Extracción y organización del conocimiento de etiquetados. Aplicación a etiquetados en repositorios digitales sobre arte[1]

Gonzalo A. Aranda Corral
Departamento de Tecnologías de la Información, Universidad de Huelva
garanda@us.es

Joaquín Borrego Díaz
Departamento de Ciencias de la Computación e Inteligencia Artificial, Universidad de Sevilla
jborrego@us.es

Juan Galán Páez
Departamento de Ciencias de la Computación e Inteligencia Artificial, Universidad de Sevilla
juangalan@us.es

INTRODUCCIÓN

Las herramientas y servicios que la Web Social ha popularizado ofrecen a la comunidad académica tecnologías para avanzar en el análisis de los recursos digitales de investigación. Una de las primeras etapas en esa adopción consiste ofrecer repositorios digitales para que cualquier investigador tenga fácil acceso a los recursos. Los recursos se disponen bajo indexación, y suelen clasificarse de acuerdo a categorías. Junto con la categorización, una de las técnicas más populares y útiles para su clasificación y exploración es el etiquetado de dichos recursos.

El etiquetado permite el uso de *folksonomías* (vocabularios con cierto consenso en la comunidad que marca un tipo de recursos). La folksonomía, que se puede entender como un tipo de ontologías, permite al usuario navegar u organizar el conocimiento que el repositorio gestiona.

Sin embargo, este tipo de *ontologías emergentes* sufren de ciertas limitaciones y deficiencias. Por un lado, al ser un vocabulario con una intencionalidad dependiente del usuario que etiqueta, es habitual que exista cierta *heterogeneidad semántica*: ciertas etiquetas representan conceptos distintos para distintos etiquetadores. La heterogeneidad semántica es uno de los problemas intrínsecos al uso de etiquetados, y que impide que un usuario se aproveche con fiabilidad del etiquetado de otro. Por otro lado, junto a etiquetas con utilidad para la comunidad, nos encontraremos con etiquetas que son usadas raramente, que no

[1] Financiado por el Proyecto de excelencia TIC-6064 de la Junta de Andalucía *Conceptos emergentes en sistemas complejos. Aplicaciones en entornos urbanos y en complejidad cultural (eComplexcity)*, cofinanciado con fondos FEDER.

guardan relación semántica con el objeto etiquetado (por ejemplo *to read* para marcar una tarea futura) o cuyo significado real no es conocido por gran parte de la comunidad. A pesar de las limitaciones, los etiquetados compartidos es una solución potencial para acercar a la comunidad el conocimiento semánticamente organizado.

El problema anteriormente descrito es muy similar al que se enfrenta un usuario cuando intenta recabar información de un repositorio etiquetado. Cuando recurrimos a repositorios ajenos los etiquetados nos ofrece una rápida navegación. Debemos "entender" el significado de las etiquetas. En este caso, debemos *reconciliar* nuestro conocimiento con el conocimiento del repositorio. Es decir, establecer relaciones semánticas entre nuestros conceptos y nuestra manera de etiquetarlos y los del repositorio. Para que esta tarea (eminentemente racional y que realizamos de manera inconsciente) tenga éxito el etiquetado que exploramos debe cumplir ciertos criterios de consistencia y robustez.

Objetivo del trabajo

En este trabajo nos planteamos el uso del Análisis Formal de Conceptos (FCA) para estudiar la adecuación de los etiquetados en repositorios digitales. Propondremos una medida para su adecuación, basada en el estudio de otros objetos similares (redes semánticas, lenguajes de descripción, etc.) y aplicamos el estudio a algunos repositorios culturales y de arte.

ETIQUETADOS

El éxito del etiquetado en la Web Social se debe a que no tiene, en principio, ningún tipo de limitación de uso personal. Etiquetar es una tarea que en diferentes servicios o portales se considerar de distinta forma. Fundamentalmente se pueden usar de cuatro formas (Smith, 2007): Para manejar información personal (navegar entre los recursos que nos parecen más interesantes y que hemos etiquetado), como método social para marcar objetos digitales y así facilitar su compartición, o incluso para mejorar la experiencia de usuario en el comercio electrónico. Así, el usuario puede explorar su *personomía* así como las personomías de otros usuarios: un usuario puede ver los recursos que otros usuarios han subido y/o etiquetado (Jäschke, 2008).

El etiquetado como representación semántica

Sin embargo, y debido a las diferentes formas de usar las etiquetas, el manejo global, fiable y automatizado del conjunto de todas las etiquetas utilizadas en una comunidad (la folksnomía) es un desafío. Cuando el usuario añade una etiqueta a un recurso es difícil atrapar todos los matices que el usuario pretendía representar. Como se argumenta en (Golder *et al*., 2006), el etiquetado es una tarea para dotar de sentido los recursos que poseemos o visitamos, y pretende categorizar de manera que existe un significado emergente que se produce con esa actividad (Weick *et al*., 2005). Como conocimiento emergente, la frontera de esos conceptos que manejamos cuando elegimos etiquetas son vagos, lo que puede producir que algunas etiquetas sean de dudosa utilidad o adecuación. Como consecuencia de eso, un etiquetado personal puede ser poco útil como bien público (Golder *et al*., 2006).

Aunque a distinta escala, la etiquetación de objetos por parte de una comunidad puede presentar el mismo problema. Sin embargo, en este último caso la dificultad es usualmente resuelta por la Inteligencia Colectiva: cuando la comunidad etiqueta globalmente se tiende a unificar el uso de la etiqueta, y el conjunto de etiquetas más comunes asociadas a un objeto o conjunto de objetos representa una descripción de un cierto concepto (Halpin *et al*., 2007). De hecho, estas etiquetaciones globales facilitan la construcción de sistemas de recomendación (Carmel *et al*., 2010).

