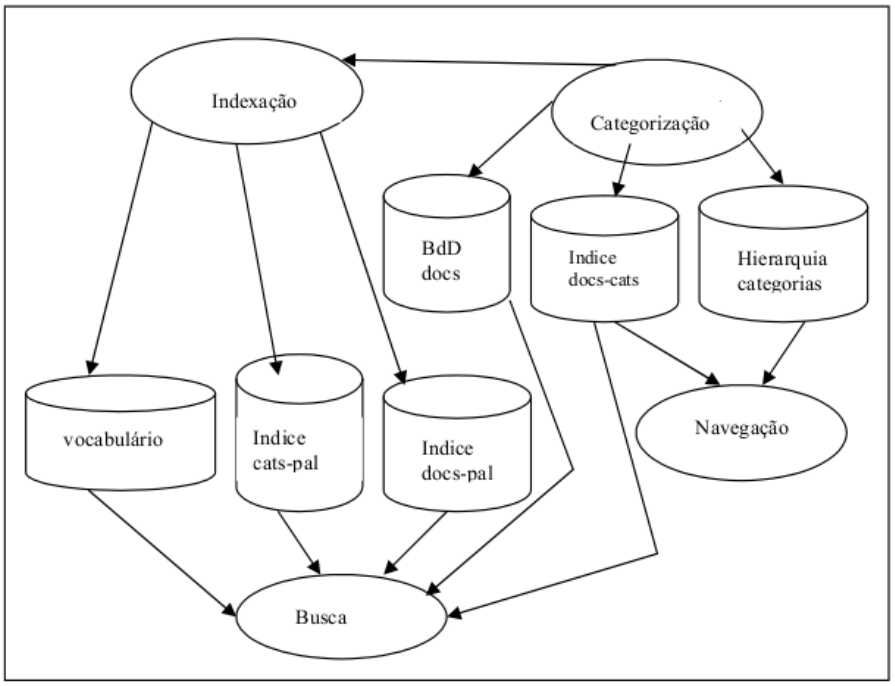

Figura 3.6: Arquitectura de alto nível de um Directório Web típico

Metabuscadores

Este tipo de sistemas de busca se caracteriza, porque enviam uma consulta inserida pelo usuário a múltiplos sistemas de busca disponíveis em Internet (tanto robôs como directórios), recebem as respostas, processam-nas e unificam para mostrá-las ao usuário.

A aparição destes sistemas de buscas que carecem totalmente de estruturas de dados locais, pode-se justificar devido a que os diferentes robôs indexam partes diferentes da Web. De facto, em [25] se estima que unicamente 1% das páginas indexadas por Altavista, HotBot, Excite e Infoseek se encontram em todos eles.

A principal contribuição deste tipo de sistemas de busca se centra na eliminação de resultados errôneos (páginas não existentes ou servidores não acessíveis), junto com uma correcta ordenação dos resultados de múltiplos motores de busca. No entanto, um possível problema consiste nos tempos de resposta proporcionados já que nestas buscas se está adicionando um passo intermédio. Ainda que possivelmente o principal problema resida no desenvolvimento de um serviço público com base ao serviço desenvolvido por outros sistemas.

3.5.3 Estudos dos acessos a um Diretório Web

Desde a aparição dos primeiros sistemas de busca na Web, sua evolução foi imparável tanto no volume da informação tratada como nas diversas melhorias tecnológicas surgidas. Em [106] o conhecimento existente sobre os usuários deste tipo de serviços em Internet é reduzido.

A maioria dos estudos realizados se centram basicamente numa análise das consultas realizadas pelos usuários, com o objectivo de confirmar as diferenças existentes entre um usuário Web e um usuário tradicional de um sistema de recuperação de informação. Por este motivo, na presente análise se estuda o comportamento dos usuários frente a um directório Web, analisando tanto as buscas realizadas pelos usuários, como os documentos conferidos e as categorias visitadas, com o fim de confirmar e contrastar as diferenças existentes e o que é mais importante, tentar obter um modelo de comportamento que se ajuste a um modelo matemático ([37], [38], [39]).

Trabalhos relacionados

Em 1998 surgiram vários estudos que examinam aos usuários de um sistema de recuperação de informação em Internet frente aos usuários de um sistema tradicional.

O primeiro destes trabalhos foi apresentado em 1998 por Steve Kirsch, onde se analisam as consultas realizadas sobre Infoseek [115]. O estudo descreve de forma fácil algumas características básicas das consultas dos usuários, como as consultas mais repetidas, a utilização de operadores lógicos e número médio de palavras por consulta. A principal conclusão contribuída se centra no facto de que as consultas realizadas são demasiadas curtas o que provoca um elevado número de resultados por consulta, impossíveis de comprovar pelo usuário.

A seguir Cansem et al. apresentaram um trabalho no SIGIR Fórum de 1.998, onde realiza uma análise detalhada das consultas realizadas a Excite [115]. Nesta análise se examinam as consultas propriamente ditas, os termos das consultas e se realiza um breve seguimento das consultas realizadas pelos usuários durante uma sessão. Esta investigação contribuiu aspectos concretos do comportamento dos usuários Web frente aos usuários tradicionais. A seguir se listam suas principais conclusões:

- As consultas dos usuários Web são curtas, utilizando aproximadamente duas palavras em média (confirmando o estudo de Kirsch) e menos de 4% das consultas empregadas mais de seis palavras.
- Os operadores lógicos se utilizam pouco frequentemente (um de cada 18 usuários) e o caso dos operadores simples, como o "+" e "-", é ligeiramente melhor já que um de cada 12 usuários tende a utilizá-los.
- Um usuário consulta em média 2,21 páginas, e mais da metade dos usuários não visitam mais para além da primeira página de resultados.
- O número médio de consultas por usuário é inferior a 3, o que indica que os usuários não reformulam as suas consultas. Isto se confirma pelo facto de que 60% dos usuários realizam uma única consulta.
- Com respeito aos termos ou palavras usadas nas consultas é interessante destacar que existe um conjunto reduzido de palavras que se repetem muitas vezes e um grande conjunto de termos que se repetem uma única vez.

Estas características fazem que o usuário de um serviço de busca na Web seja significativamente diferente de um usuário tradicional. Basicamente os usuários

Web não se encontram cómodos utilizando os operadores lógicos e não conferem à totalidade dos resultados obtidos, ficando na primeira página de resultados.

Outros estudos, em [208] apresentam o trabalho que maior número de consultas analisa (cerca de 1.000 milhões de consultas), neste caso sobre Altavista [208]. A primeira parte deste estudo se obtém resultados similares ao realizado por [106]. Em ambos os casos se observam que o número de palavras por busca é bastante reduzido (ao redor de duas palavras por busca). Assim mesmo se mantém a distribuição de palavras nas consultas, já que um grupo reduzido de consultas é repetido múltiplas vezes e ao mesmo tempo existe um grande conjunto de consultas realizadas uma vez. Com respeito às consultas realizadas durante as sessões dos usuários, observam-se que em geral as sessões são breves, aproximadamente duas consultas por sessão e o número de páginas de resultados visitados não chega a dois (este valor é ligeiramente inferior ao estudo anterior). Como aspecto inovador, a segunda parte do estudo se centra no exame das correlações existentes entre os termos utilizados nas consultas com o objetivo de examinar os pares de palavras consultadas.

O Directório e o motor de busca

Neste apartado se definem as principais características do entorno utilizado pelos usuários, tendo em conta que se trata de um Directório Web tal e como se descreveu na secção anterior , os directórios permitem a localização da informação mediante a navegação e as buscas.

No processo da navegação se percorre o gráfico de categorias examinando os documentos associados, e no caso de uma busca, se mostrarão tanto documentos Web catalogados como categorias que coincidem com os conceitos buscados.

O motor de busca está baseado no modelo vectorial, realizando uma ordenação dos resultados segundo o modelo tradicional TFxIDF[1] (term frecuencyxinverse document frecuency) modificado para o suporte de diferentes pesos em função da importância de cada palavra. A busca por defeito realizará um OR lógico das palavras, ainda que o algoritmo de ordenação garanta que as primeiras posições estarão ocupadas por aqueles documentos que se relacionem com todas as palavras (ao igual que uma operação de AND lógico).

[1]É o produto da frequência do termo no documento pela frequência do documento inverso na colecção e permite-nos quantificar a relevância de cada termo para os documentos ($W_{i,d}$) e para as consultas ($W_{i,q}$) no espaço vectorial.

Por outra parte o motor de busca suporta diferentes operadores. Os mais fáceis são os seguintes:

Operador "-": por exemplo a expressão **-informação** indica ao motor de busca que recupere todos os documentos excepto os que contenham a palavra informação

Operador "+": a expressão **+informação** indica ao motor de busca que ignore aqueles documentos que não contém a palavra informação.

Operador " ": a expressão **"recuperação de informação"** indica ao motor de busca que recupere os documentos com o literal indicado, ainda que neste caso o motor realizará um AND lógico entre as palavras. Isto é, não se garante que a ordem e a proximidade das palavras se mantenham.

Igualmente o motor de busca suporta os operadores lógicos mais comuns (and, or, not). Estes operadores podem ser combinados utilizando os parênteses.

Como se comentou anteriormente, uma característica definidora dos directórios Web é o facto de que as buscas podem ser restringidas aos documentos contidos ou descendentes de uma categoria determinada. No caso do directório analisado, por defeito a busca se realizará sobre todos os documentos, ainda que exista a possibilidade de restringir a busca a uma categoria e os seus descendentes.

A maioria dos sistemas de busca, permitem aos usuários realizarem buscas detalhadas com o objectivo de obterem os resultados, isto é permitem realizar a modificação de determinados parámetros específicos sobre os campos em que necessitam realizar a busca (título de documento, descrição de documento, URL de documento ou as palavras chaves de documento). Doravante se fará referência a cadeia de busca como a cadeia exacta que tem teclado o usuário (incluindo operadores, palavras comuns, etc.) e se denotará termo de busca a cada uma das palavras, uma vez eliminados os operadores, palavras comuns, etc. Portanto, uma cadeia de busca estará constituída por um ou mais termos de busca.

O processo de busca se inicia quando o usuário introduz a cadeia de busca, processam-se os termos de busca e se mostra uma página de resultados. Cada resultado inclui o titulo e a descrição do documento, junto com um enlace a uma página intermédia (o que permite estabelecer e controlar o número de acessos a cada página Web), que transladará automaticamente ao usuário a sua localização em Internet. O usuário pode seguir qualquer destes enlaces para conferir um documento,

ou pode seguir a outras páginas de resultados através dos botões de navegação, o que produzirá outra consulta ao motor de busca.

Avaliação do rendimento de recuperação na Web

As características da avaliação da qualidade dos buscadores na Web diferem significativamente da avaliação tradicional, precisamente pelas características do próprio World Wide Web:

- Dinamismo da Web e dos próprios motores de busca.
- Heterogeneidade dos documentos e das consultas realizadas pelos usuários.
- A estrutura de hiperenlaces existentes entre as páginas Web.
- A grande quantidade de documentos obtidos numa busca, o que faz praticamente impossível a avaliação da totalidade dos recuperados.

Com base nisto, a noção de relevância nos documentos recuperados apresenta certos matizes ao ser aplicada a páginas Web, principalmente devido à presença de enlaces entre as páginas. Portanto, pode-se considerar uma página como relevante se seu conteúdo se considera significativo pelo usuário, enquanto se pode considerar uma página útil se contém enlaces para páginas relevantes [6]. Isto gera diversas categorias de relevância-utilidade tal e como se mostra na Tabela seguinte:

PÁGINA RECUPERADA	PÁGINAS ENLAÇADAS	
	relevante	não relevante
relevante	relevante e útil	relevante
não relevante	útil	não relevante e inútil

Tabela 3.2: Categorias de relevância-utilidade das páginas recuperadas.

Em consequência, as medidas de efectividade de sistemas de recuperação usadas em experimentos de Laboratório (precisão e recall) unicamente oferecem uma informação parcial na avaliação de motores de busca, já que não consideram a existência de enlaces entre os documentos. Portanto, fez-se necessária a utilização de outras medidas desta efectividade como as estimações do recall propostas em [135], ou a novidade e o ruído. A novidade se define como a proporção de documentos relevantes obtidos numa fase concreta, mas omitidos em fases anteriores [200]. O ruído se

define como a proporção de páginas que não são relevantes e são recuperadas numa fase concreta, mas que foram omitidas em fases anteriores.

Além disto, a grande quantidade de documentos obtidos invalida a possibilidade do cálculo da precisão e do recall. Portanto, diversos estudos ([131], [195], [70], [130], entre outros) aplicaram o cálculo da precisão relativa aos primeiros dez ou vinte documentos recuperados (representada como P@10 ou P@20). Isto representa de uma maneira mais adequada o procedimento realizado pelos usuários de um sistema de busca na Web, e permite obter um valor, ao menos aproximado, da precisão do sistema. Assim mesmo, a cobertura que oferece cada motor de busca do total das páginas Web disponíveis em Internet é um factor básico para a efectividade do sistema de busca. No estudo levado a cabo por Bharat e Broder em [25] se demonstra que a cobertura que oferecem vários motores de busca amplamente utilizados em Internet, no melhor dos casos, não superava os 50%. Enquanto a parte comum entre os motores de busca estudados era inferior à 1.4%.

3.6 Indexação com aproximação linguística

Durante os anos 90, a disciplina conhecida como Processamento de Linguagem Natural (PLN) experimentou um forte impulso que possibilitou o desenvolvimento de técnicas de análises robustas, isto é, aplicáveis a textos sem restricções de domínios o que, a sua vez, permitiu ampliar os seus campos de aplicação, sendo um dos destacados o da recuperação de informação (RI).

Desde o campo do PLN não demorou observar-se como o método de indexação comummente adoptado em RI era resultado de uma análise muito superficial de texto, e que este podia aperfeiçoar-se empregando as novas ferramentas de análises desenvolvidas, para solucionar, ou quando menos, moderar os efeitos que mais se denunciavam em RI, e que ainda padecemos hoje em dia em nossa busca quotidiana em Internet, como determinantes na hora de aumentar a efectividade nos sistemas de recuperação de informação, os derivados da ambiguidade léxica, tanto a nível de categoria gramatical como a nível de significado. A representação de documentos e consultas consistia e consiste, ainda hoje em dia na maioria dos sistemas em uso, na detecção das palavras "ortográficas" ao menos as linguagens com os nossos convênios ortográficos, dos textos, a normalização das mesmas a sua forma maiúscula e minúscula (com a eliminação de acentos e diacríticos) e a supressão das que estão incluídas no que se conhece como *StopWords* ou lista de palavras comuns.

Independentemente do método de pesado adoptado e da função ou métrica de comparação de consultas e documentos que cada sistema implemente, que determinará, os documentos a recuperar e a ordem em que se devolve ao usuário, o conjunto inicial de documentos candidatos susceptíveis de ser recuperados será seleccionado entre aqueles que contenham, dependendo do sistema de recuperação todas as mesmas palavras da consulta, ou ao menos uma parte das mesmas palavras de dita consulta (caso dos sistemas baseados no modelo vectorial).

Repassamos a seguir os diferentes experimentos que se propuseram sobre colecções monolingues e que, seguindo a [219], podem dividir-se em propostas em indexação morfológica, indexação sintáctica e indexação baseada no sentido das palavras.

3.6.1 Indexação morfológica

Em RI se propuseram e experimentaram técnicas não linguísticas para tentar indexar as palavras dos documentos e das consultas pela sua raiz (*técnicas de stemming*). Estes métodos não linguísticos, fáceis e eficientes computacionalmente, simplesmente realizam uma "poda indiscriminada" de, normalmente, determinados fins de palavras. Propuseram-se métodos que vão desde um simples *s-stemmer*, isto é, aquele que, para o inglês, elimina toda palavra o caracter final "s" (com o que se procura que os plurais e singulares das palavras de documentos e consultas se indexem por um mesmo padrão), até outros mais sofisticados para tentar tratar a morfologia derivativa. Obviamente estas "eliminações cegas" de certos sufixos produzem anomalias na tentativa de obtenção da raiz tanto por excesso como por defeito.

Uma versão de conhecido algoritmo de Porter normaliza à forma *organ* as palavras *organization* e *organism*. Uma versão de um s-stemmer para o espanhol que elimina os sufixos *as*, *é*, *vos*, *a*, *e*, e *o* de todas as palavras têm, por exemplo, como efeito para transformar tanto *capa*, *capo* (e versões plurais) e *cape* em *cap* [56].

Estes problemas derivados de uma *poda cega* podem evitar-se contando com um dicionário computacional (*léxicon* computacional) e um adequado processador morfológico, isto é, com os recursos e ferramentas de PLN capazes de determinar para cada representação superficial de uma palavra, todas as possíveis categorias gramaticais, com a sua forma canónica correspondente. A morfologia computacional experimentou um forte desenvolvimento na passada década e adoptou técnicas muito eficientes desde o ponto de vista computacional (morfologia de estados finitos) [11]

para levar à cabo o reconhecimento das palavras, até o ponto de considerar-se, ao menos para as línguas dominantes, um problema praticamente resolvido.

Não obstante, utilizar este processador linguístico para levar a cabo a normalização das palavras no momento da indexação obriga a incorporar um desambiguador categorial (*Part of Speech Tagger*), isto é, uma ferramenta capaz de atribuir, para cada palavra de um texto, uma única categoria gramatical, dado o contexto, esta técnica é conhecida por Pos-Tagger. Como exemplo do que vimos dizendo, suponha-se o texto seguinte: "*o príncipe não se casa*". Centrando-nos na palavra "casa", a saída do *léxicon* e o processador morfológico nos devolveriam "casa" como "substantivo" (S), cuja forma canónica seria a mesma, e "casa" como uma flexão do verbo (V) casar. Por tanto, para indexar dito documento por suas formas canónicas, tem de determinar-se previamente qual é a correcta: casa/S ou casar/V.

Além disto, uma mesma palavra pode ter, para diferentes categorias gramaticais, a mesma forma canónica (por exemplo, "baixo" é a mesma forma canónica quando é adjectivo, substantivo e preposição), tem-se de procurar uma forma de representação, no momento de indexação, diferenciada (baixo/A, baixo/P, baixo/S). Deste exemplo que pusemos pode colegir-se facilmente que o efeito da desambiguação categorial pode ser benéfico, pois com o par canónica/categoria gramatical se discriminam diferentes usos (acepções) da cadeia de caracteres "baixo". Outros efeitos positivos que se podem obter técnicas de POS-Tagging na indexação são: uma eliminação coerente das palavras vazias (por exemplo, descartar "baixo" como "preposição" como palavra de indexação) e uma possibilidade de redução do tamanho dos índices [42].

Obviamente, também podem pensar-se exemplos nos que uma indexação de acordo a dito par acarrete efeitos negativos, por exemplo, se discriminamos "desenho" como verbo e como substantivo depois da desambiguação, ao não tratar o fenómeno morfológico da derivação, obtemos uma representação diferenciada para ambas as formas (desenho/S, desenhar/V), ainda que nos textos a indexar, as aparições de dita palavra com diferente categoria estejam sendo utilizadas para expressar o mesmo conceito [122]. Como factor em contra, além disto, ao utilizar técnicas linguísticas para acometer a indexação morfológica há que contar com o considerável aumento de recursos computacionais que implica respeito do uso das técnicas não lingüísticas antes mencionadas.

Quanto aos resultados obtidos nos diferentes experimentos de indexação morfológica no momento da recuperação, logicamente foram dependentes da linguagem da

colecção documentária, pois os diferentes fenómenos morfológicos (flexão, derivação e composição) não se manifestam com a mesma intensidade em todas as línguas. Assim, por exemplo para o inglês, a conclusão obtida é que a indexação com técnicas linguísticas não contribui melhoras aos métodos não lingüísticos, com o que não resulta aconselhável o uso das primeiras dado a diferença no custo computacional. Em [74] se comprova, inclusive, a indexação com POS-Tagging produzia melhoras inapreciáveis frente à indexação de palavras ortográficas. Seguindo com o inglês, neste caso, se supõe melhoras o uso de *stemmers* ou não, em [80] se conclui que não é, pelo contrário, em [123] se comprovam melhoras para diferentes "algoritmos de poda". Ao espanhol, os resultados obtidos em [56] parecem indicar que as técnicas de stemming produzem efeitos benéficos frente aos métodos que não realizam nenhuma normalização e, por outro lado, em [224] se realizam, entre outros, experimentos com técnicas de stemming e com ferramentas linguísticas para tratar a morfologia flexiva, obtendo melhores resultados utilizando estas últimas. Para outros idiomas, como por exemplo, o holandês e o alemão, comprovou-se que tratar a decomposição de palavras ortográficas nas correspondentes gramaticais produz efeitos benéficos, tanto utilizando técnicas linguísticas [121] e [158], como não lingüísticas [141].

Quanto à avaliação dos efeitos que pudessem derivar-se dos erros na desambiguação categorial (a precisão dos POS-Taggers se situa entre 95-97% ou inclusive superior [194], segundo se desprende de [75], não parecem relevantes).

3.6.2 Indexação sintáctica

O método de indexação por palavras isoladas implicitamente assume a independência destas respeito dos textos das que se extraem, e, por tanto, óbvia que:

- Muitos conceitos se constroem concatenando, em determinadas línguas, várias palavras ortográficas. Esse conjunto de palavras pode ter, para determinados domínios semânticos, uma grande relevância e, no entanto, isoladamente, esse conjunto de palavras, por ser muito utilizado na colecção documentária, adquirir um peso irrelevante. Além disto, a ordem das palavras na frase implica uma variação do significado ("college junior", vs. "junior in college" vs "junior college").

- Por outra parte, determinados conceitos podem expressar-se com diferentes construções sintácticas que seria conveniente, na hora de indexar, procurar

uma representação comúm ("O gato foi atacado por um cão" vs "O cão atacou um gato") [215].

Em [215] tem-se experimentado com diferentes métodos para avaliar os efeitos de uma indexação multipalavras. Basicamente, as técnicas empregadas podem agrupar-se em estatísticas e linguísticas. As primeiras se limitam a colectar *co-ocurrências* de pares de palavras nos textos (bigramas). As segundas utilizam métodos de análise sintáctica superficial ("Shallow Parsing") para reconhecer diferentes estruturas sintagmáticas, mais ou menos complexas, às vezes não só para detectar determinadas sequências de etiquetas gramaticais (com o qual são mais selectivas do que a simples recolha de pares de palavras que efectuam os métodos anteriores), senão também para normalizar diferentes árvores sintácticos em padrões comuns à hora de indexar [215].

O "analisador sintáctico superficial" se adiciona à saída de um *Pos-tagger*, o que supõe utilizar um conjunto de recursos e ferramentas que acarretam um custo computacional importante. Dos resultados obtidos pela investigação, no sentido de que supõem uma melhora ou não da indexação baseada em simples palavras ortográficas, estes se mostram, quando menos, contradictórios [176]. Ainda que fosse um tema investigado desde fins dos 70, não foi, não obstante, até a última metade dos 90 quando se pode experimentar a indexação sintagmática com grandes colecções documentários aplicando as modernas técnicas de PLN de análise robusta.

As conclusões obtidas por diversos grupos de investigação com respeito à indexação de sintagmas com técnicas linguísticas podem resumir-se como segue: na indexação por *sintagmas* ainda que se obtenha melhores resultados utilizando técnicas linguísticas que meramente estatísticas, as diferenças são escassas; as melhoras entre uma indexação por *sintagmas* com técnicas linguísticas e uma indexação por simples palavras ortográficas são inapreciáveis se as consultas são curtas, contudo se as consultas são longas sim se apreciam; a indexação por sintagmas não deve suprir à indexação dos elementos simples que os compõem; não é fácil determinar quê peso a dar aos compostos detectados [242] e [215].

3.6.3 Indexação baseada no sentido das palavras

Propuseram-se vários métodos para indexar documentos e consultas de acordo ao "significado" das palavras que os compõem, com o objectivo de medir os efeitos que pudessem produzir-se ao resolver os problemas da ambiguidade léxica semântica.

Para isso, utilizaram-se diferentes recursos, sendo os principais os "dicionários" e a rede semântica de palavras WordNet [176].

A indexação baseada nos sentidos de acordo a um "dicionário", dada a sua forma de organização, permite a representação diferenciada dos diferentes significados, isto é, possibilita o tratamento da polissemia e a homonímia. Utilizando uma rede semântica como WordNet, organizada em synsets "" (conceitos), é possível o tratamento não só dos fenómenos anteriores senão também o da sinonímia, além da meronímia, hiponímia etc. dado que na base de dados também se armazenam ditas relações entre os "synsets" [148]. Sirva como um pequeno exemplo de indexação baseada nos sentidos o seguinte: suponha-se que a palavra "carro", por simplificar, tem dois sentidos, de "veículo a motor" (1) e de "carro de bebé" (2) e a palavra "automóvel" um só, o primeiro que atribuímos a carro . Indexando de acordo a dicionários , todas as ocorrências de ambas as palavras em documentos e perguntas se indexariam, p.e., na forma, palavra#número de sentido. Isto é, "carro#1" e "automóvel#1" significaria que no documento ou consulta a indexar apareceu dita palavra com o sentido de "veículo a motor" e "carro#2" com o sentido de "carro de bebé". Com o que se adoptam representações diferenciadas para as diferentes acepções de "carro" (resolvendo o problema da polissemia), mas diferentes para as aparições de "carro" como veículo a motor e"automóvel" (com o que não resolvemos à sinonímia). Pelo contrário, indexando de acordo a WordNet, onde cada "synset" para as categorias gramaticais nome, adjectivo, verbo e advérbio tem um identificador único, a representação pode ser na forma categoria#número de synset. Isto é, todas as aparições do conceito "veículo a motor", cujo "synset", suponhamos, identifica-se na base de dados como "N#123" ficariam representadas de acordo a dito identificador e as referidas ao conceito "carro de bebê", ponhamos por caso, ao identificador "N#322".

Naturalmente, o processo de atribuir o sentido correcto às palavras dos textos deve realizar-se de forma automática. A desambiguação do sentido das palavras (Word Sense Disambiguation) é um problema computacionalmente complexo que se aborda desde diferentes propostas [223], mas que ainda precisa aperfeiçoar-se.

Quanto aos experimentos aplicados à indexação, resumindo, concentraram-se em dois aspectos principais [74]:

- Avaliar se produz melhorias e em que medida na recuperação de informação.
- Fixar um umbral de erro na precisão da desambiguação a partir do qual se produz uma degradação na efectividade da recuperação de informação.

Efectivamente, não se podia determinar se era benéfica ou não em RI a indexação por sentidos, pois não era possível estabelecer a degradação que produzia a desambiguação incorrecta. Outros experimentos utilizaram a estratégia da desambiguação manual, mas para isso recorreram a textos muito breves (p.e., pés de página) [211], com o que os resultados não podem *extrapolar-se* a colecções de grandes volumes de texto.

Diversos trabalhos se centraram no segundo aspecto. Para estabelecer a precisão mínima exigida à desambiguação em tarefas de RI, criaram-se métodos artificiais (as pseudo-palavras de M. Sanderson [202]), recorreu-se à desambiguação manual de pequenas passagens de texto (dado o custo tempo que isso supõe) ou se utilizaram "corpus" desambiguados (p.e. o "SEMCOR"). Por outro lado, ao mesmo tempo, a indexação algumas vezes se realizou segundo dicionários e outras de acordo aos "synsets". Por isso, os dados oferecidos pelos diferentes trabalhos de investigação não se aceitam como concluidos ou, inclusive, põem-se em entre dito. Assim, p.e. em [202] se conclui que se se desambigúa com uma precisão inferior ao 90%, a recuperação se degrada, mas fica por estabelecer se a ambiguidade artificialmente introduzida com as pseudo-palavras é comparável com as palavras reais, e nos experimentos de [74] se fixa a precisão sobre o 60%, concluindo-se que a diferença se deve, por um lado, a que a indexação com WordNet é mais tolerante a erros e, por outro, ao tratar com palavras reais.

No entanto, estes últimos resultados se questionam em [202], por não utilizar uma colecção padrão de avaliação e pela forma de criar documentos a partir das passagens do SEMCOR. O problema parece ainda aberto, ainda que mais bem se propusesse até que a tecnologia em desambiguação madure. Independentemente destes problemas enunciados, também se propôs o da "granularidade" dos sentidos tanto em dicionários como em WordNet. Um "grão muito fino" (trabalhar com muitas acepções diferentes para uma entrada léxica), pode ser, muitas vezes, difícil em RI, dado que ao indexar separamos "sentidos" que podem estar semanticamente muito próximos [123].

Capítulo 4

Preliminares da expansão de consultas e desambiguação

Neste capítulo se introduzirá ao leitor os antecedentes que se podem encontrar na expansão de consultas e desambiguação do sentido das palavras dentro do campo da recuperação de informação. Comentar-se-ão trabalhos anteriores que fizeram uso da informação lingüística para expandir consultas e dos resultados que se obtiveram. Também se explicará os modelos clássicos, bem como o modelo do usuário utilizado no presente trabalho.

4.1 Uso do recurso linguístico WordNet na expansão de consultas e WSD

Os usuários de sistemas de recuperação que usam a coincidência de palavras como bases de recuperação se enfrentam com o desafio de expressar as suas consultas com palavras dos vocabulários dos documentos que desejam recuperar. Esta dificuldade é grave em grandes bases de dados de texto devido a que ditas bases de dados contém muitas expressões diferentes para referir-se ao mesmo conceito. A habilidade para recuperar documentos destas bases de dados é crucial num amplo rango de aplicações: recuperar documentação para a ajuda em casos legais, facilitar a organização e recuperação de correspondência e formulários num escritório, filtrar fontes de notícias para encontrar artigos de interesse, encontrar passagens relevantes dentro de um conjunto completo de manuais de um sistema complexo para um pro-

blema particular, etc. Um método de aliviar a carga do usuário quando selecciona as palavras de uma consulta é que o sistema de recuperação expanda automaticamente a consulta adicionando-lhe termos que estejam relacionados com as palavras proporcionadas pelo usuário. Os novos termos podem estar estatisticamente relacionados com as palavras originais da consulta, isto é, os termos tendem a coincidir em documentos ou podem ser seleccionados com a ajuda dos recursos lingüísticos.

O uso da informação de ocorrência para expandir vectores de consulta é atraente devido a que as relações se podem gerar facilmente e só a partir dos documentos, obviando a necessidade de recursos linguístiõcos ou outras ferramentas adicionais, que são custosas de desenvolver e manter. Desafortunadamente, tais métodos tiveram pouco sucesso em melhorar a eficácia da recuperação quando se usam sem dados de relevância. Nota-se, no entanto, que métodos que exploram relações estatísticas mas não expandem a consulta, como o Latent Semantic Indexing [48], sim têm sido mais êxitosos.

O uso da informação linguística como fonte de termos relacionados encontrou algum sucesso em pequenos experimentos. Salton e Lesk [199] encontraram que a expansão mediante sinónimos melhora o rendimento, mas os resultados da expansão mediante termos mais gerais ou mais restringidos seleccionados de recursos linguísticos hierárquicos eram demasiados inconsistentes para ser úteis em geral, ao realizar-se sobre colecções pequenas. Wang, Vendendorpe e Evens [226] encontraram que uma variedade de relações léxico-semânticas melhoram o rendimento. No entanto, cada uma destas conclusões se atingiu em experimentos sobre colecções muito pequenas usando recursos linguísticos de um único domínio.

Quanto à utilidade da expansão de consultas mediante relações léxico-semânticas em colecções grandes que abarcam vários domínios, os trabalhos mais representativos são os que utilizam a colecção de referência TREC [218] que comentaremos mais adiante.

Os estudos prévios com estas características no uso de informação linguística ofereceram diferentes resultados. Voorhees [225] mostrou que usar WordNet para a expansão de consultas não aumenta a efectividade de recuperação de informação. Num trabalho de Mandala [138] as relações armazenadas em WordNet se combinam com outras medidas de similaridade baseadas em dependências sintácticas e informação de participação, estes sim melhoram consideravelmente a eficácia de recuperação de informação.

No trabalho de Ellen M. Voorhees [225] as consultas se expandem usando as

relações codificadas em WordNet, recurso linguístico que comentaremos no próximo capítulo, e são avaliadas com os tópicos de 101 á 150 produzidos pela segunda conferência TREC (TREC-2). Para evitar os efeitos distorsionadores a partir de uma consulta original com selecção pobre de palavras, os termos originais da consulta que se expandem foram gerados automaticamente a partir dos tópicos seleccionado além disto, os conjuntos de sinónimo de WordNet que se considerava de que resultavam os conceitos importantes do tópico. Por isso, os resultados obtidos neste trabalho representam um limite superior no rendimento que se pode esperar num procedimento de selecção totalmente automático que use a estratégia de expansão. Inclusive no melhor caso, a expansão não melhorou a efectividade das consultas que eram completas inicialmente, isto é, das consultas que continham termos suficientemente representativos e que descreviam o tópico de interesse com detalhe. Consultas menos completas, consistentes numa frase simples descrevendo o tópico de interesse, sim melhoravam significativamente com a expansão.

Em [201], a ambiguidade do sentido das palavras é mostrada em produzir somente poucos efeitos na exactidão da recuperação, confirmando aparentemente que as estratégias de emparelhamento de *query/documentos* já realizam uma desambiguação implícita. Sanderson estima também que se a desambiguação do sentido das palavras explícita fôr realizada com uma exactidão menor de 90%, os resultados são piores que não desambiguar. Na realização dos seus experimentos, a ambiguidade é introduzida artificialmente nos documentos, substituindo pares de palavras seleccionadas [209] aleatoriamente (por exemplo, *banana/kalashnikov*) com termos artificialmente ambíguos (*banana/kalashnikov*). Quando os seus resultados forem muito interessantes, em sua opinião considera que é confuso e se poderia confirmar com ocorrências reais de palavras ambíguas. Há também uma outra debilidade menor em experimentos de Sanderson. Quando ele desambigua um termo tal como *spring/bank* para [225] ter, por exemplo, banco, fez somente uma disambiguação parcial, porque banco pode ser usado em mais de um sentido na colecção do texto.

Para além da desambiguação, muitas tentativas foram feitas de explorar WordNet para finalidades da recuperação do texto. Dois aspectos foram principalmente dirigidos, o enriquecimento das consultas com termos semanticamente relacionados, e por outro lado a comparação das consultas e dos documentos através das medidas conceptuais de distância.

A expansão de consulta com WordNet tem mostrado ser potencialmente relevante para aumentar o *recall*, porque permite emparelhar os documentos relevantes que poderiam não conter alguns termos da consulta [210]. Entretanto, tem pro-

duzido poucos êxitos experimentais. Por exemplo, [225] expandiu manualmente 50 consultas sobre uma colecção TREC-1, usando os sinónimos e outras relações semânticas de WordNet 1.3. Voorhees constatou que a expansão era útil com consultas incompletas, onde outras técnicas da expansão funcionaram melhor, para consultas curtas, continuou o problema de seleccionar as expansões automaticamente; fazê-las mal podia degradar o desempenho da recuperação.

Em [210], uma combinação das técnicas bastante sofisticadas baseadas em WordNet, incluindo desambiguação automática e medidas automáticas de relações semânticas entre conceitos de *query/documentos* resultou numa pequena eficácia. Infelizmente, os efeitos dos erros da desambiguação do sentido das palavras, não podiam ser discernidos com exactidão da estratégia da recuperação. Entretanto, em [211], a recuperação numa colecção pequena de legenda de imagem, isto é, em documentos muito curtos, é melhorada razoavelmente usando medidas de distância conceptual entre as palavras baseadas em WordNet 1.4. Previamente, legenda e consultas tem sido desambiguadas manualmente com WordNet. A razão para tal sucesso é aquela com documentos muito curtos (por exemplo dos rapazes que jogam na areia "*boys playing in the sand*") que a possibilidade de encontrar os termos originais da consulta (por exemplo das crianças que correm numa praia " *children running on a beach*") é muito mais baixa do que para os documentos de tamanho médio (que incluem tipicamente muitas expressões "*phrasings*" para os mesmos conceitos). Em [225] está de acordo com estes resultados, mas continua a questão de se emparelhar à distância conceptual em documentos e consultas mais longas. Para além disto, os experimentos realizados por [211] consideram somente os substantivos, enquanto que WordNet oferece a possibilidade de usar todas as taxonomias de palavras (substantivos, verbos, adjectivos e advérbios).

A estratégia essencial da recuperação nos experimentos relatados aqui é adaptar um sistema baseado num modelo vectorial clássico que explicaremos na Secção 4.2, usando synsets de WordNet como espaço de indexação em vez de formas de palavras. Esta aproximação combina dois benefícios para a recuperação: *(i)* Que os termos são completamente desambiguados (melhoraria a precisão); e *(ii)* Que os termos equivalentes podem ser identificados (melhoraria o *recall*), o recall e a precisão são medidas de eficácia de recuperação de informação que serão explicadas claramente no capítulo 6. Anote que a expansão da consulta não satisfaz à primeira condição, porque os termos usados para expandir são palavras e portanto, estão a volta de ambíguas. Por outro lado, a desambiguação do sentido das palavras não satisfaz a segunda condição, como os sentidos equivalentes de duas palavras diferentes não

estão emparelhados. Assim, indexando com synsets se alcança um emparelhamento máximo e um emparelhamento mínimo, parecendo assim um bom ponto de partida para estudar a recuperação de texto com WordNet.

Dada esta aproximação, o objectivo foi de testar dois temas principais que não são respondidas claramente com base aos experimentos mencionados acima:

- Abstraindo desde o problema da desambiguação do sentido das palavras, quê potência oferece WordNet para a recuperação do texto. Em particular, se gostaria de estender experimentos com desambiguação de documentos e consultas manualmente para textos de tamanho médio.

- Uma vez que o potencial de WordNet é sabido para uma colecção desambiguada manualmente, se pretende testar a sensibilidade do desempenho da recuperação aos erros da disambiguação introduzida automaticamente.

Segundo os experimentos realizados por [73], utilizando distintas aproximações de indexações, tais como: indexação com synsets, indexação com palavras (SMART básico) e indexação com sentidos das plavras, os resultados mostram que a indexação com synsets de WordNet produz uma melhoria notável na colecção de prova. Uns 62% dos documentos são recuperados em primeiro lugar, contra 48% do desempenho de SMART. Isto representa mais de 14% de documentos, uma melhoria de 29% com respeito à SMART. Este é um resultado excelente, embora se deve ter em conta que é obtido com consultas e documentos manualmente desambiguados. Não obstante, mostra que WordNet pode extremamente aumentar a recuperação de texto, o problema reside em conseguir a desambiguação automática exacta do sentido das palavras.

Os experimentos de Gonzalo e outros [73], mostram por outro lado que a indexação do sentido das palavras melhora a execução ao considerar até quatro documentos recuperados para cada *query/summario*, embora seja pior do que a indexação por synsets. Isto confirma a ideia de que a indexação com synset tem vantagem sobre a desambiguação do sentido das palavras porque permite emparelhar semanticamente termos similares. Fazendo exame somente do primeiro documento recuperado para cada sumário, a colecção desambiguada dá 53.2% de sucesso contra 48% da consulta de SMART, o que representa uma melhoria de 11%. Para os níveis de recall mais altos de 0.85, a colecção desambiguada se desempenha pior. Isto pode parecer, surpreendente cômo a desambiguação do sentido das palavras aumentaria somente o

conhecimento sobre consultas e documentos. Mas se deve ter em mente que WordNet 1.5 não é uma base de dados perfeita para a recuperação de texto, e indexação de sentido das palavras, impede algum emparelhamento que podem ser útil para a recuperação. Por exemplo, WordNet 1.5 não inclui relações semânticas da parte do discurso cruzado, assim que esta relação não pode ser usada com sentido das palavras, quando simplesmente indexamos o termo não os distingue. Outros problemas de WordNet para a recuperação de texto incluem demasiadas distinções de sentidos e é demasiado fino; veja [73] na imprensa para uma discussão mais detalhada na adequação da estrutura de WordNet para a recuperação de texto.

Para além disto, os resultados de Gonzalo e outros [73] mostram, a sensibilidade do sistema da indexação do synset à degradação da exactidão da desambiguação (correspondendo os experimentos descritos acima). Desde o ponto de vista gráfico em [73], nota-se que menos de 10% de erros de desambiguação não afecta substancialmente o desempenho ou a execução. Isto está aproximadamente de acordo com [201]. Para razões de erros acima de 10%, o desempenho se degrada rapidamente. Isto está também de acordo com [201]. Entretanto, indexando com synsets é muito melhor do que a execução com Smart, acima de 30%. Desde 30% á 60%, os dados não mostram diferenças significativas com indexação de palavras com SMART padrão. Encontraram também que as consultas têm que ser desambiguadas para ter vantagem da aproximação; se não, os melhores resultados possíveis de indexação com synset não melhoram o desempenho da indexação de palavras com SMART.

O procedimento da expansão usado no trabalho de [225] foi o seguinte: A partir dos exemplos das necessidades da informação que a colecção TREC inclui se realizava uma selecção manual de termos. Para cada um desses termos, e mediante a ajuda de WordNet, seleccionava-se, também de forma manual, um conjunto de sinónimos de cada um destes termos originais para se adicionar ao vector da consulta. Esse conjunto de sinónimos podiam ser somente os sinónimos directos, ou todos os descendentes na hierarquia é-um de WordNet, ou todas as palavras em conjuntos de sinónimos a um enlace de distância do conjunto original independentemente do tipo de enlace, etc. O procedimento de expansão podia receber, além disto, diferentes parámetros para facilitar a comparação da efectividade das diferentes possibilidades destes esquemas. Este conjunto de parámetros permite especificar para uma prova determinada, e para cada relação inclusa em WordNet, a longitude máxima que se deve seguir numa cadeia desse tipo de enlaces. Todos os sinónimos contidos dentro de um conjunto de sinónimos da cadeia são adicionados à consulta. As palavras mais comuns no idioma ou *stopwords*, tiram-se, e às palavras restantes

se aplicam um processado morfológico consistente em ficar somente com a raiz da palavra (*stemming*).

