Evaluating Term Weighting Schemes for Content-based Tag Recommendation in Social Tagging Systems

E. P. Olvera and D. Godoy

***Abstract*— Social tagging systems allow users to publish different type of resources, such as Web pages or pictures, annotate them using keywords or tags and share their resources with other users. These systems achieved widespread success on the Web on account of the simplicity for organizing resources using open-ended tags. Recently, tag recommendation strategies have been proposed to alleviate the problems of ambiguity, syntactic variations and noise in tags cause by the inherent characteristics of natural language. In this work we proposed a content-based approach that generates a list of suggested tags for annotating a given resource starting from an analysis of its textual content exclusively. Thus, the proposed method can be used in situations in which there is not enough information for creating a tag-based user profile or compare the user with others. For extracting the more relevant words different term weighting approaches were evaluated, particularly considering the HTML structure of Web pages and the grammatical category of words in order to determine promising tag candidates. Experimental results of applying this technique to tag recommendation using several term weighting approaches are reported and compared.**

***Keywords*— Folksonomy, Social Tagging Systems, Term Weighting Schemes.**

I. Introducción

El etiquetado o tagging social es el proceso mediante el cual los usuarios pueden organizar un conjunto de recursos en la Web asignándoles palabras clave o términos conocidos como etiquetas o tags. Esta técnica de clasificación de información es una de las más populares en la actualidad en la Web y ha sido implementada exitosamente en distintos sitios, como Del.icio.us (http://del.icio.us/), Flickr (http://flickr.com/), CiteULike (http://www.citeulike.org/), entre otros, que permiten publicar, anotar y luego buscar páginas, publicaciones académicas, objetos multimedia, entre otros recursos en la Web.

La aparición de estos sitios sociales trajo consigo la transformación de los esquemas de clasificación tradicionales usados en la Web, como son las taxonomías, a esquemas de clasificación sociales, conocidos como folcsonomías [12]. La forma libre en que los usuarios anotan recursos, en contraposición con las estructuras rígidas propuestas por las taxonomías, es uno de los motivos fundamentales del éxito de estos sistemas, pero a su vez es una de sus principales debilidades. La falta de control sobre la terminología usada para realizar anotaciones causa muchas veces resultados poco confiables e inconsistentes en la indexación y búsqueda de contenido, inducidos por la ambigüedad, sinonimia, polisemia y falta de normalización lingüística de las etiquetas elegidas por los usuarios [5].

Los métodos de recomendación de etiquetas surgieron como una forma de aliviar estos problemas, ya que no solo reducen los esfuerzos de anotación de los usuarios, sino que permiten converger a un vocabulario común en la comunidad y facilitar así la búsqueda de contenido. Si bien existen diversas propuestas de recomendación de etiquetas en la literatura, algunas de ellas centradas en el comportamiento de otros usuarios, como las basadas en la aplicación de filtrado colaborativo a folcsonomías [9, 22, 20], algunos estudios sugieren que las etiquetas relevantes se encuentran en el contenido del recurso a etiquetar y dentro del conjunto de etiquetas previamente usadas por el propio usuario [6, 11].

En este trabajo se propone un método de recomendación de etiquetas basado en el análisis del contenido textual de los recursos y particularmente en la ponderación de la estructura HTML de las páginas Web y de las funciones gramaticales de las palabras. La recomendación en base al contenido de un recurso, es fundamental cuando se cuenta con poca información de las preferencias de etiquetado del usuario. Por ejemplo, cuando éste recién comienza a usar el sistema y aún no ha anotado contenido. En el enfoque propuesto se consideraron distintos esquemas de pesado para las palabras en el texto de un recurso, a fin de seleccionar las más descriptivas de su contenido, para ser recomendadas como posibles etiquetas.

El resto de este artículo se organiza como sigue. La Sección II discute trabajos relacionados con la recomendación de etiquetas en sistemas sociales. El método propuesto de recomendación basado en contenido, como así también las distintas funciones de pesado de términos que se evalúan, se discuten en la Sección III. Los resultados experimentales obtenidos con este enfoque se reportan en la Sección IV. Finalmente, las conclusiones alcanzadas se comentan en la Sección V.

II. Trabajos Relacionados

Los sistemas de etiquetado social se componen de un conjunto de triplas ⟨*usuario*, *recurso*, *etiqueta*⟩ conocidas como asignaciones de etiquetas. Formalmente, una folcsonomía se define como la tupla $\mathsf{F} := (U, T, R, Y, \prec)$

E. P. Olvera, Instituto de Informática de la Universidad Técnica Estatal de Quevedo, Quevedo, Los Ríos, Ecuador, eliaspo@uteq.edu.ec.

D. Godoy, ISISTAN Research Institute, Universidad Nacional del Centro de la Provincia de Buenos Aires and CONICET, Tandil, Buenos Aires, Argentina, daniela.godoy@isistan.unicen.edu.ar.

formada por el conjunto de los usuarios *U*, el de los recursos *R* y el de las etiquetas *T*, y las asignaciones de etiquetas a recursos dadas por una relación ternaria entre ellos $Y \subseteq U \times T \times R$ [7]. En esta folcsonomía, $\prec$ es un relación específica de subsunción entre las etiquetas de un usuario, $\prec \subseteq U \times T \times T$.

La personomía P_u de un usuario $u \in U$ es la restricción de F a u, es decir, $\mathsf{P}_u := (T_u, R_u, I_u, \prec_u)$ con $I_u := \{(t,r) \in T \times R | (u,t,r) \in Y\}$. Siendo $T_u := \pi_1(I_u)$ las etiquetas del usuario u, $R_u := \pi_2(I_u)$ los recursos del usuario, y $\prec_u := \{(t_1,t_2) \in T \times T | (u,t_1,t_2) \in \prec\}$ las relaciones entre etiquetas para ese usuario, donde π_i es la proyección de la i-ésima dimensión [7]. En otras palabras una personomía es la colección de recursos, etiquetas y asignaciones de etiquetas realizadas por un único usuario [9], mientras que la colección de personomías se denomina folcsonomía [25].

