D

Resumen (Spanish Summary)

"El experimentador que no sabe lo que está buscando no comprenderá lo que encuentra."
— Claude Bernard

Utilización de Folksonomías para Clasificación de Recursos

En esta tesis abordamos el problema de la clasificación automática de recursos, una tarea cada vez más frecuente e importante en nuestra vida diaria. El catalogado de libros o la organización de vídeos, entre otros, representan algunos ejemplos de actividades para las que un proceso automático de clasificación resulta cada vez más frecuente, necesario e importante. Aprovechamos la información contenida en las anotaciones que realizan los usuarios de sistemas de etiquetado social, en los cuales se recogen numerosos metadatos que detallan el contenido de diferentes tipos de recursos. Hasta el momento, son pocos los trabajos que han explotado estos metadatos con este fin, y los pocos que lo han hecho se han limitado a realizar análisis estadísticos. En esta tesis exploramos las características de estos sistemas, de los usuarios involucrados en ellos, así como de las anotaciones que aportan, con el fin de sacar el máximo partido a estas grandes colecciones, obteniendo un rendimiento lo más preciso posible de los clasificadores automáticos de recursos.

D.1 Motivación

Organizar todo tipo de recursos sobre categorías predefinidas es una tarea común en nuestra vida diaria. Tener categorías asignadas a recursos ayuda a mejorar el

posterior acceso pudiendo limitarse a la(s) categoría(s) deseada(s). Por ejemplo, los bibliotecarios suelen catalogar sus libros por temas de forma que los organizan por intereses similares. Las bases de datos de películas, los catálogos de música y los sistemas de ficheros, entre otros, suelen estar organizados también por categorías de forma que facilita su acceso en el futuro. Asimismo, directorios web como Yahoo! Directory y Open Directory Project organizan páginas web en categorías. La clasificación de páginas web es de gran interés para mejorar los resultados provistos por los motores de búsqueda, ya que ayudan a reducir el ámbito a la categoría deseada por el usuario.

El problema está en que la clasificación manual de estos recursos es muy costosa y cara cuando la colección es grande. Por ejemplo, el Library of Congress de Estados Unidos informó en 2002 de que el coste medio de catalogar cada registro bibliográfico por profesionales fue de 94,58 dólares[1]. Catalogar 291.749 registros, como hicieron en aquel año, les llegó a costar unos 27 millones y medio de dólares. Dado lo cara que resulta la tarea, la utilización de clasificadores automáticos puede ser una buena alternativa para reducir su coste, y asimismo mantener los catálogos al día con un esfuerzo manual menor.

Hasta el momento, la mayoría de los clasificadores automáticos se han basado en el contenido de los recursos para representarlos, sobre todo en tareas de clasificación de páginas web (Qi and Davison (2009)). No obstante, la falta de datos representativos en el contenido de los recursos hace que se complique la tarea. Además, puede resultar muy complicado obtener suficientes datos sobre otros tipos de recursos como libros o películas, para los cuales puede ser más difícil representar el contenido, o incluso puede que el contenido no esté disponible de forma que pueda ser procesado.

Como solución a este problema, los sistemas de etiquetado social proveen una forma más sencilla y barata de obtener metadatos sobre los recursos. Sistemas como Delicious[2], LibraryThing[3] y GoodReads[4] recopilan anotaciones de usuarios en forma de etiquetas para grandes colecciones de recursos. Estas etiquetas provistas por usuarios dan lugar a datos significativos que describen el contenido de los recursos (Heymann et al., 2008).

Por medio de estas etiquetas, los usuarios proveen una especie de organización propia de los recursos. Estas etiquetas se comparten de forma social con la comunidad. Gracias a que un gran número de usuarios contribuye en estos sistemas, las anotaciones se acumulan sobre cada recurso. Por lo tanto, esa acumulación hace que cada una de las anotaciones sea más útil. Así, la acumulación de usuarios en una comunidad activa genera un gran número de marcadores,

[1] http://www.loc.gov/loc/lcib/0302/collections.html
[2] http://delicious.com
[3] http://www.librarything.com
[4] http://www.goodreads.com

etiquetas, y por tanto, recursos anotados.

> *"Cada una de las categorizaciones individuales vale menos que la categorización de un profesional. Pero hay muchas, muchas de aquéllas."*, Joshua Schachter, fundador de Delicious, en la cumbre FOWA 2006 FOWA en Londres (Inglaterra)[5].

Los sistemas de etiquetado social son un medio para guardar, organizar y buscar recursos, todo ello anotando con etiquetas escogidas por el usuario. Creemos que estas grandes colecciones de anotaciones pueden mejorar de forma considerable la tarea de clasificación de recursos. Las anotaciones provistas por usuarios podrían ser útiles como fuente de datos que aporta información significativa que podría ayudar a inferir la categorización de los recursos.