Neste modelo de expansão de Voorhees os stems adicionados através de diferentes relações léxicas (relações de sinonímia ou relações hierárquica entre palavras) mantém-se separadas usando o modelo vectorial estendido introduzido por Fox [57]. Cada vector consta de subvectores de diferentes tipos de conceitos (*chamados ctipos*) onde cada *ctipo* corresponde a diferentes relações léxicas. Um vector de consulta potencialmente tem onze *ctipos*: um para termos originais da consulta, um para os sinónimos e um para cada um dos outros tipos de relações contidas dentro da hierarquia de substantivos de WordNet (cada metade de relação simétrica tem seu próprio *ctipo*). Um termo original da consulta que é membro de um conjunto de sinónimos seleccionados aparece nos seus dois *ctipos* respectivos. De maneira similar, uma palavra que está relacionada com um conjunto de sinónimos através de duas relações diferentes aparece em ambos os *ctipos*.

A similaridade entre um vector de documento D e um vector de consulta estendida Q se calculou como a soma ponderada de similaridades entre D e cada um dos subvectores da consulta. Voorhees usou o esquema de pesado da ferramenta SMART com os pesos *inc* sugeridos por Bckley [36] para ponderar os termos dos vectores; isto é, o peso de termos se fixa a 1,0+ln(tf) onde tf é o número de vezes que o termo aparece no documento. Normaliza-se então mediante a raiz quadrada da soma dos quadrados dos pesos no vector (normalização de cosseno). Os termos da consulta se ponderam usando o logaritmo[1] do factor da frequência dos termos multiplicando-o pela frequência inversa dos documentos e os termos no *ctipo* que representa a consulta original se normaliza mediante o cosseno [209]. Os pesos em *ctipos* adicionais se normalizam usando a longitude calculada para o *ctipo* dos termos originais. Esta estratégia de normalização permite que os pesos dos termos da consulta original não se vejam afectados pelo processo de expansão e mantém os pesos em cada *ctipo* comparáveis com os de outros *ctipos*.

De todas as possibilidades que se poderiam contemplar Voorhees usou quatro estratégias de expansão:

- Expansão só mediante sinónimos.
- Expansão mediante sinónimos mais todos os descendentes na hierarquia é-um.

[1] Geralmente, os documentos mais longos têm tido maior peso com respeito aos documentos curtos, para reduzir a vantagem aos documentos longos realiza-se a normalização utilizando diferentes técnicas.

- Expansão mediante sinónimos mais pais e todos os descendentes na hierarquia é-um.
- Expansão mediante sinónimos mais qualquer conjunto de sinónimos directamente relacionado com o conjunto dado, isto é, seguindo uma cadeia de longitude 1, para todos os tipos de enlaces existentes.

Tendo em conta que os termos do vector de consulta recebem um peso, o peso do subvector de termos originais foi normalmente maior que o peso de outros subvectores para reflectir a suposição de que os termos proporcionados pelo usuário são geralmente mais importantes do que os adicionados automaticamente. As execuções nas quais os pesos dos termos originais eram menor ou igual a outros pesos provaram esta suposição obtendo-se piores resultados.

Claramente a expansão proposta naquele trabalho não foi efectiva, nenhuma das estratégias de expansão melhorou significativamente o rendimento da consulta sem expandir. De facto, a diferença em rendimento entre execuções expandidas e sem expandir para consultas individuais foi muito pequena para a maioria das execuções expandidas. O rendimento em consultas individuais variou mais para estratégias de expansão mais agressivas, isto é, estratégias de expansão usando cadeias de enlaces longas e ponderando com mais peso os termos adicionados. Mas, para um conjunto de consultas, o rendimento global foi pior para consultas expandidas de maneira agressiva. Num conjunto inicial de experimentos, a execução expandida mais efectiva foi aquela que expandiu um conjunto de sinónimos de uma consulta mediante um conjunto de sinónimos directamente relacionado com ele e com um peso de 0,5 para todos os subvectores adicionados. Por isso, esta estratégia de expansão foi à usada para os experimentos posteriores.

Para provar a hipótese de que a expansão de termos não é de ajuda na colecção TREC devido a que a descrição do problema proporcionada por um tópico TREC é muito completa, as consultas derivadas de versões mais curtas das originais dos tópicos foram expandidas usando estratégias padrões de expansão. Nestas consultas expandidas obtidas a partir de consultas originais mais curtas sim se conseguiram melhoras significativas com respeito às originais curtas mas não com respeito às originais em sua versão mais longa.

Os experimentos de Voorhees demonstraram, por tanto, que a expansão mediante relações léxico-semânticas proporcionam um benefício pequeno quando o usuário proporciona uma consulta detalhada. Mesmo assim, os usuários frequentemente não proporcionam uma consulta detalhada. Neste caso, as relações léxico-semânticas

têm o potencial de melhorar as consultas iniciais, ainda que a consulta expandida seja pouco provável e que seja tão efectiva como uma consulta melhor formulada e proporcionada pelo usuário.

Outra investigação destas características é a de Smeaton [211] que tentou expandir a consulta da colecção TREC-4 com várias estratégias da expansão ponderada de termos, com técnicas automáticas e manuais de desambiguação de significados de palavras. Desafortunadamente todas as estratégias degradaram o rendimento da recuperação.

Ao invés que os recursos linguísticos feitos à mão, os recursos linguísticos baseados em recopilações se constrõem automaticamente a partir dessas recopilações sem a intervenção humana. Em [187] se usou um recurso linguístico construído automaticamente e melhorou a efectividade ao redor de 20% mas usando colecções de prova pequenas.

Em Schutze [205] construiu um recurso linguístico baseado em participação de termos e utilizou em duas aplicações de recuperação de informação. Usando uma colecção TREC reduzida melhorou ligeiramente o rendimento da recuperação.

Em [172] proporcionaram evidências teóricas das limitações dos dados de participação de termos para a expansão de consultas na recuperação de informação. Consequentemente, alguns buscadores tentaram construir recursos linguísticos usando métodos com uma maior base linguística. Grefenstette [78] construiu um recurso linguístico usando o contexto sintáctico e levou a cabo experimentos usando várias colecções pequenas de prova. O seu método melhorou o rendimento para algumas colecções pequenas mas não pude melhorar com outras. Jing [107] encontrou também uma melhora através da expansão de consultas usando recursos linguísticos construídos automaticamente e baseados em gramáticas.

O outro trabalho mencionado anteriormente é o de [138] no que se centram na expansão automática de consultas utilizando, para além disto, informação de coocorrência. A expansão pode incluir a todos os termos nos documentos relevantes ou algum subconjunto deles. Nesta investigação a expansão prévia à busca inicial, levou-se a cabo usando recursos linguísticos. Nestes experimentos sim se obtiveram melhorias muito significativas mas se devem sobretudo, além da combinação de vários recursos linguísticos, ao uso de informação de participação de termos nos documentos significativos o qual não é aplicável na prática devido ao seu custo computacional.

4.2 Modelos clássicos

O propósito desta secção é cobrir dois dos modelos IR mais importantes propostos ao longo dos anos. Modelos diferentes adoptam modos diferentes de formalizar os principais elementos que actuam no processo de recuperação de informação. Elegemos dois famosos modelos clássicos como exemplos ilustrativos do campo. O modelo booleano, é considerado como teoria de conjunto porque representa os documentos e as perguntas como um conjunto de termos. Por outra parte o modelo de espaço vectorial, é considerado como algébrico porque representa os documentos e as perguntas como vectores num espaço t-dimensional, onde t é o número de termos distintos existentes nos documentos da colecção. O seguinte parágrafo apresenta algumas definições utilizadas no resto da secção.

Dado um termo de índice k_i e um documento d_j; $\mathcal{W}_{i,j} \geq 0$ o peso associado com o par (k_i, d_j). Este peso quantifica a importância do termo de índice k_i para descrever o conteúdo semântica de documento d_j. A função g_i devolve o peso associado com o termo de índice k_i em qualquer vector t-dimensional, porque $g_i(d_j) = \mathcal{W}_{i,j}$.

4.2.1 Modelo booleano

O modelo booleano está baseado na teoria de conjuntos e na álgebra booleana. Considera que os termos de índice ou palavras chave estão bem presentes ou ausentes num documento. Os documentos se podem considerar como vectores pesados de maneira binária, por exemplo os pesos do termo de índice se supõem que são todos binários: $\forall i, j \in \{0, 1\}$. As perguntas se compõem de termos de índice enlaçados pelos conectores AND, OR, NOT. A estratégia de recuperação do modelo booleano está baseada num critério de decisão binária, por exemplo cada documento é bem relevante ou não relevante.

O processo de computação da similaridade entre documentos e perguntas é como segue: Dado que as perguntas são expressões boolenas convencionais, podem representar-se como conjuntos de vectores conjuntivos. Por exemplo, a pergunta q = NOT k_a AND $\langle k_b$ OR $k_c\rangle$ pode-se escrever como $q_c = \{(0, 1, 0), (0, 0, 1), (0, 1, 1)\}$, onde cada elemento do conjunto é um vector pesado de maneira binária cujo primeiro componente está associado com o termino de índice k_a, o segundo componente está associado com k_b e o terceiro componente está associado com k_c. Cada vector pesado de maneira binária se chama componente conjuntivo da pergunta. Seja q

uma pergunta e q_c a pergunta transformada num conjunto de vectores conjuntivos, a similaridade de um documento d_j com a pergunta q se define como:

$$sim(d_j, q) = \begin{cases} 1 & \text{se } \forall v \in q_c \mid \forall k_i, g_i(\vec{d_j}) = g_i(\vec{v}) \\ 0 & \text{outros casos} \end{cases}$$

O documento d_j se considera relevante para a pergunta q se $sim(d_j, q) = 1$. De outro modo, prediz-se que o documento não é relevante.

As perguntas formuladas pelos usuários têm usualmente a estrutura de termos de operações booleanas. Nos modelos booleanos se considerou primordial a tarefa de transladar estas perguntas a uma forma que possa ser avaliada de forma rápida.

As expressões booleanas se formam a partir das consultas dos usuários e representam os documentos que contém um determinado conjunto de chaves (ou termos). Por exemplo, a consulta do usuário: Encontra todos os documentos que contenham a palavra *"informação"*, poderia ser transformada na expressão booleana: *Informação*.

Igualmente podem construir-se expressões booleanas mais complexas utilizando os operadores lógicos. Assim a expressão booleana: *(informação and recuperação) or not (recuperação and ciência)* representa os documentos que contém as palavras *"informação" e "recuperação"* ou que não contenham as palavras *"recuperação" e "ciência"*.

Cada parte de uma expressão booleana tem associado um conjunto de documentos. Estes conjuntos de documentos se combinam para dar lugar ao conjunto de documentos correspondentes à resposta à consulta. Estas combinações se realizam mediante operações de conjuntos.

Como as perguntas booleanas têm uma semântica precisa e o conceito de um conjunto é bastante intuitivo, o modelo booleano proporciona um marco de trabalho claro e simples que é fácil de compreender por um usuário comum de um sistema de IR. Contudo, como os documentos são somente um conjunto de termos, o processo de emparelhamento é muito rápido. No entanto, o modelo booleano sofre alguns inconvenientes:

Primeiro, a execução da recuperação é limitada porque a decisão de relevância é binária. Isto pode levar à recuperação de demasiados ou poucos documentos, isto é os modelos booleanos não proporcionam uma ordenação dos documentos. É impor-

tante dispor de uma ordenação em ordem de relevância decrescente. Deste modo o usuário pode visualizar primeiro os documentos mais relevantes e deixar de analisar documentos quando considere de que os documentos já não são representativos para a sua consulta.

Segundo, não é fácil traduzir uma necessidade de informação numa expressão booleana. A maioria dos usuários encontram dificuldades de expressar suas petições através de expressões booleanas. Com frequência, as perguntas booleanas são representações torpes das necessidades dos usuários. Apesar destes inconvenientes, o modelo booleano foi adoptado por muitos dos primeiros sistemas bibliográficos comerciais.

4.2.2 Modelo do espacio vectorial

Apesar da simplicidade do modelo booleano, é bem sabido que o uso de pesos não binários conduz a substanciais melhorias na execução da recuperação [199]. Nesta linha, o modelo do espaço de vector atribui pesos não binários a termos em documentos e perguntas e proporciona um marco de trabalho no qual é possível o emparelhamento [209]. Muitas das aproximações à expansão de consultas se basearam no modelo vectorial. Os documentos e as perguntas se representam como vectores t-dimensionais como segue. Os documentos se representam como vectores $\vec{d} = (\mathcal{W}_{1,j}, \mathcal{W}_{2,j}, \ldots, \mathcal{W}_{t,j})$ e as perguntas se representam como vectores $\vec{q} = (\mathcal{W}_{1,q}, \mathcal{W}_{2,q}, \ldots, \mathcal{W}_{t,q})$ onde cada peso, é positivo e não binário.

Os pesos de termos se utilizam para computar o grau de similaridade entre cada um dos documentos armazenados na base de documentos e a pergunta proporcionada pelo usuário. Os documentos são armazenados em ordem decrescente de similaridade e, assim, o modelo do espaço vectorial toma em consideração os documentos que se emparelham com os termos de pergunta só parcialmente. Isto implica que o conjunto de respostas ordenadas é um conjunto mais preciso do que os conjuntos de respostas proporcionadas pelo modelo booleano.

Como os documentos e as perguntas são vectores num espaço dimensional, as medidas de correlação entre vectores se podem aplicar para obter uma medida de similaridade entre os documentos e as perguntas. Uma das medidas mais amplamente utilizadas computa a similaridade através do cosseno do ângulo entre os dois vectores [209]. Oficialmente,

$$sim(d_j, q) = \frac{\vec{d_j} \bullet \vec{q}}{|\vec{d_j}| \times |\vec{q}|} = \frac{\sum_{i=0}^{t} \mathcal{W}_{i,j} \times \mathcal{W}_{i,q}}{\sqrt{\sum_{i=1}^{t} \mathcal{W}_{i,j}^2} \times \sqrt{\sum_{i=1}^{t} \mathcal{W}_{i,j}^2}} = cos(\alpha) \tag{4.1}$$

Onde $\bullet$ representa o producto interno entre os dois vectores. O valor de $sim(d_j, q)$ varia desde 0 á 1. Com estas medidas se devolvem os documentos recuperados numa lista ordenada, sendo melhores os documentos que apresentan uma similaridade mais próxima à 1.

O processo de expansão consiste em adicionar termos ao vector original de termos da consulta.

Desenharam-se muitas outras funções de emparelhamento para computar o grau de similaridade entre um vector de pergunta e um vector de documento. Para um estudo detalhado disso nos referimos a [15] (cap: 2).

Normalmente, estabelece-se um umbral e os documentos recuperados são aqueles com um grau de similaridade por encima do umbral. Ao usuário se lhe apresenta uma lista de documentos ordenados e, assim ele pode vistoriar a lista desde o documento de cima, por exemplo o que tenha um maior grau de similaridade.

Os pesos de termo de índice se podem computar de muitas maneiras diferentes. Diferentes aproximações aplicam diferentes intuições para determinar quais são as circunstâncias que fazem que um termo seja importante. O esquema de pesado tf/idf é um dos métodos mais populares para atribuir pesos aos termos nos documentos. Duas intuições principais estão por trás da aproximação tf/idf. Primeiro, os termos que aparecem frequentemente dentro de um documento se consideram como bons representantes do documento e, assim, deveriam receber um peso alto. Segundo, os termos que aparecem em muitos documentos têm menor poder discriminatório e, assim, deveriam receber um peso baixo. Oficialmente, seja N o tamanho da colecção de documentos e n_i o número de documentos no qual aparece a palavra chave k_i. Seja $freq_{i,j}$ a frequência pura do termo k_i do documento d_j, por exemplo o número de vezes que se menciona k_i no texto do documento d_j. Então, o factor da frequência do termo no documento, factor tf, vem dado pela frequência normalizada $f_{i,j}$ da palavra chave k_i no documento d_j. Oficialmente,

$$f_{i,j} = \frac{freq_{i,j}}{max_I freq_{I,j}} \tag{4.2}$$

Onde o máximo se computa sobre todos os termos que se mencionam no texto

do documento d_j. Depois, o factor-idf (idf representa a frequência de documento inverso) e vem dada por:

$$idf_i = log\frac{N}{n_i} \tag{4.3}$$

Finalmente, o esquema de pesado tf/idf assigna o siguinte peso ao termo k_i no documento d_j.

$$w_{i,j} = f_{i,j} \times idf_i = tf \times log\frac{N}{n_i} \tag{4.4}$$

Para os pesos dos termos de pergunta se segue um procedimento similar. Diversas variações do esquema de pesado anterior se apresentaram na teoria. No entanto, em geral, o esquema tf/idf proporciona um bom método para atribuir pesos a termos em muitas colecções.

Desde um ponto de vista teórico, o modelo de vector tem a desvantagem de que os termos de índice se assume que são mutuamente independentes. Ainda que isto seja uma ultra simplificação, considerar as dependências dos termos poderia danar a total execução da recuperação. O modelo de vector produz conjuntos de respostas ordenados que são difíceis de superar sem a expansão da pergunta ou o feedback de relevância. Em geral, o modelo de vector se considera bem superior ou quase tão bom como os bem conhecidos modelos alternativos.

4.2.3 Modelos alternativos

Ao longo dos anos se propuseram outros modelos. O modelo probabilístico está baseado no uso da teoria da probabilidade para modelar documentos e pergunta. Dada uma pergunta de um usuário, há um conjunto ideal de documentos que contém exactamente os documentos relevantes e outros não relevantes. O processo de pergunta se considera como um processo de especificação das propriedades desse conjunto de respostas ideais. O cálculo inicial nos permite gerar uma descrição probabilística preliminar do conjunto de respostas ideais que se utiliza para recuperar um primeiro conjunto de documentos. A intersecção com o usuário se inicia então com o propósito de melhorar a descrição probabilística do conjunto de resposta ideal(ver em [15] (cap: 2)).

Os modelos *borrosos* e *booleano* estendidos são modelos alternativos teóricos de conjunto. O modelo de conjunto *borroso* considera que cada termo de pergunta define um conjunto *borroso* e que cada documento tem um grau de pertencer a esse conjunto. Os modelos booleanos estendidos estendem o modelo com a funcionalidade do emparelhamento parcial e o pesado de termos.

O modelo generalizado de espaço de vector é uma variação do modelo do espaço de vector no qual os vectores de termo de índice se supõem que são linealmente independentes mas não são parelhas ortogonais (pairwise orthogonal). O modelo de indexação semântica latente traça cada documento e vector de pergunta num espaço dimensional mais baixo se associa a conceitos. Os vectores de termo de índice são traçados neste espaço dimensional mais baixo. A recuperação no espaço reduzido se espera que seja superior que a recuperação no espaço original dos termos de índice.

Os modelos de rede de inferência (inference network) e de rede de confiança (belief network) são variações do modelo probabilístico. O modelo de rede de inferência associa as variáveis por azar com os termos de índice, os documentos e as perguntas do usuário. O termo de índice e as variáveis de documento se representam como nós numa rede de trabalho e os bordes são dirigidos desde o nó de um documento aos nós do seu termo para indicar que a observação do documento produz uma confiança melhorada sobre os seus nós do termo. A variável de pergunta também se representa mediante um nó na rede de trabalho. Os bordes se dirigem desde os nós de termo de índice ao nó de pergunta. O modelo de rede de confiança é uma variação do modelo de rede de inferência que adopta um espaço de prova claramente definido. Pelo contrário que o modelo de rede de inferência, a topologia do modelo de rede de confiança proporciona uma separação entre o documento e as porções de pergunta da rede de trabalho.

O nosso objectivo não é cubrir os detalhes de todos os modelos propostos na teoria e nos referimos ao livro de Baeza-Yates e Ribeiro-Neto [15] (cap: 2) para uma completa descrição de todos os modelos importantes de IR.

4.3 Modelo de usuário

Um dos problemas mais importantes na recuperação de informação consiste em formular a consulta para que plasme adequadamente a necessidade informativa do usuário. Para além dos requerimentos do sistema para formalizar a consulta, o maior problema consiste em determinar o conjunto de palavras que expressem semantica-

mente essa necessidade. O problema se agrava devido ao efeito de inconsistência na atribuição subjectiva de termos a conceitos. Figuras como a *sinonímia* ou a *polissemia* (ou outras menos importantes, como a *homonímia*, *antonímia*, a *hiperonímia*, a *hiponímia*, etc.) fazem que o mesmo conceito possa expressar-se com palavras diferentes e uma mesma palavra possa aparecer em documentos que tratam sobre temas diferentes. Nesta situação não é de estranhar que o usuário tenha que replantar sua consulta para obter melhores resultados.

Um problema fundamental em recuperação de informação é que os autores nem sempre usam as mesmas palavras que os usuários para descrever o mesmo conceito [238].

A importância deste problema tende a diminuir com o aumento do tamanho da consulta. Entretanto, em muitas aplicações, as consultas podem possuir uma pequena quantidade de termos. Um caso extremo ocorre no contexto da *Web*, onde as consultas possuem tipicamente duas palavras. A expansão de consultas é um caminho para solucionar estes problemas.

Propuseram-se diversos mecanismos para construir a nova consulta. Em geral, em todos eles se realiza uma ampliação de novos termos à consulta inicial e um recálculo da importância de cada termo na nova consulta. Isto é o que se conhece como expansão de consulta. Pretende-se ampliar o número de termos que melhor definam a necessidade informativa do usuário de acordo à colecção documentária e ao modelo de recuperação utilizado.

O interesse para agilizar a expansão de consultas se centra em consultas com poucos termos, pois as consultas longas costumam proporcionar bons resultados de recuperação. De facto, a maior parte das consultas que se realizam em buscadores e sistemas de informação em Internet têm de um á três termos [234].

Para expandir consultas devem utilizar-se palavras ou frases com significados semelhantes àqueles da consulta inicial. A ideia é que se vários termos estão semanticamente relacionados entre si, quando um usuário está interessado num deles, provavelmente também o estará nos outros, e os documentos indexados com estes também serão relevantes para o usuário.

Duas abordagens podem ser adoptadas: o uso de dicionários de sinónimos e o uso de palavras que co-ocorrem com os termos das consultas em documentos da colecção. No caso de dicionários de sinónimos os resultados obtidos não são em geral muito bons. Melhorias consideráveis foram alcançadas quando se considerou

análise automática de termos que co-ocorrem em documentos da colecção.

Outro tema importante em expansão de consultas é a quantidade de termos adicionados à consulta. Nos experimentos apresentados em [80] o melhor desempenho foi alcançado com adições entre 20 e 40 termos, mas claramente este número depende da colecção utilizada.

Para desenvolver a expansão de consultas há que resolver três aspectos importantes:

1. Em primeiro lugar, há que estabelecer a relação entre termos. O imediato é utilizar tesauros ou dicionários em que aparecem relações entre termos.
2. Depois há que seleccionar, de todos os termos relacionados com os da consulta, quais são mais adequados para ser adicionados à dita consulta.
3. Por último, tendo em conta o sistema de recuperação utilizado, há que determinar o mecanismo de pesado dos novos termos, e isso depende do critério da selecção prévia.

Um dos principais problemas encontrados pelos investigadores dos sistemas de recuperação de informação é o facto das consultas submetidas aos sistemas serem normalmente compostas de poucos termos, os quais passam pouca informação a respeito da necessidade de informação dos usuários. Além disso, a falta de conhecimento a respeito do funcionamento dos mecanismos de busca por parte dos usuários dificulta a formulação de consultas que produzam os resultados satisfatórios ([154], [207]).

Uma maneira eficaz de contornar os problemas apresentados acima é utilizar métodos que expandem automaticamente as consultas dos usuários ([187], [77]). O objectivo dos algoritmos de expansão de consultas é formular uma consulta mais elaborada a partir da consulta feita inicialmente pelo usuário, transformando-a numa consulta mais elaborada. A expansão de consultas é útil porque a linguagem natural permite que as pessoas utilizem palavras e expressões diferentes para indicar um mesmo objecto. Por exemplo, se um usuário faz a consulta “Ministério da Educação”, um algoritmo de expansão de consultas poderia incluir a palavra “MEC” na consulta expandida, pois muitos documentos fazem referência ao Ministério da Educação através da sigla MEC. Da mesma forma, a consulta por “redes de televisão” poderia ser expandida com termos como TV, Globo, Record, SBT, dentre outros.

Podemos realizar uma **classificação** de técnicas de expansão dependendo de se requerem ou não da presença do usuário. Segundo este ponto de vista se distingue dois grandes enfoques:

1. Técnicas de realimentação de relevantes de consultas utilizando critérios de relevância do usuário (*user relevante feedback*). Requer uma boa (*interfaz*) com o usuário, mas é o mecanismo que melhores resultados proporciona. Também se utiliza em motores de busca em Internet, com a opção "páginas similares" ou "more like this".

2. Técnicas de expansão automática de consultas. Não requerem da presença do usuário. Em vários trabalhos de investigação se analisou as técnicas de recuperação de informação, de maneira que poderiam aplicar-se de forma automática aquelas que melhorassem notavelmente a recuperação sem precisar de um custo computacional elevado. São técnicas totalmente automáticas, em contraposição as outras que requerem a presença do usuário. A ideia é a seguinte: As consultas que realizam os usuários costumam ser muito curtas, como qual o nível de ambiguidade é elevado, e portanto, os resultados não são sempre os esperados.

Outros estudos encontrados na literatura mostram que técnicas de realimentação de relevantes normalmente produzem resultados melhores que técnicas de expansão automática [15] (cap: 2). Contudo, as técnicas de realimentação de relevantes costumam não ser bem sucedidas em sistemas de busca disponíveis na Web, pois os seus usuários raramente fornecem informação para realimentar o sistema [154]. Nestes casos, a melhor solução para expandir as consultas é através de técnicas de expansão automática.

A expansão automática de consultas pode ser classificada como **local** ou **global** [238]. Uma expansão de consultas é localmente quando se realiza uma análise das relações entre os termos nos documentos que constituem o conjunto-resposta da consulta original. Na expansão local é necessário que primeiro se processe a consulta original, pois se faça algum tipo de análise sobre o conjunto-resposta para formular uma nova consulta, que deve também ser processada pelo sistema. Estas características fazem com que a expansão local seja extremamente cara em termos computacionais, principalmente se o método de análise das relações entre os termos utilizado tiver custo computacional elevado. Ao passo que, a expansão glGolubobal é realizada estudando-se as relações entre os termos em toda colecção de documen-

tos. Assim, não há a necessidade de se processar a consulta duas vezes, e a análise das relações entre os termos da colecção pode ser pré-computada uma única vez.

O método de **indexação semântica latente (ISL)** [66] permite encontrar relações semânticas entre os termos da colecção. O modelo ISL utiliza uma técnica estatística para identificar conceitos existentes na colecção, conhecida como decomposição do valor singular [72]. O processamento de consultas com o modelo ISL pode ser considerado um método de expansão global de consultas, onde os documentos e as consultas são representadas pelos conceitos semânticos encontrados na colecção. Uma consulta formulada pelo usuário primeiro é transformada para o espaço de conceitos e depois processada. Assim, o conjunto-resposta pode ser composto por documentos que contenham palavras que não foram incluídas pelo usuário na consulta.

Apesar da sofisticação matemática do modelo ISL, os resultados práticos obtidos têm sido modestos quanto consideramos os custos envolvidos na identificação dos conceitos semânticos. Além disso, existe um custo extra que o uso do modelo adiciona ao processamento da consulta ([66], [48], [51], [20]). Um dos principais problemas encontrados no uso do modelo ISL é que a consulta original do usuário pode ser completamente descaracterizada no processo de transformação da mesma para o espaço de conceitos. Este risco ocorre principalmente em consultas curtas onde os termos originais podem estar relacionados a muitos conceitos distintos.

Segundo os experimentos realizados por outros investigadores, tiram proveito das relações semânticas entre termos obtidos através do modelo ISL e combina esta informação com a consulta original através de um esquema que evita uma completa descaracterização da mesma. Resultados experimentais mostram que este método de expansão de consultas produz melhores resultados do que os resultados obtidos com o ISL. Além disso, o tempo para processar consultas é consideravelmente menor que o tempo gasto para se processar consultas com o modelo ISL, puro, o que viabiliza a utilização do método em sistemas onde o desempenho ou a execução da recuperação é crítica, tal como os sistemas de recuperação de informação disponibilizados na Web.

No nosso caso estudamos cômo a expansão automática de consultas baseadas em tesauros e a desambiguação do sentido das palavras utilizando o recurso linguístico WordNet, com uma metodologia TREC de avaliação baseada em simulação, pode melhorar a Recuperação de Informação. Para tal se realiza o desenvolvimento de um sistema de recuperação. Finalmente dito sistema é avaliado para obter resultados

que permitam comparar os resultados obtidos em realizar as "consultas originais" com respeito às obtidas ao utilizar a expansão de consultas, desambiguação do sentido das palavras, ou ambas. Os resultados correspondentes mostram-se no Capítulo 6.

4.3.1 Descrição do modelo

Nesta secção se descreve o modelo que permite ao usuário formular os distintos tipos de consultas para que plasme adequadamente a sua necessidade informativa para a recuperação de informação, com o objectivo de obter melhores resultados e que permita comparar os resultados obtidos em realizar os distintos tipos de consultas.

Formulação dos distintos tipos de consultas:

Consideramos um tópico[2] TREC [218] com os seguintes campos T, D e N onde T é o título, D é a descrição e N é a narrativa, que se podem representar por T, D, N, considerando somente o título T se obtém as seguintes consultas:

i) **Consulta original**: Seja $\vec{q_1} = (A_0, B_0, C_0)$ o vector da consulta não expandida só com o título, onde A_0, B_0, e C_0 são termos pertencentes ao referido vector, nos quais A_0 e C_0 são nomes.

ii) **Consulta expandida**: Suponhamos também que a partir da expansão elegendo só os nomes que aparecem na descrição textual do tópico utilizando o recurso linguístico WordNet obtemos os termos seguintes: A_1, A_2 e A_3 como sinónimos relacionados com A_0 e C_1, C_2, C_3 e C_4 como sinónimos relacionados com C_0, então o novo vector da consulta expandida ficará da seguinte maneira: $\vec{q_2} = (A_0, B_0, C_0, A_1, A_2, A_3, C_1, C_2, C_3, C_4)$.

iii) **Consulta expandida e selecção dos significados correctos**: Suponhamos agora que na consulta: $\vec{q_2}$, se seleccionou os termos A_1, C_1 y C_2 como os correctos, a consulta ficará da seguinte maneira: $\vec{q_3} = (A_0, B_0, C_0, A_1, C_1, C_2)$ com isto se pretende que se numa consulta um termo original tem diversos significados, supõe-se que aqueles que são diferentes do correcto não influem na eficácia da recuperação. Este processo é análogo com qualquer categoria sintáctica.

[2]Exemplos de petições ou necessidades de informação, chamados tópicos na nomenclatura TREC.

Ordenação dos resultados

Uma vez aplicado o modelo do usuário, o processamento das consultas após a expansão e selecção dos significados correctos, é feito utilizando o modelo de espaço vectorial. Neste modelo, os documentos da colecção e as consultas dos usuários são representados como vectores de termos dentro de um espaço vectorial.

Suponhamos que cada documento e cada consulta são representados por um vector da frequência do termo $\vec{d} = (x_1, x_2, \ldots, x_n)$ y $\vec{q} = y_1, y_2, \ldots, y_n)$ respectivamente, onde n é o número total de termos, ou tamanho do vocabulário e x_i, y_i são as frequências do termo t_i em $\vec{d}$ e $\vec{q}$ respectivamente.

Os pesos dos termos nos documentos e na consulta podem ser calculados de várias maneiras ([64], [236]), dependendo das características da colecçao e do tipo de recuperação desejada. Os pesos adoptados nos nossos experimentos foram adoptados como segue:

Dada uma colecção C, a frequência do documento inverso (idf) de um termo é dada por $log\left(\frac{N}{n_i}\right)$, onde N é número total de documentos em C e n_i é o número de documentos com o termo i. Todos os termos numa consulta ou documentos são pesados pela fórmula de pesado TFIDF heurística. Isto é os vectores pesados por $\vec{d}$ e $\vec{q}$ são:

$$\vec{d} = (tfd(x_1)idf(t_1), tdf(x_2)idf(t_2), \ldots, tfd(x_n)idf(t_n)) \tag{4.5}$$

$$\vec{q} = (tfd(y_1)idf(t_1), tdf(y_2)idf(t_2), \ldots, tfd(x_n)idf(t_n)) \tag{4.6}$$

Além disto:

$Rawtf = tf(t_i, d)$: Indica a frequência do término t_i no documento d_j, e $Idf_i = log\left(\frac{N}{n_i}\right)$ é a frequência do documento inverso. Idf_i depende da colecção de documentos. Se t_i aparece em todos os documentos , logo $idf_i = 0$. Enquanto que se t_i ocorre somente num documento, logo idf_i obtém o seu valor máximo ($idf_i = log(N)$).

Finalmente o esquema de pesado tf/idf assigna pesos ao termo t_i no documento $\vec{d}$ como segue:

$$\mathcal{W}_{i,d} = Rawtf \times idf = Rawtf \times log\left(\frac{N}{n_i}\right) \tag{4.7}$$

A função *RawTF* da consulta está definida de forma similar como segue (nota-se que neste caso, *RawTF* depende das queries ou consultas e também dos documentso).

$$\mathcal{W}_{i,q} = Rawtf \times idf = Rawtf \times log\left(\frac{N}{n_i}\right) \tag{4.8}$$

Por último, se obtém a pontuação de um documento $\vec{d}$, dada uma consulta $\vec{q}$, calculada como segue:

$$S(\vec{d}, \vec{q}) = \sum_{i=1}^{n} tfd(x_i)tfq(y_i)idf(t_i) \tag{4.9}$$

Normalmente, uma lista dos z documentos com maior similaridade é retornada ao usuário. Mais detalhes sobre este modelo podem ser encontrados na literatura ([15](cap: 2), [236]).

Modelo usado na presente tese

Utilizando o modelo de usuário como o proposto previamente, é possível comparar a bondade de uma consulta frente a outras. Por exemplo, uma vez obtida a consulta expandida e a consulta com a selecção do significado correcto das palavras, o campo seleccionado é indexado utilizando as rotinas padrões (*estándares*) do Sistema de Recuperação de Informação Lemur, computa-se um *ranking* de documentos para cada consulta. Para cada consulta os documentos estão ordenados (*ranked*) em ordem decrescente de similaridade. A pontuação de um documento $\vec{d}$ com a consulta $\vec{q}$ é dada por uma expressão matemática encima representada, e posteriormente uma vez obtidos os resultados da consulta expandida e os da consulta expandida e desambiguada, sobre os diferentes experimentos contemplados se procederá à avaliação e a comparação entre elas com os da consulta sem expandir.

Capítulo 5

Desambiguação do sentido das palavras

Neste capítulo se descrevem os conceitos relacionados com a tarefa de desambiguação do sentido das palavras (*WSD*, por suas siglas em inglês). Nas primeiras secções se exemplifica o uso de *WSD* em tarefas do processamento de linguagem natural (PLN). Posteriormente se apresentam os principais enfoques utilizados em WSD, além disto, se descrevem os recursos linguísticos que se utilizam no processo de desambiguação. Finalmente, na última parte deste capítulo, se descrevem brevemente alguns algoritmos de aprendizagem automática comummente utilizados em WSD.

5.1 A importância da WSD

A necessidade de que um computador possa interpretar a informação correctamente e desta maneira realizar o seu trabalho em forma eficaz faz com que WSD tome um papel relevante dentro do PLN, pois esta é uma tarefa fundamental da qual dependem muitas outras tarefas. A tradução automática e a recuperação de informação são duas tarefas dentro do PLN que exemplificam a necessidade da resolução da ambiguidade dos sentidos de uma palavra para cumprir com os seus objectivos.

A *tradução automática* é uma das tarefas onde a WSD joga um papel muito importante. Esta tarefa pretende realizar automaticamente uma tradução de um texto, frase ou oração que está num idioma (idioma origem) ao idioma a eleger (idioma

destino); por suposto, isto não só consiste em intercambiar as palavras do idioma origem com as do idioma destino. Basicamente o processo de tradução automática requer duas fases: (1) o entendimento da linguagem original e (2) a geração da frase (*sentença*) traduzida do texto no idioma objectivo. Em ambas as fases é necessário resolver o problema da ambiguidade do sentido das palavras. A primeira fase, o problema pode ocorrer quando há uma palavra polissêmica no idioma original gerando possíveis traduções diferentes entre si. Um exemplo é a palavra polissêmica em francês "grille", que pode ser traduzida ao inglês como "railings"(carris), "bar"(barras), "gate"(ponte), "grid"(ventilação), "scale"(escala) e "schedule"(agenda) dependendo do seu contexto. Da mesma maneira ocorre na segunda fase, o problema surge quando existe mais de uma tradução possível para uma palavra que não é ambígua no idioma original. Por exemplo, a palavra em inglês "valley" se pode referir no idioma *gaélico* como "strath" (um vale com rio muito largo) ou "glen" (um vale localizado entre colinas escarpadas). Devido a estes problemas, o trabalho em WSD incide directamente nos resultados desta área (Bar-Hillel [91] e Wilks et. al. [230]). O benefício consiste na eleição das palavras mais adequadas do idioma destino, baseando-se no contexto que as rodeia no idioma origem, permitindo que a tradução seja mais apropriada.

Outras das aplicações mais importantes dentro do PLN, é a *Recuperação de Informação*. O propósito desta área é encontrar os documentos de interesse dentro de uma colecção de documentos. No entanto, um dos problemas que se enfrenta nesta área é a precisão na busca. Isto é, os documentos resultantes da busca podem conter documentos não desejados, bem como, deixar fora documentos pertinentes. Isto é devido à forma de realizar a busca. Basicamente, utilizam-se as palavras da consulta recopilando aqueles documentos que contenham ditas palavras. Como é de imaginar-se, existe a possibilidade de que estas palavras sejam polissêmicas. Para exemplificar o problema, suponha-se que se deseja buscar sobre "*java*" em Google e se encontrou 302.000.000 documentos em espanhol com este termo; entre os documentos resultantes se terá mais de um tema devido a que "*java*" é uma palavra polissemica. Entre os possíveis sentidos de "*java*" temos: o nome de uma ilha localizada no arquipélago em Indonésia; um tipo de café; uma linguagem de programação. No entanto, se se adiciona algumas palavras que estejam relacionadas com o sentido do termo desejado, o número de documentos será menor e com mais possibilidades de que o tema de documento encontrado seja o sentido objectivo. Na Tabela 5.1 se exemplifica os resultados da busca em Google utilizando os termos mencionados, a segunda coluna indica a quantidade de documentos onde os termos estão presentes. Por esta razão, a desambiguação do sentido de uma palavra melhorará os resultados

nas buscas de documentos, ao analisar e verificar o sentido adequado das palavras utilizadas para Web a busca.

Termos	**# documentos**
java	302000000
java programação	1270000
java café	1800000
java ilha	837000

Tabela 5.1: Exemplo dos resultados da busca em Google

5.2 Metodologia básica da WSD

WSD se pode definir como o processo da eleição do sentido mais adequado de uma palavra polissêmica ajudando-se do contexto que a rodeia [214]. O contexto são as palavras que se encontram no lado direito e esquerdo da palavra a desambiguar dentro do texto, frase ou oração. Em muitas ocasiões, o tamanho do contexto deve ser definido, isto é quantas palavras se incluirão para o processo da desambiguação. Por exemplo, para desambiguar a palavra "*gato*" na oração "*uma necessidade do gato é a comida*", usando um tamanho de contexto de três palavras em ambos os lados, o contexto do lado esquerdo está composto de "*uma necessidade do*", e o lado direito de "*é a comida*". As palavras do contexto ajudam aos métodos de WSD para encontrar relações ou padrões que caracterizem o sentido de uma palavra a desambiguar, neste caso as palavras "*comida*" e "*necessidade*" determinarão que o sentido adequado da palavra "*gato*" é o de um animal.

Para o presente trabalho é de fundamental importância, abordar a questão da polissemia e as dificuldades que esta apresenta para dicionários e sistemas computacionais. Pustejovsky levanta em sua obra intitulada *Lexical Semantics* [186] a seguinte questão: qual é a representação de um item lexical que o leva a assumir diferentes sentidos em diversos contextos na composição da semântica? Isto é, o que existe na representação de um item lexical que dá origem a extensões de sentidos e ao fenómeno da polissemia lógica.

Essa questão certamente é difícil de ser respondida, uma vez que "a ambiguidade lexical é um dos problemas mais difíceis em estudos de processamento computacional de línguas, e não é surpreendente que esteja no âmbito da investigação em semântica lexical. É certamente verdade que a maioria das palavras em uma língua tem mais

de um sentido, mas as maneiras como as palavras podem carregar sentidos múltiplos podem variar" [186].

De acordo com [186], a ambigüidade lexical é um fenómeno heterogêneo, com pelo menos três factores distintos que contribuem para o aparecimento contextual de sentidos para um determinado item lexical:

i) A ambiguidade contrastiva, que normalmente é resolvida pelo contexto e pelo conhecimento do discurso.

ii) A ambiguidade complementar (ou polissemia lógica) que seria resolvida pela composição no contexto sintáctico da sentença.

iii) A extensão de sentidos, mediados por regras lexicais e condições específicas relacionadas ao falante e ao contexto.

A **ambiguidade contrastiva**, tradicionalmente conhecida como homonímia, é a situação onde um item lexical é associado com pelo menos dois sentidos distintos e não relacionados", como nos exemplos abaixo citados por [186]:

(1) a. The judge asked the defendant to approach the bar.
b. The defendant was in the pub at the bar

Figura 5.1: Ambiguidade contrastiva

Foi convencionalmente assumido que homonímias como o par acima são distribuídas em contextos diferentes e não representariam um grande problema de desambiguação em um texto. Essa é a posição defendida por autores como [21] e [93], onde a interpretação heurística do contexto lexical ajuda a estreitar a tarefa de seleccionar os sentidos de palavras ambíguas.