El problema de recomendación de etiquetas en sistemas sociales ha sido abordado en distintos trabajos, en muchos casos adaptando enfoques clásicos del área de sistemas de recomendación como los basados en contenido y filtrado colaborativo (FC) [1]. Los enfoques basados en FC hacen predicciones sobre una matriz bidimensional de opiniones de los usuarios respecto a un conjunto de ítems. Las folcsonomías, en cambio, tienen una composición tripartita de etiquetas, usuarios y recursos, por lo que inicialmente se usaron algoritmos clásicos sobre proyecciones de menor dimensionalidad de la matriz. Por ejemplo, los experimentos realizados en [9] con conjuntos de datos de BibSonomy (http://www.bibsonomy.org/) y Last.fm (http://www.last.fm/) mostraron que el enfoque de FC basado en usuarios supera las recomendaciones basadas en popularidad global. En [10] un esquema de FC se utiliza para la obtención de un conjunto candidato de etiquetas describiendo las preferencias latentes de los usuarios. En la creación de este conjunto se supone que el usuario prefiere utilizar etiquetas que usó antes, como así también etiquetas utilizadas por usuarios similares encontrados mediante un algoritmo de vecinos más cercanos. Nakamoto et al. [14] consideran diferentes contextos de etiquetado mediante la modificación del cálculo de similitud entre usuarios durante la formación del vecindario. Tso-Sutter et al. [22] proponen un mecanismo genérico para integrar las etiquetas a los algoritmos estándar de FC fusionando las predicciones basadas en usuario con las basadas en ítems sobre extensiones de las matrices que buscan capturar simultáneamente las correlaciones entre usuarios, ítems y etiquetas. En [16] se propone un método que considera las relaciones basadas en confianza entre los usuarios y en [26] se evalúan relaciones entre recursos. Otros enfoques usan descomposición en valores singulares de alto orden como CubRec [23], CubeSVD [19] y los tensores de tercer orden usados en [20]. Debido a que durante la anotación de recursos cada usuario define su propio conjunto de etiquetas que describen los recursos desde su punto de vista particular, una desventaja de este enfoque es que los usuarios que etiquetan recursos similares no necesariamente usan etiquetas parecidas y por lo tanto las etiquetas recomendadas pueden diferir en mucho de las que asignaría el usuario.

La construcción de perfiles de usuario basados en etiquetas, apareció como una alternativa para hacer recomendaciones más personalizadas y ajustadas a la realidad del usuario. Los perfiles más simples de este tipo están formados por un vector de etiquetas, donde el peso representa la importancia de cada etiqueta para el usuario en base a su frecuencia de uso. En [15] se usan estos perfiles para personalizar los resultados de una búsqueda en función de la similitud del vector de etiquetas con el contenido de las páginas Web y en [4] para buscar usuarios similares. El aprendizaje de perfiles más ricos, siempre basados en etiquetas, se propone en [24] con un algoritmo de clustering de grafos aplicado sobre la red de recursos del usuario, mientras que perfiles denotando relaciones (de co-ocurrencia, semánticas y otras) entre las etiquetas de interés del usuario se proponen en [13, 8]. El aprendizaje de perfiles de usuario requiere de un amplio historial de asignación de etiquetas a recursos a fin de lograr buenas recomendaciones. Durante la larga curva de aprendizaje de estos algoritmos es fundamental contar con un método alternativo para generar recomendaciones.

Los métodos basados en contenido ofrecen un mecanismo para obtener un conjunto de sugerencias en situaciones donde no hay suficiente información del usuario para obtener un perfil preciso o compararlo con otros usuarios del sistema. En [11] se propone un método de tres pasos. Primero, se extraen etiquetas del texto del recurso, luego se utilizan recursos léxicos para proponer etiquetas relacionadas que finalmente se comparan con la personomía del usuario. Las etiquetas más prometedoras del primero y tercer paso se recomiendan. Zhang et al. [25] combinan la extracción de términos en base a un modelo de lenguaje, que obtiene etiquetas principalmente del título del recurso, con un enfoque que utiliza un modelo conceptual de los tópicos tratados en el recurso y las etiquetas más frecuentemente asociadas a ellos en la folcsonomía. El sistema SparTag.us [6] sugiere anotaciones a partir del texto que es resaltado por el usuario y lo vuelca en una libreta electrónica que luego puede ser navegada. A diferencia de estos trabajos, el enfoque propuesto en este artículo consiste en analizar distintas formas de valorar los términos en el contenido de un recurso y ponderar tanto los elementos HTML de las páginas como las categorías gramaticales de las palabras para proveer recomendaciones aún cuando el usuario sea nuevo en el sistema o se deseen anotar recursos con etiquetas aún no empleadas por el usuario.

III. Método de Recomendación de Etiquetas

El método de recomendación que se propone se basa en la extracción de los términos más importantes del texto de un recurso o página Web y en la ponderación de tales términos según el elemento HTML en que se encuentren, así como en

la ponderación de las funciones gramaticales que cumplen en el texto. El primer paso para la recomendación basada en contenido es obtener una representación del recurso, en este caso una página Web. Para ello se utiliza el Modelo de Espacio de Vectores [18], técnica largamente estudiada en el área de Recuperación de Información (RI). En este modelo, los documentos se representan mediante vectores en un espacio t-dimensional, siendo *t* el número de términos diferentes considerados relevantes.