Las humanidades digitales pueden y deben explotar esas facilidades para, entre otras tareas, difundir repositorios digitales para que los investigadores tengan acceso a la información. El etiquetado colaborativo, como forma de indexación social, es un método natural y directo para que laboratorios y grupos de investigación permitan la navegación entre sus recursos de manera independiente a la pura clasificación o taxonomía. El conjunto de etiquetas utilizadas, en el caso de temas específicos, es una clasificación social atendiendo a criterios concretos y muy relacionados con el tema al que el repositorio está dedicado, usualmente construida por especialistas. Es decir, la inteligencia colectiva ayuda a superar las limitaciones indicadas en la sección anterior.

Desde este punto de vista, la etiquetación colaborativa ofrece una alternativa pragmática al uso de ontologías (para la Web Semántica). Las ontologías en Tecnologías de la Información son representaciones formales de la conceptualización del dominio de discurso, y su uso (al menos en las primeras etapas de digitalización y gestión de recursos culturales) presenta varias dificultades: se necesita conocer los rudimentos de ese tipo de teorías, se presupone que los usuarios comparten esa ontología (es decir, participan del consenso con el que fue construida), etcétera.

Es indudable, no obstante, que la adopción de tecnologías de la Web Semántica para organizar el conocimiento de repositorios digitales es un paso fundamental para la interoperabilidad de los repositorios y la organización del conocimiento que contiene. Fundamentalmente, resuelve el problema de la heterogeneidad semántica y provee definiciones compartidas de los conceptos. Esa fiabilidad contrasta con las organizaciones personales de la información (por ejemplo, los etiquetados de los usuarios de *Delicious* (http://delicious.com/), que necesitan reconciliar su conocimiento con el de otros usuarios para aprovechar su información. En (Aranda Corral *et al*., 2010) proponemos un método para resolver esta dificultad, basado en agentes y usando técnicas de Análisis Formal de Conceptos. Finalmente, es necesario destacar que sí existen proyectos para estandarizar la semántica de etiquetas, como por ejemplo *Faviki* (http://www.faviki.com) y *CommonTag* (http://commontag.org).

Heterogeneidad semántica dependiente del contexto

Analizando las limitaciones del etiquetado colaborativo, existen dos problemas fundamentales. El primero consiste en la dependencia de la etiqueta del contexto en la que se usa. Si en el repositorio se manejan objetos muy diversos, es posible que usuarios distintos usen la misma etiqueta para referirse a conceptos distintos. Este problema limita tanto la efectividad del etiquetado como su adecuación. Es la denominada *Heterogeneidad Dependiente del Contexto* (HDC). La segunda limitación es la que denominamos Ambigüedad Clásica (AC), inherente

del lenguaje y de los distintos niveles de abstracción/generalidad que distintos usuarios consideran al etiquetar (Golder *et al*., 2006; Tanaka *et al*., 1991).

La segunda, AC, no es una limitación crítica si los usuarios etiquetan, por ejemplo, urls (puesto que el contenido del sitio web permite la desambiguación). Utilizando técnicas de análisis clúster se pueden localizar las distintas acepciones de la etiqueta ambigua (Au Yeung *et al*., 2009). Sin embargo, la HDC está asociada a la *estructura conceptual* que el usuario utiliza pero no está explícita en ninguna parte del sistema.

Como el HDC obstruye cualquier aproximación al consenso sobre las etiquetas, debemos diseñar algún método para intentar solventarla, usando herramientas inteligentes para recuperar el conocimiento y describirlo (Jung, 2009). Llegados a este punto, existen dos perspectivas para atacar el problema. La primera trata de interoperabilidad semántica *global* entre toda la comunidad, y la segunda es facilitar puentes semánticos para que un usuario pueda aprovechar de manera fiable el trabajo de otro, por ejemplo permitiendo que entienda el etiquetado de otro. La primera requiere un tratamiento global y un análisis de la folksonomía, que es la adoptada en este artículo: como paso previo se debería estudiar la calidad del etiquetado proporcionado por la comunidad para el repositorio concreto. La segunda perspectiva se centra en encontrar la conceptualización oculta en la actividad de etiquetado de cada usuario para establecer *traducciones automáticas* de los etiquetados.

Existen proyectos para unificar los etiquetados (y convertirlos de ese modo en metadatos), como por ejemplo DTSIL (Khoo *et al*., 2012) cuyo objetivo es integrar diversos repositorios en una infraestructura de información única, accesible y que admita interfaces usables, utilizando etiquetas que son elegidas mediante extracción de información de los metadatos. Otro proyecto relacionado es *Rich Tags* (http://research.mspace.fm/projects/richtags) que se aprovecha del contexto (la comunidad que mantiene el repositorio) para facilitar la navegación/exploración usando etiquetas que, debido a la especificidad del repositorio, no tienen problemas de heterogeneidad semántica, es decir, son etiquetas *semánticas* en ese sentido.