Em [186] afirma que apesar dessa estratégia poder ser útil se se aplica à terminologia em domínios específicos, para palavras de frequência geral com sentidos comuns de usos e, possivelmente, sentidos especializados também, o problema é mais difícil. Por exemplo, a palavra "bar" tem pelo menos vinte e cinco sentidos diferentes na maior parte dos dicionários de língua inglesa.

Os sentidos contrastivos utilizados podem parecer facilmente distinguíveis, mas como [166] e [186] demonstram que, o domínio primário não é suficiente para desambiguar tais sentidos lexicais no discurso natural do dia à dia. O que seria preciso, de acordo com eles, é uma abordagem de base semântica para a selecção de sentidos, onde as palavras do contexto ajudam ajudam a realizar a desambiguação.

Na **ambiguidade complementar**, diferentemente da homonímia, os sentidos a seguir exibem uma polissemia complementar, onde as leituras alternativas são manifestações do mesmo grupo de sentidos à medida que ocorrem em diferentes contextos:

(2) a. The bank raised its interest rates yesterday. (the institution)
b. The store is next to the bank. (the building)
(3) a. John crawled through the window.
b. The window is closed.

Figura 5.2: Ambiguidade complementar

Segundo [186] afirma que nesses casos "a semântica tem, de alguma maneira, que explicar cômo um banco pode ser tanto uma instituição como um prédio, e cômo uma janela pode ser tanto uma abertura como um objecto físico". A conexão lógica entre os **sentidos lexicais** é o que motivou uma representação semântica mais rica para substantivos e adjectivos, conhecida como *qualia structure* [186]. O termo *qualia* se refere aos modos de explicação para o objecto em questão. As frases abaixo são exemplos do fenómeno da **extensão de sentidos**:

(4) a. I am parked out back.
b. Ringo squeezed himself into the parking space.

Figura 5.3: Fenómeno de extensão de sentidos

Essas frases na Figura 5.3 ilustram dois tipos de transferência referencial: um tipo de combinação entre sujeito e predicado em (a), onde é o carro que está parado, não o indivíduo, e uma combinação entre verbo e objecto, junto com uma não-identidade entre o antecedente e a anáfora na relação em (4b). Na contribuição atribuída a Nunberg para esse assunto [186], tais extensões de sentido são conhecidas como transferências de predicado (predicate transfers). Em particular, ele se diz contra uma análise metonímica, onde o sujeito I em (4a) e o objecto *himself* em (4b) são interpretados como *"my car"* e *"his car"* respectivamente. Sua posição é que existem condições pragmáticas licenciadas que permitem que o predicado estenda o seu sentido, onde é reescrito para seleccionar os sujeitos que estão presentes na sintaxe.

A polissemia não é, portanto, um fenómeno que possa ser considerado isolado. "É antes o resultado de operações composicionais na semântica, como composição, e de efeitos contextuais, como a estrutura de relações retóricas no discurso e limitações pragmáticas em co-referência" [186]. Por ser um fenômeno tão difícil de ser

abordado, a grande questão é desvendar como outros componentes no processo de interpretação da língua natural interagem com o léxico. A partir desse tipo de informação é que se torna possível desambiguar e determinar completamente a semântica de palavras em um contexto.

Retomando o trabalho de Mark [214], e o de Rada Mihalcea e Ted Pedersem [146] se apresenta nas secções subsequentes uma classificação destes enfoques para realizar a desambiguação, os quais se podem reunir em três grandes grupos:

1. Métodos baseados em conhecimentos
2. Métodos baseados em corpus
   2.1. Métodos supervisionados – corpus etiquetado
   2.2. Métodos não supervisionados – corpus não etiquetado
3. Métodos híbridos e Bootstrapping

Antes de continuar com a descrição dos enfoques, é importante aclarar os seguintes termos: *precisão*, *recall* e *cobertura*. Estes termos são medidas para determinar a eficácia dos métodos. Para avaliar este tipo de trabalho comummente se utiliza um corpus de referência, isto é, um conjunto de palavras polissêmicas em contextos específicos onde o sentido da palavra é marcado previamente por um Juíz.

- A *precisão* é a percentagem de palavras correctamente desambiguadas pelo sistema de WSD dado um corpus de referência (número de instâncias correctamente desambiguadas pelo sistema WSD/número de instância as quais o sistema de WSD propôs uma resposta).
- O *recall* é a percentagem de palavras que foram correctamente desambiguadas dentro do conjunto de todas as palavras de prova (número de instâncias correctamente desambiguadas/número total de instâncias no conjunto de prova).
- A *cobertura* é a percentagem de palavras às que o sistema WSD deu resposta (número de instâncias às quais o sistema WSD propôs uma resposta/número total de instâncias no conjunto de prova).

Por exemplo supondo que se está manejando um conjunto de prova com 100 palavras e o sistema WSD só é capaz de propor respostas para 75 palavras; e destas

75 só 50 palavras foram desambiguadas correctamente. A *precisão* neste exemplo é de $50/75 = 0.66$, e o *recall* é $50/100 = 0.50$ e pelo último a *cobertura* é de $75/100 = 0.75$.

5.2.1 Métodos baseados em conhecimento

A ideia básica destes métodos consiste em utilizar recursos externos para desambiguar as palavras, tais como dicionários, tesauros (dicionários que mostram as palavras relacionadas com o significado e sentido de uma palavra, como os sinónimos), textos sem nenhum tipo de etiquetado e inclusive a Web. O propósito destes recursos dentro de WSD é prover uma lista de significados, definições ou exemplos típicos sobre o uso das palavras. Os dicionários mais populares utilizados por estes métodos são conhecidos como *MRD (Machine Readable Dictionaries)*, porque a informação que está contida nestes dicionários pode ser utilizada por um computador. Alguns dicionários *MRD* são: *Longman Dictionary of Contemporary English (LDOC)*, *Collins English Dictionary (CED)* e WordNet. Na Tabela 5.2 se mostra um exemplo das definições de cada sentido da palavra em inglês "plant" obtidas desde WordNet.

1. Buildings for carrying on industrial labor, "they built a large plant to automobile".
2. A living organism lacking the power of locomotion.
3. Something planted secretly for discovery by another; "the police used a plant to trick the thieves"; "they claimed that the evidence against him was a plant".
4. An actor a situated in the audience whose acting is rehearsed but seems apontaneous to the audience.

Tabela 5.2: Definições dos sentidos da palavra "plant" em inglês, obtidas desde WordNet 2.1

Um dos primeiros trabalhos no uso do MRD é o de Lesk [Lesk86], quem destacou por seus resultados (50%-70%) de precisão em desambiguar os sentidos correctamente) usando um conjunto de exemplos pequenos manualmente etiquetados e as definições do dicionário "Oxford Advanced Learner´s Dictionary" para tratar de identificar o sentido mais adequado. Na Tabela 5.3 se descreve o algoritmo proposto por Lesk. O funcionamento se baseia em encontrar a quantidade de coincidências entre as palavras das definições de duas palavras que se desejam desambiguar.

Na Figura 5.4 mostra um exemplo de desambiguação das palavras em "*PINE*" e "*CONE*" usando o algoritmo de Lesk. O critério para eleger o sentido mais adequado é seleccionar o sentido onde exista o maior número de coincidências entre as palavras

1. Adquirir desde um dicionário MRD todas as definições dos sentidos das palavras a desambiguar.
2. Determinar as coincidências das palavras nas definições para todas as possíveis combinações dos sentidos.
3. Escolher os sentidos onde há maior coincidência.

Tabela 5.3: Algoritmo original de Lesk.

das definições de ambas as palavras. No caso do exemplo elegeram os sentidos 1 para "*PINE*" e 3 para "*CONE*".

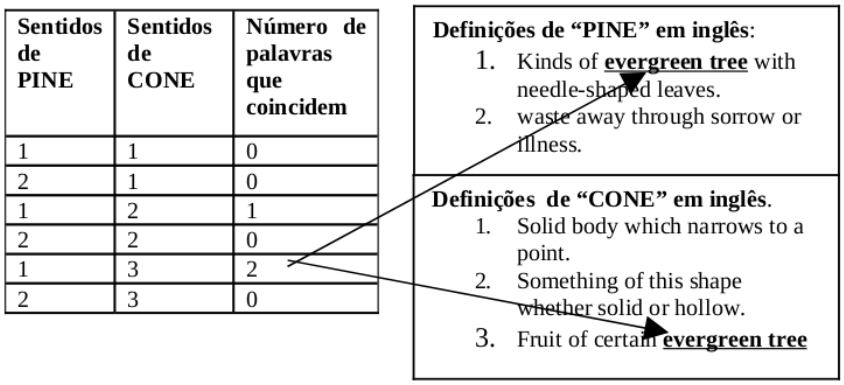

Figura 5.4: Exemplo do algoritmo original de Lesk

O problema do algoritmo de Lesk consiste em encontrar uma combinação adequada dos sentidos quando se trata de desambiguar mais de 2 palavras, ou bem, com um tamanho de contexto com mais de uma palavra num ou ambos os lados da oração ou frase. Na Figura 5.5 se exemplifica este problema, onde ao tratar de desambiguar as palavras polissemicas do texto mostrado no exemplo (9 ao total), produzem-se mais de milhões de combinações de sentidos. Encontrar uma combinação adequada pode levar muito tempo e fazer confusões.

Mas tarde Kilgariff e Rosenzweig realizaram um algoritmo simplificado de Lesk [113] para evitar o problema do algoritmo original, modificando o segundo passo do algoritmo original (ver Tabela 5.4), em lugar de buscar coincidências de palavras que ocorrem dentro das definições dos sentidos das palavras, só se realiza a busca entre as palavras da definição e as palavras que se encontram no contexto de uso da palavra a desambiguar. Na Figura 5.6 mostra um exemplo, onde a palavra objectiva é "pine", utilizando o algoritmo se determina que o sentido é 1 porque a palavra

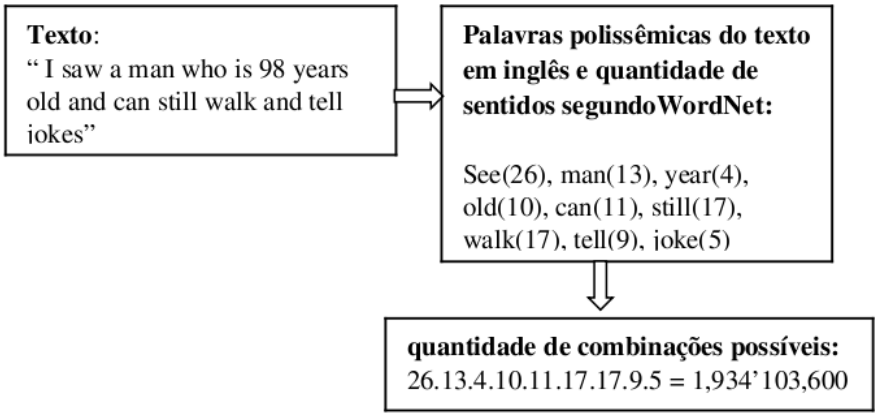

Figura 5.5: O problema de Lesk: a quantidade de combinações possíveis.

"tree" está presente no contexto e na definição.

1. Adquirir desde um dicionário MRD todas as definições dos sentidos das palavras a desambiguar.
2. Determinar as coincidências entre as palavras que se encontran nas definições dos sentidos e as palavras usadas no contexto que rodeiam a palavra a desambiguar.
3. Escolher o sentido onde há maior coincidência de palavras.

Tabela 5.4: Algoritmo simplificado de Lesk.

Nos experimentos realizados por Vasilescu, et, al. [221], mostra-se uma comparação entre o algoritmo original de Lesk e o algoritmo simplificado usando o conjunto de dados de Senseval-2 All-Words (ver Secção 5.3.2). Os resultados obtidos neste trabalho demonstram que o algoritmo simplificado é superior ao original: 58% de desambiguação correcta contra 42% de precisão gerada pelo algoritmo original.

Outro recurso potencial para este enfoque é a Web (veja-se o trabalho de Gonzalo et.al. [76] para mais detalhes) e entre os primeiros investigadores em usar a Web como um recurso léxico para desambiguar palavras se encontra Agirre et. al. [4]. O trabalho basicamente consistiu em criar colecções de documentos similares extraídos desde a Web, agrupadas pelas palavras que estão relacionadas com os sentidos das palavras a desambiguar (por exemplo, os sinónimos no WordNet), desta maneira, cada colecção representa um sentido. Depois se usaram estas colecções como fonte de conhecimento e compará-las com o contexto da palavra objectivo. Para realizar a comparação se utilizaram as colecções em lugar das definições do dicionário. Usando um conjunto de 20 palavras diferentes e 2444 palavras a desambiguar para prova, o

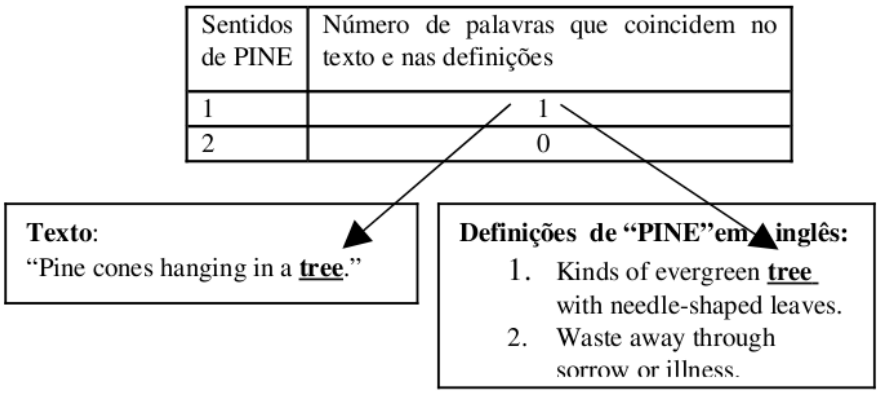

Figura 5.6: Exemplo do algoritmo simplificado de Lesk: Desambiguação do sentido das palavras "PINE".

resultado obtido nesse trabalho foi de 41% de precisão na desambiguação correcta dos sentidos. Por outro lado, Rosso et. al. [196] utilizou a Web para encontrar directamente co-ocorrências das palavras do contexto e dos sinónimos da palavra objectivo obtido em WordNet. Usando o corpus de Senseval-3 (ver Secção 5.3.2), o resultado deste método foi de 77% de precisão, no entanto o *recall* que se obteve foi de 33,7%. Apesar de ter estes resultados neste último trabalho deste enfoque, é inovador e promissor ao usar a Web como um conhecimento externo.

Os métodos deste enfoque possuem a grande desvantagem de depender da informação oferecida por um MRD. Construir estes recursos é muito laborioso e muito custoso. Se podem mencionar entre suas limitações as seguintes:

- A informação que possuem os dicionários é inconsistente (ver [94]), isto é, a quantidade de sentidos e as definições que possui uma palavra pode variar. Além disto, a maioria dos *MRD* são comerciais a excepção de WordNet.
- No algoritmo de Lesk e similares, a ausência ou a presença das palavras que há na definição do significado de um sentido pode afectar nos resultados, isto é, a qualidade das definições é de grande importância.

5.2.2 Métodos baseados em corpus

Nos últimos anos, com os avanços na área de Aprendizagem Automática (AA), tem crescido no PLN a utilização de métodos que permitem extrair conhecimento

automaticamente a partir de corpus, visando minimizar o problema de obstáculo da aquisição de conhecimento.

Um corpus provê um conjunto de exemplos que, quando submetidos a algoritmos de AA, permitem o desenvolvimento de modelos capazes de descrever esses exemplos e de predizer o comportamento de novos exemplos. Os trabalhos baseados em corpus, também chamados empíricos, realizam a desambiguação, portanto, com o uso de informações obtidas automaticamente a partir de um corpus. Nesses trabalhos, normalmente há uma etapa de treinamento, que resulta na aprendizagem do modelo de desambiguação, com base no corpus. Após o treinamento, o modelo de desambiguação é gerado (com excepção dos trabalhos baseados em instâncias) e pode ser usado para desambiguar novos casos de ambiguidades.

Alguns trabalhos de WSD tidos como baseados em conhecimento (manualmente codificado ou pré-codificado) utilizam corpus, mais especificamente, técnicas da linguística de corpus para extrair automaticamente algum conhecimento útil a ser utilizado no modelo de WSD, que extrai de um corpus um léxico de sequencias significativas de palavras que juntas criam um siginificado diferente, também chamadas colocações.

De modo geral, as principais vantagens dos trabalhos baseados em corpus são: (1) não é necessário codificar todo o conhecimento manualmente; (2) podem ser utilizados algoritmos tradicionais de AA, já implementados e disponíveis, para a aquisição automática ou semi-automática desse conhecimento; (3) os modelos criados podem ser mais facilmente generalizáveis para outros géneros e/ou domínios de textos e, com isso, aplicáveis em larga escala; e (4) os modelos gerados podem expressar algum conhecimento novo sobre o uso da palavra ambígua ou sobre as informações utilizadas para a distinção entre os seus sentidos.

Por outro lado, esses trabalhos também apresentam problemas: (1) os corpus utilizados para a criação do modelo precisam ser representativos da língua natural para o género e/ou domínio em questão. Apesar da actual disponibilidade de vários corpus de tamanho considerável, nem sempre eles apresentam as informações necessárias (como as etiquetas gramaticais das palavras); (2) no caso de trabalhos supervisionados, nas quais são necessários exemplos com as etiquetas de sentido das palavras ambíguas, o esforço para a criação do corpus é ainda maior, pois normalmente é preciso identificar essas etiquetas manualmente; e (3) não há garantia de que os resultados serão adequados, em função de várias características do processo automático de aprendizagem, como a possibilidade de inconsistências nos exemplos

(ruídos), a dificuldade de avaliação do modelo gerado.

Um corpus é um conjunto de textos recopilados, já seja de um mesmo tema ou variados. O propósito de um corpus é converter-se num conjunto de dados para prover exemplos de orações e exemplos de uso de várias palavras para ser utilizados em algoritmos de aprendizagem automática. Senseval-3 é um exemplo deste tipo de corpus (ver Secção 5.3.2 para a descrição detalhada deste corpus). Dependendo da natureza dos algoritmos e da tarefa, as palavras que há num corpus podem estar previamente desambiguadas ou não. Os métodos baseados em corpus podem subdividir-se em dois: (1) métodos supervisionados – corpus etiquetado; e (2) métodos não supervisionados – corpus não etiquetado.

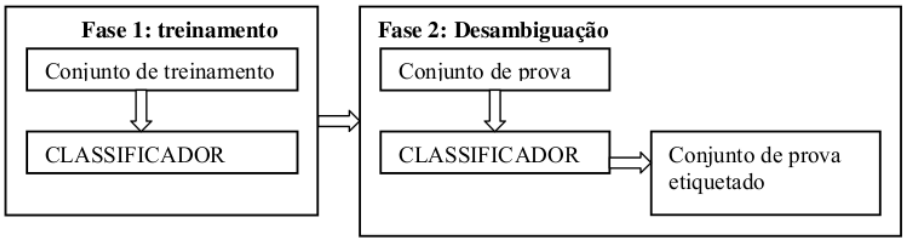

Figura 5.7: Esquema geral dos métodos supervisionados

Métodos supervisionados – corpus etiquetado

Neste enfoque, WSD se reduz a um problema de classificação onde se atribui a uma palavra objectivo o sentido mais apropriado dado um conjunto de possíveis combinações das palavras do contexto [146]. Em outras palavras, estes métodos utilizam classificadores (ou algoritmos de aprendizagem, ver Secção 5.4) para levar a cabo a desambiguação e geralmente se dividem em duas fases (ver Figura 5.7): o treinamento e a desambiguação dos sentidos (ou classificação em termos de aprendizagem automática).

- Na fase do *treinamento* é requerido um conjunto de dados etiquetados para o treinamento do classificador, isto é, um conjunto de palavras e o sentido para cada uma dessas palavras sob contextos de uso dados. (Veja-se a Secção 5.3.2 para um exemplo deste tipo de corpus *Senseval-3 English Lexical Simples*). Por esta razão, estes métodos são conhecidos como métodos supervisionados. O propósito do treinamento do classificador é permitir que o algoritmo encontre automaticamente relações ou padrões entre as palavras do conjunto de

treinamento para poder realizar a desambiguação dos sentidos dessas palavras.

- A *desambiguação*. Uma vez treinado o classificador, o classificador tentará desambiguar as palavras de um conjunto de dados etiquetados que estão destinados para a avaliação. A este conjunto se lhe chama conjunto de prova, onde a etiqueta será ignorada para propósito de classificação. Ao final da classificação se avaliam os resultados ao comparar os sentidos propostos pelo classificador e os sentidos correctos.

Em definitiva nos trabalhos baseados no modo supervisionado, um conjunto prédefinido de sentidos é especificado e cada exemplo do corpus é etiquetado com um desses sentidos. Com base nisto, os trabalhos baseados em corpus utilizam os algoritmos de aprendizagem para a desambiguação das palavras ambíguas, auxiliando-se do contexto das palavras ambíguas.

A principal vantagem dos trabalhos supervisionados é o facto de que os sentidos podem ser especificados previamente, provendo uma etiquetação mais adequada e refinada. Com isso, é possível criar modelos mais eficientes para aplicações multilingues. As duas edições do exercício de avaliação específico para a área de WSD (SENSEVAL), comprovam que as propostas supervisionadas são as que apresentam o melhor desempenho tanto para a tarefa de desambiguação de um pequeno grupo de palavras, como para a tarefa de desambiguação de todas as palavras da sentença (etiquetação de sentidos).

O problema com esses trabalhos é a necessidade de etiquetação de um corpus de treinamento, normalmente feita manualmente, por humanos. Esse problema acaba por restringir a abrangência de muitos trabalhos a poucas palavras, pois não há, ainda, corpus representativos com etiquetas de sentido, visando a uma ampla utilização para a WSD. Isso ocorre, muito provavelmente, porque o mesmo corpus não pode ser usado em diferentes aplicações que envolvam a WSD, uma vez que o nível de refinamento da etiquetação, bem como as etiquetas de sentido, propriamente ditas, dependem da aplicação e, em alguns casos, do domínio dessa aplicação.

Com relação ao domínio, [232] mencionam que um corpus etiquetado criado para um domínio específico (por exemplo, notícias de *sport*) dificilmente poderá ser utilizado para a geração de modelos em domínios distintos (por exemplo, produtos electrónicos). A solução, neste caso, parece ser a criação de corpus mais genéricos, cobrindo satisfatoriamente (ainda que não perfeitamente) diferentes domínios.

O compartilhamento de corpus entre aplicações é ainda mais complexo. Em aplicações monolingues, como a Recuperação de Informações, a etiqueta corresponde ao sentido, na língua do texto de origem, que distingue a ocorrência de cada palavra ambígua. Em aplicações multilingues, como a TA, por outro lado, a etiqueta corresponde à tradução, em outra língua, que distingue cada tradução da palavra ambígua.

Os conjuntos de dados (para treinamento e prova) para os métodos supervisionados costumam ser representados por **atributos**. Os atributos são características que contribuem informação sobre contexto para desambiguar o sentido das palavras. Os atributos comuns (ver [101] e [146]) são:

- **Co-ocorrências das palavras**. Estes atributos são palavras que estão dentro do contexto e ajudam a definir o sentido da palavra objectivo.
- **N-gramas**. Um n-grama é uma sequência de n palavras que provém de uma oração ou texto, no caso de WSD, do contexto. A ordem destas palavras é a mesma que se encontra no contexto. Por exemplo, da oração: "olá grande mundo"; "olá grande" e "grande mundo" são bigramas, que estão formados por duas palavras consecutivas.
- **Colocações**. As colocações são parecidas as n-gramas, são sequências significativas de palavras que juntas criam um significado diferente. Estas colocações podem ser similares quando se trata do mesmo sentido de uma palavra.
- **Etiquetas das partes da oração**. Em lugar de usar palavras que representam o contexto, identifica-se a categoria gramatical da palavra no fragmento (p.e.verbo, substantivo, advérbio, etc.). A categoria de uma palavra pode variar segundo o sentido da mesma. Por exemplo, na frase "a ajuda recebida" a palavra ajuda é um substantivo; pelo contrário, na frase "meu irmão ajuda a sua noiva" se trata de um verbo.

Entre os trabalhos relacionados com o métodos supervisionados se destaca o de Weiss [229] em 1974, quem demonstrou que é possível adquirir regras de desambiguação por meio de corpus etiquetado. Weiss trabalhou com um conjunto pequeno de 5 palavras e 20 orações para o treinamento e 30 orações para prova. Apesar do tamanho deste conjunto, os resultados foram alentadores (aproximadamente 90% da desambiguação foi correcta). Mais tarde, em 1998, Pedersen [174] realizou na sua tese uma comparação entre vários algoritmos para a aprendizagem supervisionada

usando um conjunto de dados para tratar de desambiguar 13 palavras. A conclusão deste trabalho mostrou que os algoritmos probabilísticos (Naive Bayes) (ver Secção 5.4.1 e variantes) tiveram mais sucessos do que os outros algoritmos de aprendizagem supervisionados com um resultado aproximado de 84% de precisão. Por outra parte G. Paliouras [169] realizou outra comparação de algoritmos supervisionados na que os algoritmos baseados em árvores de decisão por exemplo, C4.5 (ver Secção 5.5.2), foram melhores (82.6% de precisão e 77.4% de *recall*) para este último trabalho se usou um corpus com 3516 exemplos de 355 palavras, com uma média de 4.67 sentidos por palavras. Outros exemplos de métodos supervisados se podem encontrar em [26], [68], [128], [217] e [240]. Por último cabe mencionar, que em trabalhos recentes [146] se mostrou que este método de desambiguação supervisionada chegou a ser eficaz usando o corpus de SENSEVAL-3 (ver Secção 5.3.2) o qual possui exemplos sobre 57 palavras (7860 exemplos para o treinamento e 3944 exemplos para prova). Entre os melhores resultados usando este conjunto, foi o sistema desenhado por C.Grozea da Universidade de Bucarest usando o algoritmo de Naive Bayes. Este sistema gerou um resultado de 72.9% em precisão e 72.9% de *recall*.

Apesar dos bons resultados que obtêm estes métodos, sua principal desvantagem radica no uso de um corpus previamente etiquetado. Por um lado, a criação deste tipo de conjuntos de dados é de um alto custo em termos de tempo e de recursos humanos. Por outro lado, a cobertura do sistema sempre estará limitada ao conjunto de palavras do corpus de treinamento bem como aos sentidos etiquetados.

Métodos não supervisionados – corpus não etiquetado

Ao invés dos métodos supervisionados e os baseados em conhecimentos, os métodos não supervisionados identificam padrões nos conjuntos de dados sem o benefício dos dados etiquetados ou de outros recursos como são os MRD (ver [214] e [146]). Estes padrões são utilizados para dividir os dados em grupos, onde cada um dos membros de um grupo possui várias características em comum com o resto dos membros do mesmo grupo. Este enfoque se baseia na hipótese de que as palavras com significados similares tendem a ter contextos similares. Em outras palavras, este enfoque utiliza métodos que agrupam palavras baseando-se na similaridade do contexto e cada grupo representa um *sentido*. Ao não contar com informação sobre os sentidos de uma palavra a tarefa se converte em *Discriminação dos Sentidos da Palavra* e só possui um conjunto de dados não etiquetados como único recurso.

A metodologia geral deste enfoque tipicamente se baseia na selecção daquelas

palavras que se desejam discriminar com o seu respectivo contexto. A partir destas palavras seleccionadas se realizam as agrupações baseando-se no contexto das palavras seleccionadas. A Figura 5.8 ilustra este processo. Supondo que as formas são as palavras que se desejam agrupar, elege-se uma forma que represente as formas que se deseja para a agrupação (neste caso, um círculo, um losango e um quadrado), onde as formas que são iguais representam àquelas que têm um contexto similar. Uma vez eleitas as palavras, o método agrupa todas as palavras que sejam parecidas às palavras que se elegeram (no caso da Figura 5.8 são 3 grupos).

Em [173] se mostra um dos trabalhos de Discriminação dos sentidos das palavras de T. Pedersen, quem utilizou 13 palavras para criar 3 grupos (sentidos) obtendo um resultado médio de 65% de precisão, onde as palavras correctas se encontram nas agrupações apropriadas. Outros trabalhos sob este enfoque se podem apreciar em [185], [125] e [124].

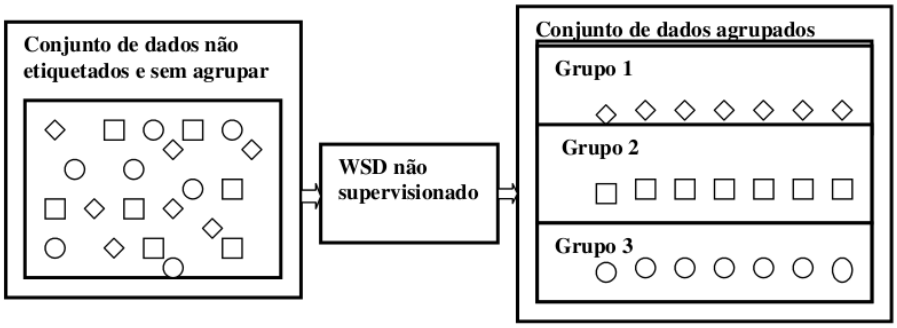

Figura 5.8: Esquema geral dos métodos não supervisionados

Nos trabalhos que utilizam métodos não-supervisionados, os exemplos do corpus consistem de características de sentenças (ou textos) cujas palavras são semanticamente ambíguas, podendo apresentar anotação de outras naturezas que não a de sentido, por exemplo, etiquetas gramaticais. Dessa maneira, não há sentidos específicos pré-definidos para uma palavra. Os algoritmos procuram identificar grupos de sentidos similares e, diante de uma nova palavra, verificar em quê grupos ela mais se assemelha.

De modo geral, a grande vantagem dos trabalhos não-supervisionados é o facto de que não há necessidade de etiquetação do corpus de exemplos. Por outro lado, uma vez que a desambiguação não é realizada com respeito a um conjunto pré-definido de sentidos, essa abordagem pode não ser apropriada para uma série de aplicações, principalmente para aquelas que necessitam de uma explícitação dos

sentidos. O conjunto de *clusters* e, com isso, o grau de generalidade/especialização dos sentidos, depende muito do algoritmo usado e dos seus parámetros. Os grupos obtidos podem ou não corresponder a diferentes níveis de uma hierarquia lexical padrão. Além disso, nem sempre os *clusters* de sentidos gerados são bem definidos.

Segundo [232], é difícil visualizar o que pode ser feito com os grupos de sentidos identificados pelos trabalhos não-supervisionados. Para a aplicação na tradução automática (TA), em especial, esses trabalhos dificilmente poderiam ser utilizados, pois como os sentidos correspondem às traduções, eles precisam ser explícitos. Certamente, isso poderia ser feito por meio de pós-edição humana, como ocorre no trabalho de [204]. Contudo, ainda assim não haveria garantia alguma de que os grupos formados corresponderiam a distinções de sentidos válidos.

De acordo [139] enfatizam o problema de que não há garantia de que os sentidos identificados distinguem adequadamente as ocorrências das palavras ambíguas, uma vez que eles podem ser muito genéricos ou muito refinados, ou, ainda, não corresponder às distinções padrão. Segundo [231], nesses trabalhos sempre haverá a atribuição das palavras ambíguas a algum grupo, de modo que a WSD sempre ocorre. O problema é como interpretar os resultados.

As vantagens deste enfoque são: A qualidade de agrupamento apesar da falta da informação dos sentidos das palavras em outros recursos e a independência para os dados etiquetados. No entanto, estas agrupações não apresentam os verdadeiros sentidos e podem variar também a quantidade de sentidos. Por tanto este enfoque não ajuda em tudo a desambiguar o sentido das palavras.

5.2.3 Métodos híbridos e Bootstrapping

Como se pode observar os métodos antes mencionados têm vantagens e desvantagens, uma maneira para reduzir as desvantagens é a combinação destes métodos resultando em métodos **híbridos**. Trabalhos híbridos para a WSD combinam características dos métodos baseados em conhecimento (codificado manualmente ou pré-codificado) e em corpus, podendo seguir modos supervisionados e não-supervisionados e empregar diferentes paradigmas de aprendizagem. Os tipos de conhecimento utilizados, bem como a interacção entre esse conhecimento e a aprendizagem automática (AA) podem variar. Um exemplo deste tipo de método é o trabalho de Luk [136], quem criou um sistema que utiliza informação das definições de dicionários (baseados em conhecimentos) e informação estatística que provê

de um corpus não etiquetado (Métodos não supervisionados). O resultado deste sistema foi de 77% de precisão e 100% de cobertura ao tratar de desambiguar 616 exemplos sobre 12 palavras diferentes. Outro exemplo é o trabalho de Mihalcea e Moldovan [143] onde a Web é utilizada como um conjunto de dados não etiquetados e ao mesmo tempo como uma fonte de conhecimento. Com ajuda de WordNet, neste trabalho extraíram exemplos desde a Web e foram etiquetados automaticamente os sentidos baseados em WordNet. Sob o juízo humano, o resultado reportado neste último trabalho foi de 92% de precisão de 658 exemplos revisados manualmente.

As vantagens dos trabalhos que seguem o método híbrido, em teoria, correspondem às vantagens dos trabalhos baseados em conhecimento e em corpus. Contudo, essa relação pode não ser tão directa: é preciso encontrar uma maneira adequada para combinar as características de ambos os métodos, de modo a minimizar seus problemas.

Bootstrapping é um enfoque que consiste em utilizar dois conjuntos de dados: Um etiquetado e o outro não etiquetado. O objectivo deste enfoque é etiquetar os sentidos das palavras que se encontram no conjunto de dados não etiquetados a partir de um pequeno conjunto inicial de dados etiquetados. O algoritmo proposto por Yarowsky em 1995 é um exemplo muito claro deste tipo de enfoque [241]. Usando 12 palavras com 2 sentidos cada um e com uma quantidade média de 3936 exemplos no melhor dos casos conseguiu obter 96.5% de precisão.

Alguns autores mencionam um terceiro modo de aprendizagem, denominado **semi-supervisionado ou pouco supervisionado**. Em determinados trabalhos, são consideradas semi-supervisionadas diversas variações desse modo de aprendizagem, incluindo o co-treinamento (*co-training*), o autotreinamento (*self-training*) e a aprendizagem activa (*active learning*), tanto para a tarefa de classificação quanto para a tarefa de *clustering*. Nos trabalhos que seguem esse modo de aprendizagem, o processo é iterativo: parte-se de um corpus com apenas alguns exemplos manualmente rotulados, que são usados para treinar o sistema para etiquetar novos casos. O objectivo é aumentar o número de exemplos de treinamento, adicionando os casos etiquetados a esse conjunto de exemplos, contudo, respeitando-se um limite mínimo de confiabilidade para a classe (ou *cluster*) atribuída. A verificação desse limite pode ser feita de diferentes maneiras. Nos trabalhos de co-treinamento para a classificação, por exemplo, são usados dois ou mais classificadores, treinados com diferentes visões (conjuntos disjuntos de características, por exemplo), sendo que todos eles (ou a maioria deles) precisam concordar com a classe atribuída para que o exemplo seja inserido no conjunto de treinamento. Esse processo, denominado muitas vezes

de bootstrapping, pode se repetir várias vezes, de modo a aumentar cada vez mais o número de exemplos de treinamento. Em alguns trabalhos, o objectivo é etiquetar a maior quantidade possível de exemplos. Em outros, o mais importante é garantir a confiabilidade na etiquetação dos novos casos.

Nos algoritmos de *clustering*, a aprendizagem semi-supervisionado também consiste em partir de um corpus de treinamento pequeno, com os exemplos já agrupados nos possíveis *clusters*, para atribuir outros exemplos, não rotulados, aos *clusters*. São usadas medidas de similaridade para identificar a proximidade entre os exemplos já anotados e os casos não anotados. Novamente, novos casos só são atribuídos a algum *cluster* se essa proximidade atinge um limite mínimo pré-estabelecido.

Analisando-se o processo de aprendizagem, tanto na tarefa de classificação quanto na de *clustering*, contudo, pode-se perceber que o treinamento ocorre sempre a partir de exemplos etiquetados, seja nas etapas intermediárias ou na etapa final, quando a etiquetação de todo o corpus de exemplos estiver concluída. Assim, os trabalhos que utilizam essas técnicas são consideradas como supervisionadas.

5.3 Corpus etiquetados manualmente e automaticamente

Para a criação de propostas baseadas em corpus, é preciso produzir um corpus de treinamento substancial. Conforme mencionado, os trabalhos dessa natureza que obtém os melhores resultados são os de aprendizagem supervisionado, que utilizam corpus anotados com sentidos, o que dificulta ainda mais a tarefa de criação do corpus. Para a desambiguação monolingue do inglês, já há alguns corpus com esse tipo de anotação. Já para a desambiguação em outras línguas e, principalmente, para a desambiguação multilingue, apenas recentemente começaram a ser desenvolvidos trabalhos de etiquetação de corpus.

Os corpus podem ser criados manual ou automaticamente. Aqui, é importante lembrar a distinção feita anteriormente entre as tarefas de WSD e a etiquetação de sentidos. As propostas de criação automática de corpus para a WSD realizam a etiquetação de sentidos.

5.3.1 Corpus etiquetados manualmente

Os principais exemplos de corpus já disponíveis e que são comumente utilizados para o treinamento e avaliação de trabalhos de WSD são os corpus DSO [164] e SEMCOR [151]. Ambos os corpus foram criados manualmente, para a desambiguação monolingue, utilizando os sentidos de WordNet.

O maior e mais significativo desses corpus é o DSO. Ele consiste de 192.800 sentenças de exemplo contendo 192.874 ocorrências dos 121 substantivos e 70 verbos mais frequentes da língua inglesa, extraídas do corpus Brown [62] e de um corpus de artigos do *Wall Street Journal*. Em média, cada verbo considerado possui 12 sentidos, enquanto cada substantivo possui 7.8 sentidos. Para cada palavra, foram extraídos até 1.500 exemplos.

Apesar da marcação manual, os autores estimam que o corpus apresenta de 10 à 20% de etiquetas que podem ser consideradas, por outros anotadores, como erros. Segundo os autores, esse corpus é bastante representativo, pois 191 palavras correspondem a cerca de 20% de todas as ocorrências de palavras em qualquer texto de inglês. O corpus SEMCOR [151] consiste de um subconjunto do corpus Brown com cerca de 200.000 palavras, das quais as palavras de conteúdo foram manualmente etiquetadas com os sentidos de WordNet. Outros corpus menores, mas também utilizados principalmente para a avaliação de sistemas, incluem os corpus criados para determinados trabalhos de WSD e disponibilizados para uso. Por exemplo, [128] e [35], cada um com pouco mais de 2.000 sentenças de exemplos com seis diferentes sentidos da palavra *line* e *interest*, respectivamente.

Outros exemplos são os corpus usados nas três edições do exercício de avaliação SENSEVAL. Com excepção da primeira edição, os demais corpus são baseados nos sentidos de WordNet. Contudo, como afirma [165], esses corpus, incluindo o DSO, são ainda muito pequenos para serem utilizados para a criação de trabalhos irrestritos de WSD. Com base no DSO, o autor examina o efeito do tamanho do corpus de treinamento, em termos do número de exemplos. Portanto, ele propõe um trabalho baseado em instâncias e realiza testes com vários subconjuntos do corpus, de modo a obter as curvas de aprendizagem nesse corpus. Os resultados do experimento mostram que a precisão aumenta à medida que o número de exemplos do corpus cresce e que todos os exemplos do corpus são efectivamente utilizados pelo algoritmo empregado.

Como conclusão desses experimentos, o autor estima que um corpus de 3.200

palavras diferentes etiquetadas com seus sentidos é suficiente para construir um sistema de Desambiguação léxica de sentido (DLS) de ampla cobertura e alta precisão, considerando-se qualquer palavra de conteúdo, em 32 textos irrestritos da língua inglesa. Assumindo uma média de 1.000 ocorrências etiquetadas por sentido por palavra, isso significa um corpus de 3.2 milhões de palavras etiquetadas.

Com base na sua experiência com a criação do DSO, segundo o autor, a criação manual desse corpus demoraria um tempo de 16 anos, considerando-se o esforço de um etiquetador humano. O autor sugere, como alternativa para minimizar o esforço de criação de corpus, o uso de técnicas de selecção de exemplos informativos, evitando a anotação redundante.

Focalizando a importância da selecção de exemplos relevantes na construção de corpus para a WSD, [65] empregam um método de amostragem selectiva de exemplos, de acordo com sua utilidade para o treinamento de um sistema de WSD. Eles desenvolvem um trabalho baseado em instâncias para a construção semi-automatizada do corpus. Apenas a desambiguação de um conjunto de verbos é contemplada.

O método apresentado pelos autores é de construção iterativa da base de exemplos de treinamento simula, em parte, a aprendizagem semi-supervisionado. É necessário um número inicial mínimo de exemplos manualmente desambiguados e um conjunto de exemplos não desambiguados de qualquer tamanho. Cada novo exemplo é submetido ao sistema, que atribui ao seu verbo uma etiqueta de sentido. Esse exemplo é então analisado pelo método de amostragem selectiva, para determinar a sua utilidade para o treinamento, com base *(a)* no conjunto de exemplos não desambiguados, analisando-se a quantidade desses exemplos que se assemelha a ele, de modo que ele possa cobrir um grande número de novos casos; e *(b)* no conjunto de exemplos já desambiguados e pertencentes à base de exemplos de treinamento do sistema, analisando-se a sua diferença com relação a esses exemplos, de modo a evitar exemplos redundantes. Os exemplos seleccionados por sua utilidade são então submetidos à revisão e/ou correcção humana da etiquetação de sentido realizada pelo sistema e, em seguida, acrescentados à base de exemplos de treinamento.

Em uma nova interação, esse exemplo já é utilizado pelo sistema. Os exemplos não seleccionados retornam para a base de exemplos não desambiguados. Outra alternativa para o problema da etiquetação manual que tem sido investigada ultimamente se mostra viável principalmente na TA é a etiquetação completamente automática dos sentidos dos exemplos.