Cada término en el vector representando un documento posee un valor numérico o peso indicando su importancia en la descripción del contenido del recurso. El peso de los términos se determina usualmente en función del número de ocurrencias en el documento y, posiblemente, en función de su frecuencia en el total de documentos de la colección. Diferentes esquemas de pesado de términos se basan en variaciones de estas funciones [17].

La representación resultante de una página Web o documento d_j es entonces equivalente a un vector t-dimensional:

$$d_j = \left(w_{1j}, w_{2j}, ..., w_{tj}\right) \tag{1}$$

donde w_{kj} representa el peso del k-ésimo término en el documento *j*

El conjunto de etiquetas candidatas para ser recomendadas se conforma por aquellos términos de mayor peso en el vector que representa un recurso. Los diferentes esquemas de pesado de términos que fueron evaluados en este trabajo a fin de determinar el más adecuado para la recomendación de etiquetas en sistemas de etiquetado social, se describen en las subsecciones III.*A* a III.*D*.

A. Frecuencia del Término

Este esquema de pesado de términos, denominado *tf* (Term Frequency), simplemente asigna un peso igual al número de ocurrencias del término t_k en un documento o recurso r_j y se denota $w_{kj} = tf\left(t_k, r_j\right)$. Así la lista de etiquetas a recomendar usando este esquema estará constituida por los términos con la frecuencia de aparición o w_{kj} más alta en el recurso actual.

B. Frecuencia del Término por su Frecuencia Inversa en los Documentos

La función *tf-idf* (term frequency x inverse document frequency) [17] es un esquema de pesado global. Mientras que el factor *tf*, definido anteriormente, mide la importancia de un término en un documento particular, el factor *idf* mide la importancia del mismo término en la totalidad de los documentos de la colección. Esta función formaliza dos observaciones empíricas: (1) cuanto más aparece un término en un documento, es más discriminante del tema del mismo; (2) cuanto más aparece el término en la colección entera de documentos, menor poder tendrá para discriminar contenidos.

Dado que la recomendación se efectuará para un usuario particular, la colección en este caso es el total de recursos en la personomía del usuario. El factor *tf* se calcula de la forma antes indicada, mientras que el factor idf se calcula mediante la siguiente fórmula:

$$idf\left(t_k\right) = \log\left(\frac{\left|R_u\right|}{1+\left|r_j : t_k \in r_j \wedge r_j \in R_u\right|}\right) \tag{2}$$

donde $\left|R_u\right|$ es el número de páginas Web o recursos del usuario u, r_j es un recurso en la personomía de este usuario, y el denominador de la fórmula corresponde al número de recursos dentro de tal personomía en donde el término t_k aparece al menos una vez.

Luego de obtener los factores $tf\left(t_k, r_j\right)$ e $idf\left(t_k\right)$ para cada término de la página Web a anotar, se multiplican obteniendo el peso $w_{kj} = tf\left(t_k, r_j\right) \cdot idf\left(t_k\right)$ para cada término k existente en el recurso r_j. Los pesos más altos corresponden a términos con alta frecuencia en una página Web, pero baja frecuencia en la personomía del usuario, dando prioridad a etiquetas más discriminantes del contenido desde el punto de vista del usuario.

C. Pesado por Componentes Estructurales

Los dos esquemas de pesado de términos antes descriptos utilizan una representación *bag-of-words* propia del área de RI que no explota la información provista por la estructura HTML de la página Web, la cual se considera que puede ayudar a mejorar la descripción del contenido de la página y consecuentemente la precisión en la recomendación de etiquetas.

La estructura de un archivo HTML permite identificar fácilmente las partes importantes de una página Web como el título o los encabezados de las secciones y se puede aprovechar para la recomendación de etiquetas. Por ejemplo, los términos en el título se pueden considerar más representativos del tema que trata una página Web que los que se encuentran simplemente dentro del cuerpo de la misma.

Una técnica de pesado orientada por la estructura [3] se utiliza para asignar pesos más altos a las palabras encerradas dentro de ciertos tags HTML que se consideran más representativos. Esta función se define como sigue:

$$w_{kj} = \sum_{e_i} weight\left(e_i\right) \cdot tf\left(t_k, r_j, e_i\right) \tag{3}$$

donde e_i es un elemento o componente estructural del HTML, $weight\left(e_i\right)$ es el peso asignado al elemento e_i y $tf\left(t_k, r_j, e_i\right)$ es el número de veces que el término t_k aparece

dentro del elemento estructural e_i en el recurso r_j.

La ponderación de términos en este esquema se realiza considerando la relevancia de los elementos HTML de las páginas Web, específicamente el título, los meta-datos, la URL, el cuerpo y los links. Los meta-datos se restringieron a aquellos con el atributo *name*. Aunque no existe ninguna especificación que defina los valores posibles para este atributo, existe cierto número de ellos que son comúnmente utilizados, como *description*, *keywords* y *robots*, mientras que otros son empleados esporádicamente, tal como se muestra en la Tabla I. Se consideraron en este trabajo los metadatos de nombre *description* y *keywords*, ya que están dedicados a la descripción del contenido del documento y aún cuando no son visibles al usuario poseen texto que puede ser una prolífica fuente de etiquetas.