Dado que un gran número de usuarios provee sus propias anotaciones sobre cada recurso, nuestro objetivo se centra en descubrir la manera de amalgamar esas aportaciones en busca de una organización que se parezca a la categorización realizada por profesionales. En este contexto, donde los usuarios aportan grandes cantidades de metadatos, nuestro reto se centra en sacar el máximo partido de ellos con el fin de mejorar el rendimiento de la tarea de clasificación de recursos.

> *"Estamos en una época en la que los datos son baratos, pero sacar partido de ellos no lo es"*, Danah Boyd, Investigadora sobre Social Media en Microsoft Research New England, en el congreso WWW2010 en Raleigh, Carolina del Norte, Estados Unidos[6].

D.1.1 Clasificación de Recursos

La clasificación de recursos se puede definir como la tarea que consiste en organizar recursos en un conjunto de categorías predefinidas. En este trabajo, utilizamos las Máquinas de Vectores de Soporte (SVM, Joachims (1998)), un método vanguardista para clasificación. Este tipo de tareas de clasificación se basan en un conjunto de instancias previamente categorizadas, con lo que se alimenta el clasificador para que adquiera el conocimiento necesario para clasificar nuevos recursos.

Un problema de clasificación de recursos puede basarse en diferentes características. Por una parte, en lo que se refiere al método de aprendizaje, puede ser *supervisado*, donde todo el conjunto de entrenamiento esta previamente categorizado, o *semisupervisado*, donde también se aprovechan instancias sin información de categoría en la fase de aprendizaje. Por otra parte, en cuanto al número de clases, la clasificación puede ser *binaria*, cuando sólo hay dos categorías que se

[5] http://simonwillison.net/2006/Feb/8/summit/
[6] http://www.danah.org/papers/talks/2010/WWW2010.html

pueden asignar a cada recurso, o *multiclase*, cuando hay tres o más categorías. El primero se utiliza habitualmente para sistemas de filtrado, mientras que el segundo suele ser frecuente en el caso de taxonomías mayores, como en el caso de la clasificación temática de recursos.

Para clasificación temática sobre grandes colecciones de recursos, como páginas web sobre la Web, o libros en bibliotecas, las taxonomías suelen estar definidas por más de dos categorías, y el subconjunto de recursos previamente categorizada suele ser muy pequeño. De esta manera, creemos que se debería considerar y analizar la aplicación de técnicas semisupervisadas y multiclase para este tipo de tareas.

En esta tesis, proponemos el análisis de varias técnicas de clasificación que utilizan SVM, con el fin de analizar su adecuación a estas tareas. Estas técnicas incluyen diferentes aproximaciones a la resolución de tareas multiclase, así como algoritmos supervisados y semisupervisados.

D.1.2 Anotaciones Sociales

Los sistemas de etiquetado social permiten a los usuarios guardar y anotar sus recursos favoritos (como por ejemplo páginas web, películas, libros, fotos o música), compartiéndolos a su vez con la comunidad. Los usuarios proveen estas anotaciones normalmente en forma de etiquetas. Se conoce como etiquetado a la forma abierta de asignar etiquetas o palabras clave a recursos, de manera que se pueden describir y organizar. Esto posibilita la posterior recuperación de los recursos de forma más sencilla, aprovechando las etiquetas como metadatos que los describen. Normalmente, no hay etiquetas predefinidas, y por lo tanto los usuarios pueden escoger libremente las palabras que deseen como etiquetas.

> *"El etiquetado es principalmente una interfaz de usuario - una manera para que la gente recuerde cosas, en qué estaban pensando en el momento en el que lo guardaron. Bastante útil para recordar, bueno para el descubrimiento, terrible para la distribución (donde los que lo publican añaden tantas etiquetas como pueden para incluirlo en el mayor número posible de cajas)."*, Joshua Schachter, fundador de Delicious, en la cumbre FOWA 2006 FOWA en Londres (Inglaterra)[7].

Mediante este proceso, se genera en los sistemas de etiquetado social una estructura de etiquetas conocida como folksonomía, es decir, una organización de recursos dirigida por usuarios. Folksonomía es una contracción de las palabras *folk* (gente), *taxis* (clasificación) y *nomos* (gestión). Es conocida también como una taxonomía basada en los usuarios, en la cual la estructura no es jerárquica, al contrario que una clasificación taxonómica básica. Por lo tanto, una folksonomía

[7] http://simonwillison.net/2006/Feb/8/summit/

tiene cierta relación con las taxonomías generadas por expertos, en cuanto a que los recursos se organizan igualmente en grupos.

Se dice que estas anotaciones pertenecen a un entorno social cuando están accesibles y utilizables para cualquier usuario. Esta característica posibilita la búsqueda de recursos aprovechando las anotaciones provistas por otros. A su vez, es uno de los motivos que anima a los usuarios a contribuir.