5.3.2 Corpus etiquetados automaticamente

Segundo [5], a criação automática de corpus é uma das estratégias mais indicadas para minimizar o problema de obstáculo da aquisição do conhecimento, contudo, é ainda muito pouco explorada. Segundo [47], além de permitir a aquisição de corpus mais representativos, a etiquetação automática permite capturar distinções diferentes das que seriam atribuídas por um anotador humano, por exemplo, distinções específicas de algum domínio ou pouco comuns. Alguns dos trabalhos recentes de criação automática de corpus etiquetados são descritos a seguir.

Continuando o trabalho de exploração de corpus paralelos para a identificação dos sentidos a serem utilizadas na WSD monolingue iniciado anteriormente [101], [102] realizam experimentos mais significativos, estendendo o número de línguas (para sete), aumentando o número de palavras ambíguas e o tamanho dos corpus paralelos. Um algoritmo de *clustering* é utilizado para criar grupos de sentidos de acordo com as diferentes traduções de cada palavra do inglês, nas diferentes línguas. As distinções de sentido são, então, adquiridas a partir do corpus.

Para avaliar seu método, os grupos resultantes são comparados a grupos formados, sobre os mesmos dados, a partir da atribuição de sentido por juízes humanos. Essa comparação mostra que o algoritmo de *clustering* usando corpus paralelos provê distinções de sentido bastante refinadas, próximas das distinções feitas pelos juízes humanos. O único problema ressaltado pelos autores é o da falta de corpus paralelos substanciais entre várias línguas.

Em [52] também destaca a relevância de corpus paralelos para a extracção de informações semânticas. Segundo o autor, corpus paralelos são importantes fontes de informações semânticas. As traduções são, em muitos sentidos, fontes mais confiáveis que descrições de significado providas por um lexicógrafo ou semanticista.

Em [49] propõem uma abordagem para a criação de um corpus de exemplos etiquetados com sentidos para ser usado, posteriormente, em aplicações de WSD supervisionada monolingue. Nessa abordagem, são utilizados corpus paralelos bilíngues e um inventário de sentidos pré-definido para uma das duas línguas. O principal objectivo é a criação de um corpus substancial anotado com sentidos para a língua da qual se dispõe do inventário de sentidos. Como consequência da etiquetação desse corpus, é possível etiquetar também o corpus da segunda língua, utilizando o mesmo inventário de sentidos. Isso seria realizado por meio da identificação das traduções das palavras da primeira língua (as quais já possuem seu sentido atribuído)

nessa segunda língua. Os autores utilizam, como exemplo, a tradução entre uma língua-fonte e uma língua-alvo, sendo que o inventário de sentidos é válido para a língua-alvo. A língua cujos textos devem ser inicialmente etiquetados é, portanto, a língua-alvo.

No mesmo trabalho, para a criação do corpus, é utilizado um sistema de TA para gerar corpus paralelos entre as duas línguas. Em seguida, os textos paralelos são automaticamente alinhados por sentenças e por palavras. Esse alinhamento permite identificar, nos textos da língua-alvo, quais as traduções correspondentes a palavras da língua-fonte. As palavras da língua-alvo que são traduções de uma mesma forma na língua-fonte são, então, agrupadas. Para cada um dos conjuntos gerados, são considerados todos os possíveis sentidos para cada palavra.

A etiqueta de sentido adequada para cada palavra é atribuída de acordo com a sua similaridade semântica com as outras palavras no grupo. Apesar da facilidade na geração do corpus paralelo alinhado, Diab e Resnik ressaltam que esses corpus podem apresentar diversos erros decorrentes de traduções automáticas ou alinhamentos automáticos inadequados. Esses erros podem se propagar pelo processo de criação do corpus etiquetado e, certamente, influenciarão no desempenho dos trabalhos supervisionados criados utilizando tais corpus como base.

Em [5] descrevem o processo de criação automática de um corpus de exemplos etiquetados, focalizando a análise do desempenho desse corpus em um trabalho supervisionado de WSD e, principalmente, a análise do papel do *bias* de distribuição dos sentidos nesse corpus, ou seja, do número de exemplos para cada sentido de cada palavra.

O método de criação do corpus de exemplos etiquetados que empregaram é o proposto por [129], que se baseia nos "parentes" não-polissêmicos dos itens ambíguos para obter exemplos etiquetados com sentidos para esses itens. Para os testes, foram considerados como itens ambíguos 29 substantivos e os seus parentes não-polissémicos indicados pela WordNet. Os parentes, nesse experimento, são os sinónimos desses itens ambíguos. São então realizadas buscas na *web*, considerando sentenças de busca com os sinónimos não-polissémicos para recuperar exemplos contendo esses sinónimos. A suposição do método é de que para um determinado sentido da palavra ambígua, se for possível encontrar um sinónimo não-ambíguo desse sentido, então os 34 exemplos que contém esse sinónimo devem ser muito similares ao sentido da palavra ambígua e podem, portanto, ser usados para treinar um modelo supervisionado para aquele sentido da palavra.

Segundo os autores, uma característica que pode ser determinante na precisão e, principalmente, na cobertura de um trabalho supervisionado é o *bias* de distribuição dos exemplos para cada sentido no corpus. Para verificar o impacto desse *bias*, eles realizam diversos experimentos, incluindo: (a) nenhum *bias*, ou seja, considerando a mesma quantidade de exemplos para cada sentido; (b) o *bias* dos exemplos adquiridos automaticamente da web; (c) o *bias* do corpus SEMCOR, ou seja, considerando a mesma distribuição de sentidos desse corpus. Para os testes, é considerado o corpus de teste disponibilizado na segunda edição do exercício de avaliação SENSEVAL e um algoritmo supervisionado simbólico. Os resultados mostram que diferentes distribuições implicam diferentes resultados, sendo que os melhores resultados, principalmente em termos de cobertura, são obtidos a partir do *bias* do SEMCOR. Os resultados utilizando o *bias* automático de distribuição dos exemplos adquiridos da *web*, contudo, não são muito inferiores. Com relação à precisão do método de criação do corpus, considerando também o *bias* da *web*, os autores concluem que ela é maior que a de outros trabalhos da mesma categoria avaliadas no SENSEVAL.

Assim como [5] também apresentam uma estratégia para a criação automática de corpus baseada na formação de sentenças de busca a partir das definições e relações da WordNet e na busca de exemplos com essas sentenças em corpus ou na *web*. Cada *synset* a que pertence uma palavra na WordNet é caracterizado, por meio de suas relações com outros *synsets* ou palavras, como uma potencial sentença de busca. Contudo, os critérios para a construção das sentenças de busca são mais elaborados e flexíveis.

No trabalho de Agirre & Martínez, bem como em outras similares [129], a estrutura das sentenças de busca é definida previamente, por exemplo, ela é constituída sempre do contexto da palavra-alvo e de mais um sinónimo não ambíguo dessa palavra. Segundo [55], por outro lado, definem uma linguagem para especificação de padrões de sentenças de busca, de modo que várias estratégias de busca podem ser previamente definidas para formar diferentes sentenças para a busca nos corpus. Com isso, a proposta torna-se mais flexível e as buscas podem retornar um número muito maior de exemplos.

A linguagem criada inclui operadores lógicos, funções para indicar que parte da WordNet deve ser usada para extrair as palavras da sentença de busca (glosas, relações, etc.) e palavras, sentidos ou relações específicas da WordNet. Assim, como a definição de uma única estratégia, podem ser geradas diversas sentenças de busca.

Os autores realizam um experimento inicial considerando apenas o corpus do

SEMCOR, localmente armazenado, mas afirmam que o método pode ser usado em qualquer buscador *web*. São criadas *seis estratégias de busca*, sendo que algumas são baseadas nas estratégias usadas em outros trabalhos similares, como os citados. Essas estratégias são aplicadas às 73 palavras ambíguas usadas na segunda edição do exercício de avaliação SENSEVAL. As sentenças de busca geradas são então utilizadas para recuperar exemplos no SEMCOR. Cada estratégia envolve um possível sentido da palavra ambígua e as sentenças de busca mantêm esse sentido. Assim, os exemplos recuperados já possuem, automaticamente, uma etiqueta de sentido, neste caso, um sentido da WordNet.

O corpus SEMCOR foi utilizado justamente porque possui etiquetas de sentido, também de WordNet, assim, os sentidos atribuídos pelo sistema podem ser comparados com os sentidos originais. A precisão e a cobertura média de todas as estratégias são baixas, entretanto, os resultados também são apresentados considerando-se cada uma das estratégias, mostrando que algumas estratégias apresentam valores bem mais altos. Para todas as palavras, as sentenças de busca de todas as estratégias recuperam, em conjunto, 48.980 exemplos (não necessariamente todos corretos de acordo com sentido buscado). Esse pode ser considerado um número alto, já que o SEMCOR é um corpus relativamente pequeno. Se as buscas forem feitas em textos da *web*, certamente, esse número pode aumentar.

Vale notar que alguns dos trabalhos descritos, apesar de voltados para o processo de desambiguação, podem ser vistos, alternativamente, como propostas para a criação automática ou semi-automática de corpus de treinamento. Podem ser citados, por exemplo, [87], [204], [45] e [47]. Contudo, nesses trabalhos, os exemplos são etiquetados durante o processo de WSD. O seu objetivo não é, portanto, a criação de corpus de exemplos, mas uma alternativa de WSD parcialmente supervisionada.

Como se comentou previamente, nos trabalhos voltados para a TA, uma estratégia simples para facilitar a criação de corpus de exemplos é a utilização de textos paralelos entre as línguas, já alinhados no nível das palavras, ou submetidos à alinhadores de textos. Contudo, os alinhamentos precisam estar correctos, ou uma revisão manual posterior é necessária. Além disso, é necessário um conjunto substancial de textos para extrair um número representativo de exemplos. Para várias línguas, entretanto, não há grandes conjuntos de textos paralelos disponíveis. Por essas razões, essa estratégia é pouco explorada, ainda.

O problema dos dados escassos

O problema de dados escassos (*sparseness data*) é comum nos trabalhos baseados em corpus, principalmente nos supervisionados. Isto ocorre quando não há exemplos para todos os sentidos no corpus ou, ainda, há uma quantidade muito pequena de exemplos para alguns sentidos, que acabam se tornando estatisticamente insignificantes.

De modo geral, a escassez de dados indica que, em um grande espaço de interpretações alternativas produzidas por palavras ambíguas, somente uma pequena parte é utilizada. Esse problema, que é comum em todas as aplicações que utilizam corpus, é especialmente grave para a WSD. Primeiramente, quantidades muito grandes de textos são necessárias para tentar garantir que todos os sentidos de todas as palavras ambíguas consideradas sejam representados. Além disso, como os trabalhos de WSD são geralmente baseados em co-ocorrências da palavra ambígua com outras palavras, muitas das possíveis co-ocorrências são improváveis ou pouco frequentes mesmo em corpus muito grandes.

Algumas estratégias têm sido empregadas para a minimização desse problema. Em geral, elas procuram estimar a probabilidade de **co-ocorrência de sentidos** que não ocorrem nos dados de treinamento, de modo que essa probabilidade não seja assumida como nula.

Em [100] dividem essas estratégias em: (a) técnicas de suavização (*smoothing*) [69]; (b) modelos baseados em classe [178]; e (c) métodos baseados em similaridade [46]. Em [100] apontam para os métodos baseados em similaridade, destacando especialmente, o método de Dagan et al., como os mais elaborados, que apresentam os melhores resultados. Dagan et al. procuram estimar a probabilidade de co-ocorrência de sentidos inexistentes no corpus por meio da analogia entre cada co-ocorrência específica não observada e outras co-ocorrências que contém palavras semelhantes, de acordo com uma medida de similaridade entre as palavras.

Em [46] apresentam o exemplo da desambiguação da palavra *chapter*, que é sucedida pela palavra *describes* em uma sentença. No corpus de exemplos, não consta o par (*chapter*, section*describes*), mas constam os pares (*book*, *describes*) e (*section*, *describes*). São utilizadas, então, métricas de similaridade para indicar que *chapter* é similar a *book* e *section* e que, portanto, *chapter* deve ser utilizado no mesmo sentido que essas palavras, se co-ocorrer com a palavra *describes*.

O método proposto é avaliado em dois cenários, o primeiro considerando a de-

sambiguação para a TA, em um sistema desenvolvido pelos autores [45], e o segundo considerando a Recuperação de Informação [45], comparativamente a outros métodos de estimação [45]. No primeiro cenário, o método aumentou em 15% a cobertura com respeito ao segundo e possibilitou uma melhoria na precisão do mecanismo de escolha da tradução mais adequada. No segundo cenário, o método proporcionou uma estimativa 27% mais precisa que a estimação baseada em frequência [45] (suavização).

Outros autores usam alternativas mais simples para o problema dos dados escassos. [217], por exemplo, permitem que o classificador gerado não atribúa nenhum sentido a palavras para as quais ele não possui evidências suficientes para a classificação. Assim, uma das classes do sistema é denominada "*do not know*". Na sua avaliação, os autores relatam uma melhoria do sistema com essa simplificação frente ao método proposto em [45]. Certamente, em alguns casos, uma resposta como essa pode ser mais indicada que uma classificação incorrecta. Contudo, na maioria das aplicações, principalmente na TA, essa resposta é de pouca validade.

5.4 Recursos linguísticos

Nesta secção se descrevem os recursos linguísticos mais utilizados na tarefa de WSD: WordNet e Senseval-3 ELS. O recurso linguístico utilizado nos experimentos desta tese na tarefa da expansão e WSD é WordNet 2.1.

5.4.1 WordNet

Nessa secção parece pertinente mencionar uma base de dados lexicais (WordNet), que apesar de não ter sido elaborada com base em corpus, teve um grande impacto na área de lexicografia e acabou sendo utilizada em diversas investigações na área de linguística de corpus. Essa base de dados acabou surgindo como uma proposta de combinação de informações lexicográficas tradicionais com os recursos modernos e velozes da computação e será brevemente comentada nesse trabalho.

Nas últimas décadas, linguístas passaram a se interessar sobre as informações que o léxico deve conter para que os componentes lexicais, sintácticos e fonológicos trabalhem juntos na compreensão e produção diária de mensagens linguísticas, e estas propostas passaram a ser incluídas no trabalho dos psicolingüistas. Come-

çando com estudos de associações de palavras no início do século e continuando com sofisticadas tarefas experimentais dos últimos vinte anos, psicolinguistas descobriram muitas propriedades sincrônicas do léxico mental que podem ser exploradas em lexicografia [149].

WordNet é um dicionário MRD para o idioma inglês (ver [147], [150], [79]) convertendo-se num dos recursos mais valiosos para o processamento de linguagem natural (PLN). O desenvolvimento de WordNet iniciou em 1985 e foi criado no Laboratório de Ciências Cognitivas da Universidade de Princeton sob a direcção do Professor de Psicologia George A.Miller. Este recurso possui uma base de dados que agrupa as palavras em conjuntos de sinónimos chamados *synsets* e provê definições, comentários e exemplos de uso destas palavras e sentido das mesmas. Desta maneira, combina os elementos de um dicionário (definições e alguns exemplos) e os de um tesauro (sinónimos), e cria um apoio muito importante para a análise automática de textos e palavras.

Actualmente, a base de dados de WordNet 2.1 contém ao redor de 155.287 palavras organizadas em mais de 117.659 synsets formando um total a mais 206.941 definições e sentidos das palavras. WordNet maneja 4 categorias léxicas (ou tipos de partes da oração) nos seus synsets: substantivos, verbos, adjectivos e advérbios, como se mostra na Tabela 5.5. Esta ontologia inclui as características de um dicionário e a potência de um tesauro, para além disto, está disponível de maneira gratuita no idioma inglês.

Em [140] afirma que, diferentemente do que ocorre usualmente nos dicionários, que têm subjacente uma concepção estática do léxico, no WordNet os sentidos das unidades lexicais não são listados e definidos através de perífrases. Os sentidos são sempre inferidos a partir das relações que estruturam a rede. Em consequência, embora a cada conceito esteja associada uma glosa, com função idêntica à das definições clássicas dos dicionários convencionais, as glosas (não se pretende que sejam definições) constituem apenas informação complementar, não tendo qualquer papel na determinação do sentido" [140].

As palavras em WordNet se organizam em grupo de sinónimos (colocações) ou synsets, cada um dos quais representa um conceito léxico diferente. Cada synset contém a lista de palavras sinónimas, além da informação das relações semânticas estabelecidas com outras palavras ou synsets. Assim, em WordNet, as relações se estabelecem fundamentalmente entre conceitos, não entre palavras, assumindo-se que um conceito vem definido pelo conjunto de formas léxicas que, num contexto

Categoria sintáctica	Palavras simples	Synsets
Substantivos	117798	82115
Verbos	11529	13767
Adjectivos	21479	18156
Advérbios	4481	3621
Total	155287	117659

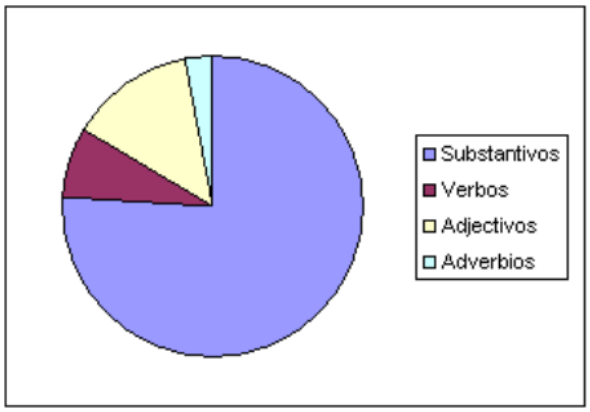

Tabela 5.5: Número de palavras e synsets em WordNet

apropriado, serve para representar a linguagem. Um sistema de desambiguação léxica que utilize WordNet como dicionário atribuirá a cada palavra ambígua um sentido de WordNet. Na Tabela 5.6 se mostra um exemplo típico onde se mostram os synsets da palavra "plant" como substantivo junto com suas definições de cada um dos sentidos desta palavra.

The noum plant has 4 senses (first 3 from tagged texts)

1. (338) **plant**, words, industrial plant - - (buildings for carrying on industrial labor; "they built a large plant to manufacture automobiles").
2. (207) **plant**, flora, plant life - - (a living organism lacking the power of locomotion).
3. (2) **plant** - - (something planted secretly for discovery by another; "the police used a plant to trick the thieves"; "he claimed that the evidence against him was a plant").
4. **plant** - - (an actor situated in the audience whose acting is rehearsed but seems apontaneous to the audience).

Tabela 5.6: Exemplos de synsets e definições da palavra "plant".

Além disto, a maior parte dos synsets estão conectados a outros sysents sob uma *rede de relações semânticas*. Entre estas relações estão as seguintes:

- **Sinónimos**. Palavras com significados idênticos ou similares.
- **Antónimos**. Palavras com significados opostos.
- **Hiperónimos**. Palavras que são mais gerais do que outras em seu significado. Por exemplo, "veículo" é o hiperónimo das palavras "trem", "automóvel", "aeroplano" e "motocicleta".

- **Hipónimos**. Palavras que são mais específicas do que outras em seu significado. Por exemplo, "tulipa", "rosa" e "girassol" são hipónimos da palavra "flor".
- **Holónimos**. Palavras que denotam ser uma parte ou membro de um todo (denotado por outra palavra). Por exemplo "carro" é o holónimo das palavras "aro", "motor" e "volante".
- **Merónimos**. Palavras que denotam ser um todo das palavras que denotam ser uma parte ou membro desse todo. Por exemplo "aro", "motor" e "volante" são merónimos de "carro" .

A modo de exemplo, na Tabela 5.7 se mostra os hiperónimos da palavra "plant" para o seu segundo sentido (planta no seu sentido de flora) desde WordNet, como se pode apreciar na Tabela, o hiperónimo mais geral é "entity" (entidade).

Sense 2
plant, flora, plant life - - (a living organism lacking the power of locomotion)

⇒ **organism, being** - - (a living thing that has (or can develop) the ability to act or function independently)

⇒ **living thing, animate thing** - - (a living (or once living) entity)

⇒ **object, physical object** - - (a tangible and visible entity; entily; an entily that can cast a shadow; "it was full of rackets, balls and other objects)

⇒ **entity, physical thing** - - (that which is perceived or know or inferred to have its own physical existence (living or nonliving))

Tabela 5.7: Exemplos de synsets hiperonímicos e as suas definições sobre o segundo sentido da palavra "plant"

Em [152], ilustram o conceito de matriz léxica, onde as formas léxicas são representadas como uma listagem de encabeçados de coluna. Nesta representação um synset é o resultado de cruzar uma fila da matriz de um lado a outro e atribuir um número arbitrário ao conjunto de palavras obtidas. Este número actuará como um identificador do conceito representado pelo conjunto de elementos léxicos que o designam. Esta representação se mostra na Tabela 5.8, onde a entrada $E_{1.1}$ implica que a forma léxica F_1 pode usar-se para expressar o significado M_1. Se há duas entradas na mesma coluna, a forma léxica é polissêmica; se há duas entradas na

mesma fila, as duas formas léxicas são sinónimas. Isto nos dá acesso à informação de duas maneiras diferentes, a primeira é acedendo a uma coluna e ir baixando até ao fim, desta forma obtemos todos os sentidos que uma palavra pode ter em diversos contextos. A segunda maneira em que temos acesso à informação é aceder por uma fila e seguí-la até ao fim, deste modo obteríamos todas as maneiras possíveis de expressar um determinado conceito. Assim, a matriz de vocabulário contempla dois dos principais problemas da semântica léxica: a polissemia e a sinonímia respectivamente.

Significados	Formas Léxicas				
	F_1	F_2	F_3	...	F_n
M_1	$E_{1,1}$	$E_{2,2}$			
M_2		$E_{1,2}$			
M_3			$E_{3,3}$		
.					
.				..	
M_n				...	$E_{m,n}$

Tabela 5.8: Matriz de vocabulário de WordNet.

1. **car**, auto, automobile, machine, motorcar – (4-wheeled motor vehicle; usually propelled by an internal combustion engine)
2. **car**, railcar, railway car, railroad car – (a wheeled vehicle adapted to the rails of railroad)
3. cable car, **car** – (a conveyance for passengers or freight on a cable railway)
4. **car**, gondola – (car suspended from an airship and carrying personnel and cargo and power plant)
5. **car**, elevator car – (where passengers ride up and down)

Figura 5.9: Sentidos da palavra car em WordNet 2.1.

As relações as podemos ver por **sentido**, por exemplo a sinonímia define os conceitos ou synsets (ver Figura 5.9), isto é, car-1 é sinónimo de auto, automobile, machine, motorcar; enquanto car-2 é sinónimo de railcar, railway car e railroad car. Enquanto a taxonomía os ordena de **forma hierárquica** (ver Figura 5.10), a seguir se mostram as relações de Hiponímia para o sentido 1 de "car", nas quais podemos ver que car se encontra relacionado taxonomicamente por hiponímia sucessiva com "motor vehicle", "vehicle", "artifact" e "object" entre outros.

Os nós superiores da estrutura taxonómica nominal constituem um conjunto de aproximadamente 30 conceitos com os que qualquer entidade do modelo do conhecimento léxico está relacionada (entidade, abstracção, lugar, forma, estado, evento,

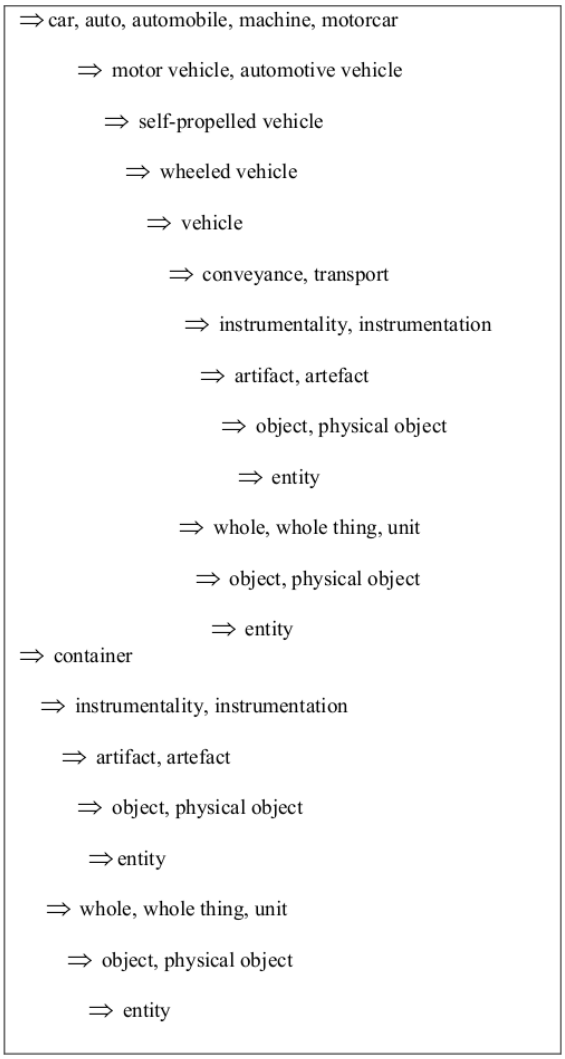

Figura 5.10: Relações de hiponímia para o sentido da palavra "car".

grupo, etc.). Na Tabela 5.9 se mostram as relações existentes em WordNet e exemplos de cada uma delas.

Relação	**Categorias em que se aplica**	**Exemplos**
Sinonímia	substantivos, verbos, adjectivos advérbios	*rápido/veloz*
Antonímia	substantivos, verbos, adjectivos, advérbios	*veloz-lento*
Hiperonímia-Hiponímia	substantivos	*carro-veículo* hiponímia *veículo-carro* hiperonímia
Meronímia-Holonímia	substantivos	o *volante* é parte do *carro* Meronímia um *carro* tem um *volante* holonímia
Implicações	verbos	*roncar-dormir*
Similaridade	adjectivos	*positivo-bom*
Atributo/Valor	Substantivos-adjectivos	*altura-alto*

Tabela 5.9: Relações existentes em WordNet.

Durante os últimos anos se desenvolveram recursos similares a WordNet, para outras línguas. Em concreto, o projecto EuroWordNet[1], que finalizou em 1999, teve como objectivo a construção de uma base de dados léxica multilíngue para vários idiomas europeus (alemão, checo, estonio,espanhol, francês, holandês e italiano).

Como consequência da construção de WordNets de outras línguas, foi criada uma associação com objectivos mais amplos, conhecida como Global WordNet Association (www.illc.uva.nl/EuroWordNet). Essa associação tem a proposta de criar uma rede mundial de interligação de todas as WordNets existentes. Uma proposta muito interessante que, segundo [140] "contribuirá de forma significativa para conferir estatuto de igualdade a nível científico, técnico e sóciopolítico das línguas representadas em tão importante projecto."

Em EuroWordNet, cada base de dados se estrutura de forma similar ao WordNet inglês, estabelecendo relações semânticas entre as diferentes palavras. Mas além disto, os diferentes WordNets estão enlaçados entre si mediante o índice Inter-Lingual-Index baseado no WordNet 1.5. Continuamente se estão melhorando e ampliando as diferentes bases de dados léxicas das línguas já existentes e desenvolvendo novas bases de dados para línguas que ainda não contam com um WordNet.

[1] http://www.illc.uva.nl/EuroWordNet/

Se encontra disponível na Web[2] um browser de WordNet que permite encontrar as relações léxicas entre palavras (sinonímia, Hiponímia, hiperonímia, etc.). A Figura 5.11 mostra a imagem do sítio Web.

É possível afirmar que o projeto WordNet tornou-se um dos recursos de maior impacto no domínio do Processamento das Línguas Naturais da actualidade. As unidades lexicais são organizadas em conjuntos de sinónimos que têm subjacente um conceito, sendo cada um desses conjuntos representados por um nó na rede. As ligações entre os diferentes nós exprimem relações semânticas diversas, como por exemplo, as de generalização e de especialização. O sentido emerge, assim, do complexo de relações que a rede permite exprimir [140].

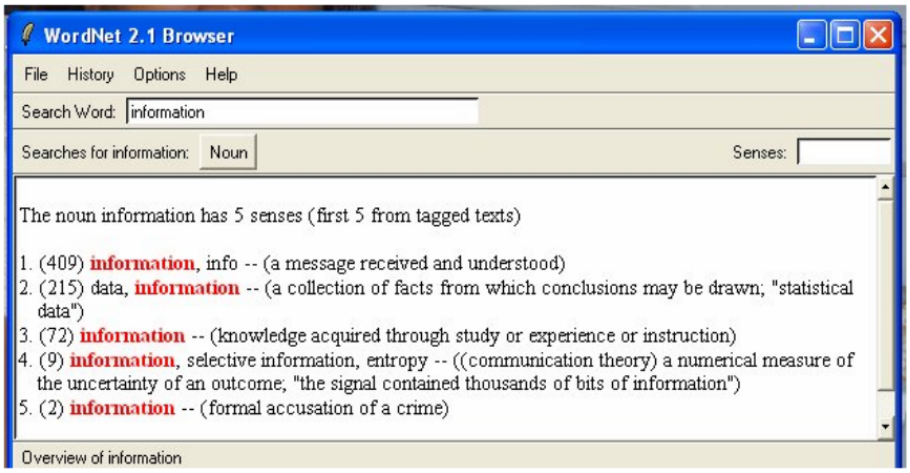

Figura 5.11: Sítio Web de WordNet para obtenção de relações léxicas entre palavras.

WordNet não é em absoluto um recurso perfeito para desambiguar o sentido das palavras devido a que a granularidade para a distinção de significados e as divisões de um sentido são em ocasiões demasiado finas para o propósito de muitos trabalhos de PLN. Isto ocasiona uma multiplicação desnecessária de sentidos além da falta de consistência nas aplicações de certas relações semânticas. Estas são algumas das razões que criam muitas dificuldades à hora de desambiguar o sentido das palavras automaticamente, devido a que há que fazer eleições quanto ao significado muitas vezes difícil inclusive manualmente.

Para os substantivos (uma das partes de WordNet usada na avaliação do sistema), as relações léxicas incluem sinonímia, antonímia, hiperonímia/hiponímia (relação é-um) e três relações diferentes do tipo meronímia/holonímia (relação parte-de).

[2]http://wordnet.princeton.edu/w3wn.html

A relação *é-um é* a relação dominante e organiza os conjuntos de sinónimos num conjunto de aproximadamente dez níveis hierárquicos.

O desenho básico de WordNet se baseia numa matriz léxica. Para sua construção se tem em conta que a palavra é uma associação por convenção entre um conceito léxico e uma forma física que joga um papel sintáctico. Se chamará, para evitar confusões no que segue, *forma de palavra* à sua morfologia, isto é, como soa ou se escreve e significado da palavra ao conceito léxico que pode expressar uma determinada forma.

A relação entre significados e formas de palavras é muitos a muitos, algumas formas têm vários significados, e alguns significados se podem expressar de diferentes formas. É o que se conhece polissemia e sinonímia respectivamente. Desta maneira a matriz léxica usada em WordNet é uma matriz formada por tantas colunas como formas de palavras inclui WordNet e por tantas filas como significados diferentes inclui WordNet. De maneira que se poderiam, em princípio, procurar palavras por seu significado ou por sua forma.

A relação de **sinonímia**, isto é, o parecido entre significados, é a mais importante em WordNet. A Habilidade para determinar esta relação entre palavras é um pré-requisito para a representação de significados na matriz léxica. De acordo com uma definição (atribuída a Leibntz) duas expressões são sinónimos se a substituição de uma por outra nunca muda o valor da verdade de uma frase na que se faz a substituição. Segundo esta definição, os autênticos sinónimos serão raros, se é que existam. Uma versão mais débil da definição, que é a utilizada, faria aos sinónimos relativos a um contexto duas expressões são sinónimos num contexto linguístico C se a substituição de uma por outra em C não altera o valor da verdade. Esta é a definição de sinonímia em termos de substituibilidade o que faz necessário partir WordNet em nomes, verbos, adjectivos, e advérbios. Isto se explica porque, se os conceitos estão representados por synsets, e se os sinónimos devem ser intercambiáveis, então as palavras em categorias sintácticas diferentes não podem ser sinónimos (não podem formar synsets) porque não são intercambiáveis. A sinonímia é o único tipo de relação que se usou na experimentação à base de expandir as consultas.

A outra relação contemplada na aplicação, das presentes em Wordnet, é a **hiponímia**, que é como já vimos uma relação semântica entre significados de palavras. Por exemplo, arce é um hipónimo de {árvore } e {árvore} é um hipónimo de {planta}. De maneira mais geral um conceito representado mediante o synset {x, x´, ...} diz-se que que é um hipónimo do conceito representado pelo synset {y,

y´, ...} se são correctas as frases construídas com os referidos conjuntos da forma Um x é uma (classe de) y. Por isso a relação de hipónimia também chama-se de subordinação ou subconjunto.

A hiponímia é transitiva e assimétrica e, como, normalmente, só há um único superconjunto, gera uma estrutura semântica hierárquica, na que se diz que hipónimo está embaixo do seu superconjunto. Estas representações hierárquicas se usam muito na construção de sistemas de recuperação da informação, onde se chamam sistemas de herança; um hipónimo hereda todas as características do conceito mais genérico e adiciona ao menos uma característica que os distingue do seu superconjunto e de qualquer outro hipónimo desse superconjunto, pelo que o nome de herança é bastante adequado.

A aplicação desenvolvida neste trabalho contempla a expansão de consultas mediante hipónimos ainda que finalmente se decidiu não os utilizar na experimentação.

Outra classe de relações léxicas são as **relações morfológicas** entre formas de palavras. Inicialmente, o interesse se limitou às relações semânticas; e não teve planos para incluir relações morfológicas em WordNet. Mas à medida que se avançou no trabalho se viu de forma clara que WordNet não se poderia utilizar de maneira prática se não tratava com morfologia. Por exemplo, se alguém busca em WordNet a palavra árvores, WordNet não deveria replicar que a palavra não está na sua base de dados, senão que um programa deveria eliminar o sufixo do plural e buscar árvore se está na base de dados.

Apesar que a morfologia do inglês é relativamente simples, um programa que realize este *stemming* é bastante complexo [182], principalmente com os verbos devido a que têm quatro formas e há muitos irregulares. No nosso caso, na experimentação se utilizou esta análise morfológica interna de WordNet. Ainda que a aplicação dê a possibilidade de realizar buscas sem aplicar nenhum tipo de transformação às palavras, o número de palavras encontradas assim é bem mais baixo e, por tanto, o número de termos que se podem expandir muito menor.

Para este acesso às diferentes funcionalidades de WordNet como buscas de palavras, e o acesso às diferentes relações léxicas e semânticas de WordNet existem multidão de interfases para diferentes linguagens de programação, ainda que não façam parte do projecto de WordNet em si mesmo. No caso desta se desenvolveu uma aplicação em Java se utilizou a API de Java. Além disto, está disponível de maneira gratuita incluindo o seu código fonte em http://www.cogsci.princeton.edu/ wn/.

Entre as possibilidades que oferece este APIs se encontram em primeiro lugar as funções de busca de palavras. As buscas se realizam indicando a que taxonomia pertence à palavra que se vai buscar, isto é, se se procura um nome ainda que existam verbos com a mesma forma de palavra a busca só devolverá os dados da palavra como substantivo. A busca devolverá além da própria palavra todos seus significados existentes na base de dados. Existem duas modalidades principais para estas buscas: (i) A primeira delas realiza uma busca simples de palavra tal qual sem aplicar nenhum tipo de algoritmo de *stemmer* nem processado morfológico.

Utilizando este método as palavras encontradas são poucas e a expansão pobre pelo que não foi a opção escolhida agora de avaliar o sistema. (ii) A segunda função se utiliza os diferentes processados morfológicos internos de Wordnet antes de buscar, pelo que a maioria das palavras, independentemente da forma em que estejam (plurais, gerúndios, etc.), são encontradas dando lugar a maiores possibilidades a expansão.

Para recuperar os conjuntos de sinónimos também se usaram as funções correspondentes de APIs de Java tendo em conta que os synsets dependem não da palavra em si, senão que para cada significado da palavra terá um synset diferente.

Para cada um destes significados se podem recuperar também outra série de relações como o é a hiponímia. Há que ter em conta que a relação de hiponímia é transitiva, e por isso é possível recuperar não só os descendentes directos de um termo senão também outros hipónimos não directos a mais de um nível de distância. No entanto, só se consideraram de utilidade para a expansão os hipónimos do primeiro nível e de facto, como se comentará no capítulo da avaliação, não se utilizaram finalmente os hipónimos à hora de realizar a experimentação devido a suas características em WordNet apesar de estar contempladas na aplicação desenvolvida.

Apesar de algumas limitações de WordNet (por exemplo as inconsistências nas definições e falta de relações semânticas [79] para mais detalhes), converteu-se num recurso indispensável para WSD, sobretudo para os enfoques baseados em conhecimentos e híbridos, já que conta com informação sobre os sentidos das palavras. Os trabalhos de Rosso e Mihalcea mencionados nas Secções 5.2.1 e 5.2.3 respectivamente, são exemplos de uso deste recurso.

A grande inovação do WordNet está no facto da rede não apenas diferenciar os conceitos entre si, como acontece nos dicionários tradicionais, mas de codificar toda a informação necessária para sua construção, organizando-a numa rede de relações que, ao serem investigadas, é capaz de fazer uma correspondência com a

organização do léxico mental. Portanto, como afirma [140], "a construção de uma WordNet assenta num pressuposto fundamental: o significado das unidades lexicais é basicamente derivado das suas relações com o sentido de outras unidades lexicais". A existência de uma estreita relação semântica entre as unidades lexicais em que este tipo de relação se regista justifica igualmente a sua especificação, por ser de grande utilidade em diversos domínios da Engenharia de Linguagem.

Uso de WordNet nesta tese.

O trabalho realizado nesta tese doutoral explorou o potencial de WordNet para esta tarefa. As vantagens de utilizar WordNet são numerosas. Primeiro, é facilmente disponível e foi desenhado para a manipulação do computador. No seu núcleo, WordNet é um gráfico de palavras conectados de acordo a várias relações que WordNet tem definido entre dois sentidos de palavras diferentes. Esta estrutura tem o potencial para discriminar o sentido da palavra em documentos e consultas e emparelhar as palavras relacionadas semanticamente.

O método de desambiguação utilizado neste trabalho é o baseado em dicionários para provar o efeito que a desambiguação do sentido das palavras pode ter sobre a recuperação da informação e implementar uma aplicação que utilize as características de WordNet, com uma tentativa de melhorar os resultados de recuperação utilizando o sistema de recuperação da informação LEMUR.

5.4.2 SENSEVAL

Como se observou nas secções prévias os resultados dos diversos enfoques são difíceis de comparar entre si. A avaliação destes métodos não só depende do enfoque mesmo (têm-se ou não exemplos de treinamento e/ou exemplos de avaliação) também depende do número de palavras a desambiguar e do algoritmo de aprendizagem utilizado, entre outros. Assim a comparação objectiva dos resultados atingidos por cada método proposto é uma tarefa difícil. Daí que se tenham proposto foros de avaliação, que sob um mesmo esquema de avaliação comprova o desempenho dos diferentes métodos propostos.

Em 1998, Adam Kilgarriff propôs um esquema de avaliação para os sistemas WSD: SENSEVAL [112]. SENSEVAL surge como uma organização dedicada à avaliação de sistemas WSD para identificar claramente as vantagens e desvantagens dos sistemas participantes. Até agora, só se realizaram 4 eventos: SENSEVAL -1 (Inglaterra, 1998, [113]), SENSEVAL-2 (França, 2001, [183]), SENSEVAL-3 (Espanha,

2004, [213]) e SENSEVAl-4 (França, 2007, [184]).

Dentro desse âmbito de desambiguação de sentidos, é pertinente comentar sobre o **SENSEVAL-2**, o segundo workshop internacional sobre avaliação de sistemas de desambiguação de sentidos que aconteceu em julho de 2001 em Toulouse. Esse evento reuniu investigadores com o intuito de discutir e solucionar problemas pertinentes ligados a sistemas de desambiguação automática de sentidos. Seu objectivo foi avaliar os pontos fortes e os problemas encontrados nos algoritmos de desambiguação automática de sentidos de palavras diferentes em contextos linguísticos diversos.

Um dos trabalhos apresentados nesse evento fazia referência à desambiguação automática de sentidos de adjectivos e substantivos da língua inglesa (English Lexical Sample Task). O corpus escolhido para esse trabalho foi o British National Corpus. O objectivo do trabalho, em princípio, era utilizar dados do inglês americano em conjunto com o BNC, porém, devido a quantidade limitada de dados disponíveis do inglês americano (em função das restrições ocasionadas por direitos autorais), os dados do corpus de inglês britânico acabaram sendo predominantes.

Para fazer a escolha dos sentidos, o WordNet acabou sendo utilizado. O maior problema em relação a essa escolha foi o facto do WordNet ter sido organizado em torno de grupos de palavras de sentidos semelhantes (os synsets), não em torno das palavras e seus sentidos variados. A desambiguação de sentido de palavras, por outro lado, é uma tarefa que exige uma distinção de sentidos muito bem definida e clara. Portanto, para resolver o problema dos casos em que o WordNet fazia uma distinção entre dois sentidos que não era totalmente clara, todas as palavras na amostra lexical e seus respectivos verbetes foram revisados por um lexicógrafo, com uma preocupação especial pelos sentidos não suficientemente distintos. Posteriormente, as revisões propostas foram incluídas na versão do WordNet (1.7).

O desempenho dos melhores sistemas de desambiguação nesse trabalho acabou ficando em torno de 14% de precisão. Os investigadores acreditam que esses resultados são justificados pela escolha do léxico, nesse caso o WordNet. Eles ressaltam que a alta exactidão da desambiguação de sentido das palavras só é possível quando o léxico faz uma distinção clara e bem definida dos sentidos e oferece informações suficientes sobre as distinções para a construção do algoritmo.