Mediante este mecanismo cada palabra es ponderada en función de su frecuencia de aparición (*tf*) dentro de cada elemento estructural de la página Web para la cual se están recomendando etiquetas. Los componentes estructurales del HTML considerados en esta función son los siguientes:

- e_{title} las palabras en el título de la página Web
- e_{url} las palabras extraídas de la dirección URL
- e_{body} las palabras extraídas del cuerpo de la página
- e_{link} las palabras dentro del texto de los links
- $e_{meta-keywords}$ las palabra extraídas de los meta-datos de tipo *keywords*
- $e_{meta-description}$ las palabra extraídas de los meta-datos de tipo *description*

D. Pesado por Categorías Gramaticales y Componentes Estructurales

Otra característica importante a tener en cuenta es la categoría gramatical de los términos, concretamente si son verbos, adverbios, sustantivos o adjetivos. Los términos en ciertas categorías pueden considerarse mejores candidatos para recomendación, por ejemplo los sustantivos pueden privilegiarse por sobre los verbos y adverbios que cumplen mayormente una función de cohesión en el texto. En contraste, las palabras consideradas stopwords (normalmente los artículos, pronombres, preposiciones y verbos muy frecuentes) son aquellas que por su frecuencia y/o semántica no poseen valor discriminatorio alguno y no se consideran términos candidatos para recomendación.

Este esquema combina las categorías gramaticales de las palabras con su ubicación en la estructura HTML de la página como sigue:

$$w_{kj} = \sum_{g_i} weight(g_i) \cdot gram(t_k, g_i) + \sum_{e_i} weight(e_i) \cdot tf(t_k, r_j, e_i) \quad (4)$$

donde g_i es una categoría gramatical, $weight(g_i)$ es el peso asignado a la categoría g_i y $gram(t_k, g_i)$ es una función binaria cuyo valor es 1 si el término cumple dicha función gramatical en el texto o 0 en caso contrario. Un mismo término puede cumplir más de una función gramatical, por ejemplo el término "*first*" puede usarse como sustantivo, adjetivo o adverbio. Las funciones gramaticales de cada término se determinaron utilizando el diccionario *WordNet* 3.0 (http://wordnet.princeton.edu/).

TABLA I
PORCENTAGE PROMEDIO DE ASIGNACIONES ENLAS CUALES LA ETIQUETA ASIGNADA SE ENCUENTRA DENTRO DEL ELEMENTO ESTRUCTURAL.

ELEMENTO ESTRUCTURAL	PORCENTAJE PROMEDIO	DESVIACIÓN ESTÁNDAR
Cuerpo	54.77	11.27
Título	19.65	7.09
Texto de los links	36.98	8.54
URL	10.96	5.45
Meta-datos (todos)	24.63	7.05
Meta-datos de tipo *keywords*	28.48	8.38
Meta-datos de tipo *description*	20.07	6.85
Otros meta-datos	3.07	1.77

IV. RESULTADOS EXPERIMENTALES

A. Descripción del Conjunto de Datos

A fin de evaluar este enfoque se utilizó el conjunto de datos DAI-Labor Del.icio.us Corpus (http://www.dai-labor.de/), particularmente las instancias de etiquetado del mes de junio del 2007 en el sistema Del.icio.us. El conjunto original está en formato de texto plano y consta de 17,660,545 instancias de etiquetado únicas. Cada instancia posee los siguientes atributos: fecha, usuario (anonimizado), URL y etiqueta. Como resultado del análisis de este conjunto de datos se observó que entre estas intancias existen 694,066 etiquetas, 3,302,932 URLs y 331,472 usuarios únicos. De ellos se tomaron los 100 usuarios más prolíficos, las 1,850 páginas más anotadas por esos 100 usuarios y sus correspondientes etiquetas.

Entre los problemas asociados a la libertad con que los usuarios crean etiquetas están: (a) las variaciones en las mismas causadas por las modificaciones sintácticas [21, 2] y (b) el uso de stopwords como etiquetas, lo cual es permitido por la naturaleza libre del etiquetado social. Por tanto, para reducir el ruido en el conjunto de datos se utilizaron las stoplists provistas por SQL Server 2008, previamente depuradas de forma que no contengan stopwords utilizadas como etiquetas en el conjunto de datos, evitando así eliminar términos que podrían ser potenciales recomendaciones a pesar

de ser stopwords. Además se detectó que una gran cantidad de etiquetas estaban iniciadas y/o culminadas por múltiples caracteres de puntuación y/o especiales. Al extraer los caracteres de puntuación y especiales iniciales y finales, quedaron sólo palabras constituidas por letras y palabras con caracteres especiales y de puntuación intermedios (ej: www.google.com). Luego de reducir el ruido se obtuvieron 3,610 etiquetas y 67,370 asignaciones de etiquetas dentro de las personomías de los usuarios que abarcó el estudio.

A partir de las asignaciones de etiquetas originales, se calculó el porcentaje de veces en promedio que cada usuario usó etiquetas provenientes de los diferentes elementos HTML de las páginas Web. En la Tabla I se muestra el porcentaje promedio, y su desviación estándar, de asignaciones en las cuales la etiqueta usada está dentro de cada elemento HTML de las páginas anotadas. Estos valores muestran que un porcentaje importante, cerca de un 50%, de las etiquetas que usan los usuarios para anotar los recursos están efectivamente en el texto de los mismos. De manera que un método de recomendación basado sólo en el contenido del recurso podría obtener un número cercano a éste de etiquetas relevantes, aún sin contar con el historial de etiquetado del usuario.

B. *Métricas de Evaluación*

Para un usuario dado $u \in U$ y un recurso $r \in R$, un recomendador de etiquetas intenta encontrar un conjunto de etiquetas $\widetilde{T}(u,r) \subseteq T$ para que el usuario anote dicho recurso [7]. La metodología de evaluación consiste en recomendar *N* etiquetas para cada par usuario-recurso, *u* y *r* en la folcsonomía, que luego se comparan con las etiquetas que el usuario *u* realmente asignó al recurso *r*.

