Enfoque UTE, V.11-N.4, Oct.2020, pp. 1-15
http://ingenieria.ute.edu.ec/enfoqueute/
e-ISSN: 1390-6542
DOI: 10.29019/enfoque.v11n4.640

Recibido (Received): 2020/03/21
Aceptado (Accepted): 2020/09/02
Publicado (Published): 2020/10/01


# Un modelo híbrido de recomendación de etiquetas para sistemas de anotación social

## *(A Tag Recommendation Hybrid Model for Social Annotation Systems)*

*E. Portilla*[1], *D. Godoy*[2]

**Resumen**

El etiquetado social consiste en clasificar recursos web, con el uso de palabras o etiquetas libremente elegidas por los usuarios. La simplicidad y apertura de los sistemas de etiquetado social para organizar recursos, es la clave de su éxito en Internet. Existen numerosos enfoques para facilitar al usuario el proceso de etiquetado, permitiéndole reutilizar etiquetas y optimizando así su limitado tiempo de lectura y escritura. Este documento propone un enfoque híbrido diferente, que resuelve de forma sencilla el problema de las recomendaciones basadas únicamente en el contenido del recurso, fusionando la lista de recomendaciones con las etiquetas más populares del historial de etiquetas del usuario, permitiéndole así reutilizar los términos asignados a otros recursos.

**Palabras clave**

Folcsonomía; *Tagging* Social; Historial de etiquetado del usuario.

**Abstract**

*Social tagging consists of classifying web resources using words or tags freely chosen by users. The simplicity and openness of social tagging systems to organize resources is the key to your success on the internet. There are numerous approaches to facilitate the user the labeling process, allowing them to reuse labels and thus optimizing their limited reading and writing time. This document proposes a different hybrid approach that simply solves the problem of recommendations based solely on the content of the resource, merging the list of recommendations with the most popular tags in the user's tag history, thus allowing them to reuse terms assigned to others resources.*

**Keywords**

*Folksonomy; Social Tagging; User's tagging history.*

## 1. Introducción

Tagging o etiquetado social es una de las técnicas de clasificación de información más populares en la actualidad y ha sido implementado exitosamente en aplicaciones web de uso masivo como BibSonomy (https://www.bibsonomy.org/), Flickr (https://www.flickr.com/), entre otras, que permiten a los usuarios etiquetar y luego localizar páginas web, publicaciones académicas, objetos multimedia, entre otros tipos de recursos. Entre las ventajas del etiquetado social se destaca, que ayuda a clasificar y organizar recursos desde el punto de vista del usuario, y optimiza el limitado tiempo de lectura que este requiere para clasificar un recurso. Según Gou, Han, Zhu, Yang y Duan (2018) el proceso de etiquetado y las posteriores búsquedas de documentos etiquetados pueden ser, a su vez, personalizados en base al perfil del usuario.

Para facilitar la tarea de etiquetar recursos y atenuar problemas causados por la ambigüedad, sinonimia, polisemia y falta de normalización lingüística de las etiquetas elegidas

1 Universidad Técnica Estatal de Quevedo. Ecuador. [eportilla@uteq.edu.ec, https://orcid.org/0000-0001-5228-1658]
2 Consejo Nacional de Investigaciones Científicas y Técnicas (Conicet). Argentina / Universidad Nacional del Centro de la Provincia de Buenos Aires. Argentina. [dgodoy@exa.unicen.edu.ar, https://orcid.org/0000-0002-5185-4570]

por los usuarios, algunos sitios web de etiquetado social han implementado el servicio de recomendación de etiquetas, que consiste en mostrar al usuario una lista de las etiquetas más relevantes para clasificar un recurso dado. Con el fin de mejorar este servicio, se ha realizado abundantes trabajos de investigación, sobre la base de enfoques clásicos de la construcción, de Sistemas de Recomendación (SR).

El aporte de este trabajo es mostrar que el historial de etiquetado del usuario puede utilizarse de una manera sencilla y aun así efectiva, para complementar modelos de recomendación basados en contenido, de manera que se obtengan modelos de recomendación híbridos basados en el contenido del recurso a etiquetar y en el historial de etiquetas del usuario. El idioma que se utiliza para el modelo de recomendación es el inglés, en el cual están también varios recursos utilizados, como el conjunto de datos de entrenamiento, el diccionario digital Wordnet 3.0[3] y la stoplist (lista de términos irrelevantes) de SQL Server.

El resto de este artículo se organiza como sigue: La sección 2 describe algunos trabajos relacionados con la recomendación de etiquetas basada en contenido y en historiales de etiquetas de los usuarios o de los recursos. La sección 3 explica brevemente algunos conceptos vinculados al trabajo propuesto. La sección 4 expone la metodología de experimentación y se refiere al conjunto de datos utilizado para ello. En la sección 5 se explica el modelo de recomendación propuesto, que incorpora el historial del usuario para realizar las recomendaciones de etiquetas, a un modelo previo basado en contenido. La sección 6 narra los resultados experimentales obtenidos, describiendo previamente la línea base o de referencia y las métricas de evaluación. En la sección 7 se proporciona una discusión sobre la interpretación de lo estudiado. Las conclusiones logradas y el planteamiento de posibles trabajos futuros se muestran en la sección 8.