No obstante, no todas las anotaciones se comparten de la misma manera. El propio sistema de etiquetado social puede definir algunas restricciones en este aspecto, principalmente estableciendo quién tiene permiso de anotar cada recurso. En este sentido, se pueden distinguir dos tipos de sistemas (Smith, 2008):

- **Sistemas de etiquetado simple:** los usuarios pueden describir sus propios recursos, así como fotos en Flickr[8], noticias en Digg[9] o vídeos en Youtube[10], pero nadie anota los recursos de otros. Generalmente, el autor del recurso es quien lo anota. Esto significa que no más de un usuario puede etiquetar cada recurso. En un sistema de etiquetado simple, hay un conjunto de usuarios (U) que anota unos recursos (R) con unas etiquetas (T). Cada usuario $u_i \in U$ puede guardar un recurso $r_j \in R$ con un conjunto de etiquetas $T_j = \{t_{j1}, ..., t_{jp}\}$, con un número p variable de etiquetas. El conjunto de etiquetas asignado a r_j seguirá estando limitado a T_j, ya que nadie más lo podrá anotar.

- **Sistemas de etiquetado colaborativo:** muchos usuarios pueden anotar cada recurso, y todos ellos pueden etiquetarlo con las etiquetas de su propio vocabulario. El conjunto de etiquetas asignado por un usuario genera una folksonomía a menor escala, conocida como personomía. Como resultado, varios usuarios tienden a anotar el mismo recurso. Por ejemplo, CiteULike.org, LibraryThing.com y Delicious se basan en anotaciones colaborativas, donde cada recurso (artículos, libros y URLs, respectivamente) puede ser anotado y etiquetado por todos aquellos usuarios que lo consideren interesante. Los sistemas de etiquetado colaborativo son algo más complejos, donde hay un conjunto de usuarios (U) que guarda sus marcadores (B) sobre unos recursos (R) anotándolos con unas etiquetas (T). Cada usuario $u_i \in U$ puede guardar un marcador $b_{ij} \in B$ de un recurso $r_j \in R$ con un conjunto de etiquetas $T_{ij} = \{t_{ij1}, ..., t_{ijp}\}$, con un número p variable de etiquetas. Después de que k usuarios guardan r_j, se describe como un conjunto pesado de etiquetas $T_j = \{w_{j1}t_{j1}, ..., w_{jn}t_{jn}\}$, donde $w_{j1}, ..., w_{jn} \leq k$ representan el número de asignaciones de cada etiqueta. Por lo tanto, cada marcador está compuesto por la tripleta de un usuario,

[8] http://www.flickr.com
[9] http://digg.com
[10] http://www.youtube.com

un recurso y un conjunto de etiquetas: $b_{ij} : u_i \times r_j \times T_{ij}$. Así, cada usuario guarda marcadores de diferentes recursos, y cada recurso tiene marcadores correspondientes a diferentes usuarios. El resultado de acumular etiquetas contenidas en los marcadores de un usuario se conoce como la personomía de ese usuario: $T_i = \{w_{i1}t_{i1}, ..., w_{im}t_{im}\}$, donde m es el número de etiquetas diferentes en la personomía del usuario.

En esta tesis nos enfocamos en los sistemas de etiquetado colaborativo. La Figura D.1 muestra un ejemplo comparativo de ambos tipos de sistemas.

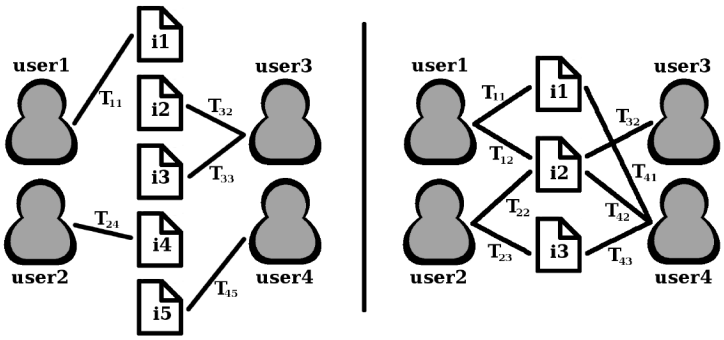

Figura D.1: Comparación de anotaciones provistas por usuarios en sistemas de etiquetado simple y colaborativo.

Generalmente, las etiquetas sobre un recurso tienden a coincidir entre usuarios, haciendo de las etiquetas acumuladas en sistemas colaborativos especialmente útiles en comparación con las de sistemas simples.