Entre las métricas de evaluación adoptadas para medir los aciertos de las recomendaciones se encuentran precisión, recall y su combinación en F-Measure, que se definen para este problema como sigue:

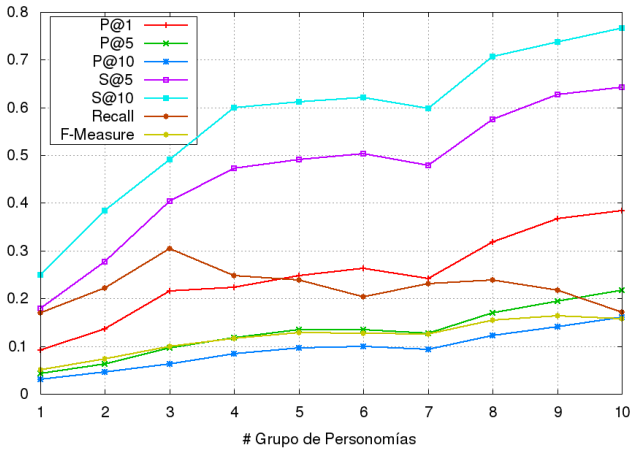
$$precision\left(\widetilde{T}(u,r)\right) = \frac{1}{\left|R_{test}\right|} \sum_{r \in R_{test}} \frac{\left|tags(u,r) \cap \widetilde{T}(u,r)\right|}{\left|\widetilde{T}(u,r)\right|} \quad (5)$$

$$recall\left(\widetilde{T}(u,r)\right) = \frac{1}{\left|R_{test}\right|} \sum_{r \in R_{test}} \frac{\left|tags(u,r) \cap \widetilde{T}(u,r)\right|}{\left|tags(u,r)\right|} \quad (6)$$

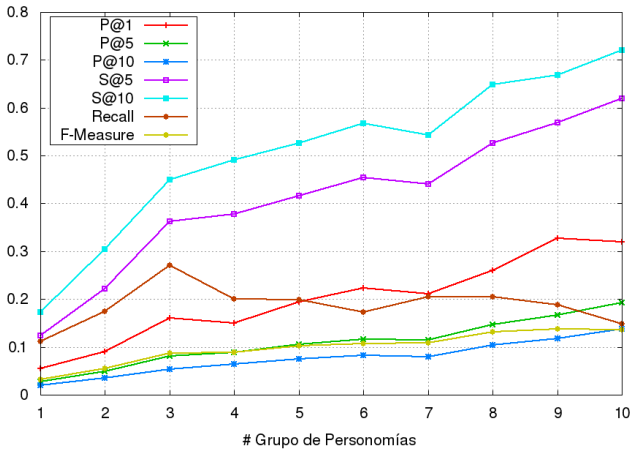
$$\mathrm{F-measure}\left(\widetilde{T}(u,r)\right) = \frac{2 \times precision \times recall}{precision + recall} \quad (7)$$

donde r es el recurso a ser anotado, $\widetilde{T}(u,r)$ es el conjunto de etiquetas recomendadas y $tags(u,r)$ el conjunto de etiquetas reales asignadas por el usuario al recurso. Es decir, precisión mide la cantidad de etiquetas recomendadas que fueron efectivamente usadas por el usuario para anotar el recurso y recall el número de etiquetas relevantes recomendadas sobre el total que debieron recomendarse.

Dado que las recomendaciones se presentan en una lista ordenada o ranking, estas medidas deben analizarse en distintos puntos del mismo. *P@k* se define como el porcentaje de etiquetas relevantes entre las *k* primeras recomendaciones. *P@*1, por ejemplo, es la cantidad de veces que la primera etiqueta recomendada fue relevante. Otra métrica utilizada para evaluar los resultados fue el éxito *S@k*, también considerado en distintas posiciones del ranking. El éxito en *k* se define como la probabilidad de encontrar una etiqueta relevante entre las *k* primeras recomendaciones, donde *S@*1 es equivalente a *P@*1 por definición.



(a) Frecuencia del Término (TF)



(b) Frecuencia del Término por la Frecuencia Inversa del Documento (TF-IDF)

Figura 1. Resultados de la recomendación de etiquetas utilizando los esquemas basados en frecuencia.

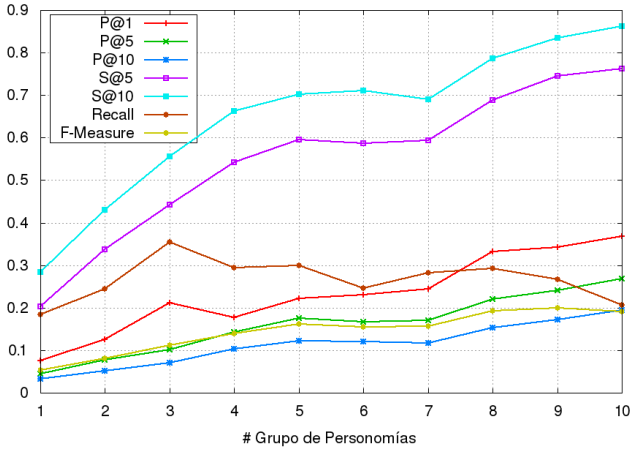
Los experimentos se desarrollaron de forma que para cada usuario $u \in U$ se utilizaron las páginas Web en su personomía. Los términos en cada recurso fueron ponderados según los distintos esquemas descriptos en las subsecciones III.*A* a III.*D* y se recomendaron un máximo de 10 etiquetas para cada recurso. En base a estas recomendaciones se calcularon las métricas mencionadas. Los resultados se promediaron por usuario o personomía.