ANÁLISIS FORMAL DE CONCEPTOS

La convergencia de la Web Social hacia una Web de conocimiento depende de cómo tratar con técnicas semánticas formas de organización del conocimiento como el etiquetado. Una herramienta muy útil para dirigir esa convergencia es el Análisis Formal de Conceptos (FCA). Según R. Wille (Ganter *et al*., 1997), FCA matematiza el concepto filosófico de concepto como una unidad de pensamiento compuesto de dos partes: la extensión y la intención. La extensión está compuesta por todos los objetos que pertenecen al concepto, mientras que la intención comprende todos los atributos comunes a los objetos de la extensión. Partiendo de esta idea, FCA se desarrolla como una rama de la matemática aplicada cuyo objetivo es descubrimiento, extracción y organización de conocimiento a partir de datos cualitativos mediante redes (jerarquías) de conceptos. La teoría provee técnicas para poder razonar con el conocimiento extraído. Es decir, FCA es una aproximación a la minería de datos con una componente semántica (conceptos formales) y una componente lógico-computacional (reglas de asociación y razonamiento automático). Introduzcamos brevemente los rudimentos de FCA y su utilización sobre etiquetados.

La unidad básica de datos es el denominado *contexto formal* Un contexto formal M = (O, A, I) está compuesto de dos conjuntos, O (objetos) y A (atributos), junto con una relación entre éstos. Los contextos se pueden representar mediante una tabla de valores 0,1. En la figura 1 (izquierda) se representa un ejemplo sencillo de *universo de peces* junto con tres atributos. Dado un conjunto X de objetos (o Y un conjunto de atributos) se define la *intensión* de X (respectivamente la *extensión* de Y) como X' = { a ∈ A | oIa para todo o ∈ X } (respectivamente Y' = { o ∈ O | oIa para todo a ∈ Y }).

El objetivo principal de FCA es el cálculo de los conceptos extraídos del contexto. Un contexto formal corresponde a la idea anteriormente esbozada: es un par (X,Y) tal que X' = Y e Y' = X. Por ejemplo, el *retículo* de todos los conceptos del contexto de peces se muestra en la figura 1 (derecha). Para leer los conceptos de la jerarquía, representados cada uno por un nodo, hay que recorrer el retículo desde el nodo hasta el concepto más general para obtener la intención y hacia el concepto más específico para obtener la extensión. Por ejemplo, el concepto más específico del retículo sobre peces es({*eel*}, { *Coast, Sea, River*}) que correspondería al concepto *pez eurihalino*. Este ejemplo ilustra claramente que el retículo de conceptos contiene conceptos para los que no disponemos de un término que lo denote. Es decir, hemos descubierto un nuevo concepto. El interés de éste debe ser estudiado por el especialista en la materia que describía el contexto.

Para aplicar FCA sobre etiquetados debemos adaptar al formato de contexto formal el entorno formado por etiquetas, usuarios y recursos etiquetados. La aproximación general se puede ver en (Jäschke, 2008), donde utilizan los denominados *conceptos triádicos*. Puesto que en nuestro caso deseamos analizar la estructura general del repositorio etiquetado, no es necesario hacer la distinción del usuario que ha etiquetado el recurso. Siguiendo con el ejemplo, en la figura 2, se describe la estrategia general para aplicar FCA a etiquetados: los ítems etiquetados (en este caso, de *Delicious*) son los objetos y sus etiquetas los atributos asociados. A partir del contexto así construido se calculan los conceptos que se extraen y se obtiene la jerarquía de conceptos, que representa una jerarquía de conceptos sobre el universo de discurso de ese recurso (en nuestro caso, de los peces).

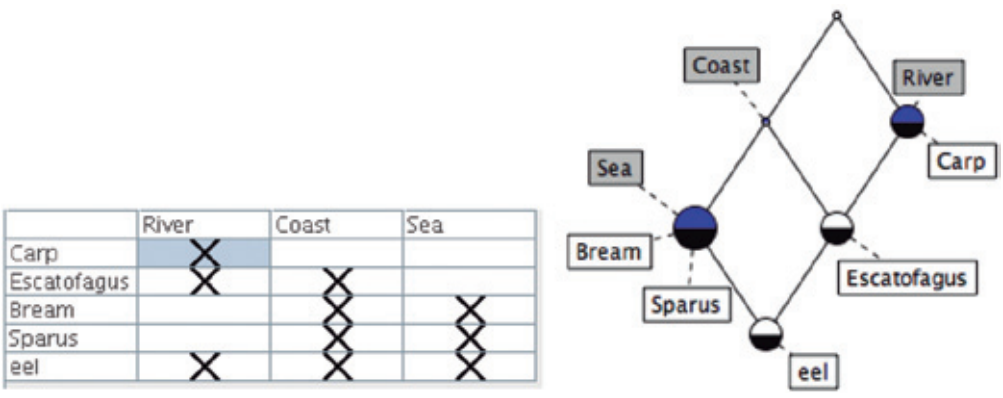

Fig. 1. Contexto formal sobre peces (izquierda) y su retículo de conceptos asociado (derecha).