## 2. 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 $\mathbf{F} := (U, T, R, Y, \prec)$, formada por el conjunto de los usuarios *U*, el de los recursos *R*, el de las etiquetas *T*, y el de las asignaciones de etiquetas a recursos, dadas por una relación ternaria entre ellos $Y \subseteq U \times T \times R$ (Hotho, Jäschke, Schmitz, & Stumme, 2006). 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, $P_u := (T_u, R_u, I_u, \succ_u)$ con $I_u := \{(t, r) \in TXR \mid (u, t, r) \in Y\}$. Siendo $T_u$ las etiquetas del usuario $u$, $R_u$ los recursos del usuario, y $\succ_u$ las relaciones entre etiquetas para ese usuario (Hotho, Jäschke, Schmitz, & Stumme, 2006). Es decir, que una personomía es la colección de recursos, etiquetas y asignaciones de etiquetas realizadas por un único usuario, mientras que la colección de personomías se denomina folcsonomía (Zhang, Zhang, & Tang, 2009). Las folcsonomías, según Peng y Zeng (Peng & Zeng, 2010), en contraparte a las taxonomías como el ODP (Open Directory Project), son una opción que aprovecha el conocimiento de la comunidad para organizar y localizar información en la Web.

La recomendación de etiquetas en sistemas de etiquetado social es un tema abordado en numerosos enfoques de investigación, incluidos en distintas revisiones del estado del arte como (Godoy & Corbellini, 2016; Zhang, Tao, & Yi-Cheng, 2011; Belem, Almeida, & Goncalves, 2017). Algunos de estos enfoques se basan en Filtrado Colaborativo (FC), con el fin de hacer predicciones sobre una matriz bidimensional de opiniones de los usuarios. Otros enfoques se

3 http://wordnet.princeton.edu/

centran solo en contenidos, diferentes modelos híbridos que aprovechan el FC y el contenido de los recursos, y otros que proponen utilizar recursos adicionales como diccionarios digitales u ontologías (Qassimi & Abdelwahed, 2019; Godoy & Corbellini, 2016). Varias investigaciones sustentan que la mayor cantidad de términos utilizados por el usuario para etiquetar recursos, existen en el contenido de los mismos y en el historial de etiquetas del usuario (Hong, Chi, Budiu, Pirolli, & Nelson, 2008; Lipczak, 2008; Ju & Hwang, 2009). Recientemente, se han propuesto trabajos que explotan las redes sociales (Liu, 2018) como recurso disponible.

Los enfoques basados en FC requieren la inclusión de una nueva dimensión a la tradicional matriz de usuarios y recursos, que está dada por las asignaciones de etiquetas entre ellos. En varios trabajos en la literatura, se han utilizado técnicas clásicas de FC, aplicadas a proyecciones de la matriz de menor dimensionalidad (Godoy & Corbellini, 2016). Otro ejemplo es TAGme (Singh, Nagwani, & Pandey, 2017), un método de recomendación de etiquetas en comunidades de pregunta-respuesta, que consta de tres etapas: se construyen folcsonomías de tópicos con la base de usuarios, etiquetas y tópicos de los *posts*; luego, estas folcsonomías se usan para construir una matriz de perfiles de tópicos y otra de perfiles de usuarios por tópico. Finalmente, se utiliza Fltrado Colaborativo para la recomendación de etiquetas sobre el historial de tópicos de los usuarios.

Un problema asociado a los sistemas de recomendación es el conocido como arranque en frío (*cold-start*), que ocurre cuando un usuario aún no da etiquetas o un recurso aún no recibe suficientes etiquetas, por lo que no es posible realizar buenas recomendaciones a partir de la matriz usuario-recurso. En tal caso, el enfoque basado en contenido u otras dimensiones adicionales, como en Wang, Jin, Wnag, Pengi & Wnag (2018, donde se modela la dinámica temporal y la popularidad de las etiquetas, permiten aliviar este problema.

Varios trabajos se enfocan en el uso de distintos elementos de contenido para la recomendación de etiquetas. Tag2Word (Wu, Yao, Xu, Tong, & Lu, 2016) es un modelo generativo basado en la coocurrencia de etiquetas en el contenido. En el trabajo presentado por Ju & Hwang (2009), los autores plantean un enfoque donde se construye una lista de recomendación a partir de tres fuentes de información:

a. El contenido del documento.
b. Las etiquetas usadas por otros usuarios para etiquetar el mismo recurso.
c. Las etiquetas usadas por el mismo usuario para anotar otros recursos.

Los términos finales son fusionados mediante una función lineal, que pondera los términos de cada fuente con coeficientes escogidos por observación directa sobre la base de cálculos preliminares. Observan que la fuente de información que mayor calidad aporta a la lista final de recomendación es el contenido del recurso a etiquetar (fuente a), seguida de las fuentes c y b, respectivamente.

Sood, Owsley, Hammond y Birnbaum (2007) proponen un sistema de recomendación de etiquetas denominado TagAssist, que proporciona recomendaciones de etiquetas para nuevos *posts* en un blog, aprovechando *posts* etiquetados previamente. TagAssist aplica el algoritmo de stemming de Porter (1980) para reducir el conjunto de términos a su raíz morfológica y Razonamiento Basado en Casos, para determinar el contexto al que se refieren las raíces obtenidas. Toman como línea base la recomendación de etiquetas de mayor frecuencia en el recurso a anotar. De manera similar, Mishne (2006) plantea, aunque sin considerar el historial de etiquetado del usuario, una herramienta llamada AutoTag, que sugiere etiquetas para *posts* en un *weblog* en un proceso en el que se destacan los siguientes pasos:

a. Selección de *posts* similares al recurso a etiquetar.
b. Extracción de los términos de mayor frecuencia en los *posts* seleccionados.
c. Incremento a un factor constante del puntaje de las etiquetas usadas en otros *posts*.
d. Estructuración de la lista de recomendación con las etiquetas de mayor peso.