C. Resultados Obtenidos

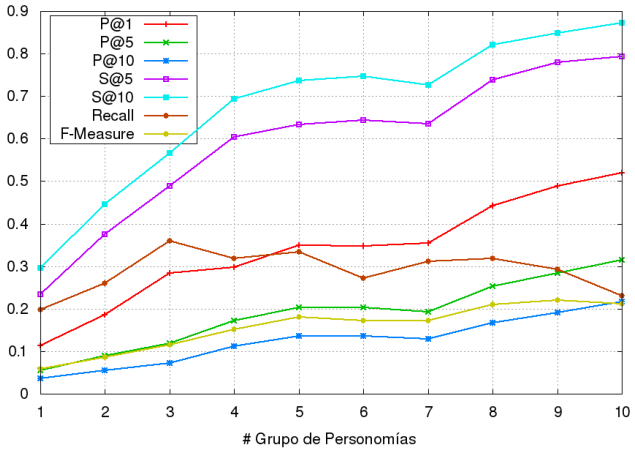
Para obtener líneas guía o puntos de comparación con el enfoque propuesto, se realizaron dos experimentos con las funciones basadas en frecuencia, uno con TF y otro con TF-IDF, en ambos casos tratando el contenido de las páginas Web como texto plano. Las Fig. 1(a) y (b) muestran los resultados promedio obtenidos para las métricas mencionadas en las recomendaciones de etiquetas. En los gráficos los resultados se muestran en 10 grupos de usuarios, los que fueron agrupados según la cantidad de etiquetas en cada personomía, con igual número de usuarios por grupo. Así, los usuarios del grupo 10 poseen las mayores cantidades de etiquetas en sus personomías, mientras que los usuarios del grupo 1 poseen las menores cantidades de etiquetas. Esta forma de mostrar los resultados permite observar que cuando los usuarios tienen más etiquetas en su vocabulario, la precisión aumenta ya que más etiquetas coinciden con las asignadas por el usuario. Si bien ambas funciones de pesado (TF y TF-IDF) mostraron un comportamiento similar, los resultados usando TF fueron levemente superiores a los de TF-IDF.

Para evaluar el esquema de pesado considerando la frecuencia de las palabras en los elementos de la estructura HTML de la página Web donde se encuentran, según la Ecuación 3, se usaron dos alternativas para asignar los pesos a cada uno de los elementos estructurales e_i considerados. La primera consistió en la observación de la ocurrencias de las etiquetas en cada uno de los elementos HTML de las páginas del conjunto de datos, tal como se muestran en la Tabla I. En este caso, los pesos fueron calculados de manera de amplificar equitativamente los porcentajes observados, por ejemplo $e_{body} = 1.55$, $e_{title} = 1.20$, $e_{link} = 1.37$, y así sucesivamente, donde por ejemplo 1.37 proviene de redondear el valor 0.3698 a dos decimales (0.37) y luego sumarle 1. La segunda alternativa fue la de aprender los pesos óptimos para dichos elementos e_i utilizando regresión lineal. Los gráficos de las Fig. 2(a) y (b) muestran los resultados obtenidos en cada caso, demostrando que los pesos para cada elemento aprendidos a través de regresión lineal mejoran levemente los resultados generales de las recomendaciones.

Por último, se experimentó combinando el esquema anterior (de frecuencia de los términos en los elementos HTML ponderados por pesos aprendidos por regresión lineal) con la ponderación de términos en base a las funciones gramaticales que cumplen en el texto. En este caso también se consideraron dos alternativas. La primera consistió en adicionar al esquema anterior la ponderación únicamente de los sustantivos, aplicando la Ecuación 4, ya que se observo que el 56% de las etiquetas en el conjunto de datos cumplían dicha función, es decir $g_{sust.} = 1.56$. La segunda alternativa consistió en ponderar todas las funciones gramaticales, aplicando la misma ecuación 4 y obteniendo los valores g_i de lo observado en el conjunto de datos ($g_{sust.} = 1.56$, $g_{verbo} = 1.19$, $g_{adj.} = 1.10$, $g_{adv.} = 1.01$). Las Fig. 3 (a) y (b) ilustran los resultados obtenidos con ambas alternativas, mostrando que los resultados son ligeramente superiores en el experimento que pondera todas las funciones gramaticales en combinación con la ponderación de los elementos HTML mediante regresión lineal, el cual es el enfoque propuesto.



(a) Pesos observados en el conjunto de datos

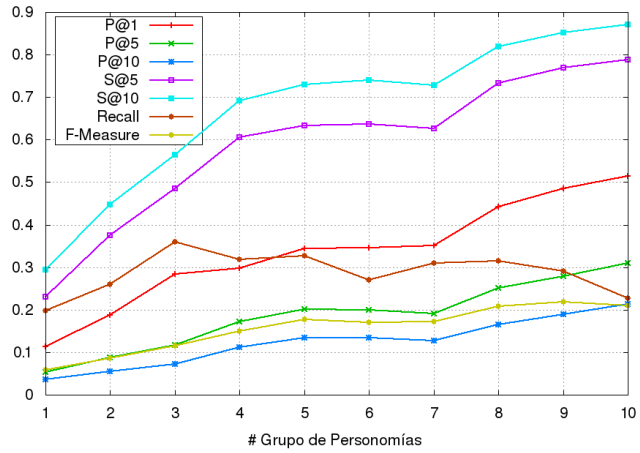


(b) Pesos aprendidos por regresión lineal

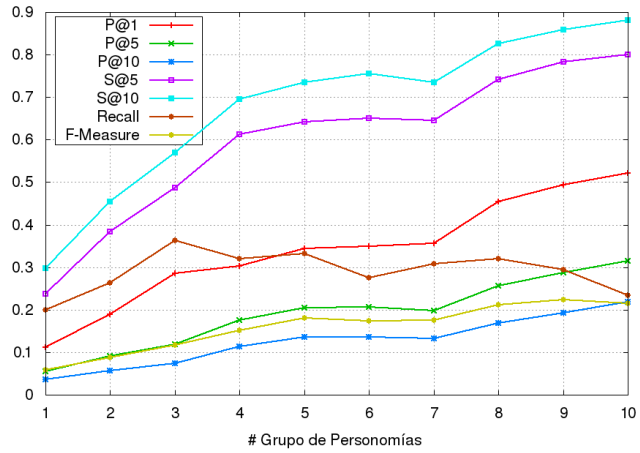
Figura 2. Resultados de la recomendación de etiquetas utilizando el esquema de pesado por componentes estructurales HTML.