En un sistema de etiquetado colaborativo, como ejemplo, un usuario podría etiquetar este trabajo como `etiquetado-social`, `investigación` y `tesis`, mientras que otro usuario podría utilizar las etiquetas `etiquetado-social`, `marcadores-sociales`, `doctorado` y `tesis` para anotarlo. El comportamiento de los usuarios puede diferir de forma considerable en estos sistemas, donde la acumulación de sus anotaciones se suele considerar como consenso. Por ejemplo, el resultado de la acumulación mediante suma de las anotaciones de arriba sería el siguiente: `tesis` (2), `etiquetado-social` (2), `marcadores-sociales` (1), `doctorado` (1) e `investigación` (1).

que guarda sus marcadores (B) sobre unos recursos (R) anotándolos con unas etiquetas (T). Cada usuario $u_i \in U$ puede guardar un marcador $b_{ij} \in B$ de un recurso $r_j \in R$ con un conjunto de etiquetas $T_{ij} = \{t_1, ..., t_p\}$, con un número p variable de etiquetas. Después de que k usuarios guardan r_j, se describe como un conjunto pesado de etiquetas $T_j = \{w_1t_1, ..., w_nt_n\}$, donde $w_1, ..., w_n \leq k$ representan el número de asignaciones de cada etiqueta. Por lo tanto, cada marcador está

compuesta por la tripleta de un usuario, un recurso y un conjunto de etiquetas: $b_{ij} : u_i \times r_j \times T_{ij}$. Así, cada usuario guarda marcadores de diferentes recursos, y cada recurso tiene marcadores correspondientes a diferentes usuarios. El resultado de acumular etiquetas contenidas en los marcadores de un usuario se conoce como la personomía de ese usuario: $T_i = \{w_{i1}t_{i1}, ..., w_{im}t_{im}\}$, donde m es el número de etiquetas diferentes en la personomía del usuario.

En esta tesis, analizamos y estudiamos las anotaciones provistas por usuarios en sistemas de etiquetas sociales. Presentamos un estudio con el fin de sacar el máximo partido de ellos con vistas a mejorar el rendimiento de una tarea de clasificación de recursos. Concretamente, nos centramos en el análisis de la utilidad de las folksonomías generadas por usuarios como aproximación a una organización parecida a las taxonomías creadas por expertos. En este contexto, estudiamos varias representaciones de anotaciones sociales, en busca de una aproximación que se parezca a la clasificación provista por expertos en la mayor media posible. Nos centramos en obtener el máximo de las etiquetas sociales, tanto buscando la mejor representación, como midiendo el impacto que puede tener en este sentido la distribución de las etiquetas sobre recursos, marcadores y usuarios. Finalmente, también estudiamos la aplicación de técnicas vanguardistas de análisis del comportamiento de los usuarios en estos sistemas, con el fin de detectar usuarios cuyas anotaciones estén más próximas a la clasificación creada por expertos.

D.2 Objetivos

El objetivo principal de esta tesis se centra en aportar nuevo conocimiento sobre el uso apropiado de la gran cantidad de datos que se pueden encontrar en los sistemas de etiquetado social. Dado el interés en clasificar recursos, y la falta de datos representativos, nos enfocamos en analizar en qué medida y de qué manera las etiquetas sociales pueden mejorar la tarea de clasificación de recursos. Al comienzo de este trabajo, vimos que no había trabajos que abordaran este problema. Por lo tanto, nos motivó a llevar a cabo este trabajo de investigación. Hacia este fin, hemos definido el siguiente planteamiento, el cual resume el objetivo principal de esta tesis:

Planteamiento del Problema

¿Cómo se pueden aprovechar las anotaciones provistas por usuarios en sistemas de etiquetado social de forma que se obtenga una clasificación de recursos más precisa?

D.3 Metodología

La metodología de investigación seguida a lo largo del trabajo se compone de las siguientes 6 partes:

1. Revisión y lectura del estado del arte, así como estudiar y comprender detalladamente el funcionamiento de los sistemas de etiquetado social.
2. Búsqueda de un clasificador SVM apropiado para llevar a cabo el trabajo.
3. Búsqueda de colecciones existentes extraídas de sistemas de etiquetado social. Como no encontramos ninguno que cumpliera nuestros requisitos, creamos tres colecciones de gran escala en su lugar.
4. Pensar y proponer aproximaciones que se ajusten a la tarea de clasificación basada en etiquetas sociales.
5. Evaluación de las aproximaciones propuestas.
6. Realización de un riguroso análisis de los resultados, con el fin de llegar a unas conclusiones sólidas.
7. Presentación de resultados parciales en congresos y talleres nacionales e internacionales, con el fin de obtener comentarios y sugerencias de otros investigadores.
8. Resumir en esta tesis la investigación, aportaciones, y conclusiones alcanzadas a lo largo de todo el trabajo.

Del paso 4 al 6, se realizó un proceso iterativo, realizándose dichos pasos de forma repetida varias veces.