Musto, Narducci, Gemmis, Lops y Semeraro (2009) exponen un sistema denominado STaR (Social Tag Recommender), que realiza recomendación basada en contenido, considerando la similitud de los recursos en la personomía y en la folcsonomía, además de la actividad de etiquetado previa del usuario. El pesado de términos se realiza aplicando un enfoque de bolsa de palabras o *bag of words*, en el que todas las palabras tienen el mismo peso sin importar su ubicación (título, subtítulo, entre otras) en el documento.

El aprovechamiento de los contenidos de los recursos y de las etiquetas utilizadas por los usuarios, se aborda también en (Godoy & Amandi, 2008); en este caso para complementar los perfiles de usuario basados en contenido con las etiquetas frecuentemente utilizadas por los usuarios, a fin de que los perfiles permitan mejorar la interacción de los usuarios con un sistema de etiquetado colaborativo. El enfoque citado se compara con dos enfoques comunes en las folcsonomías que son: *tags* (o etiquetas) más populares entre usuarios (MPTU) y *tags* más populares por recurso (MPTR). Demuestran que el perfilado híbrido otorga mejores resultados que MPTR y MPTU.

En el trabajo de recomendaciones basadas en contenido descrito en (Portilla Olvera & Godoy, 2012), se evalúan seis esquemas de pesado para las palabras del texto de un recurso, a fin de recomendarlas como posibles etiquetas. El estudio se centra en la ponderación de términos según su frecuencia en diferentes componentes del HTML (título, body, URL, entre otros), las funciones gramaticales (sustantivo, verbo, adverbio, entre otras), que cumplen los términos y pesos aprendidos por regresión lineal, para establecer la importancia de los elementos del HTML y de las funciones gramaticales. La lista de recomendación se construye con los términos de mayor peso en el recurso a etiquetar. En tal enfoque resultó como segundo mejor esquema de pesado de palabras, al que pondera la estructura HTML y la función gramatical sustantivo, destacando que los términos que cumplen tal función gramatical son sustancialmente mejores candidatos para ser incluidos en la lista de recomendación.

En este trabajo se propone un método de recomendación de etiquetas basado en dos fuentes de información:

a. El contenido del recurso a etiquetar.
b. El historial de etiquetado del usuario.

La primera fuente permite abordar el arranque en frío del lado del usuario, realizando recomendaciones incluso cuando el usuario sea nuevo en el sistema; es decir, que no haya anotado aún recursos o solo una pequeña cantidad que no permita determinar sus preferencias. Para obtener las recomendaciones del contenido del recurso se implementa el enfoque de pesado de sustantivos y estructura HTML, descrito en Portilla Olvera & Godoy (2012). La segunda fuente se utiliza para enriquecer la lista de recomendaciones extraídas del contenido, con las etiquetas más populares entre los documentos más similares al recurso a etiquetar, dentro de la personomía del usuario. De esta forma el método planteado puede utilizarse también para abordar el arranque en frío del lado recurso, que ocurre cuando el recurso a etiquetar no contiene suficiente información que permita construir una lista de recomendación.

## 3. Conceptos relacionados

### *A. El Modelo de Espacio de Vectores*

El Modelo de Espacio de Vectores (Salton, Wong., & Yang, 1975) (VSM), actualmente muy utilizado en el área de Recuperación de Información, permite representar textos de forma que cada documento $d_i$ se corresponde con un vector de términos $t_1, t_2,...,t_n$ donde $n$ es el total de términos en el vocabulario. Para cada término perteneciente a un documento, se almacena un peso en el vector que representa al documento, en la dimensión correspondiente al término. El peso almacenado puede ser un valor entre 0 y 1, u otro que represente la importancia del término en el documento. Generalmente, el vocabulario de términos se construye extrayéndolos del total de documentos, de forma que ningún término del vocabulario se repita. La representación de un documento o página web $d_j$ sería:

$$d_j=(w_{1j}, w_{2j},...,w_{tj}) \quad (1)$$

donde $w_{ij}$ es el peso del término $i$ en el documento $j$. Dados los vectores de dos documentos, es posible computar una medida de similitud entre ambos mediante los pesos asignados a sus elementos. Tal medida de similitud podría ser el producto interno de los dos vectores o una función inversa del ángulo entre ambos, produciendo el máximo grado de similitud cuando el ángulo sea cero.

### *B. Métrica de similitud del coseno y técnica TF*

Una medida de similitud comúnmente aplicada en el VSM es la métrica del coseno definida como:

$$\mathrm{sim}(v_a, v_b) = \cos\theta = \frac{v_a v_b}{\|v_a\|\|v_b\|} = \frac{\sum_{i=1}^{n} w_{ia} * w_{ib}}{\sqrt{\sum_{i=1}^{n} w_{ia}^2}\sqrt{\sum_{i=1}^{n} w_{ib}^2}} \quad (2)$$

Donde $\theta$ es el ángulo entre los vectores $v_a$ y $v_b$ , y los elementos $w$ son los pesos asociados a los términos en cada vector. Entre las formas comunes de pesado de términos se destacan aquellas basadas en frecuencia (Salton, Wong., & Yang, 1975), como lo es TF (Term Frecuency), que mide la frecuencia de un término en un documento, y se normaliza mediante el total de términos del documento, sin contar las *stopwords*.

### *C. Algoritmo de regresión lineal*

El algoritmo de regresión lineal de Microsoft deriva del algoritmo de árboles de decisión (Albarran & Eduard, 2008). Para lograr la regresión lineal, los parámetros del algoritmo se controlan para restringir el crecimiento del árbol y mantener todos los datos en un nodo único y sin ninguna bifurcación, de modo que el conjunto de datos completo se utiliza como un único grupo, para calcular las relaciones en el paso inicial. Todos los datos residen en el nodo raíz (Veerman, Lachev, & Sarka, 2009). Así, el algoritmo no crea nunca una división y, por tanto, efectúa una regresión lineal que calcula una relación lineal entre una o más variables independientes y una dependiente y, a continuación, utiliza esa relación para la predicción.