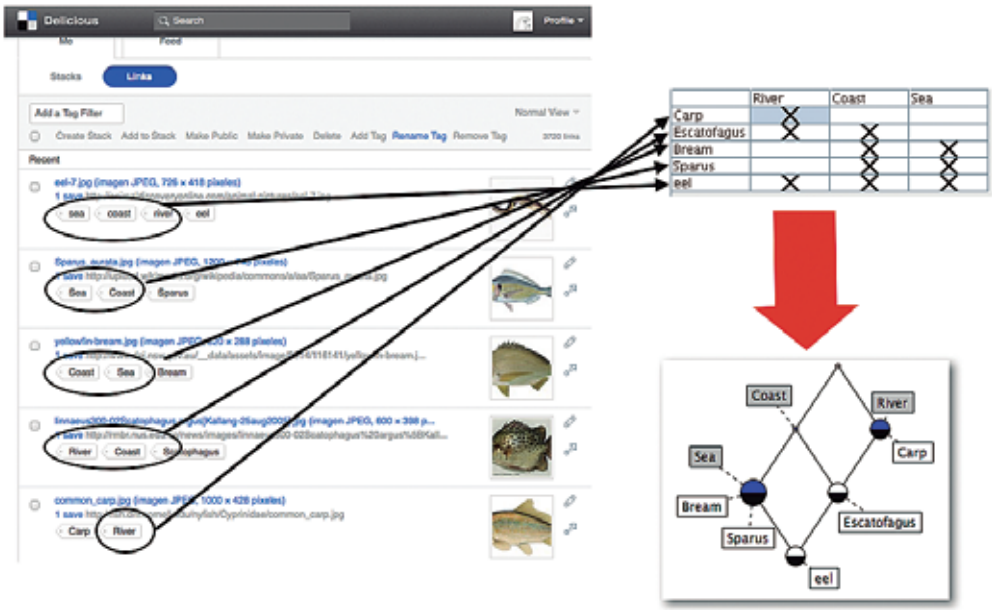

Fig. 2. Extracción del retículo de conceptos a partir del etiquetado de un usuario.

CONCEPTUALIZACIÓN LIBRE DE ESCALA

Llegados a este punto, y antes de analizar repositorios concretos mediante FCA, es necesario encontrar criterios que nos indiquen hasta qué punto el etiquetado es adecuado. Téngase en cuenta que el análisis debe ser independiente del área que abarca el repositorio y del área de conocimiento donde se enmarca el contexto. Recuérdese que FCA produce incluso conceptos que la propia etiquetación no es capaz de asignarle un nombre. Por tanto, cualquier medida de adecuación debe considerar la estructura de los conceptos que abarca una vez aplicado el proceso de conceptualización descrito anteriormente. Para obtener ese test de adecuación debemos analizar la estructura del retículo de conceptos como grafo (red) semántica.

El primer y fundamental análisis se basa en estudiar la distribución de los grados de conexión que posee cada concepto, es decir, con cuántos conceptos está relacionado. Esa distribución es un factor característico de las redes semánticas y ha sido extensamente estudiada para gran variedad de casos. Veamos primero la distribución que caracteriza la complejidad de la red de conceptos. Una vez estudiada, debemos compararla con otras redes semánticas existentes como por ejemplo la asociada a *WordNet*, que representa un caso de éxito, nuestro lenguaje. Parece natural conjeturar que si la estructura de los conceptos que se extrae es similar a la de los casos de éxito, entonces es adecuada (y útil).

Al contrario que otras redes semánticas (Motter *et al.*, 2003), nos enfrentamos a una red con relaciones que no están acotadas por el propio lenguaje, pues FCA extrae todos los conceptos, incluso los que no están definidos en el lenguaje original (puede que ni siquiera dispongamos de una definición explícita). Es tarea del especialista en el tema del repositorio el que debe interpretar los que

considere interesantes. Por tanto, el retículo será una red semántica compleja. De hecho, durante el resto del artículo veremos que en el caso de repositorios con etiquetaciones adecuadas, su topología responde a la de una *red compleja libre de escala*. Una distribución libre de escala es una distribución de los grados de conexión de tipo $P(k) \sim ck^{-\gamma}$, donde P(k) denota la proporción de nodos de la red que tienen k conexiones y c es una constante. Es decir, el número de nodos con alta conectividad es muy pequeño, y existen una gran cantidad de nodos con pocos vecinos. El parámetro γ suele verificar $2 < \gamma < 3$, aunque ocasionalmente pueden encontrarse valores fuera de ese intervalo.

En (Aranda Corral *et al*., 2012a) estudiamos la siguiente característica, que parafraseamos aquí para analizarlo en el caso de etiquetados, denominada *Hipótesis de la conceptualización libre de escala* (SFCH):

El conjunto de etiquetas asociado al repositorio es adecuado si induce una red libre de escala.

Es decir, si el conjunto de etiquetas induce una estructura de conceptos similar a la de una red semántica de las que emergen en el estudio del lenguaje. Por tanto, podemos considerar FCA como una forma de estudiar el lenguaje emergente de la etiquetación. La hipótesis SFCH se ha testeado en casos muy diversos, y se ha comprobado que los retículos provenientes de contextos aleatorios no exhiben esa distribución (Aranda Corral *et al*., 2012a).

Es necesario en este momento analizar un poco más la existencia de conceptos que no tienen nombre concreto. Es natural pensar que el uso de términos específicos por parte del especialista tiene influencia en las relaciones entre conceptos. Como el retículo de conceptos abarca todo potencial nuevo término del lenguaje (aquel que representaría a un nuevo concepto), la topología de la red será compleja. En (Motter *et al*., 2003) se estudia la red semántica entre conceptos expresados por términos (en el idioma inglés), mostrando que la topología es libre de escala. En (Motter *et al*., 2003) se muestra que la conectividad sigue una ley libre de escala $P(k) \sim ck^{-\gamma}$, con $\gamma = 3.5$. En las redes construidas en FCA se tiene que $\gamma > 3.5$ debido a que no está acotada por el propio lenguaje. En (Barabasi *et al*., 2002) se analiza la red que representa las palabras y la relación de sinonimia, estimando en este caso $\gamma = 2.8$. Otro ejemplo es la red semántica asociada el Tesauro de Roget (Roget, 1911) estudiada en (Steyvers, 2005), para el que $\gamma = 3.19$.

LA ESTRUCTURA CONCEPTUAL DEL LENGUAJE COMPLETO: WORDNET

Para analizar la bondad de un conjunto de etiquetas, es interesante observar cómo se comporta, desde el punto de vista de los conceptos que define, el propio lenguaje que usamos. Con este objetivo usaremos la base de datos léxica *WordNet* (http://wordnet.princeton.edu/).