D.4 Estructura de la Tesis

Esta tesis esta compuesta de 8 capítulos. A continuación resumimos brevemente el contenido de cada uno de estos capítulos:

Capítulo 1 en la página 21

Introducción

Presentamos la motivación para el estudio sobre el uso de anotaciones sociales para clasificación de recursos. Formalizamos el problema, y motivamos la necesidad de realizar dicho estudio.

Capítulo 2 en la página 33

Trabajo Relacionado

Ofrecemos un resumen de los trabajos previos en el campo de investigación. Resumimos los avances en campos relacionados, tanto en cuanto al uso de anotaciones sociales como en cuanto a la clasificación de recursos.

Capítulo 3 en la página 47

Máquinas de Vectores de Soporte para Clasificación a Gran Escala

Realizamos un estudio de diferentes aproximaciones SVM para resolver

el problema de la clasificación de grandes colecciones de recursos sobre taxonomías multiclase. Damos con la mejor aproximación SVM para estos casos, el cual utilizamos a lo largo del trabajo para realizar las tareas de clasificación.

Capítulo 4 en la página 59

Creación de Colecciones de Etiquetado Social

Describimos y analizamos en detalle las colecciones de etiquetado social que creamos para utilizar en este trabajo. Detallamos el proceso de generación de dichas colecciones, y analizamos las principales características de sus correspondientes folksonomías.

Capítulo 5 en la página 75

Representando la Acumulación de Etiquetas

Proponemos y evaluamos diferentes representaciones de recursos utilizando etiquetas sociales para la tarea de clasificación de recursos. Estudiamos la utilidad de las etiquetas sociales en comparación a otras fuentes de datos, y proponemos la representación que saca el máximo partido de ellas. También abordamos el problema combinando las etiquetas sociales con las otras fuentes de datos para obtener un mejor rendimiento.

Capítulo 6 en la página 95

Analizando la Distribución de Etiquetas para Clasificación de Recursos

Abordamos la novedosa idea de considerar la representatividad de las etiquetas dentro de una colección de anotaciones de un sistema de etiquetado social. Estudiamos la aplicación de funciones de pesado adaptadas a estos sistemas, y analizamos su adecuación teniendo en cuenta las configuraciones de cada sistema.

Capítulo 7 en la página 111

Analizando el Comportamiento de Usuarios para la Clasificación

Exploramos el efecto del comportamiento de usuarios en sistemas de etiquetado social con vistas a la tarea de clasificación de recursos. Basándonos en trabajos previos que sugieren la existencia de ciertos usuarios que tienden a categorizar recursos, estudiamos si realmente se ajustan en mayor medida a la clasificación de recursos.

Capítulo 8 en la página 125

Conclusiones y Trabajo Futuro

Resumimos y comentamos las principales conclusiones y aportaciones del trabajo. Presentamos las respuestas a las preguntas de investigación formuladas al inicio, y planteamos el trabajo futuro.

Además, la tesis contiene los siguientes apéndices al final, con información adicional y resúmenes en otros idiomas:

D.5 Preguntas de Investigación Resueltas

Pregunta de Investigación 1

¿Qué tipo de clasificador SVM debería utilizarse para llevar a cabo este tipo de tareas de clasificación: un clasificador multiclase nativo, o una combinación de clasificadores binarios?

Se ha demostrado una clara superioridad de los clasificadores SVM multiclase nativos sobre las otras aproximaciones que combinan clasificadores binarios. Los resultados muestran que basarse en un conjunto de clasificadores binarios no es una buena opción cuando se trata de taxonomías multiclase. Por lo tanto, los clasificadores multiclase nativos, los cuales consideran todas las clases al mismo tiempo y tienen más conocimiento de la tarea completa, funcionan mejor para estos casos.

Pregunta de Investigación 2

¿Qué método de aprendizaje rinde mejor para este tipo de tareas de clasificación: uno supervisado o uno semisupervisado?

Los métodos semisupervisados podrían rendir mejor cuando el subconjunto etiquetado es muy pequeño, pero los métodos supervisados, computacionalmente menos costosos, consiguen un rendimiento muy similar con unas pocas instancias más etiquetadas. Por lo tanto, hemos mostrado también que a diferencia de las tareas de clasificación binarias como ya demostró Joachims (1999), un método supervisado obtiene unos resultados muy similares a los de un semisupervisado para estos casos. Parece razonable pensar que predecir la clase de las instancias no etiquetadas es mucho más difícil con el incremento del número de clases, y por tanto el incremento de errores en las predicciones se refleja también en la fase de aprendizaje del clasificador.

Por lo tanto, basándonos en estas conclusiones, decidimos utilizar un clasificador SVM multiclase supervisado a lo largo de esta tesis.

Pregunta de Investigación 3

¿Cómo afecta la configuración de los sistemas de etiquetado social en las anotaciones de los usuarios y las folksonomías resultantes?