En un análisis comparativo de los resultados de los experimentos, la Fig. 4 muestra la performance promedio de los distintos esquemas de pesado de términos. Es posible ver en dicha figura que los valores logrados por los esquemas basados solamente en frecuencia (experimentos con TF y TF-IDF sobre texto plano) son ampliamente superados cuando se tiene en cuenta la ubicación de las palabras en los elementos estructurales HTML de la página Web y aún más cuando se consideran las funciones gramaticales de las palabras. Esta forma de pesado que privilegia palabras en el título, meta-datos y otros componentes, así como sus funciones gramaticales, permite obtener un vector más descriptivo del contenido de la página y, por ende, mejores etiquetas candidatas para recomendación. El análisis de las categorías gramaticales de las palabras, aunque aporta mejoras, estas no son significativas a la recomendación e insumen un costo

computacional alto al tener que consultar el diccionario por cada palabra en el texto.



(a) Pesando sólo sustantivos y los elementos estructurales HTML



(b) Pesando todas las categorías gramaticales y los elementos estructurales HTML

Figura 3. Resultados de la recomendación de etiquetas pesando los componentes estructurales HTML y las funciones gramaticales de las palabras.

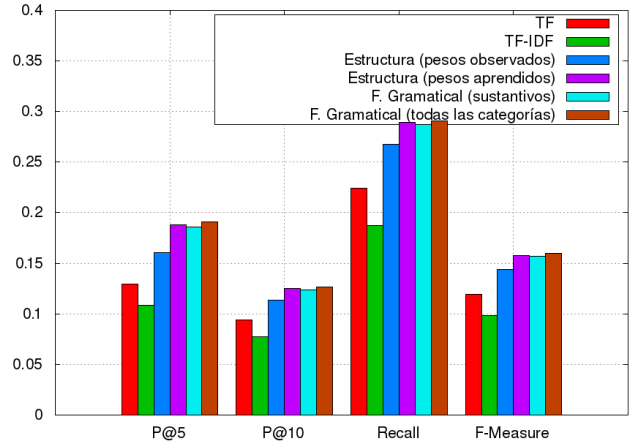


Figura 4. Comparación de los resultados alcanzados por los distintos esquemas de pesado de términos.

V. Conclusiones

En este trabajo se propuso un método de recomendación de etiquetas basado en contenido, que pondera los términos según su ubicación en la estructura HTML del recurso Web y según las funciones gramaticales que cumplen. Este método es aplicable a sistemas de etiquetado social y puede ser utilizado aún cuando no se disponga de información suficiente del usuario, ya sea porque recién empieza a usar el sistema o porque no ha anotado una cantidad de recursos que permita aprender un perfil preciso.

Primero se evaluaron los esquemas de pesado de términos tradicionales basados en frecuencia (TF y TF-IDF). Dado que TF resultó mejor, se lo consideró para combinarlo con la ponderación de las palabras según su ubicación en la estructura HTML de la página Web (título, cuerpo, etc.) y según la función gramatical que cumplen (sustantivo, verbo, adjetivo o adverbio). Estos esquemas de pesado fueron aplicados a la representación de recursos y recomendación de etiquetas usando un conjunto de datos extraído de Del.icio.us., uno de los sistemas de etiquetado social más populares. Los resultados mostraron que el método basado puramente en contenido es capaz de extraer etiquetas relevantes, pero considerando los elementos de la estructura HTML de las páginas Web, esta capacidad mejora en forma significativa por sobre los esquemas clásicos de pesado de términos. Los aciertos en las recomendaciones crecen ligeramente si además se contemplan los atributos gramaticales de las palabras, aunque en este caso también crece el costo computacional requerido.

Agradecimientos

Los autores quisieran agradecer el apoyo financiero brindado por la Universidad Técnica Estatal de Quevedo. Este trabajo ha sido parcialmente financiado por el Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET) a través del proyecto PIP Nº 114-200901-00381.

Referencias

[1] G. Adomavicius and A. Tuzhilin, “Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions”. *IEEE Transactions on Knowledge and Data Engineering*, 17(6):734-749, 2005.

[2] J. J. Astrain, F. Echarte, A. Córdoba and J. Villadangos, “Clustering Method for Social Network Annotations”. *IEEE Latin America Transactions*, 8(1):88-93, 2010.

[3] B. Choi and Z. Yao, “Web Page Classification”. In W. Chu and T. Young Lin, editors, *Foundations and Advances in Data Mining in Studies in Fuzziness and Soft Computing*, pp. 221-274. Springer Berlin / Heidelberg, 2005.

[4] J. Diederich and T. Iofciu, “Finding Communities of Practice from User Profiles Based on Folksonomies”. *Proceedings of the 1st International Workshop on Building Technology Enhanced Learning solutions for Communities of Practice (TEL-CoPs'06)*, 2006.

[5] S. Golder and B. Huberman, “Usage patterns of collaborative tagging systems”. *Journal of Information Science*, 32(2):198-208, 2006.