### *D. Stopwords y Stoplists*

Los vectores que representan documentos pueden alcanzar una muy alta dimensionalidad en el espacio de vectores. Una forma de reducir la dimensionalidad consiste en la eliminación de *stopwords*, que son palabras que por su alta frecuencia de repetición en un documento, y bajo valor discriminatorio, no contribuyen a la determinación del contexto o del tema del recurso, y por tanto deben ser retiradas durante la indexación del conjunto de datos y el procesamiento de consultas (Lo, He, & Ounis, 2005). La remoción de esta clase de palabras se realiza contrastando cada término del contenido de un recurso contra una lista de *stopwords* o *stoplist*.

## 4. Metodología

Como se indicó antes, en este trabajo se propone un método de recomendación de etiquetas basado en dos fuentes de información:

a. El contenido del recurso a etiquetar.
b. El historial de etiquetado del usuario.

Para obtener las recomendaciones del contenido del recurso, se implementa el enfoque de pesado de la función gramatical sustantivo y de la estructura HTML, descrito en Portilla Olvera & Godoy (2012. La incorporación de recomendaciones en base asobre la base del historial de etiquetado del usuario, constituye la idea principal de este trabajo, cuyo objetivo es complementar las recomendaciones basadas en contenido, con recomendaciones basadas en el historial del usuario, mediante un método sencillo pero efectivo, que provoque mejoras en la lista de recomendaciones.

Para evaluar el enfoque propuesto y la línea base, se utilizaron las instancias de etiquetado de diciembre del 2007 del conjunto de datos DAI-Labor Delicious Corpus[4], pre-procesado de la forma descrita en (Portilla Olvera & Godoy, 2012), obteniendo así 90,618 instancias de etiquetado, 8,963 recursos y 40,586 *tags* entre las personomías de los 120 usuarios más prolíficos.

## 5. Modelo de recomendación propuesto

El modelo híbrido de recomendación de etiquetas propuesto, ilustrado en la figura 1, construye una lista de recomendación basada en dos fuentes:

a. El contenido del recurso a etiquetar.
b. El historial de etiquetado del usuario.

Los ítems obtenidos de la primera fuente se ponderan mediante un esquema propuesto en Portilla Olvera & Godoy (2012), pesando los términos que cumplen la función gramatical sustantivo y la ubicación de los términos en la estructura del HTML. Los ítems obtenidos de la segunda fuente, es decir del historial del usuario, corresponden a las etiquetas más populares entre los documentos más similares al recurso a etiquetar dentro de la personomía del usuario, por lo cual se toma el texto del recurso como el vector de consulta. El modelo de recomendación contempla dos fases que son:

4 http://www.dai-labor.de/. El dataset DAI-Labor Delicious Corpus completo se describe y analiza en (Wetzker, Zimmermann, & Bauckhage, 2008).

a. Aprendizaje.
b. Predicción o recomendación.

En la fase de aprendizaje, se obtiene tanto los pesos requeridos para ponderar la estructura HTML de los recursos web, como para ponderar la función gramatical sustantivo de los términos. En la fase de predicción o recomendación, un grupo de procesos construye una sublista de recomendación basada en el contenido del recurso a etiquetar, mientras que paralelamente otro grupo de procesos recupera, del historial de etiquetado del usuario, los *k* documentos más similares al recurso a anotar, con el fin de extraer de tales documentos otra sublista con las etiquetas más populares. Ambas sublistas se fusionan luego mediante tres pasos:

a. Impulsión.
b. Reemplazo.
c. Complementación.

Conformando así la lista final de recomendación mostrada al usuario. Las fases de aprendizaje y predicción utilizan recursos comunes, como las *stoplists*, mediante las cuales se filtran las palabras irrelevantes y la base de datos léxica Wordnet 3.0, utilizada para determinar los términos que son sustantivos. A continuación se explica con mayor detalle las fases de aprendizaje y de predicción o recomendación.

### *A. Fase de aprendizaje de pesos*

En la fase de aprendizaje, que se realiza en modo *offline*, se determinan por regresión lineal, los pesos para ponderar los elementos HTML y la función gramatical sustantivo de los términos. Como se observa en la figura 1, para la fase de aprendizaje se debe contar con un conjunto de datos de entrenamiento, que luego de ser descargado desde Internet, pasa por un proceso de extracción de términos, donde al mismo tiempo que se extrae cada término de cada tipo de componente HTML (título, URL, *body, links*, metadatos *keywords* y metadatos *description*), se calcula el valor TF del término respectivo en cada tipo de componente. Los términos extraídos se guardan en una base de datos relacional.

El proceso A1 ilustrado en la figura 1 se realiza mediante un componente de software implementado con la clase Term Extraction de Microsoft Integration Services 2008. Este componente toma como entrada una *stoplist* de referencia para omitir términos irrelevantes. En cuanto los términos se han almacenado en la base de datos, se determina mediante el proceso A2, si cumplen la función gramatical sustantivo, utilizando para ello la base de datos léxica WordNet.

El proceso A3 actúa sobre el conjunto de datos relacional, que almacena los términos clasificados por componente HTML y por ser o no sustantivos. Se utiliza el algoritmo de minería de datos de regresión lineal de Microsoft Analysis Services 2008 para aprender los pesos para ponderar los componentes HTML y los términos que cumplen la función gramatical sustantivo. Se concuerda con lo expuesto por Levandoski, Ekstrand, Ludwig, Eldawy, Mokbel, y Riedl (2011), en donde el uso de un conjunto de datos relacional es viable para la recomendación basada en filtros sobre los metadatos (documento etiquetado y el usuario que realizó el etiquetado), de la etiquetas. Al estar indexadas por usuario y documento, las recomendaciones se pueden recuperar rápidamente con veloces funciones de búsqueda de texto embebidas en el motor de base de datos.