Con este fin, hemos analizado diversas características que se encuentran en la configuración de los sistemas de etiquetado social. Entre las características analizadas, hemos mostrado el gran impacto de las sugerencias de etiquetas, lo cual altera de forma considerable la folksonomía resultante. En los sistemas de etiquetado social que hemos estudiado, todos presentan alguna característica diferente en este aspecto:

- **Sugerencias basadas en recursos (Delicious):** cuando el sistema sugiere etiquetas asignadas por otros usuarios al recurso que se está guardando, reduce la probabilidad de utilizar nuevas etiquetas que aporten nueva información. En este caso, los usuarios dedican poco esfuerzo a pensar por ellos mismos, y prefieren basarse en las sugerencias provistas por el sistema.
- **Sugerencias basadas en la personomía (GoodReads):** cuando el sistema sugiere etiquetas que el mismo usuario ha utilizado previamente, el vocabulario de su personomía tiende a ser mucho más reducido. No obstante, los usuarios no saben qué es lo que otros han anotado sobre cada recurso, y por tanto es muy probable que aporten nuevas etiquetas que anteriormente no se habían anotado sobre el recurso.
- **Ausencia de sugerencias (LibraryThing):** cuando el sistema no sugiere etiquetas al usuario, el vocabulario de su personomía tiende a ser mayor, así como las etiquetas asignadas a cada recurso son más diversas.

Pregunta de Investigación 4

¿Cuál es la mejor manera de acumular las anotaciones de los usuarios sobre un recurso con el fin de obtener una representación?

Hemos demostrado que es mejor tener en cuenta todas las etiquetas anotadas sobre un recurso que basarse sólo en aquéllas que han sido anotadas por más usuarios. Las etiquetas más anotadas han demostrado ser las más importantes, y aportan la información más importante sobre la temática del recurso. No obstante, las etiquetas menos populares también pueden ser útiles en menor medida, aportando información útil que mejora el rendimiento del clasificador.

En cuanto a los pesos que se asignan a las etiquetas al representar el recurso, los mejores resultados se obtienen considerando el número de usuarios que anotan cada etiqueta. El uso de este valor ha producido los mejores resultados en nuestros experimentos. Ha superado a otras aproximaciones que ignoran estos pesos, y ha demostrado que no hace falta considerar el número total de usuarios que anota el recurso.

Por lo tanto, la mejor representación que concluimos de nuestros experimentos es aquélla que aprovecha todas las etiquetas, asignando como peso el número de usuarios que las anota.

Pregunta de Investigación 5

A pesar de la utilidad de las etiquetas sociales para estas tareas, ¿merece la pena considerar otras fuentes de datos como el contenido de los recursos para mejorar aún más los resultados?

Utilizando técnicas de unión de clasificadores, los cuales combinan las predicciones de diferentes clasificadores, hemos demostrado que las etiquetas aportan criterios fiables a tener en cuenta. Estos criterios son muy útiles para combinar dichas etiquetas con otras fuentes de datos. No obstante, no todas las fuentes de datos son útiles para combinar, y se debe seleccionar con cautela las que obtienen unos resultados sólidos, y además ofrecen unas predicciones fiables. Cuando las fuentes de datos se escogen de manera apropiada, la mejora de rendimiento es considerable.

Pregunta de Investigación 6

¿Son las etiquetas sociales también útiles y suficientemente específicas para clasificar recursos en categorías a nivel más bajo?

Hemos analizado la utilidad de las etiquetas sociales para la clasificación sobre dos niveles diferentes de las taxonomías. Además de las categorías de más alto nivel, también hemos explorado la clasificación sobre categorías más precisas del segundo nivel. En este aspecto, las etiquetas sociales han sido superiores a otras fuentes de datos para aquellos sistemas de etiquetado social que animan a los usuarios a aportar anotaciones (Delicious y LibraryThing). La superioridad es muy clara en estos casos, sobre todo para Delicious, donde la diferencia es aún mayor cuando se trata del segundo nivel. Esta diferencia es muy similar para LibraryThing. Por último, las etiquetas de GoodReads no superan a las otras

fuentes de datos, ni siquiera para el primer nivel, ya que el sistema no anima a los usuarios a anotar los libros, con lo que muchos de los marcadores se quedan sin etiquetas.

Estos descubrimientos proveen una conclusión diferente a la que dan Noll and Meinel (2008a), donde los autores lanzan la hipótesis de que las etiquetas sociales podrían no ser útiles para niveles más bajos de las taxonomías, y que deberían utilizarse otros tipos de datos para estos casos.

Pregunta de Investigación 7

¿Podemos tener en cuenta la distribución de etiquetas a lo largo de la colección para así medir la representatividad general de la etiqueta?