Em julho de 2004, foi sediado o **SENSEVAL-3**. Esse evento foi realizado em conjunto com a Associação para Linguística Computacional (Association for Computacional Linguistcs) e teve grande repercussão.

Neste penúltimo evento, definiu-se 2 tipos de tarefas para a WSD com o seu respectivo corpus de referência desenhado especificamente para estas tarefas: ALL-Words Task e Englisk Lexical Simples. A seguir descreve-se cada uma destas tarefas.

- **All-Words Task**. A meta desta tarefa é desambiguar todas as palavras polissêmicas possíveis que há dentro de um texto [213]. O conjunto de dados para esta tarefa possui todas as palavras etiquetadas que se desejam desambiguar para propósitos de treinamento e avaliação dos sistemas. Na Tabela 5.10 se mostra uma frase deste corpus "*His shout had been involuntary*", onde as palavras polissêmicas marcadas com a cor negra são as palavras que se desejam desambiguar. As etiquetas *<head id=...>* se utilizam para identificar as palavras a desambiguar. É importante remarcar que este corpus dentro da modalidade no idioma inglês (*English All-words*) possui 2212 palavras polissêmicas etiquetadas com o sentido apropriado.

```
<?xml version="1.0"?> <!DOCTYPE corpus SYSTEM "all-words.dtd">
<corpus lang="en>
<text id = "d000">
⋮
His <head id="d000.s014.t001">shout</head> had <head id="d000.s014.t003">been</head>

<head id="d000.s014.t004">involuntary</head>
⋮
</text>
⋮
```

Tabela 5.10: Extracto do corpus da tarefa All-Words em inglês.

- **Lexical Sample Task**. O objectivo desta tarefa é desambiguar somente uma palavra polissêmica (ver [144]). Esta palavra é a única unidade léxica que está etiquetada com o seu sentido correspondente (ver Tabela 5.8). Para o idioma inglês, o corpus é conhecido como SENSEVAL-3 English Lexical Sample (ELS). Este conjunto contém mais de 11.000 exemplos sobre 57 palavras especificadas (7.860 exemplos para treinamento e 3.944 para prova). Na Tabela 5.11 se mostra o conteúdo deste corpus. A palavra objectivo se mostra em cor negra e marcada pela etiqueta <head>. A etiqueta <answer> contém o sentido correcto da palavra a desambiguar, onde o parâmetro instante é uma chave para identificar o texto e *senseid* indica o sentido. Os valores de *senseid* estão codificados e SENSEVAL proporciona os respectivos sentidos de WordNet equivalentes. Os exemplos deste corpus foram extraídos do corpus *Bristish*

National Corpus (BNC), o qual possui mais de 100 milhões de palavras. Os exemplos foram etiquetados manualmente ([144] e [40]).

```
<lexelt item="image.n">
⋮
<inatance id="image.n.bnc.00003870" docsrc="BNC"> <answer
instance= "image.n.bnc.00003870"  senseid="image%1:06:00::"/> <content> After
all, it is an original work in itself. Nevertheless, the differences between
painting and print may be instructive , and help the interpretation of both. By
other forms of reproduction an <head>image</head> may be more or less degraded,
so that nothing can be learnt from them. There remains the courtesy paid by one
art to another, as in the poem which Baudelaire suggested might be a high from of
critical writing . Or it may be music, as in Mussorgsky ´s composition which be
entitled Pictures from an Exhibition.
< / context>
< / instance>
⋮
< / lexelt >
```

Tabela 5.11: Extracto do corpus da tarefa English Lexical Samples em inglês.

Dentro da tarefa de Lexical "Simples", participaram 47 sistemas, onde o melhor sistema (ver [144]) entre os sistemas baseados em métodos supervisionados se obteve 72.9% de precisão e *recall*; enquanto o pior sistema obteve nesta categoria 78.2% em precisão e 31% em *recall*. Por outra parte, na categoria dos sistemas baseados em métodos não supervisionados ou em conhecimentos o melhor foi de 66.1% em precisão e 65% em *recall*; enquanto o pior sistema só conseguiu atingir o 19.7%. e 11.7% em *recall*.

A Tabela 5.12 mostra o inventário das palavras utilizadas para a tarefa "*Lexical Simple*". Para cada palavra se indica a sua categoria gramatical ou parte da oração (P. O); o número de sentidos (# sent.); o número de exemplos de treinamento (# ex.ent) e o número de exemplos de prova para a sua avaliação (#ex. prova). O inventário contém 57 palavras, onde 20 são substantivos, 32 são verbos e 5 são adjectivos; a quantidade promédio dos sentidos é de 6.47 sentidos por cada palavra. Os conjuntos para cada palavra se dividem em dois tipos: para o treinamento e para a prova.

Como já se refereiu anteriormente, foram apresentados 14 trabalhos diferentes entre os 47 sistemas que participaram no evento. Dentro desses trabalhos estavam incluídos trabalhos de desambiguação automática de sentidos para diversas línguas, como vasco, catalão, chinês, inglês, italiano, romeno e espanhol.

Palavra	P.O	#sent.	#ex.tren.	#ex. prova	Palavra	P.O sent.	#sent.	#ex.tren.	#ex. prova
activate	V	5	228	114	miss	V	8	58	30
add	V	6	263	132	note	V	3	132	67
appear	V	3	265	133	operate	V	5	35	18
argument	S	5	221	111	organization	S	7	112	56
arm	S	6	266	133	paper	S	7	232	117
ask	V	6	261	131	party	S	5	230	116
atmosphere	S	6	161	81	performance	S	5	172	87
audience	S	4	200	100	plan	S	3	166	84
bank	S	10	262	132	play	V	12	104	52
begin	V	4	181	79	produce	V	6	186	94
climb	V	5	133	67	provide	V	6	136	69
decide	V	4	122	62	receive	V	9	52	27
degree	S	7	256	128	remain	V	3	139	70
difference	S	5	226	114	rule	V	4	59	30
different	A	5	98	50	shelter	S	5	196	98
difficult	S	4	46	23	simple	A	7	36	18
disc	S	4	200	100	smell	V	7	108	55
eat	V	7	181	87	solid	A	14	58	29
encounter	V	4	130	65	sort	S	4	190	96
expect	V	3	156	78	source	S	7	64	32
express	V	4	110	55	suspend	V	7	128	64
hear	V	7	63	32	talk	V	9	146	73
hot	A	22	86	43	treat	V	9	112	57
image	S	7	146	74	use	V	5	26	14
important	A	5	36	19	wash	V	12	66	34
interest	S	7	185	93	watch	V	7	100	51
judgment	S	7	62	32	win	V	7	78	39
lose	V	9	71	36	write	V	8	44	23
mean	V	7	80	40	TOTAL DE EJEMPLOS			7860	3944

Tabela 5.12: Inventário de palavras que contém SENSEVAL-3 ELS

Um dos trabalhos apresentados nesse evento foi o "Word-Sense Disambiguation of WordNet Glosses", ou seja, Desambiguação Automática das Glosas do WordNet. Segundo [134] esse trabalho, foi criado para fomentar o desenvolvimento de tecnologia que fizesse uso de padrões de recursos lexicais. O trabalho foi baseado na disponibilidade de glosas de desambiguação de sentidos feitos à mão no projecto de extensão do WordNet.

De acordo com o autor, "o processo de anotação seguido do projecto XWN [134], com as etiquetas usadas nesse trabalho, mais uma vez indicaram dificuldades com o conjunto de sentidos de WordNet. Esse facto permanece devido ao WordNet não ter tido o benefício de recursos lexicográficos suficientes na construção de suas glosas e na aquisição de outras informações lexicográficas em seus verbetes. O projecto WordNet continua com o seu esforço para adicionar informação, mas com recursos limitados.

Aproveitando os corpus utilizados no SENSEVAL, [175] desenvolveu uma abordagem de desambiguação através do princípio colocacional. Ele apresentou uma abordagem de base em corpus onde uma árvore de decisões atribui um sentido para uma palavra ambígua baseada nos n-gramas de que faz parte. Nessa abordagem,

a árvore de decisões é determinada a partir de um número de sentenças onde cada instância da palavra ambígua foi anotada manualmente com uma etiqueta de sentido que denota o sentido mais apropriado para aquele contexto.

Os bigramas são sequências de duas palavras que ocorrem num texto. O contexto em que a palavra ocorre é representado por um número de traços binários que indicam se um determinado bigrama ocorreu em aproximadamente 50 palavras à direita ou à esquerda da palavra a ser desambiguada" [175]. Segundo [175] justifica essa abordagem argumentando que "traços lexicais como bigramas, colocações e co-ocorrências geralmente contribuem muito para um bom desempenho dos programas de desambiguação."

Depois de todas as questões abordadas, parece evidente que a Linguística de Corpus acabou exercendo uma maior influência na área de lexicografia, tanto que praticamente todos os grandes dicionários da língua inglesa são feitos com base nesse tipo de abordagem hoje em dia. Esse fenómeno reflete também um interesse no âmbito empresarial de investir em estudos baseados em corpus. Um bom exemplo disso foi o Cobuild [89] uma parceria entre a Universidade de Birmingham (Grã-Bretanha) e a editora Collins. No âmbito do Cobuild foram produzidos não apenas dicionários com base em corpus, mas também gramáticas e livros didácticos para o ensino do inglês [18].

Outro exemplo de projecto que, nesse caso, envolve o uso de corpus da língua portuguesa é o "Dicionário UNESP do Português Contemporâneo" [22], publicado pela Editora UNESP. Esse dicionário, segundo seus organizadores, foi desenvolvido com base em um corpus de 90 milhões de palavras.

Em [22] afirma que o conjunto das entradas foi estabelecido pelo critério de ocorrência do corpus que faz parte do banco de dados do Laboratório de Lexicografia da Faculdade de Ciências de Letras de Araraquara. No entanto, resulta que, como em todo dicionário de língua, a principal informação é de natureza semântica, sendo que o sistema definitório do dicionário não é diferente do adotado pela maioria dos dicionários. O autor diz que "para patentear ou esclarecer melhor as acepções, utilizou-se um sistema de contextualização constituído por frases e expressões extraídas de textos reais do corpus, com adaptações e condensações para melhor cumprir os seus objectivos."

O autor não esclarece se os verbetes do dicionário foram estabelecidos com base nos sentidos encontrados no corpus. Ao que parece, o corpus só foi utilizado para definir o conjunto das entradas do dicionário e prover exemplos. Tradicionalmente,

na área de lexicografia que faz utilização de corpus, os verbetes são definidos a partir das ocorrências encontradas no mesmo, ou seja, "os dados são o ponto de partida para a construção do dicionário" [193]. Em suma como já foi referido acima a desambiguação automática de sentido consiste no processo de seleccionar o sentido mais apropriado para uma palavra baseado no contexto em que ela ocorre. Para esse propósito, é pressuposto que um conjunto de sentidos possíveis seja previamente determinado.

Por exemplo, ao se pressupor que o adjectivo *pesado* tem o seguinte conjunto de sentidos possíveis: algo que tem muito peso; árduo; e indigesto, quando utilizado num contexto "*Cabe a ele o serviço pesado todos os dias, em jornadas de até 14 horas*", o leitor humano entende imediatamente que *pesado* está sendo usado com o sentido de *árduo*. Entretanto, um programa de computador tentando realizar a mesma tarefa enfrenta um grande problema, uma vez que ele não pode contar com um conhecimento lingüístico prévio, como acontece com seres humanos. Por esse motivo, a criação de programas de desambiguação de sentidos acaba se tornando uma tarefa árdua e complexa, que exige um estudo detalhado dos contextos em que a palavra polissêmica está inserida.

5.5 Algoritmos de Aprendizagem Automática

A aprendizagem automática é a disciplina que estuda cômo construir sistemas computacionais que melhoram automaticamente mediante a experiência. Em outras palavras, diz-se que um programa adquiriu conhecimento (ou "aprendido") para realizar uma tarefa específica T se depois de proporcionar-lhe a experiência E e mediante um conjunto de exemplos de T o sistema é capaz de desempenhar-se quando se apresentam novas situações da tarefa. O desempenho é medido usando uma métrica de qualidade P (precisão e *recall*). Portanto, um problema de aprendizagem bem definida requer que T, E e P estejam bem especificados (referir-se a [155] para um estudo mais detalhado). Nesta disciplina, desenvolveram-se vários algoritmos para realizar o processo de aprendizagem, por esta razão recebem o nome de algoritmos de aprendizagem.

Em geral, os algoritmos são utilizados para o processo de classificação. Este processo pode ser formalizado como a tarefa de aproximar uma função objectivo desconhecida $\Phi : I \times C \rightarrow \{V, F\}$ (que descreve como as instâncias do problema devem ser classificadas de acordo a um experto no domínio) por meio de uma fun-

ção $\Theta : I \times C \rightarrow \{V, F\}$ chamada o classificador, onde $C = \{c_I, ..., c_{|c|}\}$ é um conjunto predefinido de categorias ou classes, I é um conjunto de instâncias do problema, e $\{V, F\}$ são os valores de Verdadeiro e Falso respectivamente. Comumente cada instância $i_j \in I$ é um exemplo que está representado como uma lista $A = (a1, a2, ..., a_{|A|})$ de valores característicos, conhecidos como atributos, por exemplo $i_j = (a_{1j}, a_{2j}, ..., a_{|A|j})$. Se $\Phi : i_j \times c_i \rightarrow V$, então i_j é chamado um exemplo positivo de c_i, enquanto se $\Phi : i_j \times c_i \rightarrow F$ este é chamado um exemplo negativo de c_i.

Para gerar automaticamente o classificador de c_i é necessário realizar um processo indutivo, chamado o *aprendiz*, o qual para observar os atributos de um conjunto de instâncias preclassificadas sob c_i o $\bar{c_i}$, adquire os atributos que uma instância desconhecida deve ter para pertencer à categoria. Por tal motivo, na construção do classificador se requer a disponibilidade inicial de uma colecção Ω de exemplos tais que o valor de $\Phi(i_j, c_i)$ é conhecido para cada $(i_j, c_i) \in \Omega \times C$. À colecção usualmente se chama conjunto de treinamento (Tr). Em resumo, ao processo anterior se identifica como aprendizagem supervisionada devido à dependência de Tr. É por esta razão, que os métodos de WSD que usem a aprendizagem supervisionada se conhecem como métodos WSD supervisionados.

5.5.1 O classificador Naive Bayes

A classificação Bayesiana é um método baseado em estatística. O Seu funcionamento usa o cálculo de probabilidades a partir do teorema de Bayes, que se apresentará a seguir.

O classificador Naive Bayes (NB) considera-se como parte dos classificadores probabilísticos, os quais se baseiam na suposição que as quantidades de interesse se regem por distribuições de probabilidade, e que a decisão óptima pode tomar-se por meio de raciocinar a respeito dessas probabilidades junto com os dados observados [155]. Dentro de WSD, este algoritmo se encontra entre os mais utilizados (ver [145]). Em [Lewis98] se apresenta uma guia básica das diferentes direcções que tomaram as investigações sobre Naive Bayes, as quais se caracterizam pelas modificações realizadas ao algoritmo. A seguir descreveremos o Naive Bayes.

Neste esquema o classificador é construído usando o conjunto de treinamento Tr para estimar a probabilidade de cada classe. Então, quando uma nova instância i_j é apresentada, o classificador lhe assigna a cate‘00’goria $c \in C$ provável aplicando a

regra:

$$c = \arg\max_{c_i \in C} P(c_i \mid i_j) \tag{5.1}$$

Utilizando o teorema de Bayes para estimar a probabilidade temos

$$c = \arg\max_{c_i \in C} \frac{P(i_j \mid c_i)P(c_i)}{P(i_j)} \tag{5.2}$$

O denominador na equação anterior não difere entre categorias e se pode omitir

$$c = \arg\max_{c_i \in C} P(c_i \mid i_j)P(c_i) \tag{5.3}$$

Tendo em conta que o esquema é chamado "naive" devido ao suposto de independência entre atributos, por exemplo, assume-se que as características são condicionalmente independentes dadas as classes. Isto simplifica os cálculos produzindo

$$c = \arg\max_{c_i \in C} P(c_i) \prod_{k=1}^{n} P(a_{kj} \mid c_i) \tag{5.4}$$

onde P(ci) é a fracção do exemplo em Tr que pertencem à classe c_i, e $P(a_{kj} \mid c_i)$ calcula-se de acordo ao teorema de Bayes.

Em resumo, a tarefa de aprendizagem no classificador Naive Bayes consiste em construir uma hipótese por meio de estimar as diferentes probabilidades $P(c_i)$ e $P(a_{kj} \mid c_i)$ em termo das suas frequências sobre Tr. Em [174] e [2] se apresentam uma descrição detalhada dos cálculos. A Figura 5.12 reflecte a estrutura gráfica de um modelo Naive Bayes.

Segundo o que acabamos de comentar, o modelo Bayesiano apresenta problemas. Assim que se não há independência condicional o modelo Bayesiano fracassa. Não pretendemos abordar uma demonstração rigorosa e formal; no entanto, sim trataremos de ilustrar este facto por meio de um exemplo que se mostra a seguir na Tabela 5.13 abaixo na que não se cumpre o critério de independência.

Tratemos de encontrar $P(k_1 \mid m_1)$, com os dados da Tabela, de duas formas diferentes:

a) Utilizando o teorema de Bayes

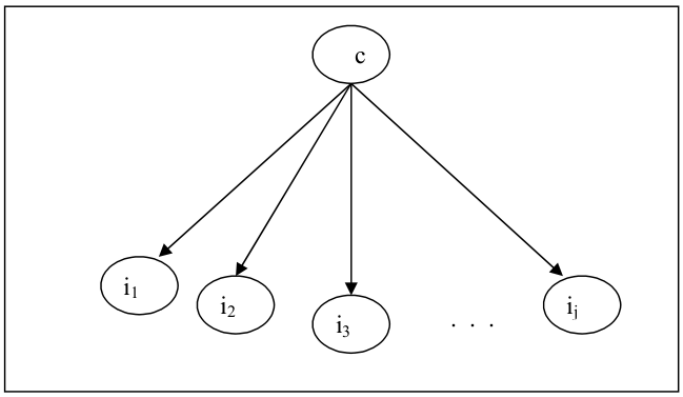

Figura 5.12: Estrutura gráfica de um modelo Naive Bayes

	só K_1	K_1 e K_2	só K_2	K_3	Totais
só m_1	$b1$	c_1	d_1	e_1	n_1
m_1 e m_2	$b2$	c_2	d_2	e_2	n_2
só m_2	$b3$	c_3	d_3	e_3	n_3
m_3	$b4$	c_4	d_4	e_4	n_4
Totais	B	C	D	E	N

Tabela 5.13: Problemas sobre o modelo Bayesiano.

b) Directamente a partir do conceito de probabilidade condicional

Para calcular $P(k_1 \mid m_1)$ a partir do teorema de Bayes necessitamos:

$$P(k_1) = \frac{B + C}{N} \tag{5.5}$$

$$P(k_2) = \frac{C + D}{N} \tag{5.6}$$

$$P(k_3) = \frac{E}{N} \tag{5.7}$$

$$P(m_1 \mid k_1) = \frac{b_1 + b_2 + c_1 + c_2}{B + C} \tag{5.8}$$

$$P(m_1 \mid k_2) = \frac{c_1 + c_2 + d_1 + d_2}{C + D} \tag{5.9}$$

$$P(m_1 \mid k_3) = \frac{e_1 + e_2}{E} \tag{5.10}$$

Aplicando directamente a equação do teorema de Bayes:

$$P(k_1 \mid m_1) = \frac{P(m_1 \mid k_1)P(k_1)}{P(m_1 \mid k_1)P(k_1) + P(m_1 \mid k_2)P(k_2) + P(m_1 \mid k_3)P(k_3)} \tag{5.11}$$

Substituindo valores:

$$P(m_1 \mid k_1)P(k_1) = \frac{b_1 + b_2 + c_1 + c_2}{N} \tag{5.12}$$

$$P(m_1 \mid k_2)P(k_2) = \frac{c_1 + c_2 + d_1 + d_2}{N} \tag{5.13}$$

$$P(m_1 \mid k_3)P(k_3) = \frac{e_1 + e_2}{N} \tag{5.14}$$

E, em consequência:

$$P(k_1 \mid m_1) = \frac{\frac{b_1+b_2+c_1+c_2}{N}}{\frac{b_1+b_2+c_1+c_2}{N} + \frac{c_1+c_2+d_1+d_2}{N} + \frac{e_1+e_2}{N}} \tag{5.15}$$

Operando:

$$P(k_1 \mid m_1) = \frac{b_1 + b_2 + c_1 + c_2}{b_1 + b_2 + 2c_1 + 2c_2 + d_1 + d_2 + e_1 + e_2} \tag{5.16}$$

$$P(k_1 \mid m_1) = \frac{b_1 + b_2 + c_1 + c_2}{n_1 + n_2 + c_1 + c_2} \tag{5.17}$$

Este resultado é manifestamente falso. Aplicando o conceito de probabilidade condicional directamente sobre os dados da Tabela obtemos que:

$$P(k_1 \mid m_1) = \frac{b_1 + b_2 + c_1 + c_2}{n_1 + n_2} \tag{5.18}$$

Esta limitação do modelo propõe problemas quando se pretende sua aplicação em domínios do mundo real, em que os requisitos da independência quase nunca se cumprem.

Mas esta não é a única deficiência do modelo Bayesiano. Nos problemas interessantes para a aplicação de técnicas de inteligência artificial, a informação costuma aparecer progressivamente, sequencialmente e, geralmente, de forma pouco ordenada. Nestes casos, adequar a aproximação Bayesiana à interpretação sequencial supõe considerar que a informação aparece incrementalmente e, portanto, terá que adaptar as equações correspondentes.

A principal vantagem dos métodos Bayesianos reside em que estão fortemente fundados na teoria da probabilidade, no entanto sua principal dificuldade reside na grande quantidade de probabilidades que é necessário obter para construir uma base de conhecimentos a prior e o alto custo computacional.

Assim, ainda supondo hipóteses mutuamente excluentes, evidências condicionalmente independentes e variáveis restringidas a dois valores (verdadeiro e falso). Além disto, desafortunadamente a suposição de independência condicional raramente é válida e a suposição de mútua exaustividade das hipóteses costuma ser falsa, sendo o mais corrente a aparição de hipóteses concorrentes e superpostas, os métodos Bayesianos não permitem uma explicação clara das suas conclusões e permitem que uma mesma evidência apóie ao mesmo tempo, a uma hipótese e a sua negação.

Em geral, a assunção de independências entre os atributos é demasiado restritiva, pelo que uma aproximação para resolver estes problemas são as redes de crenças ou redes Bayesianas (*belief networks*). Uma rede de crença é um gráfico acíclico dirigido em que cada nó representa uma variável e cada arco uma dependência probabilística, na qual se especifica a probabilidade condicional de cada variável dados os seus nós pais, a variável à que aponta o arco é dependente (causa-efeito) da que está na origem deste. A topologia ou estrutura da rede nos dá informação sobre as dependências probabilísticas entre as variáveis mas também sobre as independências condicionais de uma variável (ou conjunto de variáveis) dada outra ou outras variáveis, ditas independências, simplificam a representação do conhecimento (menos parámetros) e o raciocínio (propagação das probabilidades). A obtenção de uma rede Bayesiana a partir de dados, é um processo de aprendizagem e se divide em duas etapas: a aprendizagem estrutural e a aprendizagem paramétrica [171] e [90]. (i) A aprendizagem estrutural, consiste em obter a estrutura da rede Bayesiana, isto é, as relações de dependência e independência entre as variáveis envolvidas. (ii) A

aprendizagem parámetrica, tem como finalidade obter as probabilidades a prior e condicionais requeridas a partir de uma estrutura dada.

Estas redes provêem uma forma compacta de representar o conhecimento e métodos flexíveis de raciocínio baseados nas teorias probabilísticas capazes de predizer o valor de variáveis não observadas e explicar as observadas. Entre as características que possuem as redes Bayesianas, pode-se destacar que permitem aprender sobre relações de dependência e casualidade, permitem combinar conhecimento com dados [85], [50] e [90] e podem manejar bases de dados incompletas [85], [86], [190] e [90]. Em [171] demonstrou que o uso de redes de crenças permite construir bases de conhecimento probabilísticos consistentes, sem impor desnecessárias assunções de independência condicional. Estas redes também asseguram que a evidência a favor de uma hipótese não será construída por suporte parcial da sua negação, e que explicações consistentes podem ser obtidas mediante o rastreamento das crenças até os pontos iniciais da rede.

As redes Bayesianas nos permitem a definição de dependência entre os atributos dos dados usados para a classificação, além de oferecer uma estrutura visual de melhor compressão no classificador obtido.

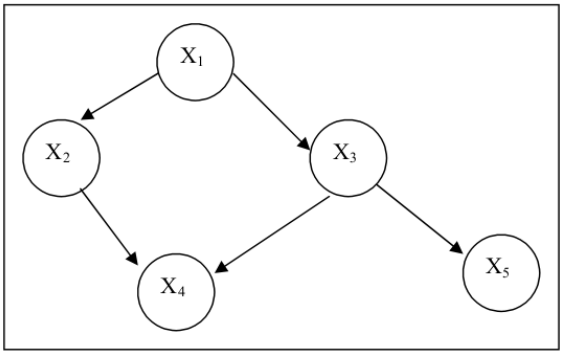

Figura 5.13: Um exemplo de uma rede Bayesiana

A Figura 5.13 representa um exemplo de uma rede Bayesiana para uma distribuição de probabilidade conjunta $P(x_1, x_2, x_3, x_4, x_5)$. Neste caso, a dependência declarada na rede permite a expressão natural da distribuição de probabilidade conjunta em termos de probabilidades condicionais locais (uma vantagem chave da rede Bayesiana) como se segue:

$$P(x_1, x_2, x_3, x_4, x_5) = P(x_1)P(x_2 \mid x_1)P(x_3 \mid x_1)P(x_4 \mid x_2, x_3)P(x_5 \mid x_3) \quad (5.19)$$

A probabilidade $P(x_1)$ é a probabilidade a prior para a rede Bayesiana.

5.5.2 O Algoritmo C4.5

O esquema C4.5 foi desenhado como uma extensão do algoritmo ID3 [188], este último faz parte dos classificadores conhecidos como *árvores de decisão*, os quais são árvores onde os seus nós internos são etiquetados como atributos, os ramos salientes de cada nó representam provas para os valores dos atributos, e as folhas da árvore identificam às categorias. Estes algoritmos proporcionam um método prático para aproximar conceitos e funções com valores discretos. Em [169], [240] são exemplos onde o algoritmo é usado. A seguir se apresenta a descrição do algoritmo ID3 com objectivo de facilitar a posterior descrição de C4.5 (para mais detalhes refira-se a [189]). Para construir a árvore, ID3 usa uma aproximação descendente que, dá preferência às árvores pequenas sobre as grandes. O nó raiz é seleccionado por possuir o atributo mais valioso no conjunto de treinamento, isto é, aquele atributo com que melhor classifica as instâncias; a busca se realiza por meio de uma prova estatística que mede que tão bem um atributo separa o conjunto de treinamento de acordo às classes. Uma vez que a raiz é seleccionada, agrega-se um ramo desde a raiz para cada possível valor do atributo correspondente, e o conjunto de treinamento é ordenado nos nós apropriados, por exemplo para cada nó contém os exemplos que cumprem a restrição do ramo anterior. Para seleccionar o atributo mais valioso em cada ponto da árvore, repete-se o processo inteiro usando o conjunto de treinamento associado com o nó. De maneira que quando uma nova instância precisa ser classificada, os atributos especificados pelos nós são avaliados iniciando pelo nó raiz, então de maneira descendente se percorrem os ramos da árvore que correspondem aos valores dos atributos na instância dada, o processo se repete até que uma folha é atingida, e é este ponto onde a etiqueta associada à folha é atribuída à nova instância como a sua categoria.

Na Tabela 5.14 se resume o algoritmo ID3, onde a medida tradicional para encontrar o atributo mais valioso é a *ganância* da informação (GI), que mede cômo um atributo dado separa o conjunto de treinamento conforme às classes. Uma definição intuitiva da ganância de informação é a diferença entre a entropía de um nó e de um dos seus descedentes. Em definitiva não é mais que uma heurística, que serve para a eleição do melhor atributo em cada nó.

Na equação seguinte se apresenta a forma de calcular a *ganância* na informação do atributo a com respeito ao conjunto de treinamento Tr. Formalizando a defi-

ID3 (tr, ci, A)

1. Criar um nó raiz para a árvore
2. Se todos os exemplos em **Tr** são positivos, regressar a árvore com o único nó raiz etiquetado como $\mathbf{c_i}$
3. Se todos os exemplos em **Tr** são negativos, regressar a árvore com o único nó raiz etiquetado como $\bar{\mathbf{c}}_\mathbf{i}$
4. Se a lista A está vazia, regressar a árvore com o único nó raiz etiquetado com o valor de $\mathbf{c_i}$ mais frequente em **Tr**.
5. Em outro caso começar
6. Seja **a** o atributo em **A** que melhor classifica a **Tr**
7. Etiquetar á **raiz** como **a**
8. Para cada possível valor v_i de a (por exemplo , $vi \in valores(a)$)
   - 8.1. Agregar um ramo abaixo o nó raiz correspondente á prova $a = vi$.
   - 8.2. Seja Tr_{vi} o subconjunto de exemplos para os que $a = vi$
   - 8.3. Se Tr_{vi} está vazio
     - 8.3.1. Agregar debaixo da rama um nó com o valor de ci más frecuente en Tr como etiqueta
   - 8.4. Pelo contrário
     - 8.4.1. $ID3(Tr_{vi}, ci, A - (a))$
9. Terminar
10. Regressar á raiz.

Tabela 5.14: Algoritmo de ID3.

nição diremos que, para o nó com o conjunto de treinamento Tr e o atributo a, a Ganância(Tr, a) será dada como segue:

$$GI(Tr, a) \equiv Entropia(Tr) - \sum_{vi \in Valores(a)} \frac{|Tr_{vi}|}{|Tr|} Entropía(Tr_{vi}) \tag{5.20}$$

Sendo Tr_{vi} o subconjunto de Tr formado por aquelas instâncias que no atributo a tomam o valor vi e a $Entropía(S)$, não é mais que a medida da incerteza que há num sistema, quer dizer, antes uma determinada situação, a probabilidade de que ocurra um dos possíveis resultados e se calcula da siguinte maneira:

$$Entropía(S) = \sum_{i=1}^{|C|} -P(C_i)log_2P(C_i) \tag{5.21}$$

Finalmente, uma vez introduzido ID3 os passos a seguir em C4.5 são :

1. Separa Ω em conjunto de treinamento e conjunto de validação.
2. Construir a árvore de decisão para o conjunto de treinamento (aplicar ID3)
3. Converter a árvore num conjunto de regras equivalente, onde o número de regras é igual ao número de possíveis rota desde a raiz aos nós folha.
4. Podar cada regra eliminando precondiciones que resultem em melhorar a exactidão no conjunto de validação.
5. Ordenar as regras descendentemente de acordo a sua exactidão, e usá-las nessa ordem para classificar futuros exemplos.

Tabela 5.15: Passos a seguir em C4.5.

5.5.3 K- Vizinhos Mais Próximos

K-Vizinhos mais próximos (k-NN, por suas siglas em inglês) é um dos métodos de aprendizagem baseados em instâncias mais básicos, mas com resultados aceitáveis em tarefas que envolvam na tarefa de WSD (ver [169] e [170]). Este algoritmo não tem uma fase de treinamento fora da linha, portanto, o principal cálculo se dá em linha quando se localizem os k vizinhos mais próximos. A ideia no algoritmo é armazenar o conjunto de treinamento, de modo tal que para classificar uma nova instância, buscam-se nos exemplos armazenados casos similares e nestes se atribui a classe mais provável.

Na Tabela 5.16 se resume o algoritmo, aqui uma maneira comum de encontrar os k exemplos mais próximos à instância i_q é por meio da distância Euclidiana, onde a distância entre as instâncias i_j e i_q é definida pela seguinte equação:

$$d(i_j, i_q) = \sqrt{\sum_{k=1}^{|A|}(a_{jk} - a_{qk})^2} \tag{5.22}$$

Sendo $|A|$ o número total dos pesos das classes dos vizinhos, a_{jk} e a_{qk} os pesos das classes dos vizinhos.

Treinamento:

1. Para cada exemplo em Tr, agregar o exemplo na lista exemplos_treinamento.

Classificação:

1. Dada uma instância de prova i_q a ser classificada,
2. Sejam $i_1, ..., i_k$ os k exemplos da lista_treinamento que são mais próximos a i_q.
3. Regressar

$$c = \arg\max_{c_i \in C} \sum_{j=1}^{k} \delta(c_i, c_{ij})$$

onde $\delta(a, b) = 1$ si $a = b$ y $\delta(a, b) = 0$ em outro caso

Tabela 5.16: Algoritmo de k vizinhos mais próximos.

5.5.4 Máquinas de Vectores de Suporte

As máquinas de vectores de suporte (SVM, por suas siglas em inglês) mostraram conseguir bom desempenho de generalização sobre uma ampla variedade de problemas de classificação, destacando recentemente em problemas de classificação de textos (ver [108], [109], [51]), onde se aprecia que SVM tende a minimizar o erro de generalização, por exemplo os erros do classificador sobre novas instâncias. Em termos geométricos, SVM pode ser visto como a tentativa de encontrar uma *superfície (δ_i)* que separe aos exemplos positivos dos negativos por margem mais ampla possível (ver [88] para uma descrição detalhada do algoritmo).

A busca de δ_i que cumpre que a distância mínima entre δ_i e um exemplo de treinamento seja máxima, realiza-se através de todas as superfícies δ_1, δ_2, ..., , no espaço $|A|$ -dimensional que separa aos exemplos positivos dos negativos no conjunto de treinamento (conhecidas como superfícies de decisão). Para entender melhor a ideia por trás do algoritmo SVM tomaremos o caso em que os exemplos positivos e negativos são linealmente separados, portanto as superfícies de decisão são ($|A|$-1)-hiperplanos. Por exemplo, no caso de duas dimensões várias linhas podem ser tomadas como superfícies de decisão, então o método SVM elege o elemento médio do conjunto mais largo de linhas paralelas, por exemplo, desde o conjunto em que a distância máxima entre dois dos seus elementos é a maior. Cabe ressaltar que a melhor superfície de decisão é determinada unicamente por um conjunto pequeno de exemplos de treinamento, chamados vectores de suporte.

Uma vantagem importante de SVM é que permite construir classificadores não lineares, por exemplo, o algoritmo representa dados de treinamento não lineares num

espaço de alta dimensionalidade (chamado o espaço de características), e constrói o hiperplano que tem a margem máxima, isto permite uma aprendizagem correcta de problemas difíceis, proporcionando erros muito baixos. Além disto, tem uma elevada capacidade de generalização, não sofre a maldição da dimensionalidade, etc. Os inconvenientes de SVM radicam no facto de que exige elevados requerimentos em tempo de computação e por outro lado, não permite a aprendizagem incremental (existem versões de SVM incremental, mas possivelmente não funcionam tão bem)[3].

[3]http://www.siam.org/meetings/sdm02/proceedings/sdm02-15.pdf

Capítulo 6

Experimentação com expansão de consultas e desambiguação

Para atingir os objectivos deste trabalho é necessário: a avaliação da execução ou desempenho da recuperação, utilizando o sistema da recuperação de informação Lemur 3.1, uma colecção de referência TREC e medidas de avaliação.

A colecção de referência de prova consiste numa colecção de documentos, num conjunto de petições de informação chamadas tópicos em TREC e um conjunto de documentos relevantes (proporcionados por especialistas).

Como já foi mencionado anteriormente para avaliar a eficácia se usarão os procedimentos padrões no campo de recuperação de informação, utilizando, portanto, as medidas estatísticas comuns na matéria e também as colecções de documentos de prova que mais se usam neste âmbito.

Neste capítulo se começará a ver na Secção 6.1 os elementos que podem ser utilizados para avaliar a qualidade do processo de recuperação de informação. A continuação, na Secção 6.2 se detalhará como se vão realizar as provas de avaliação, para nas seguintes secções se incluir os resultados obtidos à hora de realizar consultas que incluem a possibilidade de usar expansão de consultas e/ou desambiguação.

6.1 Avaliação em recuperação de informação

Antes da implementação final de um sistema de recuperação de informação é necessário levar a cabo uma avaliação do sistema. O tipo de avaliação a considerar depende dos objectivos do sistema de recuperação. Claramente, qualquer sistema software tem que prover a funcionalidade para que foi concebido. Mas independentemente da análise funcional e da busca de erros na aplicação, num sistema de recuperação da informação é necessário avaliar o seu rendimento.

As medidas mais comuns do rendimento do sistema são o tempo e o espaço. Quanto mais curto seja o tempo de resposta, e menor o espaço usado, melhor será considerado o sistema. Há um compromisso inerente entre a complexidade espacial e a complexidade temporária que frequentemente permite trocar um por outro.

Num sistema desenhado para a recuperação de dados, o tempo de resposta e o espaço requerido são, normalmente, as medidas de maior interesse e as que se adoptam para avaliar o sistema. Neste caso, busca-se a eficiência das estruturas da indexação (que se usa para acelerar a busca), e as sobrecarregas produzidas por muitas capas de software que normalmente estão presentes. Mas este tipo de avaliação é uma avaliação de rendimento a secas.

Num sistema desenhado para a recuperação da informação outras medidas, para além do tempo e o espaço, são também de interesse. De facto, já que a petição da consulta do usuário é inerentemente pouco concreta, os documentos recuperados não são respostas exactas e têm que ser ordenados de acordo com sua relevância com respeito à consulta.

Esta classificação de relevância introduz um componente que não está presente nos sistemas de recuperação de dados e que joga um papel central na recuperação de informação. Assim, os sistemas de recuperação de informação requerem a avaliação de como é necessário o conjunto de documentos de resposta. Este tipo de avaliação se conhece como avaliação da eficácia da recuperação e se baseia normalmente numa colecção de prova de referência e em algumas medidas de avaliação. Falar-se-á das colecções de referência e as medidas.

6.1.1 Colecções de referência

A investigação em recuperação de informação se enfrentou tradicionalmente a dois problemas fundamentais: O primeiro é a carência de um meio de trabalho sólido

e formal como base fundamental, isto é ainda que alguns grupos de investigação tivessem utilizado as mesmas colecções, os grupos não tinham feito uma tentativa coordenada para trabalhar com os mesmos dados, utilizar as mesmas técnicas de avaliação e em geral comparar os resultados nos diferentes sistemas. O segundo é a carência de uns bancos de provas robustos e consistentes, isto é uma grande colecção de prova. A primeira destas críticas é difícil de rebater inteiramente devido ao grau de subjectividade associado à tarefa de decidir a relevância de um documento determinado (o qual caracteriza a recuperação da informação como diferença com a recuperação de dados). A segunda destas críticas, no entanto, é completamente refutável actualmente.

Durante três décadas, a experimentação em recuperação de informação se baseava em colecções de prova relativamente pequenas que não reflectiam os principais aspectos presentes em grandes meios bibliográficos. Inclusive, as comparações entre vários sistemas de recuperação eram difíceis de fazer porque os experimentos dirigidos por diferentes grupos se centravam em aspectos diferentes da recuperação, inclusive quando a mesma colecção de prova usada, e não tinha meios de prova amplamente aceitados. A princípios dos anos 90 se iniciou um movimento para defrontar a este estado de desordem sob a liderança de *Donna Harman no NIST (Nacional Institute of Standards and Technology)*, em Maryland. Este movimento consistiu em promover uma conferência anual de recuperação TREC (*Text Retrieval Conference*), dedicada à experimentação com uma grande colecção de prova que compreende ao redor de um milhão de documentos. Em cada conferência TREC, desenha-se um conjunto de experimentos de referência. Os grupos de investigação que participam na conferência usam estes experimentos de referência para comparar os seus sistemas de recuperação.

Uma explicação clara do objectivo das conferências TREC se podem encontrar na Web do NIST [Sítio Web TREC] e vem a dizer o seguinte:

A série das conferências TREC está patrocinada pelo NIST e o Escritório de Tecnologias da informação do DARPA (*Defense Advanced Resseca Projects Agency*) como parte do programa de textos TIPSTER. O objectivo das conferências é avançar na investigação em recuperação de informação para aplicações em textos grandes proporcionando grandes colecções de prova, homogeneizar os procedimentos da pontuação e classificação e comparação, e ser um fórum para organizações interessadas em comparar os seus resultados. A assistência das conferências TREC está restringida aqueles investigadores e desenvolvedores que realizaram com sucesso tarefas de recuperação TREC e ao pessoal seleccionado do governo de Estados Unidos desde

as agências patrocinadoras.

Os participantes numa conferência TREC empregam uma ampla variedade de técnicas de recuperação, incluindo métodos usando recursos linguísticos automáticos, ponderações sofisticadas de termos, técnicas de linguagem natural, retroalimentação de relevância, e concordância de padrões avançada. Cada sistema trabalha com a mesma colecção de prova consistente em arredor de 2 Gigabytes de texto (sobre um milhão de documentos em inglês obtidos desde uma variedade de fontes incluindo tais como: jornais, extractos de artigos técnicos, etc.) e um conjunto dado de necessidades de informação chamadas tópicos. Os resultados são executados através de um pacote comum de avaliação de maneira que os grupos podem comparar a efectividade de diferentes técnicas e podem determinar a diferença entre os sistemas afectos ao rendimento.

Como a colecção foi construída sob o programa TIPSTER se encontra frequentemente sob o nome de colecção de prova TIPSTER ou TIPSTER/TREC. Ainda que, no entanto, o nome mais comum é o da colecção TREC.

Basicamente para a avaliação do sistema de recuperação se utilizará a colecção de documentos TREC. Como tem sido com a maioria das colecções de prova, a colecção TREC está constituida por três partes: (i) os documentos, (ii) os exemplos de petições ou necessidades de informação, chamados tópicos na nomenclatura TREC, (iii) um conjunto de documentos relevantes proporcionados por especialistas, para cada exemplo de petição da informação. Por outro lado, as conferências TREC também incluem outros conjuntos de tarefas para serem usados como provas.