*Preparación del conjunto de datos para el aprendizaje de pesos:* Los pasos A1 y A2 ilustrados en la figura 1, constituyen la preparación del conjunto de datos para el aprendizaje de pesos que se efectúa en el paso A3. El modelo contempla que para tal proceso se requiere un conjunto de datos relacional, constituido por documentos web, *tags*, usuarios e instancias de *tagging*. A su vez, los documentos web deben estar divididos en sus diferentes elementos HTML y estos deben estar atomizados por términos de los cuales también se requiere conocer si son sustantivos en el texto. La preparación de datos inicia con la descarga de los *datasets* de referencia, que están en texto plano y poseen las direcciones URL de los recursos web, pero no su contenido, el cual es recuperado por un *crawler* (desarrollado en C#.Net 2008), que de cada dirección web visitada, extrae el texto de cada elemento HTML, para almacenarlo en la base de datos.

*Aprendizaje de pesos para componentes estructurales y función gramatical sustantivo:* Sobre el conjunto de datos relacionales, que como parte de los casos de etiquetado conocidos almacena los términos clasificados por componente HTML junto a un campo que determina si son sustantivos, opera el algoritmo de minería de datos de regresión lineal descrito previamente, para aprender los pesos que mejor se ajusten a la probabilidad de que un término sea una etiqueta que el usuario potencialmente asignaría a un recurso. Los datos introducidos a la herramienta de minería de datos para el aprendizaje de pesos son los siguientes:

a. Los valores de TF en los elementos HTML en que existe el término en una página.
b. Un valor booleano que indica si el término es un sustantivo.
c. Un atributo conocido que identifica si un término ha sido utilizado como etiqueta para anotar un documento web.

En cuanto el algoritmo de minería de datos proporciona los pesos de ponderación de las TF y de la función gramatical sustantivo, tales pesos se guardan en la base de datos para luego ser utilizados en la recomendación de etiquetas candidatas.

### *B. Fase de predicción o recomendación de etiquetas*

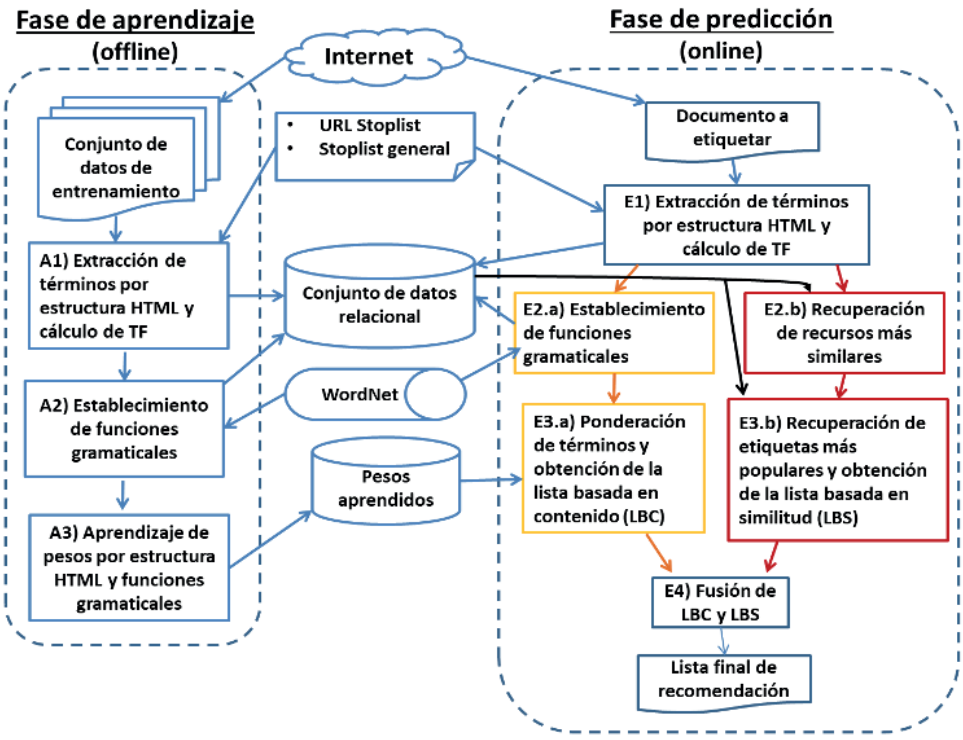
En esta fase el modelo recomienda un máximo de 10 etiquetas. La fase se efectúa cuando el usuario ha navegado hacia el recurso web que desea anotar y este se ha descargado desde Internet. Se extraen los términos de cada componente HTML (título, URL, *body, links*, metadatos *keywords* y metadatos *description*) y se calcula su TF respectiva.

Similar al paso A1 de la fase de aprendizaje, la extracción de términos y cálculo de los valores de TF, que comprende el paso E1, involucra el uso de dos *stoplists*. Para omitir términos irrelevantes de la URL, se utiliza la URL *stoplist*, mientras que para los demás componentes del HTML se establece la *stoplist* general. Ambas listas se describen más adelante.