[6] L. Hong, E. H. Chi, R. Budiu, P. Pirolli and L. Nelson, “SparTag.us: A Low Cost Tagging System for Foraging of Web Content”. *Proceedings of the Working Conference on Advanced Visual Interfaces (AVI '08)*, pp. 65-72, Napoli, Italy, 2008.

[7] A. Hotho, R. Jäschke, C. Schmitz and G. Stumme, “Information Retrieval in Folksonomies: Search and Ranking”. *The Semantic Web: Research and Applications, 3rd European Semantic Web Conference (ESWC 2006)* in LNCS, pp. 411-426, 2006. Springer.

[8] Y-C. Huang, C-C. Hung and J. Yung-jen Hsu, "You Are What You Tag". *AAAI Spring Symposium on Social Information Processing (AAAI-SIP)*, pp. 36-41, 2008.
[9] R. Jäschke, L. Marinho, A. Hotho, L. Schmidt-Thieme and G. Stumme, "Tag Recommendations in Folksonomies". *Knowledge Discovery in Databases (PKDD 2007)* in LNCS, pp. 506-514, 2007.
[10] H.-N. Kim, A.-T. Ji, I. Ha and G.-S. Jo, "Collaborative Filtering based on Collaborative Tagging for Enhancing the Quality of Recommendation". *Electronic Commerce Research and Applications*, 9(1):73-83, 2010.
[11] M. Lipczak, "Tag Recommendation for Folksonomies Oriented towards Individual Users". *Proceedings of ECML PKDD Discovery Challenge (RSDC08)*, pp. 84-95, 2008.
[12] A. Mathes, "Folksonomies - Cooperative Classification and Communication Through Shared Metadata". *Computer Mediated Communication*, 2004.
[13] E. Michlmayr and S. Cayzer, "Learning User Profiles from Tagging Data and Leveraging them for Personal(ized) Information Access". *Proceedings of the Workshop on Tagging and Metadata for Social Information Organization*, 2007.
[14] R. Nakamoto, S. Nakajima, J. Miyazaki, S. Uemura and H. Kato, "Investigation of the Effectiveness of Tag-Based Contextual Collaborative Filtering in Website Recommendation". *Advances in Communication Systems and Electrical Engineering in Lecture Notes in Electrical Engineering*, pp. 309-318. Springer, 2008.
[15] M. G. Noll and C. Meinel, "Web Search Personalization via Social Bookmarking and Tagging". *Proceedings of 6th International Semantic Web Conference (ISWC) and 2nd Asian Semantic Web Conference (ASWC)* in LNCS, pp. 367-380, 2007.
[16] A. D. R. Oliveira, J. S. Sichman, L. N. Bessa, L. V. L. Filgueiras and T. R. Andrade, "Trust-based Recommendation for the Social Web". *IEEE Latin America Transactions*, 10(2):1661-1666, 2012.
[17] G. Salton and C. Buckley, "Term Weighting Approaches in Automatic Text Retrieval". *Information Processing and Management*, 24(5):513-523, 1988.
[18] G. Salton, A. Wong and C. S. Yang, "A Vector Space Model for Automatic Indexing". *Communications of the ACM*, 18:613-620, 1975.
[19] J-T. Sun, H-J. Zeng, H. Liu, Y. Lu and Z. Chen, "CubeSVD: A Novel Approach to Personalized Web Search". *Proceedings of the 14th International Conference on World Wide Web (WWW '05)*, pp. 382-390, 2005.
[20] P. Symeonidis, A. Nanopoulos and Y. Manolopoulos, "A Unified Framework for Providing Recommendations in Social Tagging Systems Based on Ternary Semantic Analysis". *IEEE Transactions on Knowledge and Data Engineering*, 22(2):179-192, 2010.
[21] E. Tonkin and M. Guy, "Folksonomies: Tidying up tags?". D-Lib, 12(1), 2006.
[22] K. H. L. Tso-Sutter, L. B. Marinho and L. Schmidt-Thieme, "Tag-Aware Recommender Systems by Fusion of Collaborative Filtering Algorithms". *Proceedings of the 2008 ACM Symposium on Applied Computing (SAC '08)*, pp. 1995-1999, 2008.
[23] Y. Xu, L. Zhang and W. Liu, "Cubic Analysis of Social Bookmarking for Personalized Recommendation". *Frontiers of WWW Research and Development (APWeb 2006)* in LNCS, pp. 733-738. Springer, 2006.
[24] C. M. Au Yeung, N. Gibbins and N. Shadbolt, "A Study of User Profile Generation from Folksonomies". *Social Web and Knowledge Management, Social Web 2008 Workshop at WWW2008*, 2008.
[25] N. Zhang, Y. Zhang and J. Tang, "A Tag Recommendation System for Folksonomy". *Proceeding of the 2nd ACM Workshop on Social Web Search and Mining (SWSM '09)*, pp. 9-16, Hong Kong, China, 2009.
[26] S. D. Zorzo, P. R. M. Cereda and R. A. Gotardo, "Adaptive Automata in Recommendation Systems". *IEEE Latin America Transactions*, 9(2):152-159, 2011.

Elías Portilla Olvera received the Engineering degree in Systems Engineering from Universidad Técnica Estatal de Quevedo, Los Ríos, Ecuador in 2005, and is a Master student in Systems Engineering in the Universidad Nacional del Centro de la Provincia de Buenos Aires, Tandil, Buenos Aires, Argentina. His current research interest are recommender systems in social networks and Web mining.

Daniela Godoy is a professor in the Computer Science Department at UNICEN University, member of the ISISTAN Research Institute and researcher at CONICET, Argentina. She received her PhD in computer science from the UNICEN University in 2005. Her research interests include intelligent agents, user profiling and Web mining.