A colecção de documentos TREC

A colecção TREC tem estado crescendo constantemente ao longo dos anos. Na conferência TREC-2, o tamanho da colecção era aproximadamente 2 gigabytes enquanto na conferência TREC-6 ultrapassou os 5,8 gigabytes. Ao princípio, as restrições dos direitos legais impediram a livre distribuição da colecção e, por isso, os discos 1, 2 e 3 da distribuição tinham que se vender já que dependem do LDC [127] (Linguistic Data Consortium). Em 1998, atingiu-se um acordo que permitiu o acesso livre aos documentos usados nas conferências TREC mais recentes. Como consequência disto, os discos 4 e 5 de TREC estão agora disponíveis no NIST pagando unicamente para cobrir os custos da distribuição. A colecção TREC se distribui em cinco CD-ROM de quase 1 gygabyte de texto comprimido cada um. Os

documentos que inclui provem das seguintes fontes: 1)Wall Street Journal, 2)Associated Press, 3)Selecção de Informática (artigos), Ziff-Davis, 4)Federal Register, 5)Publicações de US DOE, 6)San Jose Mercury News, 7)US Patentes, 8)Financial Times, 9)Congressional Record, 10)Foreign Broadcast Information service, 11) LA Times.

Aos documentos originais se lhes dá formato seguindo o padrão (*estándar*) SGML para permitir uma leitura fácil (o que permite escrever um código simples aos usuários dos documentos TREC). Há uma série de campos principais como são: Um campo para o número de documentos (identificado por <DOCNO>) e um campo para o texto do documento (Identificado por <TEXT>) que são comuns a todos os documentos. Campos menos importantes ou menos gerais poderiam ser diferentes nas diferentes subcolecções para preservar partes da estrutura do documento original. Esta foi à filosofia para as decisões de formato no NIST: preservar no possível a estrutura original e proporcionar ao mesmo tempo um meio de trabalho comum que permitisse uma dosificação simples dos dados.

O propósito de uma colecção de prova é ser capaz de avaliar a efectividade dos sistemas de recuperação. Proporcionando um esquema de avaliação comum é um elemento importante de TREC.

Nas conferências TREC se utilizam quatro medidas de avaliação básicas: resumo de estatísticas, medidas de *recall*-precisão, medidas de nível de documentos e o histograma de precisão média. A nível global, através destas diferentes medidas se apresenta os resultados de precisão e *recall* obtidos para as diferentes solicitações de informação propostas ao sistema.

A conferência TREC considera a colecção TREC como referência para sistemas de recuperação de informação hoje em dia, apesar que existem outras colecções como por exemplo a colecção CACM, com dimensões mais reduzidas e que consideram os documentos estruturados em campos. A colecção CACM está formada por 3.204 artigos das comunicações da ACM (Association for Computing Machinery, [3]).

No entanto, a medida em que foi aumentando o desenvolvimento de sistemas de busca orientados ao Web, também as colecções de documentos se foram adaptando, criando-se a TREC-8 Web track ou WEBTREC [83]. Esta colecção está baseada no conjunto de dados do VLC2 (Very Large Collection, second edition) obtidos a princípios de 1.997 a partir de Internet Arquive [104]. Ao todo se obtiveram aproximadamente 18,5 milhões de páginas, constituindo mais de 100 gigabytes de dados e conformando uma representação do estado de World Wide Web.

Os objectivos pretendidos se concentram tanto na obtenção de medidas de eficiência como da velocidade obtida nestes sistemas. As medidas de efectividade se seguem ponderando em base à precisão e o *recall*, tomando valores relativos ao número de documentos analisados, se bem, apreciados a pouca importância do *recall* nos documentos em Internet [83].

Os exemplos de necessidades de informação (Tópicos)

A colecção TREC inclui um conjunto de exemplos de petições de informação que podem ser usados para provar um determinado sistema de recuperação. Cada petição é uma descrição de uma necessidade de informação em linguagem natural. Na nomenclatura TREC, cada necessidade da informação de prova é conhecida como um tópico.

O texto de um tópico TREC tem um conjunto de *campos* com marcadores especiais. Por exemplo, o campo *Narrativa* proporciona uma descrição particularmente detalhada de que deve conter um documento relevante; o campo conceito lista normalmente palavras e frases que o criador do tópico considera relacionadas com o tópico. Também há versões disponíveis mais curtas do tópico. Esta versão mais curta, o sumário, é normalmente uma frase simples descrevendo as petições de busca. O Título é uma versão mais reduzida do que o Sumário. De todos os modos nem todos os tópicos têm os mesmos campos.

A tarefa de converter uma petição de informação (tópico) num sistema de consultas, por exemplo, num conjunto de termos indexados, numa expressão booleana, numa expressão difusa, etc. deve ser feita pelo próprio sistema de recuperação e se considera uma parte integral do processo de avaliação.

O número de tópicos preparados pelas seis primeiras conferências TREC passa de 350. Os tópicos de 1 à 150 se prepararam para o seu uso nas conferências TREC-1 e TREC-2. Foram escritos por pessoas que eram usuários expertos de sistemas reais e representavam necessidades de informação de muitos anos de antiguidade. Os tópicos de 151 à 200 se prepararam para o seu uso na conferência TREC-3, são mais curtos, e têm uma estrutura mais simples que inclui só três campos chamados *Título*, *Descrição* e *Narrativa*. Os tópicos de 201 à 250 foram preparados para a conferência TREC-4 e são ainda mais curtos. Na conferência TREC-5 (onde se incluíram os tópicos 251-300) e na TREC-6 (onde se incluíram os tópicos 301-350), os tópicos se prepararam com uma composição similar aos tópicos da conferência

TREC-3 e foram expandidos com respeito aos tópicos da conferência TREC-4 que se consideraram a posterior como demasiados curtos.

Os documentos relevantes para cada exemplo de petição de informação

Nas conferências TREC, o conjunto de documentos relevantes para cada tópico se obtém desde um conjunto de possíveis documentos relevantes. Este conjunto se cria tomando os K documentos mais relevantes (normalmente para K = 100) na classificação gerada por vários sistemas de recuperação participantes. Os documentos nesse grupo se mostram a uns assessores humanos que são os que finalmente decidem sobre a relevância de cada documento.

Esta técnica de assessoramento da relevância se conhece como método de sondagem e se baseia em duas suposições. Em primeiro lugar, que a grande maioria de documentos relevantes se recopila no conjunto final. A segunda, que os documentos que não estão no conjunto podem ser considerados como não relevantes. Ambas as suposições têm que ser verificadas para demonstrar sua exactidão nas provas feitas nas conferências TREC.

Medidas de avaliação nas conferências TREC

Como já se referiu anteriormente, nas conferências TREC se usam quatro tipos de medidas de avaliação: Tabela de resumo estatístico, medidas recall-precisão, medidas a níveis de documentos, e histogramas de precisão média.

Tabela de resumo de estatístico: Consiste numa Tabela que resumem as estatísticas relativas a uma tarefa dada. As estatísticas incluem: o número de tópicos usados nas tarefas, o número de documentos recuperados em todos os tópicos, o número de documentos relevantes que são efectivamente recuperados para todos os tópicos e o número de documentos relevantes que poderiam ter sido recuperados para todos os tópicos. Entre estas medidas cabe mencionar:

- Medidas *recall-precisão*: Consiste numa Tabela ou Gráfico com a precisão média para todos os tópicos a 11 níveis padrão de *recall*.

- Medidas a *níveis de documento*: Neste caso, a precisão média para todos os tópicos se calcula em documentos com valores específicos de recall (em

lugar dos níveis padrão de recall). Por exemplo, a precisão média deveria ser computada quando 5, 10, 20, 1000 documentos relevantes foram vistos.

- *Histograma de precisão média*: Consiste num gráfico que inclui uma medida simples separada para cada tópico.

6.1.2 Métricas a utilizar

Para começar com a avaliação do rendimento da recuperação, deveríamos considerar primeiro a tarefa de recuperação a ser avaliada. Por exemplo, a tarefa de recuperação podia constar simplesmente de uma consulta processada por lotes (isto é, o usuário envia uma consulta e recebe uma resposta de volta) ou de uma sessão interactiva inteira (isto é, o usuário especifica a sua necessidade da informação através de uma série de passos interactivos com o sistema). A tarefa de recuperação poderia também compreender uma combinação destas estratégias. As tarefas de consultas interactivas ou por lotes são processadas bastante diferentes e suas avaliações também são diferentes. De facto, numa sessão interactiva, o usuário valoriza características do desenho da interface, a guia proporcionada pelo sistema e a duração da sessão que são aspectos críticos que deveriam ser observados. Numa sessão de processamento por lotes, nenhum destes aspectos é tão importante como à qualidade do conjunto de resposta gerado. Neste caso, há que ter em conta que o que nos interessa é *comparar as consultas entre si*, isto é, diferentes versões de uma mesma necessidade da informação inicial que se transforma em consultas sem nenhum tipo de expansão e desambiguação do sentido das palavras ou com expansão e desambiguação do sentido das palavras dos diferentes tipos possíveis, sendo neste último caso necessária certa iteracção com o usuário. Suponhamos que numa consulta $\langle A, B\rangle$ o termo A tem como significados possíveis A_1, A_2 e A_3 e supondo que se seleccionou o termo A_1 como correcto, a consulta ficaria da seguinte forma: $\langle A, B, A_1\rangle$. No entanto, para comparar as consultas ante um mesmo algoritmo e esquema de recuperação bastará em aplicar as medidas que se explicam a seguir, supondo que a selecção realizada pelo usuário à hora de efectuar a expansão é mais adequada.

Recall e precisão

Tradicionalmente as principais medidas da qualidade dos documentos recuperados por um sistema de busca sempre foram *à precisão* e *o recall*. Estas medidas

permitem realizar comparações a partir de um exemplo de solicitação de informação ou documentos. Consideramos um exemplo de petição da informação I de uma colecção de referência de prova e o seu conjunto R de documentos relevantes. Seja $|R|$ o número de documentos neste conjunto. Suponhamos que uma estratégia de recuperação dada, que está sendo avaliada, processa a petição da informação I e gera um conjunto de documentos de resposta A. Seja $|A|$ o número de documentos neste conjunto. Adicionalmente, seja $|R_a|$ o número de documentos na intersecção dos conjuntos R e A, isto é o número de documentos recuperados que são relevantes (ver Figura 5.1) [15] (cap: 3). As medidas de *recall* e precisão se definem da seguinte forma:

- Recall: é a fracção dos documentos relevantes (no conjunto R) que foram recuperados isto é:

$$\text{Recall} = \frac{\text{número de documentos relevantes recuperados}}{\text{número de documentos relevantes na colecção}} = \frac{|R_a|}{|R|} \tag{6.1}$$

- Precisão: é a fracção dos documentos recuperados (no conjunto A) que é relevante isto é:

$$\text{Precisão} = \frac{\text{número de documentos relevantes recuperados}}{\text{número de documentos recuperados}} = \frac{|R_a|}{|A|} \tag{6.2}$$

As medidas de *recall* e precisão, definidas anteriormente, supõem que todos os documentos do conjunto de resposta A foram examinados (ou vistos). No entanto, num sistema de recuperação, ao usuário não se lhe apresenta todos os documentos no conjunto de resposta A de uma vez. Os documentos de A se ordenam primeiro de acordo com um grau de relevância, isto é, gera-se uma ordem. O usuário examina então esta lista ordenada começando pelo primeiro documento. Nesta situação, as medidas de *recall* e precisão variam segundo como o usuário proceda com o exame do conjunto A. Por isso, uma avaliação apropriada requer desenhar um gráfico de *precisão* em contraposição ao *recall* (gráfico de *precisão/recall*).

Como antes, consideramos uma colecção de referência e os seus conjuntos de exemplos de petições de informação. Centraremos-nos num exemplo de necessidade da informação dado, para o qual se gera uma consulta q. Suponhamos que se definiu um conjunto Rq contendo os documentos relevantes para q.

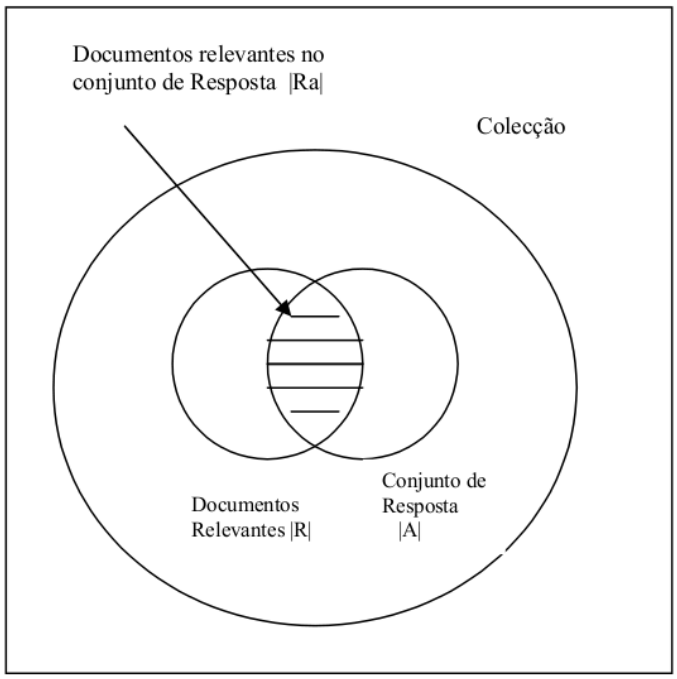

Figura 6.1: Precisão/*recall* para um exemplo de solicitação de informação.

Vejamos a continuação cômo se obtém os valores da *precisão e racall*, assim como os Gráficos de *precisão não interpolada* e *interpolada* para tal, suponhamos que uma colecção de documentos tem 20 documentos, quatro dos quais são relevantes para o tema t. Suponhamos por outro lado de que um sistema de recuperação ordena os documentos relevantes primeiro, segundo, quarto e décimo quinto.

Seja $R_q = \{d3, d5, d9, d20\}$ o conjunto dos documentos relevantes, os valores de precisão e *recall* correspondentes mostram-se na Tabela 6.1.

Rank	DocId	Recall	Precisão
1	d3	1/4	1/1
2	d5	2/4	2/2
3			
4	d9	3/4	3/4
⋮			
15	d20	4/4	4/15

Tabela 6.1: Exemplo de Precisão/*Recall*.

Uma situação mais comum é que os sistemas proporcionam conjuntos de respostas ordenadas (*ranked*) utilizando uma medida não binária de similaridade entre

os documentos e as perguntas. Neste caso à medida que se recuperam mais documentos, o *recall*, aumenta enquanto a precisão normalmente decresce. Então uma avaliação adequada tem que produzir valores de Precisão/Recall em pontos dados do *ranking*. Isto proporciona uma visão incremental das medidas de execução ou desempenho de recuperação quando encontramos todos os documentos relevantes. A Tabela 6.1 apresenta um exemplo. Os valores de Precisão/Recall se computam depois de encontrar um documento relevante (1ª, 2ª, 4ª e 15ª posição no ranking). Como o conjunto de respostas tem quatro documentos relevantes, a Figura final se compõe de quatro troços e se obtém o gráfico dos valores de **precisão não interpolada** (ver Figura 6.2).

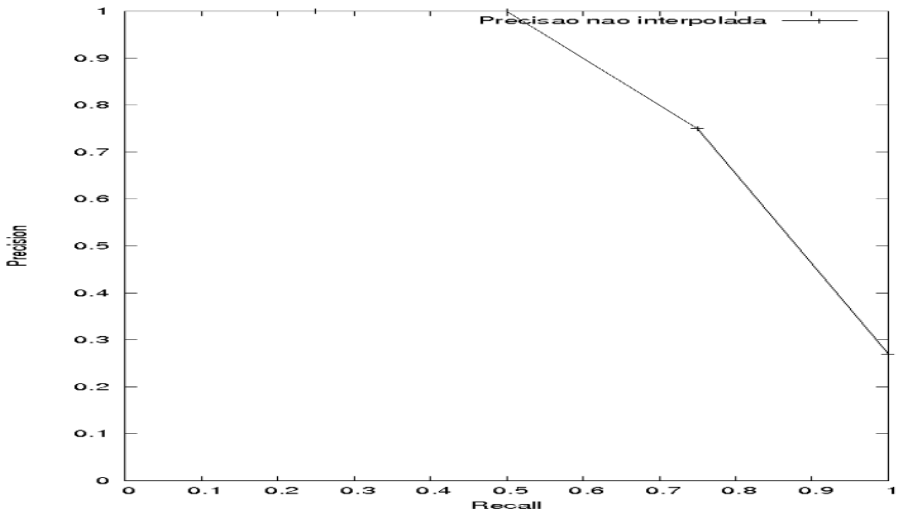

Figura 6.2: Exemplo do gráfico de precisão não interpolada.

As Figuras de Precisão/Recall se baseiam normalmente em onze níveis de *recall* padrão que são 0%, 10%, 20%, 30%, ..., 100% que é a curva padrão. Para cada um desses níveis de *recall* se calcula a precisão a esse nível e se representa em forma de gráfico situando o *recall* no eixo x e a precisão no eixo y (ver Figura 6.4).

Para uma avaliação adequada do sistema de recuperação não é suficiente, como é natural, com a prova de uma só consulta, tão pouco é adequado avaliá-lo mediante um conjunto de consultas se o seu rendimento se mede de maneira individual gerando para cada consulta se gera uma curva de precisão em contraposição ao recall (precisão-recall). Para avaliar o rendimento que produz a expansão de maneira global sobre todas as consultas de prova, calculamos a média das Figuras de precisão para cada nível de *recall* da seguinte forma:

$$\bar{P}(r) = \sum_{i=1}^{N_q} \frac{P_i(r)}{N_q} \tag{6.3}$$

Onde $\bar{P}(r)$ é a precisão média no nível de recall r, N_q é número de perguntas utilizadas e $P_i(r)$ é a precisão no nível de recall r para a (i-ésima) pergunta. Como resultado se obtém uma *Figura geral de precisão/recall*.

Os níveis de *recall* padrão facilitam a medição e traçado dos resultados de recuperação. Como os níveis de *recall* para cada consulta poderiam ser diferentes dos 11 níveis *recall* padrão, é muitas vezes necessário utilizar um *procedimento de interpolação*. A Figura de precisão aos 11 níveis de *recall* r se obtém com o seguinte processo de interpolação (ver Figura 6.3).

Seja r_j, $j \in \{0, 1, 2, ..., 10\}$, uma referência ao j-ésimo nível padrão de *recall*, por exemplo, r_5 é uma referência ao nível de *recall* de 50%. Então,

$$P(r_j) = \max r_j \leq r \leq r_{j+1} P(r) \tag{6.4}$$

Valores de Prec./Rec. não Int.		Valores de Precisão/Recall interpolados	
Recall	Precisão	Recall	Precisão
0.27	1.00	0.00	$\max\, r_0 \leq r \leq r_1 P(r) = 1.00$
0.50	1.00	0.10	$\max\, r_1 \leq r \leq r_2 P(r) = 1.00$
0.75	0.75	0.20	$\max\, r_2 \leq r \leq r_3 P(r) = 1.00$
1.00	0.27	0.30	$\max\, r_3 \leq r \leq r_4 P(r) = 1.00$
		0.40	$\max\, r_4 \leq r \leq r_5 P(r) = 1.00$
		0.50	$\max\, r_5 \leq r \leq r_6 P(r) = 1.00$
		0.60	$\max\, r_6 \leq r \leq r_7 P(r) = 0.75$
		0.70	$\max\, r_7 \leq r \leq r_8 P(r) = 0.75$
		0.80	$\max\, r_8 \leq r \leq r_9 P(r) = 0.27$
		0.90	$\max\, r_{10} \leq r \leq r_{11} P(r) = 0.27$
		1.00	$P(r_{10}) = P(100\%) = 0.27$

Figura 6.3: Processo de interpolação para uma Figura de Precisão/*Recall*.

Utilizando a regra de interpolação, a **precisão interpolada** para todos os níveis de *recall* padrão até 0.5 é 1, a precisão interpolada para os níveis de recall 0.6 e 0.7 é 0.75 e a precisão interpolada para os níveis de *recall* 0.8 ou maiores é 0.27. Isto pode observar-se nas Figuras 6.3 e 6.4.

O que estabelece que a precisão interpolada no nível padrão j-ésimo de *recall* é a precisão máxima conhecida para qualquer nível de *recall* entre o nível de *recall*

j-ésimo e o nível de recall $(j + 1)$-ésimo. A curva de precisão em contraposição ao *recall* que resulta de promediar os resultados de várias consultas se conhecem normalmente como *Figuras de precisão* em contraposição ao *recall*.

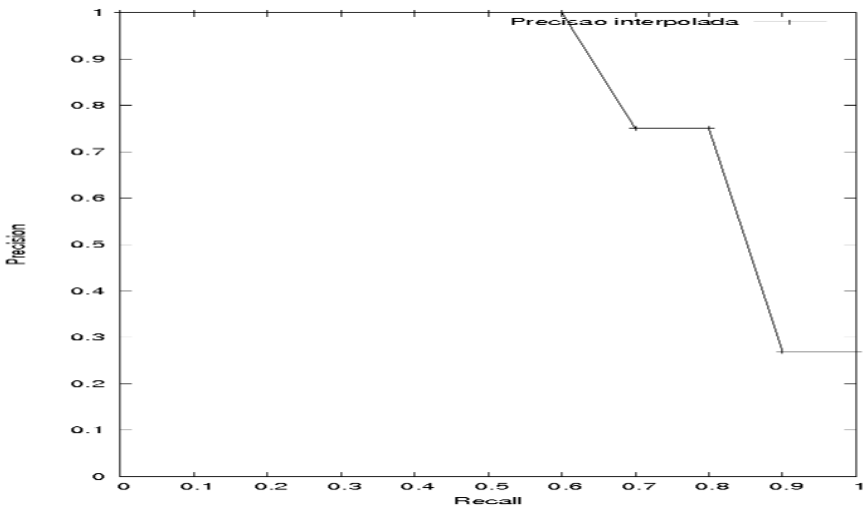

Figura 6.4: Figura da precisão interpolada a 11 níveis de Recall padrão.

Estes gráficos de médias se usam normalmente para comparar o rendimento da recuperação de diferentes algoritmos de recuperação unindo as diferentes curvas num mesmo gráfico. Por exemplo, se poderia comparar o rendimento da recuperação de um algoritmo de recuperação novo com o rendimento da recuperação do clássico modelo vectorial. Nesta tese, como o algoritmo de recuperação vai ser o mesmo, o que realmente se vai comparar serão as melhorias que podem produzir as consultas expandidas e desambiguadas com respeito às consultas originais, resultantes da colecção de termos.

Uma aproximação adicional costuma ser calcular a precisão média a uns valores de corte de documentos dado. Por exemplo, podemos calcular a precisão média quando 5, 10, 15, 20, 30, 50, ou 100 documentos relevantes foram vistos. O procedimento é análogo ao cálculo da precisão média em 11 níveis padrão de recall, mas proporciona informação adicional no rendimento do algoritmo de ordenação, mas nesta avaliação em concreto, ao utilizar o mesmo algoritmo de ordenação para todas as provas, não é de interesse, já que o que estamos comparando são as consultas expandidas com as consultas sem expandir e não o algoritmo.

A Figura de precisão em contraposição ao recall são estratégias de avaliação padrão para os sistemas de recuperação da informação e se usam externamente

na literatura sobre recuperação de informação. São úteis porque nos permitem avaliar quantitativamente tanto a qualidade do conjunto de resposta global como o alcance do algoritmo de recuperação. Além disto, são simples, intuitivos, e se podem combinar numa única curva. No entanto, as Figuras de precisão em contraposição ao recall também têm suas desvantagens e o seu uso estendido foi criticado na literatura, se falará brevemente sobre isto na Secção 6.1.3. Seguidamente se discutirão técnicas para resumir as Figuras de precisão com relação as de um único valor numérico.

Resumo de valor único

As Figuras de precisão média em contraposição ao recall são úteis para comparar o rendimento da recuperação de diferentes algoritmos de recuperação sobre um conjunto de consultas. No entanto, há situações nas quais nos agradaria comparar o rendimento de recuperação dos nossos algoritmos de recuperação para as consultas individuais. As razões são duas: (i) A precisão média sobre muitas consultas poderia ocultar anomalias importantes nos algoritmos de recuperação em estudos que se desse em consultas concretas. (ii) Quando comparamos dois algoritmos, poderíamos estar interessados em investigar se um deles melhora ao outro para todas as consultas num exemplo dado de consultas, nota-se que este facto se pode ocultar facilmente ao realizar um cálculo de precisão média. Nestas situações, se pode usar um único valor de precisão para cada consulta. Este valor único deveria interpretar-se como um resumo da curva de precisão em contraposição ao recall correspondente. Normalmente, este valor único se toma como a precisão a um nível específico de *recall*. Por exemplo, poderíamos avaliar a precisão quando vemos o primeiro documento relevante e tomar esta precisão como o valor de resumo único. Por suposto, como resulta óbvio, esta não é uma boa aproximação, porque ao avaliarmos a precisão de documentos relevantes sobre um determinado valor de corte ou nível específico no *ranking*, não estariamos a investigar globalmente a qualidade do conjunto de resposta do algoritmo de recuperação. Se podem adoptar estratégias mais interessantes que se discutirão a continuação.

Precisão média em documentos relevantes vistos

A ideia é gerar um valor de resumo único da classificação mediante a média da Figura de precisão obtida depois de que cada novo documento relevante é observado na ordenação. Esta medida favorece aos sistemas que recuperem os documentos rele-

vantes rapidamente, isto é, pronto na ordenação. Por suposto, o algoritmo apresenta uma média em documentos relevantes vistos.

Precisão R

A ideia aqui é gerar um valor de resumo único da ordenação mediante o cálculo da precisão R da lista, onde R é o número total de documentos relevantes para a consulta actual, isto é, um número de documentos no conjunto R_q. A medida de precisão R é um parámetro útil para observar o comportamento de um algoritmo para cada consulta individual num experimento. Adicionalmente, pode-se também calcular uma Figura de precisão R média para todas as consultas. No entanto, usar um número único para resumir o comportamento completo de um algoritmo de recuperação sobre várias consultas poderia ser bastante impreciso.

Histograma de precisão

As medidas de precisão R para várias consultas se podem usar para comparar a história de recuperação de dois algoritmos de modo seguinte. Seja $RP_A(i)$ e $RP_B(i)$ os valores de precisão R dos algoritmos de recuperação A e B para a consulta i-ésima. Definamos, por exemplo, a diferença:

$$RP_{A/B}(i) = RP_A(i) - RP_B(i) \tag{6.5}$$

Um valor de $RP_{A/B}(i)$ igual à zero indica que ambos os algoritmos têm rendimento equivalente, em termos de precisão R, para a consulta i-ésima. Um valor positivo $RP_{A/B}(i)$ indica um melhor rendimento de recuperação do algoritmo A (para a consulta i-ésima) enquanto um valor negativo indica um rendimento de recuperação melhor do que o algoritmo B. Com este tipo de valores se gera um histograma com uma barra para cada consulta. Chama-se *histograma de precisão* e nos permite comparar rapidamente a história do rendimento de recuperação de dois algoritmos através da inspecção visual.

Tabela de resumo de estatística

As medidas do valor único se podem armazenar também numa Tabela para proporcionar um resumo estatístico com respeito ao conjunto de todas as consultas

numa tarefa de recuperação. Por exemplo, estas Tabelas de resumo de estatísticas poderiam incluir: o número de consultas usadas na tarefa, o número total de documentos recuperados por todas as consultas, o número total de documentos relevantes que foram recuperados efectivamente quando se consideraram todas as consultas, o número total de documentos relevantes que poderiam ter sido recuperados por todas as consultas, etc.

6.1.3 Adequação da precisão e recall

A precisão e o *recall* são usados extensamente para avaliar o rendimento de algoritmo de recuperação. No entanto, uma reflexão mais cuidadosa revela problemas com estas duas medidas: (i) A própria estimação do *recall* máximo para a consulta requer conhecimento detalhado de todos os documentos na colecção. Com colecções grandes, este conhecimento não está disponível, já que o que se faz é fixar um número máximo de documentos relevantes considerando-se o resto de documentos como não relevantes, podendo sê-lo, ainda que em menor medida do que os anteriores. Isto implica que o *recall* não se pode estimar de maneira precisa. (ii) O *recall* e a precisão são medidas relacionadas que capturam aspectos diferentes de conjunto de documentos recuperados. Em muitas situações, o uso de uma medida única que combinasse *recall* e precisão poderia ser mais apropriado. (iii) O *Recall* e precisão medem a efectividade sobre um conjunto de consultas processadas por lotes. No entanto, com sistemas modernos, a interactividade é o aspecto chave do processo de recuperação. Assim, as medidas que quantificam a quantidade da informação do processo de recuperação poderiam ser agora mais apropriadas. (iv) O *recall* e a precisão são fáceis de definir quando se faz cumprir um ordenamento linear dos documentos recuperados. No entanto, para sistemas que requerem um ordenamento débil, o *recall* e a precisão poderiam não ser adequados.

6.1.4 Avaliação do rendimento do sistema

Outros aspectos básicos de um sistema de recuperação de informação é o tempo empregado no processo de busca e o tempo empregado na construção das estruturas de dados necessários para a realização da busca. De facto, em [117] se reconhece uma relação entre as três medidas típicas dos sistemas de recuperação de informação: velocidade (ou tempo de resposta), precisão e *recall* (ver Figura 6.5).

O objetivo primordial de todo sistema de busca consiste em obter um equilíbrio

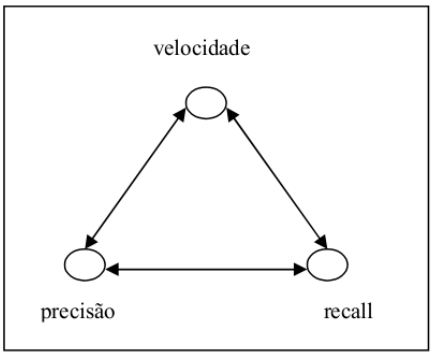

Figura 6.5: Relação entre velocidade, precisão e *Recall* num sistema de recuperação de informação.

entre estas três medidas, aspecto que se converte em mais complicado quando se produz um aumento no número de documentos e/ou de usuários do sistema. Em [243], se descreve de forma detalhada os critérios a ter em conta na hora de realizar comparações entre vários sistemas de indexação, dos quais, os mais relevantes são os seguintes: Escalabilidade, tempo de reposta, espaço usado e tempo de CPU.

Mais nesta tese os experimentos estão relacionados com a qualidade das respostas do sistema, perante distintas consultas que representam a necessidade de informação do usuário, e não em medir a eficiencia do sistema.

6.2 Desenho experimental e realização de exercícios

Para a realização dos experimentos, se utilizou o sistema de recuperação Lemur-3.1 [132] e a colecção de referência TREC [218]. Concretamente, em todos os experimentos realizados que se comentarão a seguir, tomaram-se os tópicos TREC numerados de 401 á 450 de TREC-8 para obter cinquenta consultas de provas em cada experimento. Quanto à colecção de documentos se usou a colecção Small Web de TREC-8 "WT2g" que contém 250.000 documentos aproximadamente, um conjunto de documentos relevantes proporcionados por especialistas, e, além disto, se utilizaram as medidas de *recall* e precisão para determinar se há melhorias na recuperação dos documentos, relacionando as consultas: original, expandida e expandida desambiguada.

A partir destes tópicos as consultas se geraram na maioria dos experimentos de forma automática a partir das palavras dos títulos, seleccionando todas excepto aquelas muito habituais que figuravam no ficheiro de *StopWords*. Uma consulta gerada sem expandir é simplesmente um conjunto dos termos do título.

Para a busca dos significados dos termos seleccionados em WordNet [235] se utilizou além da *taxonomia de substantivos (nomes)* outras taxonomias para comprovar a efectividade das mesmas no processo de recuperação de informação. Mas durante os experimentos realizados nesta tese doutoral que se comentarão a continuação, verificou-se que as outras taxonomias existentes não são úteis na expansão de consultas: a *taxonomia de advérbios* é claramente inservível já que em qualquer documento se pode utilizar qualquer advérbio independentemente da temática; *a taxonomia dos verbos* igualmente não serve apesar de se ter incluido na experimentação porque, em geral, um verbo não determina uma necessidade de informação e as possibilidades de expansão são poucas; por último *a taxonomia de adjectivos* não é aconselhável porque estes apenas aparecem nos títulos dos tópicos e, além disto, um adjectivo pode ser aplicado a múltiplos substantivos e, portanto, aparecer, em documentos tratando qualquer tema. Em definitiva, a busca nestas taxonomias não contribuiram para nada na efectividade da recuperação e só provocam demasiado ruído na expansão de consultas. Uma vez obtida a consulta expandida e a desambiguação do sentido das palavras, o campo seleccionado "título" (do tópico TREC) foi indexado utilizando as rotinas padrões do sistema de recuperação de informação Lemur, depois se computou um *rank* de documentos para cada tópico. Para cada tópico os documentos se ordenam em ordem decrescente de similaridade.

Uma vez obtidos os resultados da consulta expandida e desambiguada do sentido das palavras dos diferentes experimentos contemplados se procede a sua avaliação e a comparação entre elas com as consultas sem expandir.

A partir dos resultados do algoritmo se calcularam, para cada conjunto de consultas, os valores médios de precisão aos 11 níveis padrões de *recall* que, como se disse, é a medida mais habitual usada para comparar sistemas de recuperação. Estes valores se comentarão a seguir partindo desde gráficos conjuntos de precisão em contraposição ao *recall*.

Outro aspecto incluido no desenvolvimento do prototipo usado nos estudos realizados é a expansão mediante hipónimos de primeiro nível além da expansão mediante sinónimos. No entanto, decidiu-se não utilizar finalmente esta funcionalidade na experimentação. A causa é a natureza da relação de hiponímia inclusa em WordNet.

Isto deve-se a que em muitas palavras o número de hipónimos é muito elevado. Por exemplo, a palavra *oil* tem aproximadamente 30 hipónimos somente no primeiro nível e um grande número dos termos dos títulos dos tópicos sobre os que se decidiu efectuar as provas, tem um número ao redor de 15 hipónimos. Este número tão elevado introduziria demasiado ruído, podendo-se reduzir este unicamente se leva a cabo uma laboriosa selecção manual pelo que grande parte do interesse do experimento se perderia. Inclusive naqueles termos nos que o número de hipónimos não é elevado, ou supondo que se pudesse filtrar de maneira automática e adequada um número de hipónimos limitado, observou-se que os hipónimos costumam ser palavras que fazem demasiado concreto o termo original restringindo demasiado os possíveis documentos relevantes. Portanto, quando muito se poderia seleccionar um hipónimo por cada termo original, e este em lugar de expandir a consulta mais bem a restringiria a um tema mais concreto. Voltando ao exemplo de *oil*, que é representativo, a maioria dos seus hipónimos são palavras compostas de *oil* e outro termo designando múltiplos tipos de azeites ou derivados do petróleo. Convém recordar que o objectivo da expansão é adaptar o vocabulário do usuário aos vocabulários dos diferentes documentos, com este tipo de expansão não se conseguiria este objectivo.

Os experimentos realizados consistiram em seleccionar automaticamente os termos relacionados para cada uma das palavras das consultas originais e efectuar duas expansões: (i) uma com os sinónimos para o significado correcto de cada palavra e (ii) outra com os sinónimos para os dois tipos de expanssões (com sinificado correcto e sem o significado correcto). Os termos relacionados se adicionam, neste e em todos os experimentos, directamente ao termo original e não como termos independentes ao final da consulta. Isto é, dada uma consulta se o termo A_1 está relacionado com A e o termo B_1 está relacionado com B a consulta expandida, ficará de seguinte modo: (A, A_1, B, B_1, C). Os termos relacionados seleccionados foram os sinónimos para o significado correcto de cada palavra.

6.3 Expansão de consultas com e sem a WSD

Nesta secção apresentamos os resultados e a interpretação da expansão de consultas com e sem a WSD, utilizando somente a categoria sintáctica de nomes. Nos centramos aos efeitos de uso da expansão de consultas com e sem a WSD num sistema de recuperação de informação. Na Figura 6.6 e Tabelas 6.2 e 6.3 respectivamente apresentamos os valores de *recall* e precisão obtidos ao expandir a consulta original utilizando somente a categoria sintáctica de nomes. Neste experimento a

WSD foi aplicada (WSD). Na prática quando a expansão de consultas foi aplicada se obteve uma precisão média não interpolada de 16.49%. Realmente a precisão média não interpolada da consulta original é de 24.31% e consequentemente os resultados obtidos pela consulta original são melhores com respeito às consultas expandidas. O método de WSD utilizado durante os experimentos é o baseado em dicionário, utilizando o recurso línguistico WordNet. Basicamente está baseado em ([179] e [180]) e contém os seguintes passos: (i) Adquirir desde WordNet os conjuntos de sinónimos das palavras a desambiguar, (ii) Determinar a coincidência entre o contexto das palavras a desambiguar e os conjuntos de sinónimos, (iii) Escolher o significado correcto das palavras num texto, dado o contexto de forma automática.

Nas Figuras 6.6 e 6.7 e nas Tabelas 6.2 e 6.3 respectivamente apresentamos igualmente os resultados obtidos quando a expansão de consultas é combinada com WSD. Como se podem ver os valores de precisão média não interpolada nas Figuras e Tabelas referenciadas comparando os resultados observa-se que aplicando a desambiguação melhoram os resultados de recuperação com relação à consulta expandida sem a desambiguação. Concretamente utilizando a expansão de consulta sem a desambiguação se obteve uma precisão média não interpolada de (16.49%), incluindo a desambiguação se obteve uma precisão média não interpolada de 21% aproximadamente confirmando o visto nos Gráficos e Tabelas. De facto a precisão da consulta expandida e desambiguada é notavelmente maior para todos os níveis de *recall* com respeito a consulta expandida e não desambuguada. Igualmente os resultados obtidos pela consulta original são melhores com respeito às consultas expandidas com WSD.

6.4 Expansão de consultas sem WSD com distintas categorias sintácticas

Nesta secção apresentamos os resultados da expansão de consultas sem a WSD, utilizando as distintas categorias sintácticas. Como se podem comprovar na Figura 6.8 os resultados dos experimentos são claramente melhores quando se utilizou a expansão somente com nomes. De facto a expansão é notavelmente melhor e maior para todos os níveis de *recall*. Como valor de resumo se pode extrair a precisão média não interpolada, visto que as precisões médias não interpoladas das consultas expandidas somente com nomes, adjectivos, verbos e advérbios são de 16.49%, 9.64%, 6.76% e 4.47% respectivamente. A precisão média não interpolada

Recall - Precisão	Consulta Expandida Sem WSD	Consulta Expandida Com WSD	Consulta original
0.00	0.5203	0.5929	0.6714
0.10	0.3800	0.4415	0.5276
0.20	0.2737	0.3226	0.3994
0.30	0.2090	0.2535	0.3365
0.40	0.1726	0.2239	0.2570
0.50	0.1499	0.1982	0.2275
0.60	0.1144	0.1558	0.1746
0.70	0.0878	0.1229	0.1409
0.80	0.0732	0.0963	0.1066
0.90	0.0579	0.0781	0.0759
1.00	0.0363	0.0485	0.0454
Precisão Média Não Interpolada	0.1649	0.2051	0.2431
% Câmbio de Precisão Média		+24.38%	+47.42%

Tabela 6.2: Valores de Recall-Precisão dos resultados das consultas: Original, expandida sem WSD e expandida com WSD.

Consultas	Expandida sem WSD	Expandida Com WSD	Original
Tópicos	50	50	50
Número de Doc.Rel.	2279	2279	2279
Núm.Doc.Rel. Recuperados	1677	1688	1652

Tabela 6.3: Resumo estatístico dos resultados da consultas: Original, expandidas sem WSD e com WSD.

da consulta original é 24.31%.

A diferença percentual da precisão média não interpolada da consulta original com respeito à consulta expandida somente com os nomes é apenas de 7.82%, mas com relação às outras consultas as diferenças são abismais que são 14.67% com relação à expansão utilizando somente os adjectivos, 17.55% com os verbos e 19.84% com relação aos advérbios, neste caso a expansão somente com nomes é a melhor de todas, registando uma melhoria de +268.90% com relação à consulta expandida somente com advérbios que é a pior de todas, confirmando o visto nos gráficos da Figura 6.8 e na Tabela 6.4 e nota-se que as taxonomias de adjectivos, verbos e advérbios provocam demasiado ruído na expansão de consultas.

Segundo a Tabela 6.5 a expansão somente com os nomes também é melhor em termos de *recall* com respeito a todas as consultas, já que recupera mais documentos

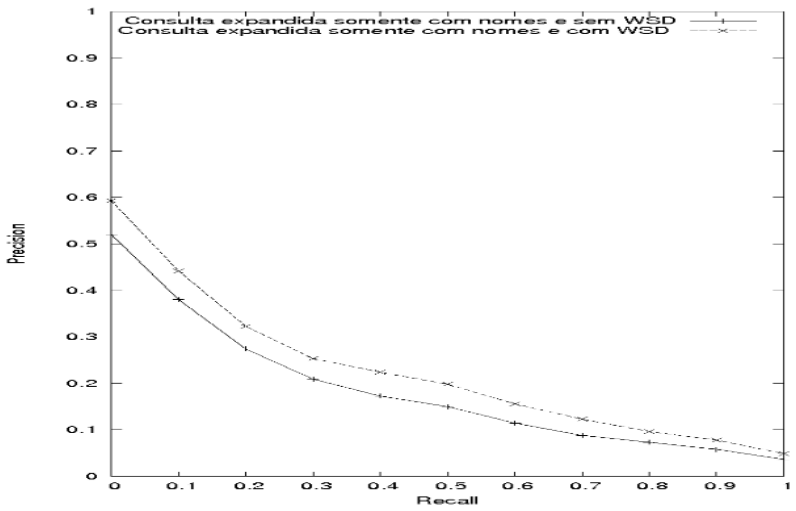

Figura 6.6: Gráfico de Recall-Precisão das consultas: expandida sem WSD e expandida com WSD.

relevantes implicando o aumento de *recall*, a expansão de consultas pode melhorar o *recall* desde que não haja nenhum ruído na consulta a expandir. Fazendo uma retrospectiva verifica-se que a expansão de consultas com as categorias sintácticas de adjectivos, advérbios e verbos são inservíveis na expansão de consultas, visto que somente provocam ruído baixando assim a precisão e o *recall* pelas seguintes razões mencionadas anteriormente: a taxonomia de advérbios é claramente inservível, pois que em qualquer documento se pode utilizar qualquer advérbio independentemente da temática; a taxonomia dos verbos igualmente é inservível apesar de se incluir na experimentação porque, em geral, um verbo não determina uma necessidade de informação e as possibilidades de expansão são poucas; por último a taxonomia de adjectivos não é aconselhável porque estes apenas aparecem nos títulos dos tópicos e, além disto, um adjectivo pode ser aplicado a múltiplos substantivos e, portanto, aparecer, em documentos tratando qualquer tema. Em definitiva, a busca nestas taxonomias não contribuem para nada e só provoca demasiado ruído na expansão das consultas.