Como resultado del proceso de extracción de términos, además de almacenarlos en la base de datos relacional, como historial de navegación del usuario, se obtienen dos representaciones del recurso a etiquetar:

a. Una lista de términos con los sustantivos identificados y con un valor de TF por cada componente HTML.
b. Una representación vectorial aplicando el VSM de Salton et al. (1975), que se utiliza como una consulta sobre la cual se recuperan los K recursos más similares en la personomía del usuario. En los vectores se utilizan valores binarios para identificar los términos que existen en cada documento, así como en la consulta.

***Figura 1.*** Funcionamiento general del modelo híbrido de recomendación




*Lista Basada en Contenido (LBC):* En cuanto los términos han sido clasificados por componente HTML, se inician en paralelo dos grupos de procesos. El primero, enumerado como E2.a y E3.a, obtiene una Lista Basada en Contenido (LBC). Esta lista preliminar de recomendación se consigue ponderando, mediante los pesos obtenidos en la fase de aprendizaje, los valores de TF y la función gramatical sustantivo de los términos. Los *N* términos más probables de ser etiquetas son incluidos en la LBC.

*Lista Basada en Similitud (LBS):* El segundo grupo de procesos (E2.b y E3.b), genera una lista de recomendación con las N etiquetas más populares en los K recursos más similares al recurso a etiquetar, en la personomía del usuario. Así se aprovecha el historial de etiquetas del usuario. Para el cálculo de similitud entre documentos se utiliza la métrica del coseno, con pesos binarios que adoptan el valor 1, si un término existe en un recurso, y cero en caso contrario.

*Fusión de LBC y LBS:* Mediante el proceso E4 se fusionan las listas LBC y LBS, en tres pasos (impulsión, remplazo y complementación), descritos a continuación: primero se impulsan en LBC los ítems que coexisten en LBS, para que ocupen los primeros lugares de LBC. En la tabla 1 se muestra el resultado de la impulsión de términos descrita.

En el segundo paso (remplazo), los 5 últimos ítems de LBC se remplazan por los primeros 5 ítems de LBS, que no coexistan en los primeros 5 ítems de LBC. A la derecha de la tabla 2 se muestra un ejemplo del remplazo de términos explicado.

Se optó por tomar los últimos 5 ítems de LBS, porque constituyen las etiquetas del historial del usuario, con las cuales el mismo ha anotado otros recursos similares al recurso a etiquetar. Se consideró que de esa forma habría equilibrio entre las etiquetas obtenidas del contenido del recurso y las etiquetas obtenidas del historial del usuario, logrando así una lista de recomendación híbrida.

***Tabla 1.*** Primer paso: Impulsión

| LBS | LBC | LBC (con términos impulsados) |
|---|---|---|
| Internet | Column | Resource |
| Video | Paper | Printer |
| Word | Woman | Column |
| Car | Mother | Paper |
| Dog | Resource | Woman |
| Resource | English | Mother |
| Facebook | Champion | English |
| Printer | Printer | Champion |
| Mouse | | |
| Photo | | |

***Tabla 2.*** Segundo paso: Remplazo

| LBS | LBC (con términos impulsados) | LBC (últimos términos remplazados) |
|---|---|---|
| Internet | Resource | Resource |
| Video | Printer | Printer |
| Word | Column | Column |
| Car | Paper | Internet |
| Dog | Woman | Video |
| Resource | Mother | Word |
| Facebook | English | Car |
| Printer | Champion | Dog |
| Mouse | | |
| Photo | | |

Si en este punto LBC no contiene 10 ítems, se efectúa el tercer paso (complementación), con el cual se adicionan a LBC más ítems de LBS, hasta que LBC contenga 10 ítems o hasta que en LBS no haya más ítems disponibles. Luego de este último paso, se obtiene la lista final de recomendación, ilustrada en la tabla 3, como LBC (complementada).

***Tabla 3.*** Tercer paso: Complementación

| LBS | LBC (complementada) |
|---|---|
| Internet | Resource |
| Video | Printer |
| Word | Column |
| Car | Internet |
| Dog | Video |
| Resource | Word |
| Facebook | Car |
| Printer | Dog |
| Mouse | Facebook |
| Photo | Mouse |

### *C. Conjunto de datos relacional*

La base de datos relacional, almacena el contenido del *dataset* de entrenamiento y el historial de etiquetado del usuario, y agiliza la búsqueda de palabras por componente HTML, y de documentos similares a un documento específico que va a ser etiquetado. La herramienta de minería de datos utilizada en el aprendizaje, se conecta con la base de datos para consultar los datos de entrenamiento. En concordancia con Levandoski et al. (2011), se considera que las funciones propias del motor de base de datos para las búsquedas de texto, pueden aprovecharse para acelerar las búsquedas de documentos.

### *D. La URL* stoplist *y la* stoplist *general*

Se observa en la figura 1, que se utilizan dos *stoplists*. La URL *stoplist*, sirve para filtrar términos irrelevantes de la URL y la *stoplist* general, se emplea en el procesamiento de los demás elementos HTML y además es parte de la URL *stoplist*. La URL *stoplist* posee además otros términos que aparecen con mucha frecuencia en las direcciones web, como HTTP, PHP, WWW, INDEX, por lo cual se los considera *stopwords* en la URL de las páginas web, pero no en los demás elementos estructurales del HTML. La *stoplist* general utilizada es la provista por el motor de base de datos SQL Server 2008, en inglés.

## 6. Resultados

### *A. Línea base*

En Portilla Olvera & Godoy (2012), se describen seis enfoques de pesado de términos para la recomendación de etiquetas basada en contenido. En tales enfoques se destacó como segundo mejor, el que genera recomendaciones ponderando los valores TF, elementos de la estructura HTML y los términos que cumplen la función gramatical sustantivo, con pesos aprendidos por regresión lineal. A fin de evaluar si la incorporación del historial de etiquetado del usuario en la lista de recomendación genera mejoras, se utilizó el referido enfoque basado en contenido de dos formas:

a. Como componente basado en contenido al que se incorporan las recomendaciones basadas en el historial del usuario, formando así el modelo híbrido propuesto.
b. Como línea base, de comparación o referencia para determinar si el modelo híbrido genera mejoras respecto al modelo de recomendación basado solo en contenido.