A través de la experimentación llevada a cabo, hemos demostrado la utilidad de considerar las distribuciones de etiquetas por medio de una función de pesado inversa como la ofrecida por IDF. Estas funciones han servido para determinar la representatividad de las etiquetas para cada colección con el fin de mejorar el rendimiento de la tarea de clasificación de recursos. No obstante, hemos mostrado que la configuración del sistema de etiquetado social tiene mucho que ver con esas distribuciones. Entre las características en la configuración de los sistemas de etiquetado social, se ha visto que las sugerencias basadas en los recursos influyen en gran medida la estructura de las folksonomías resultantes. Aquellos sistemas que sugieren etiquetas al usuario basándose en anotaciones previas sobre el recurso producen unas distribuciones de etiquetas muy diferentes a aquéllos que no sugieren etiquetas y dejan a los usuarios que hagan su propia elección. Esta característica ha sido determinante también para la aplicación con éxito de las funciones de pesado sobre estas distribuciones.

Hemos descubierto que las funciones de pesado de etiquetas superan claramente a la aproximación basada en TF cuando el sistema no sugiere etiquetas basadas en los recursos (es decir, en LibraryThing y GoodReads), tanto cuando se utilizan por sí solas como cuando se combina con otras fuentes de datos. En realidad, es mejor considerar simplemente la aproximación basada en etiquetas que combinarla con otras fuentes de datos, ya que por sí sola ofrece los mejores resultados, los cuales no son mejorados cuando se combinan.

No obstante, cuando el sistema sugiere etiquetas basadas en el recurso, las folksonomías generadas son muy diferentes al resto. Esto afecta a las distribuciones de etiquetas en gran medida y, por lo tanto, a las funciones de pesado que hemos estudiado. Debido a ello, el uso de funciones de pesado de etiquetas obtiene peores resultados que no tenerlos en cuenta, y necesitan ser combinadas con otras fuentes de datos para funcionar mejor. En este último caso, pueden llegar a mejorar a la aproximación basada en TF, gracias a las buenas predicciones que aporta, que ayuda a alimentar de forma adecuada la combinación de clasificadores.

Pregunta de Investigación 8

¿Cuál es la mejor aproximación para establecer la representatividad de las etiquetas en la colección?

Entre las funciones de pesado que hemos estudiado, aquél que se basa en las frecuencias en marcadores ha demostrado ser la mejor para los sistemas sin sugerencias de etiquetas basadas en recursos. En estos casos, IBF es la mejor opción, seguida por IRF e IUF. Todos ellos superan con claridad a TF, tanto cuando se utilizan por sí solas como cuando se combinan con otras fuentes de datos.

Por otro lado, cuando el sistema sugiere etiquetas basadas en el recurso, es mejor basarse en la frecuencia en usuarios. IUF funciona mejor que IBF e IRF para estos casos, debido a la importancia de aquéllos usuarios que tienden a escoger sus propias etiquetas en lugar de basarse en las sugerencias. Aunque ni siquiera IUF supera a TF, cuando se combina con otras fuentes de datos llega a ser la mejor opción. No obstante, los resultados de este último son sólo ligeramente superiores a los obtenidos por la combinación que utiliza TF, por lo que cualquiera de ellas podría emplearse para llegar a unos resultados parecidos.

Pregunta de Investigación 9

¿Podemos discriminar diferentes perfiles de usuario de manera que encontremos un subconjunto de usuarios que proporciona anotaciones que se ajustan en mayor medida a la tarea de clasificación?

Hemos demostrado que dicho tipo de usuario, llamado Categorizador, en realidad existe. Según nuestros experimentos, estos es verdad sobre todo cuando se trata de sistemas sin sugerencias de etiquetas como en LibraryThing, donde la clasificación de recursos realizada utilizando etiquetas de los usuarios Categorizadores obtiene mejores resultados. Cuando las sugerencias existen, la detección de usuarios que se adecúan a la tarea se complica, como hemos demostrado que ocurre con GoodReads y Delicious. Sin embargo, la utilización de la medida apropiada puede producir una selección exitosa de usuarios que se ajustan a las características de un Categorizador.

Pregunta de Investigación 10

¿Cuáles son las características que identifican a un usuario como apropiado para la tarea de clasificación de recursos?

De las dos características que hemos considerado en este trabajo, hemos visto que si se diferencian los usuarios por su nivel de verbosidad, se puede encontrar un conjunto de usuarios que se ajustan más a la tarea de clasificación. Por otra parte, hemos visto que separando usuarios por la diversidad de su vocabulario no se consigue una buena discriminación para este fin, sino para encontrar otro tipo de usuarios llamados Descriptores. Además de esto, hemos visto que aquéllos

usuarios que no utilizan datos descriptivos en sus anotaciones ofrecen etiquetas que se ajustan mejor a la clasificación de recursos.