WordNet es una gran base de datos léxica de palabras del Inglés, de uso libre. En esta base de datos, nombres, verbos, adjetivos y adverbios se agrupan en conjuntos de sinónimos cognitivos (llamados *synsets*), donde cada uno de estos *synsets* expresa un concepto diferente del lenguaje. Es decir, en esta base de datos las palabras se relacionan entre si mediante su significado y por tanto dos palabras estarán relacionadas si comparten una o más acepciones.

Para realizar el análisis de la estructura conceptual, primero construimos el contexto formal asociado a la base de datos. En este paso, consideraremos las

palabras y los *synsets* como objetos y atributos (respectivamente) del contexto. Una vez creado el contexto, calculamos a partir de este, el retículo de conceptos asociado (la estructura conceptual bajo estudio).

Se han considerado tres conjuntos de datos diferentes relacionados con WordNet para ser analizados. Por un lado se ha tomado WordNet al completo, y por otro dos subconjuntos de este, el de los adverbios y el de los verbos (en la tabla 1 se muestran las principales características de los conjuntos de datos usados). De esta forma veremos que las propiedades de todo el conjunto se mantienen en cada una de sus partes.

Tabla 1. Principales parámetros para cada conjunto de datos. Donde |O|, |A| y |I| se refieren al número de objetos, atributos y relaciones (respectivamente) del contexto. *Densidad* indica la densidad del contexto, |CL| el número de conceptos del retículo y <k> el grado medio de los nodos (conceptos) del retículo.

	|O|	*|A|*	*|I|*	*Densidad*	*|CL|*	*<k>*
WordNet completo	155.287	117.659	206.941	0,001%	110.663	2,406
Adverbios sólo	4.481	3.621	5.580	0,03%	3.529	2,187
Verbos sólo	11.529	13.767	25.047	0,02%	12.222	2,967

En la figura 3 se muestra la distribución de grados asociada a los tres conjuntos de datos descritos. Como se puede observar todos ellos cumplen la hipótesis SFCH. Estas estructuras nos servirán para a continuación estudiar los conjuntos de datos relacionados con etiquetados. Además es interesante mencionar que los *conceptos formales* extraídos de WordNet coinciden en gran parte con los conceptos del lenguaje.

ANÁLISIS DE ETIQUETADOS DE REPOSITORIOS DIGITALES DE ARTE E HISTORIA

Para ilustrar el análisis, hemos elegido dos repositorios digitales de humanidades, el de *Baroque Art* (<http://baroqueart.cultureplex.ca/>) del laboratorio *CulturePlex* (<http://www.cultureplex.ca/>) y el de *Gothic Past* (<http://www.gothicpast.com/>). Compararemos su estructura con la conceptualización asociada a subconjuntos notables de WordNet.

El Barroco Hispánico

El proyecto *The Hispanic Baroque: Complexity in the first Atlantic culture* (<http://www.hispanicbaroque.ca/>) es un proyecto multidisciplinar desarrollado por un grupo de investigadores de universidades de diferentes países, financiado por el *Social Sciences and Humanities Research Council of Canada.* Con una duración de siete años, el proyecto pretende estudiar el origen, evolución, transmisión y efectividad de los patrones de comportamiento y representación del barroco en el mundo hispánico.

Los objetivos de este proyecto son: identificar y describir los patrones más comunes y resistentes en entornos diversos; establecer su relación con procesos de identidad y organización sociales; analizar las tecnologías de ámbito cultural