### *B. Métricas de evaluación*

Para evaluar la línea base y el modelo propuesto, se utilizaron las métricas estándar de precisión, *recall* (cobertura) y *f-measure* (medida f mejor conocida como F1) (Baeza-Yates & Ribeiro-Neto, 1999). La precisión mide la cantidad de etiquetas recomendadas que fueron usadas por el usuario para anotar el recurso, *recall* mide el número de etiquetas relevantes recomendadas sobre el total que debieron recomendarse y F1 es una medida que balancea precisión y *recall*.

Dado que las recomendaciones se presentan en una lista ordenada o ranking, las métricas se analizan en distintos puntos del mismo. P@k es 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 fue el éxito S@k, que es la probabilidad de encontrar una etiqueta relevante entre las k primeras recomendaciones, donde S@1 es equivalente a P@1 por definición.

En los experimentos, para cada usuario se utilizaron los recursos de su personomía. Para calcular las métricas se realizaron simulaciones de recomendaciones sobre la línea base y el modelo propuesto, recomendando un máximo de 10 etiquetas para cada recurso. Se decidió trabajar con máximo 10 ítems, tomando como referencia los trabajos relacionados expuestos en Ju & Hwang (2009); Yu-Ta, Shoou-I, Tsung-Chiej & Yung-jen (2009); Musto et al, (2009); y, Zhang, Zhang & Tang (2009), donde se experimenta con listas de recomendación de máximo 10 ítems.

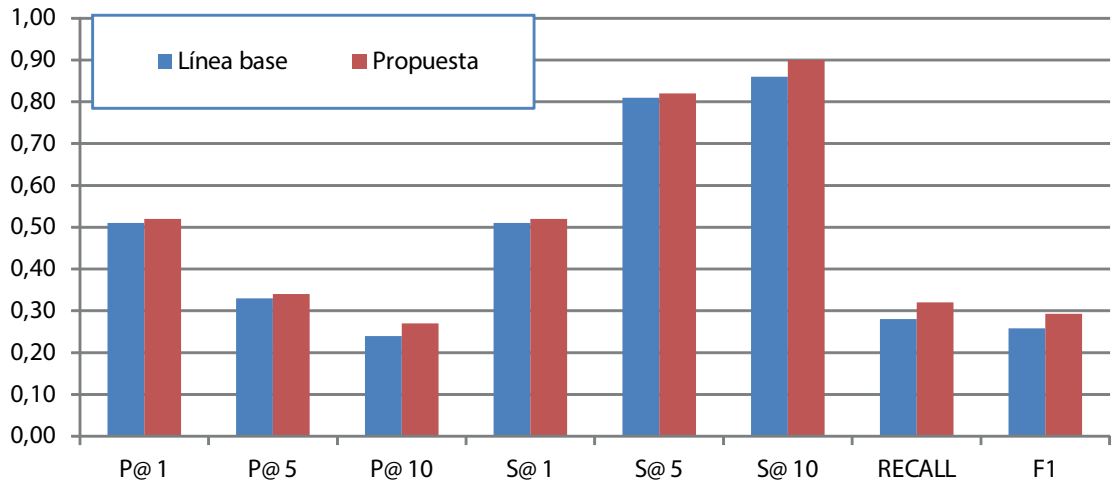
### *C. Resultados obtenidos*

La figura 2 ilustra que, en general, el enfoque basado en contenido es ligeramente superado por el enfoque híbrido propuesto, en todas las métricas. Los mejores resultados del modelo híbrido propuesto respecto a la línea base, se pueden apreciar en la tabla 4, en las métricas P@10, Recall y consecuentemente en F1, en comparación con las demás métricas.

***Tabla 4.*** Resumen de métricas de evaluación sobre la línea base y el modelo propuesto

| Modelo | P@1 | P@5 | P@10 | S@1 | S@5 | S@10 | RECALL | F1 |
|---|---|---|---|---|---|---|---|---|
| Línea base | 0.51 | 0.33 | 0.24 | 0.51 | 0.81 | 0.86 | 0.28 | 0.258 |
| Propuesta | 0.52 | 0.34 | 0.27 | 0.52 | 0.82 | 0.90 | 0.32 | 0.293 |

***Figura 2.*** Resumen de métricas de evaluación sobre la línea base y el modelo propuesto




En la figura 3 se muestran los resultados por separado para cada métrica, en 12 grupos que se construyeron según la cantidad de etiquetas de cada personomía, de forma que los usuarios del grupo 1 poseen menores cantidades de etiquetas que los usuarios del grupo 2. La figura 3 (a), ilustra que en la P@1, el modelo propuesto mejora a la línea base en 9 de los 12 grupos. En los grupos 6.8 y 9 resultó mejor la línea base. En cuanto a la P@5, ilustrada en la figura 3 (b), el modelo propuesto mejoró a la referencia en todos los casos, aunque por un margen de diferencia muy pequeño. La figura 3 (c) muestra que en la P@10, la línea base resultó mejor que el modelo propuesto en el grupo 8. En los otros 11 grupos, el modelo híbrido propuesto mejoró a la referencia.