D.6 Principales Contribuciones

La idea novedosa de este trabajo de investigación se basa en la utilización de anotaciones sociales para enriquecer una tarea de clasificación de recursos. Hasta donde nosotros sabemos, el primer trabajo de investigación que llevó a cabo experimentos con tareas de clasificación reales fue nuestro primer trabajo en este campo (Zubiaga et al., 2009d). Previamente, sólo Noll and Meinel (2008a) habían realizado un análisis estadístico que comparaba etiquetas sociales con una clasificación hecha por expertos. Teniendo en cuenta la carencia de trabajos en el área, el trabajo recogido en esta tesis aporta nuevo conocimiento hacia el uso y modo de representación apropiados de etiquetas sociales para la clasificación de recursos. Concretamente, nuestras aportaciones principales al área de investigación son las siguientes:

- Hemos creado 3 colecciones de gran escala que incluyen tanto etiquetas sociales como información de la categoría correspondiente para una serie de recursos. Éstas pueden considerarse como unas de las mayores colecciones utilizadas en el área de investigación y, por lo que nosotros sabemos, las mayores utilizadas para clasificación de recursos. Algunas de estas colecciones, junto con otras más pequeñas que hemos creado a lo largo del trabajo, se han hecho públicas para fines de investigación[11]. Entre otros, Godoy and Amandi (2010) y Strohmaier et al. (2010b) han utilizado alguna de nuestras colecciones para su investigación.

- Nuestro trabajo es el primero que compara diferentes representaciones de etiquetas sociales. Además, es el primer trabajo que realiza tareas de clasificación comparando etiquetas sociales con otros tipos de fuentes de datos. Hemos demostrado que las etiquetas sociales son también útiles para categorías más precisas de más bajo nivel. Al contrario de lo que indican que Noll and Meinel (2008a), donde los autores realizan un estudio estadístico con el que concluyen que las etiquetas sociales podrían no ser útiles para categorías más precisas, hemos demostrado que son aún más útiles que para categorías más generales.

- Hemos analizado las distribuciones de etiquetas sociales en folksonomías, y hemos realizado un riguroso estudio de cómo la configuración de un sistema de etiquetado social afecta tales distribuciones. En este aspecto,

[11] http://nlp.uned.es/social-tagging/datasets/

hemos adaptado funciones de pesado basadas en la consolidada TF-IDF al ámbito del etiquetado social y las folksonomías.

- Hemos mostrado la existencia de un grupo de usuarios, llamados Categorizadores, cuyas anotaciones se parecen más que las de otro grupo de usuarios llamados Descriptores a la clasificación hecha por expertos. Aunque la aproximación para diferenciar Categorizadores y Descriptores ya estaba consolidada de previos trabajos, en éste hemos llevado a cabo la novedosa tarea de demostrar que los Categorizadores se ajustan más a la clasificación de recursos.

La utilización de anotaciones sociales para el beneficio de tareas de clasificación de recursos era una línea de investigación nueva al comienzo de esta tesis. Sin embargo, el crecimiento de interés de los investigadores sobre contenidos generados por usuarios en medios de comunicación social, y concretamente en los sistemas de etiquetado social, ha ocasionado recientemente la aparición de numerosos trabajos en el área. Junto con este crecimiento, más investigadores han mostrado su interés en utilizar anotaciones sociales para clasificación de recursos, y el número de trabajos relacionados ha aumentado considerablemente. Godoy and Amandi (2010), por ejemplo, presentan un estudio de clasificación basada en etiquetas que se inspira en un trabajo nuestro (Zubiaga et al., 2009d).

D.7 Trabajo Futuro

La utilización de anotaciones sociales para la clasificación de recursos es un campo de investigación que está aún en sus inicios, y se ha realizado relativamente poco trabajo hasta el momento. El trabajo presentado en esta tesis concluye con la manera de representar etiquetas sociales en busca de una clasificación de recursos lo más precisa posible. Además, da lugar a diversos trabajos futuros.

A lo largo de esta tesis, hemos considerado cada etiqueta como un símbolo diferente, sin tener en cuenta su significado semántico. En este aspecto, nuestros planes para trabajo futuro incluyen el análisis del significado de las etiquetas para tratar de descubrir palabras sinónimas y relaciones entre ellas. Bien utilizando técnicas de procesamiento de lenguaje natural, o bien mediante aproximaciones semánticas, esto podría ayudar a entender el significado de cada etiqueta, pudiendo explorar más allá el conocimiento que aportan las folksonomías.

Las tres funciones de pesado que hemos empleado en el Capítulo 6 se basan en la conocida TF-IDF, que fue diseñada inicialmente para colecciones de texto. Pensamos que probar otras funciones de pesado, así como explorar la posible definición de una nueva función que se ajuste a las necesidades de estas estructuras sociales, pueden resultar en interesantes aportaciones como trabajo futuro. Esto ayudaría sobre todo para sistemas que dan sugerencias de etiquetas basadas en

recursos, como pasa con Delicious, donde las funciones de pesado que hemos experimentado no han dado buenos resultados.