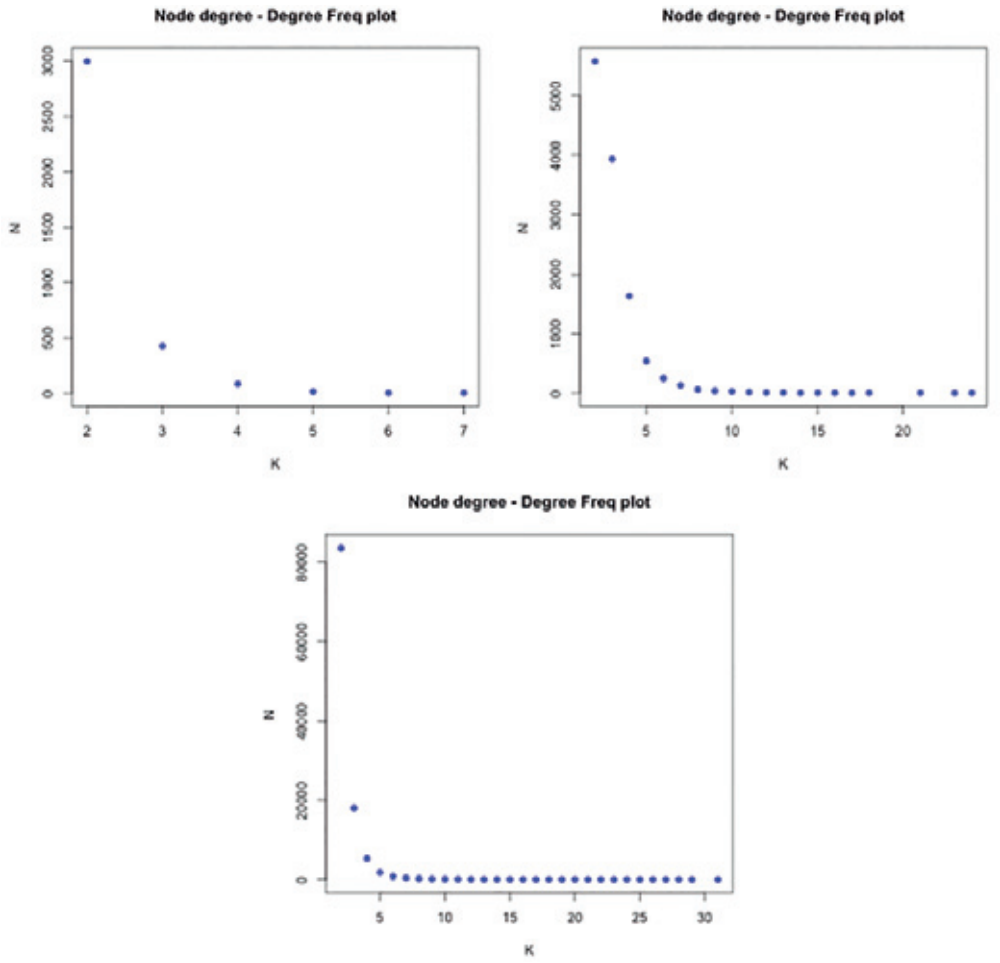

Fig. 3. Distribución de grados de la estructura conceptual asociada al los conjuntos de datos de WordNet: adverbios (arriba a la izquierda), verbos (arriba a la derecha) y WordNet completo (abajo).

que han hecho posible esta adaptación del barroco; determinar su efectividad basada en la reaparición del Neo-barroco en el mundo contemporáneo; desarrollar nuevas herramientas, con la intervención de otras disciplinas, que permitan reforzar los métodos de investigación en humanidades. Este proyecto cubre tres dimensiones fundamentales: investigación, formación de estudiantes y difusión del conocimiento obtenido.

Aunque el proyecto es de gran envergadura, con secciones de diverso ámbito. En este trabajo hemos usado como recurso a analizar la base de datos de obras de arte *Baroque Art*. Esta base de datos proporciona un etiquetado de obras de arte, que es uno de los productos del proyecto *Hispanic Baroque*. De este etiquetado destacamos que ha sido desarrollado por un equipo de personas cerrado y reducido, siguiendo unas directrices comunes preestablecidas y

partiendo de un vocabulario (etiquetas) también cerrado y preestablecido (ontología). El conjunto de datos bajo estudio esta formado por aproximadamente 11.000 obras y 200 etiquetas.

En la figura 4 (izquierda) se muestra la distribución de grados de la estructura conceptual asociada al etiquetado de obras del barroco, que también cumple la hipótesis SFCH. Es interesante observar la diferencia que hay en la forma de la distribución, al inicio de esta, con respecto a las de WordNet. Esto se debe a que el lenguaje por sí sólo presenta una estructura muy ordenada, como se comentó en el apartado anterior, mientras que el etiquetado no.

Gothic Past

Gothic Past: Visual Archive of Gothic Architecture and Sculpture in Ireland es un archivo de acceso libre para el estudio de las construcciones medievales de Irlanda. El archivo incluye entre otros elementos, imágenes, etiquetas, descripciones detalladas, etc. Es un proyecto colaborativo entre varias instituciones de Irlanda y financiado por el *Irish Research Council for the Humanities and Social Sciences (IRCHSS)* y el *Irish Heritage Council*. Incluye imágenes y recursos de las principales colecciones de historia, arte y arquitectura de Irlanda. El proyecto está construido sobre la plataforma de software libre *Omeka* (<http://omeka.org/>) que facilita la gestión de la información además de su visualización. Además permite de forma sencilla la colaboración con el proyecto de usuarios externos.

Este aspecto colaborativo, por lo que es sumamente inteligente trabajar con su etiquetado de monumentos históricos, ya que nos permitirá compararlo con el caso anterior, de diferente naturaleza. El archivo *Gothic Past* permite que cualquier usuario añada nuevos monumentos o modifique existentes, por tanto el vocabulario de etiquetas es abierto y dinámico, en contraposición con el vocabulario de *Baroque Art* que era cerrado y por tanto estático. Esto implica además, que el número de personas que intervienen en el etiquetado es muy elevado lo que conlleva gran heterogeneidad en criterios de etiquetado, mientras

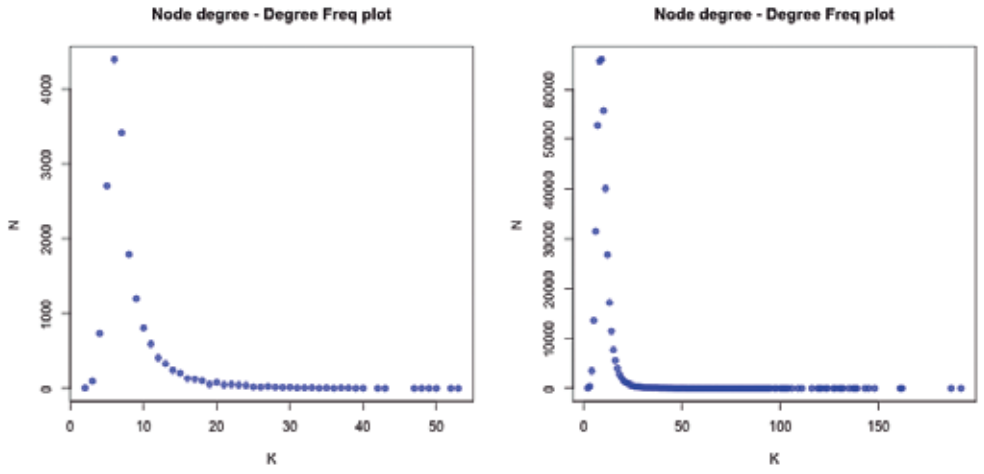

Fig. 4. Distribución de grados de la estructura conceptual asociada al los conjuntos de datos de Baroque Art (izquierda) y Gothic Past (derecha).

que en el caso del *Baroque Art* el equipo encargado del etiquetado es reducido y organizado. En este caso el conjunto de datos bajo estudio está formado por aproximadamente 3.250 obras y 1.700 etiquetas.