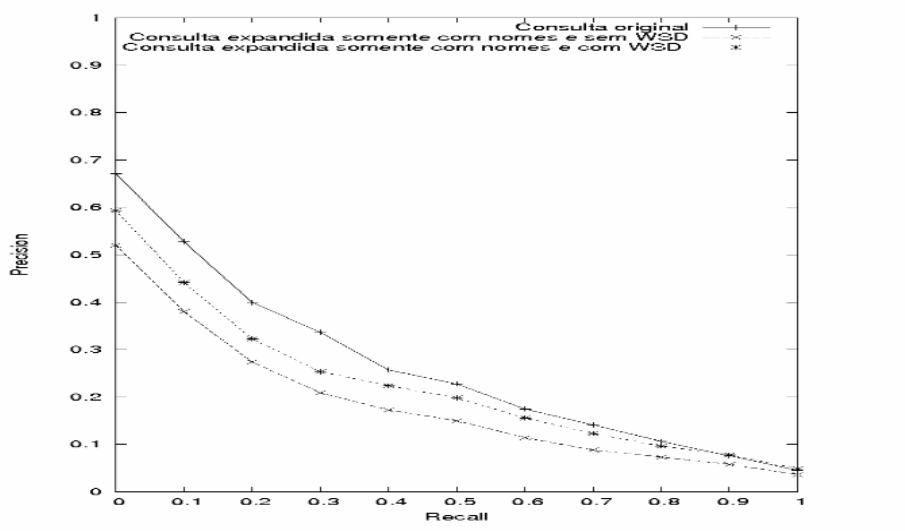

Figura 6.7: Gráfico de Recall-Precisão das consultas: Original, expandida sem a WSD e expandida com WSD.

6.5 Expansão de consultas e WSD com distintas categorias sintácticas

Nesta secção apresentamos os resultados e interpretação da expansão de consultas aplicando WSD, com as distintas categorias sintácticas. De acordo a Figura 6.9, a Tabela 6.6 e a 6.7, os resultados demonstram que sejam quais forem os resultados obtidos na expansão de consultas com qualquer categoria sintáctica, a WSD melhora a eficácia da recuperação de informação. Concretamente aplicando o algoritmo de desambiguação proposto utilizando somente a categoria sintáctica de nomes se obteve uma precisão média não interpolada de 21% aproximadamente e com respeito à consulta expandida sem a WSD se obteve 17% aproximadamente. Na categoria de adjectivos se obteve uma precisão média não interpolada de 11.95% com respeito à consulta expandida e WSD e 9.64% com relação a consulta expandida sem a WSD. Na categoria de verbos se obteve uma precisão média não interpolada de 8.53% com respeito à consulta expandida e WSD e 6.76% da consulta expandida sem a WSD, e na categoria de advérbios uma precisão média não interpolada de 6.28% com respeito à consulta expandida e WSD e 4.47% da consulta expandida sem a WSD. Apesar das melhorias em todas as categorias na expansão de consultas e WSD, a categoria de nomes é a melhor em ambos os aspectos, visto que se obteve a precisão média não interpolada mais alta tanto na expansão da consulta sem a WSD como

Recall -Precisão	Consulta expandida somente com:				Consulta original
	advérbios	verbos	adjectivos	nomes	
0.00	0.2449	0.3230	0.4187	0.5203	0.6714
0.10	0.1471	0.2152	0.2533	0.3800	0.5276
0.20	0.0868	0.1142	0.1746	0.2737	0.3994
0.30	0.0533	0.0893	0.1401	0.2090	0.3365
0.40	0.0406	0.0677	0.0985	0.1726	0.2570
0.50	0.0315	0.0496	0.0729	0.1499	0.2275
0.60	0.0229	0.0392	0.0544	0.1144	0.1746
0.70	0.0112	0.0392	0.0347	0.0878	0.1409
0.80	0.0056	0.0104	0.0234	0.0732	0.1066
0.90	0.0007	0.0031	0.0140	0.0579	0.0759
1.00	0.0002	0.0010	0.0063	0.0363	0.0454
Precisão Média Não Interpolada	0.0447	0.0676	0.0964	0.1649	0.2431
%Câmbio Prec. Média		+51.23%	+115.60%	+268.90%	+443.87%

Tabela 6.4: Valores de *Recall*-Precisão dos resultados das consultas expandidas com distintas categorias sintácticas e com a original.

Consulta	Expandida somente com:				Original
	advérbios	verbos	adjectivos	nomes	
Tópicos	50	50	50	50	50
Núm. Doc. Rel.	2279	2279	2279	2279	2279
Núm.Doc. Rel. Recuperados	761	912	1043	1677	1652

Tabela 6.5: Resumo estatístico dos resultados das consultas expandidas com distintas categorias sintácticas e com a original.

na expansão da consulta com WSD e além disto, nesta categoria houve melhorarias em termos de *recall* em ambos os casos, superando no entanto o *recall* da consulta original.

6.6 Expansão de consultas combinando todas as categorias sintácticas sem e com WSD

Nesta secção apresentamos os resultados e interpretação sobre a expansão de consultas, combinando todas as categorias sintácticas sem e com a WSD. Atendendo a Figura 6.10, as Tabelas 6.8 e 6.9, os resultados da expansão de consultas

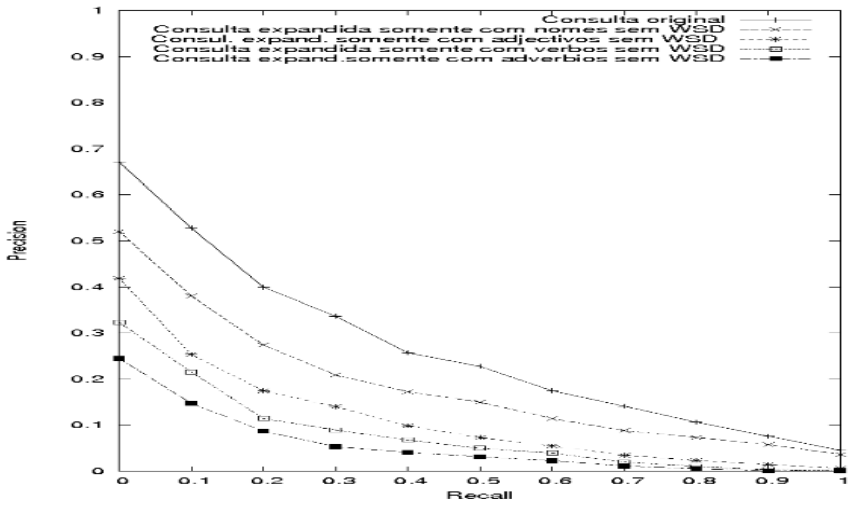

Figura 6.8: Gráfico de Recall-Precisão das consultas expandidas sem a WSD com distintas categorias sintácticas e com a consulta original.

combinando todas as categorias sintácticas (nomes, adjectivos, verbos e advérbios) aplicando a desambiguação das palavras se obteve uma precisão média não interpolada de 10.41% e sem a desambiguação uma precisão média não interpolada de 7.67%. Estes resultados não melhoraram os resultados da consulta expandida e a expandida com desambiguação do sentido das palavras utilizando a categoria de nomes que obteve as precisões médias não interpoladas de 16.49% e 20.51% respectivamente, as diferenças são muito grandes. Comparando todos estes resultados se observam que nenhuma estratégia de expansão melhorou a precisão das consultas sem expandir em todas as categorias sintácticas, concretamente o experimento que piores resultados obteve é o da categoria de advérbios e os melhores resultados foram da categoria de nomes. Realizando por outro lado uma comparação consulta a consulta destes experimentos, se observa que existe um determinado número das mesmas em que as consultas expandidas e WSD na categoria de nomes obtém uma precisão maior com respeito à consulta sem expandir.

Também o facto de que um determinado número de consultas melhorou o rendimento da recuperação da consulta original e o facto de que as consultas expandidas com a desambiguação do sentido das palavras melhoraram o rendimento das consultas expandidas sem a WSD, isto demonstra que a expansão com uma boa selecção de termos a adicionar na consulta original pode melhorar os resultados de recuperação.

Recall -Precisão	Consulta expandida e WSD somente com:				Consulta original
	advérbios	verbos	adjectivos	nomes	
0.00	0.3259	0.3875	0.4702	0.5929	0.6714
0.10	0.1880	0.2609	0.3006	0.4415	0.5276
0.20	0.1136	0.1580	0.2141	0.3226	0.3994
0.30	0.0771	0.1094	0.1729	0.2535	0.3365
0.40	0.0566	0.0780	0.1234	0.2239	0.2570
0.50	0.0403	0.0572	0.0958	0.1982	0.2275
0.60	0.0359	0.0461	0.0719	0.1558	0.1746
0.70	0.0239	0.0282	0.0509	0.1229	0.1409
0.80	0.0096	0.0154	0.0362	0.0963	0.1066
0.90	0.0044	0.0065	0.0197	0.0781	0.0759
1.00	0.0016	0.0029	0.0130	0.0485	0.0454
Precisão Média Não Interpolada	0.0628	0.0853	0.1195	0.2051	0.2431
%Câmbio Prec. Média		+35.83%	+90.29%	+226.59%	+287.10%

Tabela 6.6: Valores de Recall-Precisão dos resultados das consultas expandidas e WSD com distintas categorias sintácticas e com a original.

Consulta	Expandida e WSD somente com:				Original
	advérbios	verbos	adjectivos	nomes	
Tópicos	50	50	50	50	50
Núm. Doc. Rel.	2279	2279	2279	2279	2279
Núm.Doc.Rel.Rec.	892	991	1185	1688	1652

Tabela 6.7: Resumo estatístico dos resultados das consultas expandidas e WSD com distintas categorias sintácticas e com a original.

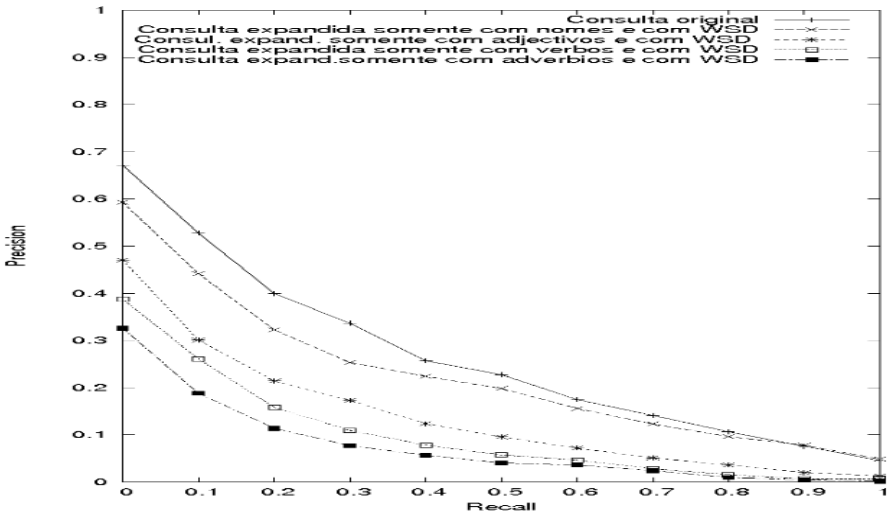

Figura 6.9: Gráfico de Recall-Precisão das consultas Expandidas e WSD com distintas categorias sintácticas e com a consulta original.

Recall -Precisão	Consulta com todas categorias sintácticas:		Original
	Expandida	Exp. e WSD	
0.00	0.3490	0.3689	0.6714
0.10	0.2296	0.2569	0.5276
0.20	0.1311	0.1683	0.3994
0.30	0.0999	0.1458	0.3365
0.40	0.0768	0.1068	0.2570
0.50	0.0572	0.0865	0.2275
0.60	0.0440	0.0680	0.1746
0.70	0.0240	0.0523	0.1409
0.80	0.0127	0.0346	0.1066
0.90	0.0037	0.0250	0.0759
1.00	0.0012	0.0131	0.0454
Precisão Média Não Interpolada	0.0767	0.1041	0.2431
%Câmbio Prec. Média		+35.72%	+216.94%

Tabela 6.8: Valores de Recall-Precisão dos resultados das consultas expandidas sem WSD e com WSD com todas as categorias sintácticas e com a original.

Consulta	Consulta com todas categorias sintácticas:		Original
	Expandida	Exp. e WSD	
Tópicos	50	50	50
Núm. Doc. Rel.	2279	2279	2279
Núm.Doc. Rel. Recuperados	1000	1315	1652

Tabela 6.9: Resumo estatístico dos resultados das consultas expandidas sem WSD e com WSD com todas as categorias sintácticas e com a original.

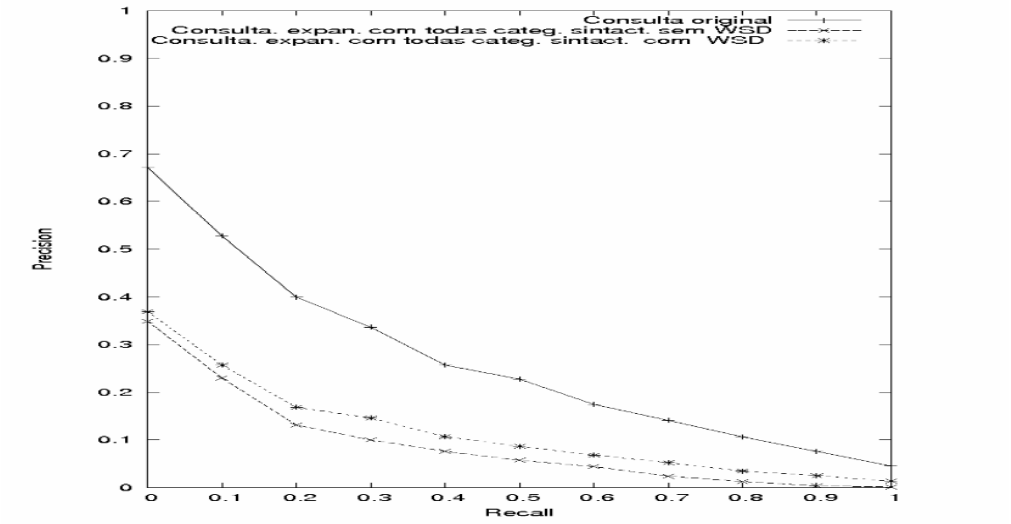

Figura 6.10: Gráfico de Recall-Precisão das consultas expandidas com todas as categorias sintácticas e com a consulta original.

Capítulo 7

Conclusões e trabalho futuro

Nesta tese tratou-se um tema de recuperação de informação, que basicamente pode ser resumido como o processo em que um usuário solicita uma informação num sistema de RI (por meio de uma consulta) e dito sistema devolve um conjunto de dados ou documentos que deveriam ser os mais relevantes possíveis. Dito processo envolve passos que vão desde cômo se armazena a informação (para tal são necessários estruturas de indexação que auxilíam o acesso aos dados) até a forma em que os usuários façam as suas consultas e cômo as ditas consultas são interpretadas pelo sistema.

Com base aos estudos realizados sobre as técnicas de indexação, no decorrer deste trabalho se descreve o índice invertido como uma das principais técnicas de indexação empregadas para a localização da informação numa Base de Dados textual ou na Web em geral.

Por outra parte se analisaram as principais diferenças que apresentam os sistemas de recuperação de informação na Web frente aos sistemas de recuperação de informação tradicionais: volume de dados, dinamismo, heterogeneidade, distribuição, redundância e falta de estrutura, qualidade e finalmente, os usuários.

Por sua vez, os sistemas de busca na Web se dividem em três categorias: robôs ou motores de busca, directórios e metabuscadores. Os robôs se caracterizam por indexar uma grande quantidade de informação (idealmente a totalidade da World Wide Web) com base à técnica de ficheiros invertidos, com respeito aos directórios Web que se caracterizam pela catalogação de documentos Web na sua hierarquia de categorias, o que implica uma grande qualidade nos seus conteúdos.

Tradicionalmente, os sistemas de busca na Web em geral, e os directórios Web em particular, têm uma presença bastante limitada nas publicações técnicas relacionadas com a recuperação de informação, baseando-se na maioria dos casos em adaptações ou migrações das técnicas de recuperação de informação tradicional à recuperação de informação na Web. Em consequência, os directórios Web se caracterizam por estar baseados em estruturas de ficheiro invertido que englobam diferentes aspectos: documentos, categorias e palavras chave, e as inter-relações entre eles, e que, portanto apresentam uma complexidade inerentemente maior do que o caso dos robôs ou motores de busca.

Enquanto ao objectivo principal perseguido com a presente tese doutoral deve mencionar-se o estudo da influência da expansão de consultas e da WSD (*Word Sense Desambiguation*) como técnicas que permitem melhorar a recuperação de informação que um usuário solicita, por meio de uma consulta. Mas especificamente este trabalho se centra na avaliação do sistema de recuperação de informação Lemur mediante medidas de precisão e *recall* considerando diferentes modos de consultas: i) Mediante a consulta original, ii) Expandindo a consulta com sinónimos com o uso do recurso linguístico WordNet, iii) Desambiguando previamente o sentido das palavras através de uso do recurso linguístico WordNet.

Em concreto com este trabalho realizamos um estudo que nos permitiu fazer a experimentação com expansão de consultas e desambiguação do sentido das palavras no âmbito da recuperação de informação mediante o uso do recurso linguístico WordNet e sob o modelo de usuário, utilizando uma metodologia de avaliação TREC, baseada na simulação. Além disto, se estudou cômo a expansão automática de consultas baseadas em tesauros e WSD utilizando o recurso lingüístico WordNet pode melhorar a recuperação de informação. Para tal realizamos o desenvolvimento de um sistema de recuperação que permite a utilização com expansão de consultas, assim como a desambiguação. Finalmente dito sistema foi avaliado para se obter os resultados que permitam comparar os resultados obtidos ao realizar as "consultas originais" com relação aquelas obtidas ao utilizar a expansão de consultas, WSD ou ambas.

Para implementar estas funcionalidades de formulação e expansão de consultas se acedeu à informação armazenada em WordNet. Uma vez obtidas as consultas expandidas dos diferentes experimentos contemplados se procedeu a sua avaliação e a comparação entre elas e com as consultas sem expandir. Os resultados da avaliação

são positivos com respeito à consulta expandida e com WSD e permitiram extrair algumas conclusões importantes:

- Utilizando diferentes técnicas de expansão os resultados melhoraram aplicando a desambiguação.
- Os resultados empíricos obtidos sobre grandes colecções de textos de referência (TREC) demonstram que a desambiguação de palavras permite obter o máximo proveito na expansão de consultas com WordNet.
- Por outro lado em algumas consultas concretas nos experimentos com expansão de sinónimos, o resultado da expansão melhorou de maneira significativa. Isto nos indica que quando a selecção de termos é adequada à expansão pode fornecer resultados positivos.

A WSD não é uma tarefa fácil dentro da recuperação de informação. Muitos sistemas tentaram seleccionar o sentido mais apropriado usando técnicas estatísticas e/ou aprendizagem automática. No presente trabalho da tese doutoral se utilizou o método de desambiguação baseado no dicionário aproveitando o enorme potencial de Wordnet sob o seguinte algoritmo: *i)* Adquirir desde WordNet os conjuntos de sinónimos das palavras a desambiguar, *ii)* Determinar a coincidência entre o contexto das palavras a desambiguar e os conjuntos de sinónimos, *iii)* Escolher o significado correcto das palavras num texto, dado o contexto de forma automática.

Aplicando o algoritmo acima citado os resultados da expansão e desambiguação melhoraram utilizando as diferentes categorias sintácticas com respeito aos resultados das consultas somente expandidas, visto que ao realizar a expansão sem a desambiguação se obteve os seguintes resultados: 16.49%, 9.64%, 6.76% e 4.47% de precisão média não interpolada com as categorias sintácticas de nomes, adjectivos, verbos e advérbios respectivamente com respeito aos 20.51%, 11.95%, 8.53% e 6.28% de precisão media não interpolada realizando a expansão e desambiguação na mesma ordem (nomes, adjectivos, advérbios e verbos). Por outro realizando a expansão e a desambiguação com a consulta completa (combinando todas as categorias sintácticas) os resultados são igualmente melhores desambiguando obtendo (7.67% realizando somente a expansão) e (10.41% realizando a expansão e a desambiguação). Com base aos resultados alcançados se obtém as seguintes conclusões:

- Utilizando qualquer categoria sintáctica ou combinando todas no processo de

expansão de consultas, os resultados melhoram aplicando a desambiguação do sentido das palavras.

- Os resultados aplicando a expansão com ou sem a desambiguação são melhores utilizando somente a categoria sintáctica de nomes, que quando se usou verbos, adjectivos, etc.

7.1 Trabalho futuro

Consideramos como principal linha de trabalho futuro, analisar a problemática sobre o impacto da expansão de consultas e desambiguação do sentido das palavras nos sistemas de recuperação de informação com buscadores reais, implementando um interfaz, sobre um motor de busca actual que utilize características de WordNet, com propósito de melhorar a relevância dos documentos que dito sistema devolve. A presente tese doutoral trata uma temática de actualidade e útil no caminho para a elaboração futura de um buscador semântico que resolva os problemas dos actuais buscadores sintácticos. Conscientes de algumas das limitações que apresentam a aproximação tratada nesta tese doutoral, cremos que duas interessantes linhas de trabalho futuro deveriam ser: (i) Experimentar o método sob o algoritmo proposto com outros idiomas; e (ii) Explorar a Web como recurso linguístico para a WSD.

Apêndice A

Lista de palavras comuns

Nosso sistema *prototype* utiliza a seguinte lista de palavras comuns na *phrase stopword*.

A lista é a mesma utilizada pelo sistema *SMART*.

Lista de palavras comuns

a a's able about above according accordingly across actually after afterwards again against ain't all allow allows almost alone along already also although always am among amongst an and another any anybody anyhow anyone anything anyway anyways anywhere apart appear appreciate appropriate are aren't around as aside ask asking associated at available away awfully b be became because become becomes becoming been before beforehand behind being believe below beside besides best better between beyond both brief but by c c'mon c's came can can't cannot cant cause causes certain certainly changes clearly co com come comes concerning consequently consider considering contain containing contains corresponding could couldn't course currently d definitely described despite did didn't different do does doesn't doing don't done down downwards during e each edu eg eight either else elsewhere enough entirely especially et etc even ever every everybody everyone everything everywhere ex exactly example except f far few fifth first five followed following follows for former formerly forth four from further furthermore g get gets getting given gives go goes going gone got gotten greetings h had hadn't happens hardly has hasn't have haven't having he he's hello help hence her here here's hereafter hereby herein hereupon hers herself hi him himself his hither hopefully how howbeit however i i'd i'll i'm i've ie if ignored immediate in inasmuch inc indeed indicate indicated indicates inner insofar instead into inward is isn't it it'd it'll it's

its itself j just k keep keeps kept know knows known l last lately later latter latterly least less lest let let's like liked likely little look looking looks ltd m mainly many may maybe me mean meanwhile merely might more moreover most mostly much must my myself n name namely nd near nearly necessary need needs neither never nevertheless new next nine no nobody non none noone nor normally not nothing novel now nowhere o obviously of off often oh ok okay old on once one ones only onto or other others otherwise ought our ours ourselves out outside over overall own p particular particularly per perhaps placed please plus possible presumably probably provides q que quite qv r rather rd re really reasonably regarding regardless regards relatively respectively right s said same saw say saying says second secondly see seeing seem seemed seeming seems seen self selves sensible sent serious seriously seven several shall she should shouldn't since six so some somebody somehow someone something sometime sometimes somewhat somewhere soon sorry specified specifyspecifying still sub such sup sure t t's take taken tell tends th than thank thanks thanxthat that's that's the their theirs them themselves then thence there there's thereafterthereby therefore therein theres thereupon these they they'd they'll they're they've thinkthird this thorough thoroughly those though three through throughout thru thus to together too took toward towards tried tries truly try trying twice two u un under unfortunately unless unlikely until unto up upon us use used useful uses using usuallyuucp v value various very via viz vs w want wants was wasn't way we we'd we'll we'rewe've welcome well went were weren't what what's whatever when whence wheneverwhere where's whereafter whereas whereby wherein whereupon wherever whether which while whither who who's whoever whole whom whose why will willing wish with within without won't wonder would would wouldn't x y yes yet you you'd you'll you're you've your yours yourself yourselves z zero.

Apêndice B

Desenvolvimento de aplicação para formulação de consultas

Neste apêndice se descreve o desenvolvimento de uma aplicação que nos permite realizar a formulação da expansão de consultas e desambiguação do sentido das palavras utilizando o recurso linguístico WordNet, no âmbito da recuperação de informação. Na primeira secção se desenvolve uma base de dados que nos permite satisfazer as funcionalidades necessárias. Posteriormente se descreve a sua implementação. Finalmente se explica o funcionamento do desenvolvimento da aplicação

B.1 Introdução

Um primeiro passo na fase de análise é a identificação das funcionalidades necessárias para atingir o objectivo da aplicação que é a experimentação com a expansão de consultas. Para atingir os objectivos do projecto será necessário, em primeiro lugar, poder desenhar diferentes tipos de experimentos de expansão. Além disto, se adicionam outras funcionalidades secundárias como, por exemplo, mostrar as consultas geradas para um experimento determinado. Para ver melhor as funcionalidades se pode observar a Figura de diagrama de casos de uso da aplicação.

Se pode observar que dentro da realização de um experimento existem diferentes possibilidades segundo a selecção dos termos da consulta seja automática ou manual. Além disto, durante a realização do experimento será gerado e guardado uma consulta por cada tópico do exercício uma vez seleccionados os significados adequados

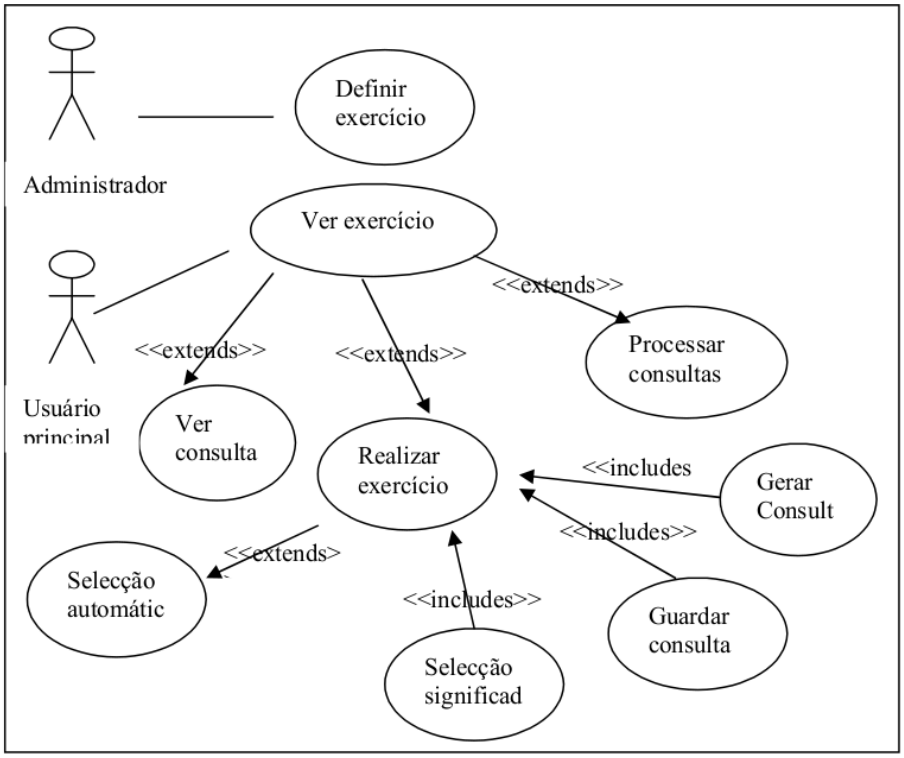

Figura B.1: Diagrama de casos de uso

de cada palavra.

B.2 Dados que definem um experimento

Para satisfazer as funcionalidades necessárias, se precisará armazenar informação sobre os exercícios criados e as consultas expandidas criadas de maneira persistente. Desenhar-se-á para isso uma pequena base de dados, de tal forma que, para cada experimento se armazenarão os seguintes dados:

- O conjunto de tópicos que se usarão no experimento.
- O algoritmo de *stemmer* para processado de palavras.
- O algoritmo de *stemmer* para acesso a WordNet.
- As fontes de listas de palavras comuns (stopwords).
- A lista de possíveis *tags* dos tópicos.
- As taxonomias existentes no recurso linguístico, isto é, sobre quê tipo de categorias se realizarão os experimentos.
- O tipo de consultas que gerará o exercício.
- Os conjuntos de palavras que se usarão na expansão.

B.3 Implementação

Para implementar a aplicação se usaram as tecnologias associadas à plataforma Visual Basic.Net. Para a realização do interfaz de usuário se optou por um interfaz de janelas sob o sistema operativo Windows. O seu interfaz mostra-se na Figura B.2.

B.4 Funcionamento

O funcionamento do programa basicamente consiste em filtrar Títulos, Descrições e Narrativas dos tópicos, filtrar as palavras comuns, etc.

Uma vez feito isto nos encontramos com as duas funcionalidades básicas da aplicação: *(i)* por um lado, a definição ou criação de um experimento, e por outro lado, *(ii)* a realização de ditos experimentos mediante a selecção dos significados e expansão de consultas.

No momento de definir o exercício ou experimento existe um grande número de possibilidades cujo significado se detalha a seguir:

Em primeiro lugar, é necessário indicar o *rango* de tópicos sobre os que se realizará o experimento. Para isso, no primeiro dos campos do formulário (Introduzir tópicos (151-550) se pode introduzir um rango na formulação $min - max$, onde min será o número de tópico de começo e max o número de tópico a usar no exercício). Os números dos tópicos devem estar compreendidos entre 151 e 550.

Depois se selecciona o algoritmo de *Stemmer* para consulta a aplicar sobre as palavras, para passar a fazer parte das consultas. Pode-se seleccionar não aplicar nenhum, ou bem aplicar o algoritmo de Porter [182]. Se elege não aplicar nenhum, as palavras das consultas irão completas, se elege o algoritmo de Porter aparecerão unicamente as raizes obtidas por este algoritmo.

A seguinte possibilidade *Seleccionar Stemmer* para acesso a WordNet nos permite indicar se queremos que na hora de buscar uma palavra em WordNet se busca de maneira exacta e só se encontra se está incluída exactamente ou se deseja aplicar o processado morfológico de WordNet encontrando assim também as palavras cuja raiz aparece em WordNet ainda que não apareça em si mesma.

Segundo a opção que se escolha em *Eleja a taxonomia a buscar*, a busca de significados devolverá os resultados de busca das palavras somente como nomes, somente como verbo, somente como adjectivo, somente advérbio ou buscando todos os significados de todas as palavras em todas as taxonomias.

A seguinte secção *Palavras a usar na expansão* permite indicar quê palavras se adicionarão a cada consulta na expansão podendo-se expandir com sinónimos, hipónimos de primeiro nível, ou com ambos os tipos ao mesmo tempo, já que se permite a selecção múltipla.

Como as consultas se geram a partir dos tópicos dos títulos a seguinte opção é *Tags* a usar, onde se deve indicar a partir de quê partes de um tópico se devem gerar, podendo-se escolher entre o Título (Headline), a Descrição (Head) ou a Narrativa (Text) ou qualquer combinação delas.

Com estes *tags* dos tópicos *termos a usar antes da expansão* se seleccionam as

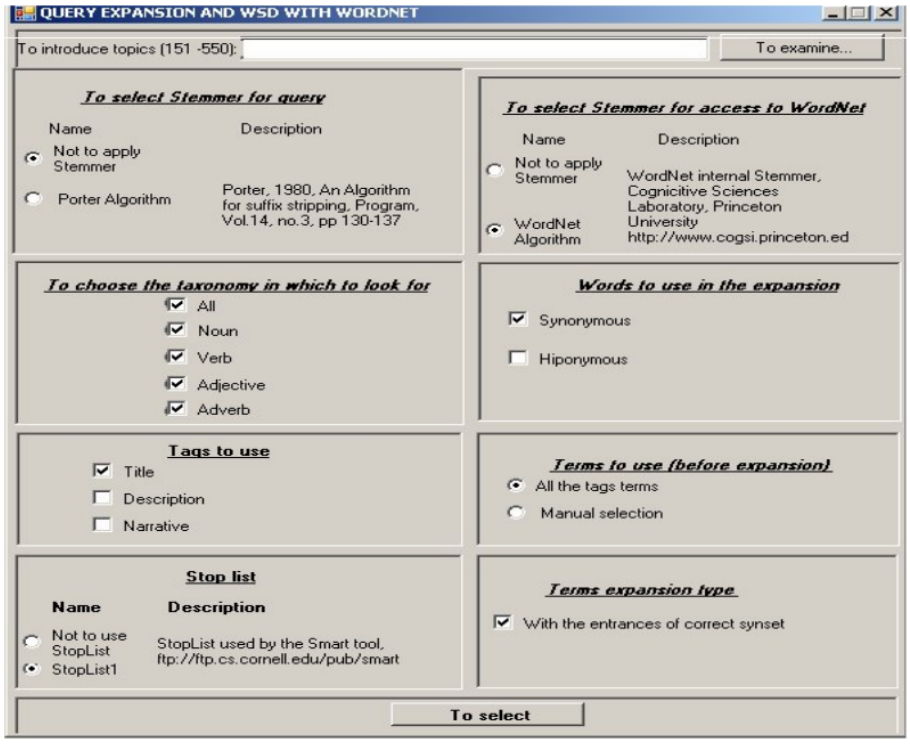

Figura B.2: Interface da aplicação que nos permite definir as opções da expansão.

partes que formam a consulta sem expandir, por isso, se pode marcar uma selecção manual dos termos ou uma selecção automática indicando que se usem todos os termos dos tags.

Aos termos seleccionados se lhes pode aplicar um filtro ou não segundo o escolhido na selecção *StopList*. Nela indicaremos se não se aplicará nenhum tipo de filtrado, ou bem se eliminarão aquelas palavras consideradas como muito comuns pelas ferramentas de recuperação de informação Smart.

A última opção *Tipo da expansão dos termos* permite escolher o tipo de expansão podendo-se expandir mediante termos dos synsets escolhidos como correctos.

Ao terminar de cumprimentar o formulário, pulsando o botão *Seleccionar* criaremos o experimento de que nos aparecerá um breve resumo com os seus dados e que já estará disponível para ser realizado.

Este formulário será o mesmo para todos os tópicos que tenha que processar no experimento. Em primeiro lugar, a página situa ao usuário nos tópicos que vão processar, identificando-os com os seus números e mostrando os campos do título, descrição e narrativa. Isto se faz assim independentemente de que a selecção se efectue a partir de um só campo do tópico. A partir desta informação o usuário poderá conhecer o significado das palavras seleccionadas dentro do contexto do tópico.

Embaixo destes campos aparece uma lista com estas palavras seleccionadas e os seus possíveis significados e sinônimos obtidos desde WordNet. Segundo o tipo de exercício, o usuário poderá seleccionar somente o significado correcto para cada palavra.

Uma vez seleccionados todos os significados oportunos em cada caso, para gerar a consulta se pulsa o botão *Gerar consulta*. Uma vez feito isto aparecerá um quadro de diálogo com a consulta escrita em área do texto. Esta área de texto é editável, isto permitirá ao usuário modificar a consulta manualmente segundo as suas necessidades, doptando assim o sistema de maior flexibilidade e também será possível mudar algum significado e voltar a gerar a consulta de novo se considera oportuno.

Modificando a consulta ou não, quando se dê por válida, pulsando o botão *Guardar consulta*, armazena-se a consulta na base de dados. Ao guardar a consulta aparecerá um quadro de diálogo, com a informação e o usuário elegerá o lugar para guardar a consulta com o nome desejado.

Este processo de realização de experimentos é o mesmo tanto se escolheu selecção manual como automática. Mas no caso da selecção manual de termos dos tópicos

será necessário seleccionar previamente e para cada tópico as palavras originais que se quer que formem a consulta.

Estas funcionalidades básicas que vão permitir-nos desenhar um experimento, realizar a expansão dos experimentos definidos de diversas maneiras e, a partir das consultas geradas, avaliar a eficácia da expansão em cada experimento. A avaliação é o tema abordado anteriormente no Capítulo 6.

Bibliografia

[1] *Aarts, J.; Meijs, L.* **Corpus Linguistics.** Amsterdam Rodoppi, 1984.

[2] *Aas, K.; Eikvil, L.* **Text Categorization.** A survey, Norwegian Computing Center, reporte técnico, 1999.

[3] *ACM.* **Sítio Web de Association for computing Machinery**, htt://www.acm.org, página principal, 2007.

[4] *Agirre, E.; Ansa, O.; Hovy, E. y Martinez, D.* **Enriching very large ontologies using the WWW.** En Proceedings of the Ontology Learning Workshop, ECAI, Berlin, Alemania, 2000.

[5] *Agirre, E.; Martínez, D.* **Unsupervised WSD Based on Automatically Retrieved Examples: The Importance of Bias.** In Proceedings of the Conference on Empirical Methods in Natural Language Processing, Barcelona, 2004.

[6] *Agostín, M.; Melucci, M.* **Information Retrieval on the Web.** En Agostín, M.; Crestani, F.; Pasi, G., editores, "Lectures on Information Retrieval: Third European Summer-School", ESSIR 2000. Revised Lectures, Springer-Verlang, Berlin/Heidlberg, 2001, página 242-285, 2001.

[7] *Agrawal, R.; Chakrabarti, B.; Dom, B.; Raghavan, R.* **Using taxonomy, descriminants, and signatures for navigating intext databases.** The 23th International Conference on very Large Databases, VLDB, página 446-455, 1997.

[8] *Agrawal, R.; Chakrabarti, B.; Dom, B.; Raghavan, R.* **Scalable feature selection, classification and signature generation for organizing large text**

databases into hierarchical topic taxonomies. VLDB Journal,volume 7, no.3, página 163-178, Agosto 1998.

[9] *Aho, A.V; Corasick, M.J.* **Fast pattern matching: an aid to bibliographic search.** Communication of ACM (CACM), 18(6): página 333-340, Junio 1975.

[10] *Ala.* **Committe on Cataloging:Description and Access. Task Force on Metadata. Summary report, June 1999. On-line**, Disponível em http://www.ala.org/alcts/organization/ccs/ccda/tfmeta3. html Acesso em : 24 de Dezembro de 2002.

[11] *Alegria, I.* **Morfología de estados finitos.** SEPLN (18), página 1-26, 1996.

[12] *Alvarenga, L.* **A teoria do conceito revisitada em conexão com ontologias emetadados no contexto das bibliotecas tradicionais e digitais.** DataGramaZero – Revista de Ciência da Informação, v. 2, n.6, dez. 2001,18p. [on-line]. Disponível em: http://www.dgzero.org/Atual/Art_05.htm, Acesso em: 26 de Dezembro de 2001.

[13] *Atkins, D.***Vision for digital libraries.** An International research agenda for digital libraries, página 11-14, Outubro 1998.

[14] *Avram, H.D.* **MARC: Its history and implications.** Washington: Library of Congress, ISBN 0-8444-0176-5, 1975.

[15] *Baeza-Yates, R.; Ribeiro-Neto, B.* **Modern Information Retrieval**. Addison Wesley, ACM press, 1999.

[16] *Bayer, R.;Unteruer.* **Prefix B-tries.** ACM Transactions on Database Systems, volume 2, no. 1, página 11-26, 1977.

[17] *Bell, T.; Moffat, A.; Nevill-Manning, c.; Witten, I.; Zobel, J.* **Data compression in full-text retrieval systems**. Journal of the American Society for Information Science, volume 44, no. 9, página 508-531, 1993.

[18] *Berber Sardina, T. B.* **Lingüística de Corpus**. Editora Manole Ltda, 2004.

[19] *Berners-Lee, T.; Cailliau, R.; Luotonen, A.; Nielsen, H.F.; Secret, A.* **The Wold Wide Web**. Communication of the ACM, volume 37, no.8: página 76-82, 1994.

[20] *Berry, M. W.; Drmac, Z. and Jessup,E.r.* **Matrices, vector spaces, and information retrieval**. Siam Review, 41(2): página 335-362, 1999.

[21] *Boguraev, B.K.* **Automatic Resolution of Linguistic Ambiguities**. Technical Report No. ii, University of Cambridge Computer Laboratory, Cambridge,1979.

[22] *Borba.* **Dicionário UNESP do Português Contemporâneo**. Editora UNESP, 2004.

[23] *Bowman, C.M.; Danzing, P.B.; Hardy, D.R.; Manber, U.; Schwartz, M.F.* **The Harvest information discovery and accesss sytem**. The 2nd International World Wide Web Conference, página 763-771, Outubro 1994.

[24] *Boyer, R.S.; Moore, J.S.* **A fast string searching algorithm**. Communications of ACM (CACM), volume 20, no. 10, página 762-772, Outubro 1977.