En la figura 4 (derecha) se muestra la distribución de grados de la estructura conceptual asociada al etiquetado de monumentos del gótico en Irlanda, que también cumple la hipótesis SFCH. Vemos que también aparece esa sensible diferencia al inicio de la distribución, por los motivos ya mencionados en el caso del barroco.

Por último es interesante mencionar la diferencia en la complejidad de la estructura conceptual asociada a cada uno de los etiquetados descritos. Aunque en las gráficas de las distribuciones de grados es más difícil observar este hecho, se aprecia claramente en la tabla 2 donde se muestran las principales características de estos conjuntos de datos. Partiendo de contextos formales con dimensiones similares, se deducen de estos retículos de conceptos de tamaño muy diferente. En el caso del etiquetado de obras del barroco, el retículo tiene casi 18.000 conceptos, mientras que en el caso del gótico el retículo tiene casi 420.000 conceptos, esto es, más de veinte veces más.

Esta importante diferencia estructural se debe a la diferente naturaleza de cada etiquetado. Como se ha comentado el etiquetado del barroco usa un vocabulario preestablecido y cerrado en el que intervienen pocos etiquetadores, mientras que en el del gótico el vocabulario es libre y dinámico y en el intervienen todos los etiquetadores que se ofrezcan. Esto hace que el etiquetado del gótico sea mucho más heterogéneo y complejo, si bien, el hecho de que también verifique la hipótesis SFCH indica que es igualmente válido.

Tabla 2. Principales parámetros para cada conjunto de datos. El significado de las columnas se corresponde con las de la tabla 1

	$\|O\|$	$\|A\|$	$\|I\|$	*Densidad*	$\|CL\|$	$<k>$
Baroque Art	11.062	221	74.993	3.067%	17.817	7,949
Gothic Past	3.246	1.781	66.432	1,149%	416.896	9,834

CONCLUSIONES Y TRABAJO FUTURO

En este trabajo proponemos el estudio de la estructura de las redes semánticas asociadas a repositorios digitales como medida de la adecuación de su etiquetación. Medir la calidad del etiquetado, una actividad que el usuario o comunidad que construye el repositorio puede hacer con cierto grado de libertad, puede indicarnos cómo de útil es ésta en dos tareas fundamentales: 1) Para facilitar la exploración del repositorio con éxito. 2) Para poder combinar distintos repositorios.

Existe otro aspecto, no tratado en este trabajo por cuestiones de espacio, que consiste en utilizar etiquetados robustos como sistemas de recomendación de etiquetas para los nuevos objetos digitales que deseamos integrar en el repositorio. Si bien los expertos introducen los objetos con etiquetas, el análisis global del repositorio provee de sistemas basados en conocimiento que completarían éste (y, en última instancia, conservaría la estructura del retículo de conceptos). Esta idea, utilizada con éxito en otros campos (Aranda Corral *et*

al., 2013), sería muy útil cuando el repositorio es suficientemente completo como para extraer conocimiento en forma de sistema experto.

BIBLIOGRAFÍA

Aranda Corral, Gonzalo A.; Borrego Díaz, Joaquín y Galán Páez, Juan, "Complex Concept Lattices for Simulating Human Prediction in Sport", *Journal of Systems Science and Complexity*, 26 (2013), pp. 117-136.

Aranda Corral, Gonzalo A.; Borrego Díaz, Joaquín y Galán Páez, Juan, "Scale-free structure in concept lattices associated to complex systems", *Proceedings International Conference on Complex Systems,* 2012a, pp. 1-6.

Aranda Corral, Gonzalo A.; Borrego Díaz, Joaquín y Giráldez Cru, Jesús, "Conceptual-Based Reasoning in Mobile Web 2.0 by means of Multiagent Systems", *Knowledge Engineering Notes,* 2012b, pp. 176-183.

Aranda Corral, Gonzalo A.; Borrego Díaz, Joaquín y Giráldez Cru, Jesús, "Agent-mediated shared conceptualizations in tagging services", *Journal of Multimedia Tools and Applications*, 65 (2012c), pp. 5-28.

Aranda Corral, Gonzalo A. y Borrego Díaz, Joaquín, "Reconciling Knowledge in social tagging web services", *Proceedings International Conference on Hybrid Artifial Itelligence Systems,* 2010, pp. 383-390.

Au Yeung, Ching-man; Gibbins, Nicholas y Shadbolt, Nigel, "Contextualising tags in collaborative tagging systems", *Proceedings of the 20th ACM conference on Hypertext and Hypermedia,* 2009, pp. 251-260.

Barabasi, Albert-Lazslo y Albert, Reka, "Statistical mechanics of complex networks", *Reviews of Modern Physics*, 74 (2002), pp. 47-97.

Carmel, David; Roitman, Haggai y Yom Tov, Elad, "Social bookmark weighting for search and recommendation", *The International Journal on Very Large Data Bases*, 19 (2010), pp. 761-775.

Clauset, Aaron; Rohilla Shalizi, Cosma y Newman, M.E.J., "Power-law distributions in empirical data", *SIAM Review*, 51 (2009), pp. 661-703.

Eklund, Peter W. y Wray, Tim, "Social Tagging for Digital Libraries using Formal Concept Analysis", *Proceedings 7th International Conference on Concept Lattices and Their Applications,* 2010, pp. 139-150.

Ganter, Bernhard y Wille, Rudolf, *Formal Concept Analysis: Mathematical Foundations*, New York, Springer-Verlag, 1997.

Golder, Scott y Huberman, Bernardo A., "Usage Patterns of Collaborative Tagging Systems", *Journal of Information Science*, 32 (2006), pp. 198-208.

Halpin, Harry; Robu, Valentin y Shepherd, Hana, "The complex dynamics of collaborative tagging", *Proceedings of the 16th International Conference on World Wide Web,* 2007, pp. 211-220.

Jäschke, Robert, *Formal Concept Analysis and Tag Recommendation in Collaborative Tagging Systems*, Heidelberg, IOS Press, 2011.

Jäschke, Robert *et al*., "Discovering shared conceptualizations in folksonomies", *Journal of Web Semantics*, 6 (2008), pp. 38-53.

Jung, Jason J., "Knowledge distribution via shared context between blog-based knowledge management systems: A case study of collaborative tagging", *Journal of Expert Systems with Applications*, 36 (2009), pp. 10627-10633.

Khoo, Michael *et al*., "Towards digital repository interoperability: the document indexing and semantic tagging interface for libraries (DISTIL)", *Proceedings of the Second international conference on Theory and Practice of Digital Libraries,* 2012, pp. 439-444.

Lee, Sangjin y Park, Jonghun, "Topic based photo set retrieval using user annotated tags", *Journal of Multimedia Tools and Applications*, 64 (2013), pp. 7-26.

Motter, E.A. *et al*. "Topology of the conceptual network of language", *Physical Review E*, 65 (2003).

Onifade, Olufade; Falade, Williams *et al*., "Fuzzontology: Resolving Information Mining Ambiguity in Economic Intelligent Process", *Communications in Computer and Information Science*, 54 (2010), pp. 232-243.

Roget, Peter Mark, *Roget's Thesaurus of English Words and Phrases*, 1911.

Smith, Gene, *Tagging: People-Powered Metadata for the Social Web, Safari*, Berkeley, New Riders, 2007.

Steyvers, Mark y Tenenbaum, Joshua B., "The large-scale structure of semantic networks: statistical analyses and a model of semantic growth", *Journal of Cognitive Science*, 29 (2005), pp. 41-78.

Tanaka, James W. y Taylor, Marjorie, "Object Categories and Expertise: Is the Basic Level in the Eye of the Beholder", *Proceedings of the Fourth International Workshop on Machine Learning,* 1991, pp. 457-482.

Weick, Karl E.; Sutcliffe, Kathleen M. y Obstfeld, Oscar, "Organizing and the Process of Sensemaking", *Journal of Organization Science*, 16 (2005), pp. 409-421.

RESUMEN

El análisis formal de conceptos (FCA) es una rama de la matemática aplicada cuyo objetivo es el descubrimiento, extracción y organización de conocimiento a partir de datos cualitativos. La teoría provee técnicas para poder razonar con el conocimiento extraído. Es decir, FCA es una aproximación a la minería de datos con una componente semántica (conceptos formales) y una componente lógico-computacional (reglas de asociación y razonamiento automático).

El objetivo de este trabajo es comprobar la aplicabilidad de FCA para tratar conocimiento a partir de repositorios digitales que mantienen la información etiquetada. Desde el punto de vista de FCA, se pretende estudiar la estructura de los conceptos que subyacen en repositorios digitales de arte. La idea es detectar patrones en la estructura conceptual de este tipo de etiquetados, con el objetivo a largo plazo de establecer medidas que estimen la calidad de éstos.

Entre otros, analizaremos etiquetados como los de *Baroque Art* del laboratorio *CulturePlex* (<http://baroqueart.cultureplex.ca/>) y el *Visual Archive of Gothic Architecture and Sculpture in Ireland* (<http://www.gothicpast.com/>). Compararemos su estructura con la conceptualización asociada a subconjuntos notables de WordNet (<http://wordnet.princeton.edu/>).

Bajo la denominada *hipótesis de la conceptualización libre de escala* (Aranda Corral *et al*., 2012a), se estima la potencia de los etiquetados. También se presentarán técnicas automatizadas para enriquecer el etiquetado mediante sugerencias basadas en razonamiento automático adaptación de las utilizadas en *Delicious* (Aranda Corral *et al*., 2012b).

Palabras clave: Repositorios digitales, semántica emergente, análisis formal de conceptos, etiquetación de contenidos, humanidades digitales.

ABSTRACT

Formal Concept Analysis (FCA) is a branch of Applied Mathematics whose aim is to discover, extract and organize knowledge from data. FCA provides techniques and tools to reason with such knowledge.

The aim of this paper is to analyse the applicability of FCA to cultural digital repositories that use tagging. It is possible to study the structure of concepts implicit in the

tags by means of FCA, in such way that it is possible to detect some patterns allowing the estimation of the soundness of these tag set.

Among other examples, we apply results from (Aranda Corral *et al*., 2012a) to analyse two repositories about art: *Baroque Art* from *CulturePlex* Lab (http://baroqueart.cultureplex.ca/) and the Visual *Archive of Gothic Architecture and Sculpture in Ireland* (http://www.gothicpast.com/). The conceptual structure extracted from both repositories is compared with the one from WordNet (http://wordnet.princeton.edu/), which is an example of successful semantic representation.

Under the called *Scale-Free Conceptualization hypothesis* (SFCH) (Aranda Corral *et al*., 2012a), the soundness of tagging sets (folksonomies) is estimated.

Keywords: Digital Repositories, Emergent Semantics, Formal Concept Analysis, Content tagging, Digital Humanities.