[25] *Bharat, K.; Broder, A.* **A technique for measuring the relative size and overlap of public Web search engines**. The 7th International World Wide Web Conference, página 379-338, Brisbane, Australia, Abril 1998.

[26] *Black, E.* **An experiment in computational discrimination of English word senses in a large corpus**. En IBM Journal of Research and Development, Volume 32, No. 2, página 185-194, 1998.

[27] *Blattman, Ursula.* **Modelo de gestão da informação digital online em bibliotecas acadêmicas na educação à distância**. Biblioteca virtual. Florianópolis, página 187, 2001.

[28] *Blum, A. y Mitchell, T.* **Combining labeled and unlabelled data with cotraining**. En Proceedings of the 11th Annual Conference on Computational Learning Theory, página 92-100, 1998.

[29] *Brin, S.; Page, L.* **The anatomy of a large-scale hypertextual web search engine**. The 7th International World Wide Web Conference, página 102-117, Brisbane, Australia, Abril 1998.

[30] *Brisaboa, N. R. ; Places, A. S. ; Rodríguez, F. J.* **Arquitectura para Federacióńde Bases de datos documentales basada en Ontologías**. En Torres Rojas, F. ; ArayaMonge, J. E. ; Sandoval Sánchez, Y. (ed.): 4º Workshop Iberoamericano de Ingenharia de Requisitos e Ambientes de Software (IDEAS'2001): 252-262. (Cartago, Costa Rica: Instituto Tecnológico de Costa Rica), 2001a.

[31] *Brisaboa, N. R.; Penabad, M. R.; Places, A. S.; Rodríguez, F. J.* **Using ontologies for federation of Web accessible databases**. En 13th international conference on software engineering & knowledge engineering (SEKE'2001): 87-94 (Skokie, IL,USA: Knowledge Systems Institute), 2001b.

[32] *Brisaboa, N. R.; Places, A. S.; Pérez-Sanjulián, C. F.; Rodríguez, F. J.* **An Arquitectural Proposal for a Cross-Language System to FederMultilingual Digital Libraries**. En Russian Academy of Sciences (org): DigitaLibraries: Advanced Methods and Technologies, Digital Collections (Petrozavodsk: Institute of AppliedMathematical Research), 2001c.

[33] *Brisaboa, N.R.; Fariña, A.; Navarro, G.; Paramá, J.R.* **Lightweight natural language text compression**. Information Retrieval.10(1): página 1-33, 2007.

[34] *Brisaboa, N.R.; Fariña, A.; Ladra, S.; Navarro, G.* **Reorganizing Compressed Text**. 31st ACM SIGIR Conference, Singapore, 2008.

[35] *Bruce, R.; Wiebe, J.* **Word-sense disambiguation using decomposable models**. In Proceedings of the 32nd Annual Meeting of the Association. for Computational Linguistics, página 139-145. Las Cruces, 1994.

[36] *Buckley, C.; salton, G.; Allan, J.* **Automatic retrieval with locality information using SMAR**. Proceedings of the First Text retrieval Conference (TREC-1), Nist Special Publication, página 59-72, 1993.

[37] *Cacheda, F.; Viña, A.* **Experiencies retrieving information in the World Wide Web**. 6th IEEE Symposium on Computers and Communications, ISBNs: 0-7695-1177-5, 0-7695-1178-3 (cse), 0-7695-1179-1 (microfiche), página 72-79, Tunisia, Julio 2001.

[38] *Cacheda, F.; Viña, A.* **Understanding how people use search engines: a statistical análisis for e-Business**. e-2001 (e-Business and e-Work Conference and Exhibition), ISBNs: 1-58603-205-4 (IOS Press) y 4-274-90469-5-C3055 (Ohmsha), Volume1, página 319-325, Venecia, Italia, Outubro 2001.

[39] *Cacheda, F.; Viña, A.* **Superimposing Codes Representing Hierarchical Infomation in Web Directories**. 3rd International Workshop on Web Information and Data Management (WIDM 2001) en Tenth International Conference on Information and Knowledge Management (ACM-CIKM 2001), ACM ISBN: 1-58113-444-4, página 54-60, Atlanta, usa, Noviembre 2001.

[40] *Chklovski, T. y Mihalcea, R.* **Building a sense tagged corpus with Open Mind Word Exper**. En Proceedings of the ACL 2002 Workshop on Word Sense Disambiguation: Recet Successes and Future Directions, Philadelphia, 2002.

[41] *Cho, J.; García-Molina, H.; Page, L.* **Efficient crawling through URL ordering**. The 7th International World Wide Web Conference, página 161-172, Brisbane, Australia, Abril 1998.

[42] *Chowdhury, A. e McCabe, MC.* **Improving Information Retrieval using Part of Speech Tagging**. (url:citeseer.ist.psu.edu/256084.html), 1998.

[43] *Clarke, C.; Cormack, G.; Burkowski, F.* **An algebra for structured text search and a framework for its implementation**. The Computer Journal, volume 38, página 43-56, 1995.

[44] *Cutting, D.; Persen, J.* **Optimizations for dynamic inverted index maintenance**. The 13th International ACM SIGIR Conference on Research and Development in Information Retrieval, página 405-411, 1990.

[45] *Dagan, I.; Itai, A.; Schwall, U.* **Two Languages are More Informative than One**. In Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics, página 130-137, Berkeley, 1991.

[46] *Dagan, I.; Marcus, S.; Markovitch, S.* **Contextual Word Similarity and Estimation from Sparse Data**. In Proceedings of the 31st Meeting of the Association for Computational Linguistics. Columbus, 1993.

[47] *Dagan, I.; Itai, A.* **Word Sense Disambiguation Using a Second Language Monolingual Corpus**. Computational Linguistics, página 563-596, 1994.

[48] *Deerwester, S.; Dumais, S. T.; Furnas, G. W.; Landewer, T. K. and R. Harshman.* **Indexing by latent semantic analysis**. Journal of the American Society of Information Science, página 1-407, 1990.

[49] *Diab, M.; Resnik, P.* **An Unsupervised Method for Word Sense Tagging using Parallel Corpora**. In Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics, Philadelphia, 2002.

[50] *Diaz, F.; Corchado, J.M.* **Rough sets bases learning for bayesian networks**. International workshop on objetive bayesian methodology, Valencia, Spain, 1999.

[51] *Dumais, S.; Platt, J.P; Heckerman, D. y Sahami, M.* **Inductive learning algorithms and representations for text categorization**. En proceedings of the seventh International Conference on Information Retrieval and Knouledge Management (ACM-CIKM' 98), página 148-155, 1998.

[52] *Dyvik, H.* **Translations as semantic mirrors: From Parallel Corpus to Wordnet**. In Proceedings of the 23rd International Conference on English Language Research on Computerized Corpora of Modern and Medieval English, Gothenburg, 2002.

[53] *Elias, P.* **Universal codeword sets and representations of the integers**. IEEE Transactions on Information Theory, volume 21, página 194-203, Marzo 1975.

[54] *Falousos, C.; Oard, D.* **A survey of Information Retrieval and filtering methods**. Techical Report CS-TR3514, Department of Computer Science, University of Maryland, 1995.

[55] *Fernández, J.; Castilho, M.; Rigau, G.; Atserias, J.; Turmo, J.* **Automatic Acquisition of Sense Examples using ExRetriever**. In Proceedings of the International Conference on Language Resources and Evaluation, página 25-28. Lisbon, 2004.

[56] *Figuerola, C.G.; Gómez, R.; Zazo, A.F. e Alonso, J.L.* **Spanish Monolingual Track: The Impact of Stemming on Retrieval**. En Carol P. (ed): Working notes for the CLEF 2001 workshop, Springer, 2001.

[57] *Fox, E. A.* **Extending the Boolean and Vector Space Models of Information Retrieval with P-norm Queries and Multiple concept Types**. Tesis doctoral, 1993.

[58] *Fox, E.A.; Lee, W.C.* **FAST-INV: A fast algorithm for building large inverted files**. Techical Report TR-91-10, Virgilia Polystechic Institute Department of Computer Science, Marzo Translations as semantic mirrors: From Parallel Corpus to Wordnet, 1991.

[59] *Fox, E. A.; Marchionini, G.* **Toward a worldwide digital library.Communications of the ACM**. Volume 41, página 29-32, Abril 1998.

[60] *Fox, E. A.; Sornil, O.* **Modern Information Retrieval**. Digital Libraries. En Baeza-Yates, R.; Ribeiro Neto, B. (New York: Adison Wesley/ACM Press), 1999.

[61] *Fox, E.A.;URS, S. R.* **Digital libraries**. Annual Review of Information Science and Technology, volume 33, página 503-589, 2002.

[62] *Francis, W. M.; Kucera, H.* **Brown Corpus – Manual of Information**. Department of Linguistics,Brown University, 1979.

[63] *Francis, N.* **Language corpora B.C.I**. In Directions in Corpus Linguistics. Proceedings of Nobel Symposium, 1992.

[64] *Frankes, W. B. and Baeza-Yates, R.* **Data Structures and Algorithms**. Editors. Information Retrieval. Prentice-Hall, Englewoods Cliff, N..J. 1992.

[65] *Fujii, A.; Inui, K.; Tokunaga, T.; Tanaka, H.* **Selective Sampling for Example based Word Sense Disambiguation**. S. Computational Linguistics, página 573-597, 1998.

[66] *Furnas, G.; Deer Wester, S.; Dumais, S.; Landauer, T.; Harshman, R.; Streeter, L. and Locbbaum, K.* **Information retrieval using a singular value decomposition model of latent semantic structure**. In Proceedings of the 11th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,ACM, página 465-480, 1988.

[67] *Galaxy.* **Sítio Web de Galaxy, http://www.galaxy.com/**, página principal, 2007.

[68] *Gale, W. A.; Church, K.W. y Yarowsky, D.* **A method for disambiguating word senses in a large corpus**. In Computers and the Humanities, Volume 26, página 415-439, 1993.

[69] *Gale, W.A.; Church, K.W.* **A Program for Aligning Sentences in Bilingual Corpora**. In Proceedings of ACL-91, Berkeley CA, 1991.

[70] *Gauch, S.; Wang, G.* **Information Fusion With Profusion**. The World Conference of the Web Society, WebNet´96, página 174-179, Outubro 1996.

[71] *Gils* **overview-ideas behind the GILS approach. On-line**. Disponível em : <http://www.gils.net/overview.html > Acesso em: 26 de novembro de 2001.

[72] *Golub, G.; Klema, V. and StreWart, G.* **Rank degeneracy and least squares problems**. Techical report, University of Maryland, College Park, MD, 1976.

[73] *Gonzalo, J.; Verdejo, M. F.; Chugar, I. and Cigarrán, J.* **Indexing with WordNet synsets can improve text retrieval**. Proceedings of the COLING/ACL '98 Workshop on Usage of WordNet, 1998.

[74] *Gonzalo, J.; Peñas, A. e Verdejo. F.* **Lexical ambiguity and Information Retrieval revisited**. En Proceedings of the 1999 Joint SIGDAT Conference on EMNLP and VLC, Maryland, página 195-202, 1999.

[75] *Gonzalo, J.; Peñas, A. e Verdejo, F.* **La indexación con técnicas lingüísticas en el modelo clásico de recuperación de información**. En Primeras Jornadas de Tratamiento y Recuperación de Información, JOTRI-2002, página 97-106, 2002.

[76] *Gonzalo, J.; Verdejo, F. e Chugar, I.* **The Web as a resource for WSD**. En 1st MEANING Workshop, España, 2003.

[77] *Greenberg, J.* **Automatic query expansion via lexical-semantic relationships**. Journal of American Society For Information Science and Technology (JASIS), página 402-415, 2001.

[78] *Grefenstette, G.* **Use of syntactic context to produce term association lists for text retrieval**. Proceedings of the 15th ACM-SIGIR Conference, página 89-97, 1992.

[79] *Harabagiu, S. M.; Miller, G. A. y Moldovan, D.I.* **WordNet–A Morphologically and Semantical Enhanced Resource**. En Proceedings of the SIGLEX Workshop, 1999.

[80] *Harman, D. K.* **Ranking Algorithms**. En Frakes, W. B. y Baeza-Yates, R., eds.: Information Retrieval: Data Structures and Algorithms, Englewood Cliffs (NJ): Prentice-Hall Inc., página 363-392, 1992.

[81] *Harper, D.J.* **Relevance Feedback in Automatic Document Retrieval Systems: An Evaluation of Probabilistic Strategies**. Disertación Doctoral, Jesus College, Cambridge, England, 1980.

[82] *Haskin, R.L.* **Special-purpose processors for text retrieval**. Database Engineering, volume 4, página 16-29, Septiembre, 1981.

[83] *Hawking, D., Craswell, N.; Thistlewaite, P.; Harman, D.* **Results and challenges in Web search evaluation**. The 8th World Wide Web Conference, página 243-252, Mayo, 1999.

[84] *Heaps, J.* **Information retrieval – Computational and theoretical Aspects**. Academic Press, New York, 1978.

[85] *Heckerman, D.; Chickering, M.; Geiger, D.* **Learning bayesian networks, the combination of knowledge and statistical data**. Machine learning, página 197-243, 1995.

[86] *Heckerman, D.; Chickering, M.* **Efficient approximation for the marginal likelihood of incomplete data given a bayesian network**. Techical report MSR-TR-96-08, Microsoft Research, Microsoft Corporation, 1996.

[87] *Hearst, M.* **Noun Homograph Disambiguation using Local Context in Large Text Corpora**. Proceedings of the 7th Annual Conference of the Centre for the New OED and Text Research: Using Corpora. Oxford, 1991.

[88] *Hearst, M.; Schölkopf, B.; Dumais, S.; Osuna, E. y Platt, J.* **Trends and controversies - Support vector machines**. IEEE Intelligent systems, Volume 13, página 18-28, 1998.

[89] *Henry, H.; Collins, Jr.* **English language Dictionary**. Birminghman University International Database. John Sinclair, 1998.

[90] *Hernandez Orallo, J.* **Extracción automática de conocimiento de bases de datos e ingeniería de software**. Programación declarativa e ingeniería de la programación, 2000.

[91] *Hillel, B.* **The present status of automatic translation of languages**. En Advances in computers, Vol.1, Academic Press, New York, volume 1, página 91-163, 1960.

[92] *Hirschberg, D.; Lelewer, D.* **Efficient decoding of prefix codes**. Communications of the ACM, volume 33, página 449-459, Abril, 1990.

[93] *Hirst, G.* **Semantic Interpretation and the resolution of ambiguity**. Cambridge University Press, 1987.

[94] *Hobbs, J.R.* **World knowledge and word meaning**. En Proceedings of the Third Workshop on Thoretical Issues in Natural Language Processings (TINLAP-3), Las Cruces, página 20-25, 1987.

[95] *Hodge, G.* **Sítio Web de Galaxy, http://www.galaxy.com/**. Metadata made simpler: a guide for libraries, Bethesda, MD: National Information Standards Organizations, 2001.

[96] *Hollaar, L.A.; Smith, K.F.; Chow, W.H.; Emrath, P.A. and Haskink, R.L.* **Architecture and operation of a large, full-text information-retrieval system**. Advanced Database Machine Architecture, página 256-299. Ed. Prentice-Hall, 1983.

[97] *Hooper, R. S.* **Indexer consistency test - origin, measurements, results and utilization.** Bethesda, MD, 1965.

[98] *Huang, L.* **A survey on web information retrieval technologies**. Research Proficiency Exam Report, Experimental Computer System Lab, 2000.

[99] *Hyman, R.J.* **Information Access: Capabilities & Limitations of Printed & Computerized Sources**. American Library Association, ISBN 0-838-90512-9, 1989.

[100] *Ide, N. y Véronis, J.* **Word Sense Disambiguation: the state of the art"**. En Computacional Linguistics, volume 24, página 15-23, 1998.

[101] *Ide, N.* **Parallel Translations as Sense Discriminators**. In Proceedings of the SIGLEX99 Workshop: Standardizing Lexical Resources, página 52-61. Maryland, 1999.

[102] *Ide, N.; Erjavec, T.; Tufi, D.* **Sense Discrimination with Parallel Corpora**. In Proceedings of ACL'02 Workshop on Word Sense Disambiguation: Recent Successes and Future Directions, página 54-60. Philadelphia, 2002.

[103] *Indyk, P.; Chakrabarti and Dom, B.* **Enhanced Hypertext categorization using Hyperlinks**. ACM SiGMod International Conference on Management of Data, página 307-318, 1998.

[104] *InternetArchive.* **Internet Archive Building a digital library for the future, http://www.archive.org/**, 2007.

[105] *Iso.* **International Standartization Organization.** Introduction to ISO, 1999.

[106] *Jansen, B.; Spink, A.; Bateman, J.; Saracevic, T.* **Real Life Information Retrieval: A Study Of User Queries On the Web**. SIGIR Forum, vol.32, página 5-17, 1998.

[107] *Jing, Y.; .Croft, B.* **An association thesaurus for information retrieval proceedings of RIAO**, página, 146-160, 1994.

[108] *Joachims, T.* **Text categorization with support vector machines: learning with many relevant features**. En Proceedings of the Tenth European Conference on Machine Learning (ECML'98), Lecture Notes in Computer Science, No. 1398, página 137-142, 1998.

[109] *Joachims, T.* **A statistical learning model of text classification with support vector machines**. En Proceedings of the 24th ACM International Conference on Research and Development in Information Retrieval, ACM Press, 2001.

[110] *Kahle, B.* **Archiving the Internet**. Scientific American, Marzo, 1997.

[111] *Kilgarriff, A.* **I Don´t Believe in Word Senses, Computers and the Humanities**, 1997.

[112] *Kilgarrif, A.* **SENSEVAL: An exercise in evaluating Word Sense Disambiguation programs**. En Proceedings of LREC, Volume 1, Granada, España, página 581-588, 1998.

[113] *kilgarriff, A. y Palmer, Eds., M.* **SENSEVAL. Evaluation Word Sense Disambiguation programs**. En Computer and the Humanities, Volume 34, 2000.

[114] *Kilgarriff, A. y Palmer, M.* **Introduction to the Special Issue on SENSEVAL, Computers and the Humanities**, 2001.

[115] *Kirch, S.* **Infoseek´s experiences searching the internet**. ACM SIGIR forum, volume 32, página 3-7, 1988.

[116] *Kanuth, D.E.; Morris, J.H.; Pratt, V.R.* **Fast pattern matching in strings**. SIAM Journal of Computing, volume 6, página 323-350, Junio, 1977.

[117] *Kobayashi, M. and Takeda, K.* **Information Retrieval on the Web**. ACM, computing Surveys, volume 32, página 144-173, Junio, 2000.

[118] *Kosala, R. y Blockeel, H.* **Web mining research: a survey, SIG KDD Explorations**. Volume 2, página 1-15, 2000.

[119] *Koster, M.* **A standard for robot exclusion**. Disponible en http://www.robotstxt.org/wc/norobots.html, 1994.

[120] *Kowalski, G.J.; Maybury, M.T.; Luwer, Ed.* **Information storage and retrieval systems: Theory and implementations**. Academic Publishers, ISBN: 0-7923-7924-1, 2000.

[121] *Kraaij, W. e Pohlmann, R.* **Comparing the effect of syntactic vs.statistical phrase index strategies for Dutch**.En Proceedings ECDL'98, página 605-617, 1998.

[122] *Krovetz , R. e Croft, B.W.* **Lexical Ambiguity and Informationm Retrieval**. ACM Transactions on Information Systems, 10(2), página 115-141, 1992.

[123] *Krovetz, R.* **Homonymy and Polisemy in Information Retrieval**. En Proceedings of the 33th Meeting of the ACL, página 72-79, 1997.

[124] *Kulkarni, A.* **Word Sense Discrimination by clustering similarity contexts**. Tesis de Maestria, Departmento de Ciencias Computacionales, Universida de Minnesoata, Duluth, Agosto, 2004.

[125] *Kulkarni, A. y Pedersen, T.* **Sense Lusters: Unsupervised clustering and labelling of similar contexts**. En the Proceedings of the Demonstration and Interactive Poster Session of the 43rd Annual meeting of the Association for Computational Linguistics, Ann Arbor, Michigan, 2005.

[126] *Lancaster, F. W.; Sandore, B.* **Technology and management in library and information services**. Champaign: University of Illinois, 1997.

[127] *LDC.* **Sítio Web de LDC**. En http://www.ldc.upenn. Edu, 2004.

[128] *Leacock, C.; Towell, G.; Voorhees, E.M.* **Corpus-Based Statistical SenseResolution**. In Proceedings of the ARPA Human Language Technology Workshop, Morgan Kaufmann Publishers, página 260-265, San Francisco, 1993.

[129] *Leacock, C.; Chodorow, M.; Miller, G.A.* **Using Corpus Statistics and WordNet Relations for Sense Identification**. Computational Linguistics, página 147-165, 1998.

[130] *Leighton, H.V.* **Performance of four World Wide Web (WWW) index services: Infoseek, Lycos, WebCraweler, and WWWorm**. Disponible en http://www.winona.msus.edu/library/webind.htm, 1996.

[131] *Leighton, H.V.; Srivastava, J.* **Precision among World Wide Web Search Service (Search Engines): AltaVista, Excite, Hotbot, Infoseek, Lycos**. Disponible en http://www.winona.msus.edu/is-f/library-f/webind2/webind2.htm, 1997.

[132] *Lemur*. **Sítio Web de SRI LEMUR**. En http://www.lemurproject.org, 2004.

[133] *Lifantsev, M.* **OpenGrid (Open Global Ranking Search Engine and Directory)**. Disponible en http://www.ecsl.cs.sunysb.edu/ maxin/OpenGRID/, 1998.

[134] *Litkowski, K.C.* **Senseval-3 Tarea: Word Sense desambiguación de WordNet Glosses**. Proceedings of Senseval-3: El Tercer Taller Internacional sobre la Evaluación de Sistemas para el análisis semántico del texto, Asociación de Lingüística Computacional: Barcelona , España, página 13-16, Julio 2004.

[135] *Lowley, S.* **The evaluation of WWW search engines**. Journal of Document, volume 56, página 190-211, 2000.

[136] *Luk, A. k.* **Semantical sense disambiguation with relatively small corpora using dictionary definitions**. En Proceedings of the 33rd Meeting of the Associationfor Computational Linguistics (ACL), Cambridge, Massachustts, página 181-188, 1995.

[137] *Lyons , J.* **Semantics**. Cambridge University Press, 1997.

[138] *Mandala, R.; Tokunaga, T.; Tanaka, H.* **Combining multiple evidence from different types of thesaurus**. Proceedings of ACM-SIGIR'99, página 191-197, 1999.

[139] *Manning, C.D.; Schütze, H.* **Foundations of Statistical Natural Language Processing**. The MIT Press, Cambridge, 2001.

[140] *Marrafa, P.* **WordNet do Português: uma base de dados de conhecimento linguístico**. Instituto Camões, 2001.

[141] *McName, P. e Mayfield, J.* **A Language-Independent Approach to European Text-Retrieval**. En Peters, C. (ed): Cross-Language Information Retrieval and Evaluation, Workshop, CLEF 2000, Springer, página 129-139, 2000.

[142] *Méndez Rodrigues, E.M.; Merlo Vega, J.A.* **Localización, identificación y descripción de recursos web: tentativas hacia la normalización**. 1999. [on-line]. Disponível em: <http://rayela.uc3m.es/ mendez/ > Acesso em:15 de janeiro de 2003.

[143] *Mihalcea, R. y Moldovan, D.* **Automatic acquisition of sense tagged corpora**. Proceedings of Florida Artificial Intelligence Research Society Conference (FLAIRS 1999), Orlando, FL, May, 1999.

[144] *Mihalcea, R. y Edmonds, P.* **Proceedings of Senseval-3**. The 3rd. Int. Workshop on the Evaluation of Systems for the semantic Analysis of Text. Barcelona, Spain, 2004.

[145] **Advances on Word sense Disambiguation**. Notas del Tutorial, IBERAMIA-2004, Puebla, México, 2004.

[146] *Mihalcea, R.; Chklovski, T. y Kilgarriff, A.* **The Senseval-3 english lexical sample task**. En Proceedings of Senseval-3: The Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text, página 25-28, 2004.

[147] *Miller, A.; Beckwith, R.; Fellbaum, C.; Gross, D. y Miller, K. J.* **Introduction to WordNet: An on- line lexical database**. En International Journal of Lexicography, Volume 3, página, 235-312, 1990.

[148] *Miller, G.* **WordNet: A lexical database for english**. Communications of the ACM, 1991.

[149] *Miller, G. A.* **Introduction to WordNet: An On-Line Lexical Database**. NILC Núcleo Institucional de Lingüística Computacional, 1993.

[150] *Miller, George A.; Beckwith, R.; Fellbaum, C.; Gross, D.; Miller, K.***Introdution to WordNet: An On-line Lexical Database**, 1993.

[151] *Miller, G.A.; Chorodow, M.; Landes, S.; Leacock, C; Thomas, R.G.* **Using a Semantic Concordancer for Sense Identification**. In Proceedings of the ARPA Human Language Technology Workshop - ACL, página 240-243, Washington, 1994.

[152] *Miller, A.* **Wordnet: A lexical Database for English**. Comunications of the ACM, página 39-41, 1995.

[153] *Milstead, J.; Feldaman, S.* **Metadata: cataloging by any other name**, página 20, Fevereiro 2001.

[154] *Miltra, M.; Singhal, A. and Buckley, C.* **Improving Automatic query expansion**. In Proceedingss of ACM SIGIR Conference on Research and Development in information Retrieval, página 206-214, 1998.

[155] *Mitchell, T.* **Machine Learning**, McGraw-Hill, 1997.

[156] *Moffat, A.; Culpepper, J.S.* **Enhanced Byte Codes with Restricted Prefix Properties**. SPIRE, página 1-12, 2005.

[157] *Montoyo, A.* **Método basado en marcas de especificidad para WSD**. Procesamiento del Lenguaje Natural, página 24, 2000.

[158] *Monz, C. e Rijke, M.* **Shallow Morphological Analysis in Monolingual Information Retrieval for Dutch, German and Italian**. En Peters, C.: Cross-Language Information Retrieval Systems, Springer, página 262-277, 2002.

[159] *Moore,N.* **A sociedade da informação: tendências para o novo milénio**. Brasilia: IBICT, página, 94-108, 1999.

[160] *Navarro, G.* **Approximate text searching**. Phd Thesis, Department of Computer Science, Universidad de Chile, Diciembre, 1998.

[161] *Navarro, G.* **Indexing and Searching**. En Baeza-Yates, R.; Ribeiro-Nrto, B. "Modern Information retrieval", Addilson Wesley, ISBN 0-201-39829-X, capítulo 8, página 191-228, 1999.

[162] *Navarro, G.; Tarhio, J.* **Boyer-Moore string matching over Ziv-Lempel compressed text**. In Proceedings of CPM'2000, Lecture Notes in Comput. Sci, Springer verlag, Berlin, volume 1848, página 166-180, 2000.

[163] **NetSizer http://www.netsizer.com/**, página principal, 2005.

[164] *Ng, H.T.; Lee, H.B.* **Integrating Multiple Knowledge Sources to Disambiguate Word Senses: An Exemplar-Based Approach**. In Proceeedings of the 34th Annual Meeting of Association for Computational Linguistics, página 40-47, Somerset, 1996.

[165] *Ng, H. T.* **Getting Serious about Word Sense Disambiguation**. In Proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How?, Washington, página 1-7, 1997b.

[166] *Nicholas, A.; Locarides, A.* **Logics of converstion**. Cambbridge: cambridge University Press, 2003.

[167] *Open Directory Project* **http://www.dmoz.org**, página principal, 2007.

[168] *Page, L.; Brin, S.; Motwani, R. ; Winograd, T.* **The page rank citation ranking: Bringing order to the Web**. Annual Meeting of the American Society for Information Science, ASIS´98, 1998.

[169] *Paliouras, G.; Karkaletsis, V.; Androutsopoulos, I. y Spyropoulos, C. D.* **Learning rules for large-vocabulary Word Sense Disambiguation: a comparison of various classifiers**. En Proceedings of the 2nd International Conference on Natural Language Processing, Patra, Grecia, 2000.

[170] *Pancardo-Rodríguez, A.; Montes-y-Gómez, M.; Villaseñor-Pineda, L. y Rosso, P.* **A mapping between Classifiers and Training Conditions fow WSD**. En Computational Linguistics ant Intelligent Text Processing (CICLing 2005), Ciudad de México, México, página 246-249, 2005.

[171] *Pearl, J.* **Probabilistic reasoning in intelligent Systems**. Morgan Kaufmann, San Mateo, CA, 1988.

[172] *Peat, J.H; Willet, P.* **The limitations of term coocurrence data for query expansion in document retrieval systems**. Journal of American society for Information science, página 378-383, 1991.

[173] *Pedersen, T. y Bruce, R.* **Distinguishing word sense in untagged text**. En Proceedings of the second Conference on Empirical Methods in Natural Language Processing, Providence, RI, página 197-207, 1997.

[174] *Pedersen, T.* **Learning Probabilistic Models of Word Sense Disambiguation**. Tesis Doctoral, Universidad Metodista del Sur, Dallas, 1988.

[175] *Pedersen, T.* **A Decision Tree of Bigrams is an Accurate Predictor of Word Sense**, 2001.

[176] *Penas Padilla.* **Técnicas lingüísticas aplicadas a la búsqueda textual multilingüe: ambigüedad, variación terminológica y multilingüismo**, SEPLN, 2004.

[177] *Peón Espantoso, J. J.* **O arquiteto da informação e o bibliotecário do futuro**. Revista de Biblioteconomia de Brasília, volume 23/24, página, página 135-146, 2000.

[178] *Pereira, F.; Tishby, N.; Lee, L.* **Distributional Clustering of English Words.**. In Proceedings of the 31st Annual Meeting of Association for Computational Linguistics, página, 183-190, 1993.

[179] *Pinto, F. J.* **Uso de Recurso Linguístico WordNet en la Expansión de Consultas con un Modelo de Usuario de Recuperación de Información**. CEDI 2007, II Congreso Español de Informática, Zaragoza-España, Septiembre, 2007.

[180] *Pinto, F. J.* **Evaluación del Sistema de Recuperación de Información Lemur con Distintos Tipos de Indexación Automática**. Zoco'07/CAEPIA, XII Conferencia de la AEPIA, Salamanca-España, 12 de Novembro de 2007.

[181] *Platt, J. C.* **Fast training of SVMs using sequential minimal optimization**. En Advances in Kernel Methods-Support Vector Learning, B. Scholkopf, C. Burges, y A. Smola, Eds. MIT Press, Cambridge, Mass, 1998.

[182] *Porter, M.F.* **An algorithm for siffix stripping, Readings in information retrieval**. Morgan Kauffmann Publishers, página 313-316, 1997.

[183] *Preiss, J. y Yarowsky, Eds., D.* **Proceedings of SENSEVAL**. Association for Computational Linguistics Workshop, Toulouse, France, 2001.

[184] *Preiss, J. y Yarowsky, Ed., D.* **Proceedings of SENSEVAL**. Association for Computational Linguistics Workshop, Praga, France, 2007.

[185] *Purandare, A. y Pedersen, T.* **SenseClusters – Finding clusters that represent word senses**. En Proceeddings of Nineteenth National Conference on Artificial Intellegence (AAAI-04), San José, California y en Proceedings of

5th Annual Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL-04), Boston, Massachusetts, 2004.

[186] *Pustejovsky, J. y Boguraev, B.* **Introduction: Lexical Semantics in Context, Lexical Semantics: The Problem of Polysemy**. Oxford University Press, Oxford, 1996.

[187] *Qiu, Y. and Frei, H.* **Concept-based query expansion**. In Proceeding of SIGIR-93, 16th ACM International Conference Research and Development in Information Retrieval, Pittsburgh, US, página, 160-169, 1993.

[188] *Quinlan, J. R.* **Induction of decision trees**. En Machine Learning, Volume 1, página 81-106, 1966.

[189] *Quinlan, J. R.* **C4.5: Programs for Machine Learning**. Morgan Kaufmann, 1993.

[190] *Ramoni, M.; Sebastiani, P.* **Learning bayesian networks from incomplete databases**. Technical report KMI-TR-43, knowledge Media Institute, The Open University, 1996.

[191] *Ramachandran, V.* **Design Patterns for Building Flexible and Maintainable J2EE Applications**, 2000.

[192] *Rijsbergen, K. Van.* **Information Retrieval**. London: Butterwoths, 1979.

[193] *Rocha, M. E.* **O uso de corpora computadorizados no ensino de língua portuguesa: metodologia e avaliação**. In Grimm Cabral, L. et all (orgs). Linguística e ensino: novas tecnologias, Blumenal: Nova Letra, 2001.

[194] *Rodríguez Hontoira , H.* **Técnicas estadísticas en el tratamiento del lenguaje natural**. En Blecua, J.M. (ed): Filología e informática: nuevas tendencias en los estudios filológicos, Barcelona: UAB, página 111-140, 1999.

[195] *Rosenthal, M.; Chu, H.* **Searching engines for the World Wide Web: A comparative study and evaluation methodology**. American Society for Information Science, ASIS, página 127-135, Outubro, 1996.

[196] *Rosso, P.; Montes-y-Gómez, M.; Buscaldi, D.; Pancardo-Rodriguez, A. y Villaseñor Pineda, L.* **Two Web-based approaches for noum sense disambiguation**. En Computacional Linguistics and Intelligent Text Processing (CICling 2005), Ciudad de México, 2005.

[197] *Rowley, J.* **A biblioteca eletrônica. Brasília: Briquet de Lemos/Livros**, página, 2002.

[198] *Sabastiani, Marcelo* **Publicações científicas eletrônicas na Internet: modelos, padrões, e tendências**. São Bernardo do Campo. Dissertação apresentada ao Curso de Pós- Graduação em Comunicação Social da Universidade Metodista de São Paulo, página 256, 1999.

[199] *Salton, G., Lesk, M.* **Computer evaluation of indexing and text processing, The Smart Retrieval System: Experiments in Automatic**. Document Processing, 143-180, Prentice-Hall, Inc.Englewood Cliffs, New Jersey, 1971.

[200] *Salton, G.; Wong, A.* **Generation and search of clustered files**. ACM Transactions on Databases Systems, volume 3, página 321-346, 1978.

[201] *Sanderson, M.* **Word sense disambiguation and information retrieval**. In Proceedings of 17th International Conference on Research and Development in Information Retrieval, 1994.

[202] *Sanderson , M.* **Retrieving with good sense**. Information Retrieval, página 49-69, 2000.

[203] *Lancaster, F.W., Sandore, B.* **Technology and Management in Library and Information Services**. Library Association, London, 1997.

[204] *Schütze, H.* **Dimensions of Meaning**. In Proceedings of Supercomputing'92. IEEE Computer Society Press, Washington, página 787-796, 1992.

[205] *Schutze, H.; Pederson, J.O.* **A coocurrence-based thesaurus and two applications to information retrieval**. Proceedings of RIAO Conference, página 266-274, 1994.

[206] *Senseval.* **Sítio Web de Senseval**, http://citeseer.ist.psu.edu/, 2007.

[207] *Silva, I.; Ribeiro-Neto, B.; Calado, P.; Moura, E. and Ziviani, N.* **Link-based and content-based evidential information in a belief network model**. In Proceedings of the 23rd International ACM SIGIR Conference on Research and Development in Information Retrieval, página 96-103, 2000.

[208] *Silverstein, C.; Henzinger, M.; Morais, H.; Moriez, M.* **Analysis of a very Large Web Search Engine Query Log**. SGIR Forum, volume 33, página 6-12, 1999.

[209] *Singhal, A.; Buckley, C.; Mitra. M.* **Pivoted Document Length Normalization**. Procedings of ACM-SIGIR'96, página 21-29, 1996.

[210] *Smeaton, A.; Kelledy, F. and O'Donnell, R.* **TREC-4 experiments at dublin city university: Thresolding posting lists, query expansion with Wordnet and POS tagging of spanish**. In Proceedings of TREC-4, 1995.

[211] *Smeaton, A.F. and Quigley, A.* **Experiments on using semantic distances between words in image caption retrieval**. In Proceedings of the 19th, 1996.

[212] *Smrz, P.* **Finding Semantically Related Words in Large Corpora**. FIMU Report Series:Masaryk University, 2001.

[213] *Snyder, M. y Palmer, M.* **The English all-words task**. En Proceedings of Senseval-3: The Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text, página 41-43, 2004.

[214] *Stevenson, M.* **Word Sense Disambiguation. The case forCombinations of Knowledge Sources**. CSLI Publications, 2003.

[215] *Strazalkowski, T.; Lin , F.; Wang, J.; e Pérez-Carballo, J.* **EvaluatingNLP techniques in Information Retrieval. A TREC Perspective**. En : Strazalkowski (ed.): Natura Language Information Retrieval, Kluwer Academic Press, Elsevier, página 113-145, 1999.

[216] *Strohman, T.; Croft, W.B.* **Efficient Document Retrieval in Main Memory**, SIGIR 2007 Proceedings, página 175-182, 2007.

[217] *Towell, G. y Voorhees, E. M.* **Disambiguating highly ambiguous words. Computational Linguistics**. En Computational Linguistics, Volume 24, página 125 – 146, 1998.

[218] *TRECNIST.* **Sítio Web de TREC NIST**. En http://trec.nist.gov, 2004.

[219] *Tzoukerman, E.; KLAVANS, J.L.; Jacquemin, C.* **Efective Use of Natural Language Processing of Multi-Word Terms: The Role of Derivational Morphology, Part of Speech Tagging, and Shallow Parsing**. En Proceedings of 20th ACM/SIGIR, página 148-155, 1997.

[220] *Ullmann.* **Semântica: uma Introdução à Ciência do Significado**. Fundação Calouste Gulbenkian, Lisboa, 1964.

[221] *Vasilescu, F.; Langlais, P.y Lapalme, G.* **Evaluation variants of the Lesk approach for disambiguating words**. En Proceedings of LREC 2004, página 633-636, 2004.

[222] *Velluci, Sherry L.* **Metadata. Annual Review of Information Science and Technology**. Volume 36, página 503-389, 2002.

[223] *Veronis, J.* **A study of polysemy judgements and interannotator agreement**. Programme and advanced papers of the Senseval workshop, 1998.

[224] *Vilares, F.J.; Barcala, R.F.M.B; Fernández, L. S. e Pombo, O.J.* **Manejando la variación morfológica y léxica en Recuperación de Información Textual**, página 99-106, 2003.

[225] *Voohees, E. M* **Query Expansion using Lexical-Semantic Relations**. Proceedings of ACM-SIGIR'94, página 61-69, 1994.

[226] *Wang, Yih-Chen.; Vandendorpe, J.; Evens, M.* **Relational thesauri in information retrieval, Journal of the American Society for Information Science**, página 15-27, 1985.

[227] *Weaver, W.* **Translation in Machine Translation of Languages: Fourteen Essys**. The MIT Press, Cambridge, Massachusetts, página 15-23, 1949.

[228] *Weibel, S.; Godby, J.; Miller, E.* **OCLC/NCSA metadata workshop report**, Outubro, 2002.

[229] *Weiss, S.* **Learning to disambiguate**. En Information Storage and Retrieval. Volume 9, página 33-41, 1973.

[230] *Wilks, Y. A.; Fass, D.; Gruo, C.M.; MacDonald, J. E.; Plate, T. y Slator, B. A.* **Providing machine tractable dictionary tools**. In Semantics and the Lexicon, Pustejovsky, James (ED), MIT Press, Cambridge, Massachusetts, 1990.

[231] *Wilks, Y.* **Senses and Texts. Computers and the Humanities**, 1997.

[232] *Wilks, Y.; Stevenson, M.* **Sense Tagging: Semantic Tagging with a Lexicon**. In Proceedings of the SIGLEX Workshop "Tagging Text with Lexical Semantics: What,why and how". Washington, 1997a.

[233] *Wilks, Y. y Catizone, R.* **Can we make information extraction more adaptive?**. University of Sheffield, Computer Science Department, Memoranda en Computer and Cognitive Science, 2000.

[234] *Wolfram, D.; Spink, A.; Jansen, B.J.M.and Saracevic, T.* **Searching the web: The public and their queries**. Journal of the American Society for Information Science and Technology, página 226-234, 2001.

[235] *WordNet.* **Sítio Web de WordNet**, http://www.cogsci.princetion.edu/ wn/, 2005.

[236] *Witten, I.; Moffat, A. and Bell, T. Bell.* **Managing Gigabytes**. Morgan Kaufmann Publishers, San Francisco, California, USA, 2nd Edition, 1999.

[237] *Wu, S. and Manber, U.* **Agrep – a fast approximate pattern searching tool**. Usenix Conference, página 153-162, Janeiro, 1992.

[238] *Xu, J. and Croft , W. B.* **Improving the effectiveness of information retrieval with local context analysis**. ACM Transaction on Information Systems CACM Tois), página 79-112, 2000.

[239] *Yahoo!.* **Sítio Web de Yahoo!**, Página principal http://www.yahoo.com/, 2007.

[240] *Yarowsky, D.* **Decision list for lexical ambiguity resolution: Aplication to accent restoration in Spanish and French**. En Proceedings of the Annual Meeting of the Association for Compuatational Linguistics, página 88-95, 1994.

[241] *Yarowsky, D.* **Unsupervised word-sense disambiguation rivalling supervised methods**. En Proceedings of the 33rd Annual Meeting of the Association for Computational Lingistics (ACL-95). Cambrige, MA, 1995, página 189-196.

[242] *Zhai, C.; Tong, X.; Mili-Frayling, N. e Evans, D.* **Evaluation ofsyntactic phrase indexing CLARIT TREC5 NLP track report**. En The Fifth Text Retrieval Conference TREC-5, NIST Special Publication, 1997.

[243] *Zobel, J.; Moffat, A. and Ramamohanarao, K.* **Guidelines for Presentation and Comparision of indexing Techniques**. ACM SIGMOD Record, Volume 25, no. 3, página 10-15, Setembro 1996.