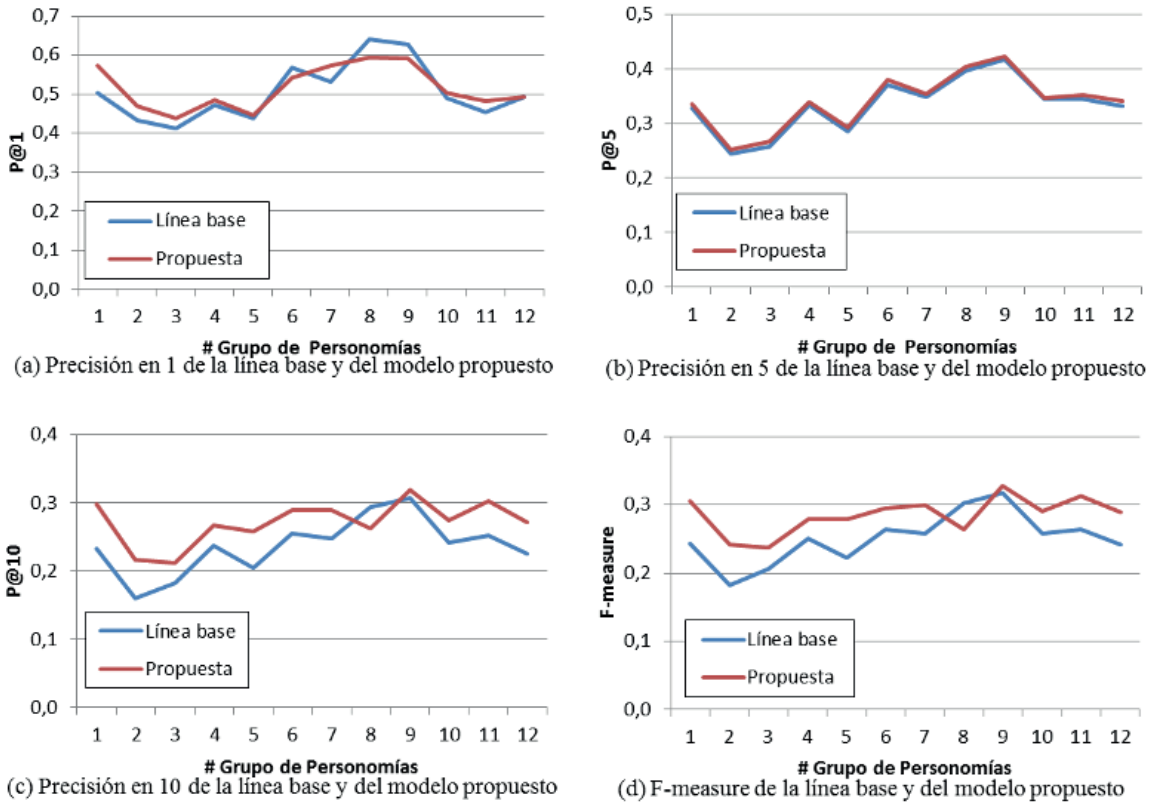
## 7. Discusión

Según lo ilustrado en la figura 3(a) y figura 3(b), al recomendar listas de 1 ítem y 5 ítems respectivamente, las diferencias entre la línea base y la propuesta son mínimas, mientras que la figura 3(c) y la figura 3(d) muestran que al recomendar listas de 10 ítems las diferencias entre la línea

base y el modelo propuesto se amplían, resultando mejor el modelo propuesto. Esto podría ser un indicador de que el mejor rendimiento del modelo propuesto se logra al recomendar listas de 10 ítems. Los resultados del modelo propuesto mejoran levemente a los resultados de la referencia. Este comportamiento concuerda con lo expuesto por Ju y Hwang (2009), quienes observan que la fuente de información que mayor calidad aporta a la lista de recomendación final es el contenido del recurso a anotar.

Hay que destacar que las recomendaciones del modelo híbrido incluyen etiquetas que no existen en el contenido del recurso a anotar ni en ningún diccionario digital, debido a que generalmente son escritas por el usuario y tienen un significado comprensible solo por él. Por tal razón, un modelo basado en contenido jamás recomendaría estas etiquetas, mientras que un modelo híbrido que incluya el historial del usuario si lo haría, en casos de que el usuario requiera anotar otros recursos con las mismas etiquetas, evitándole volver a escribirlas y con ello logrando que el usuario perciba al sistema de recomendación amigable y cómodo.

***Figura 3.*** Gráficos individuales de las métricas sobre la línea base y el modelo propuesto




(a) Precisión en 1 de la línea base y del modelo propuesto

(b) Precisión en 5 de la línea base y del modelo propuesto

(c) Precisión en 10 de la línea base y del modelo propuesto

(d) F-measure de la línea base y del modelo propuesto

## 8. Conclusiones y trabajos futuros

Los resultados de los experimentos sustentan que al fusionar la lista de recomendación basada en contenido, con las etiquetas más populares por personomía mediante el sencillo enfoque de impulsión, remplazo y complementación, el modelo propuesto aporta al enfoque planteado en Portilla Olvera & Godoy (2012), pequeñas mejorías en todas las métricas evaluadas.

Si bien los resultados del modelo propuesto no distan significativamente de los resultados de la línea base, sí constituyen mejoras. Considerando que el modelo híbrido propuesto permite reutilizar etiquetas escritas manualmente por el usuario, evitándole volver a escribirlas al etiquetar otros recursos, lo cual resulta en mayor comodidad para el usuario, tales mejoras justifican plenamente la implementación del modelo híbrido en sistemas de recomendación de etiquetas.

Un posible trabajo futuro podría ser experimentar con modelos que adicionen otros pasos, además de los indicados como impulsión, remplazo y complementación, o varíen el orden de los mismos en busca de mejores resultados en las listas de recomendación.

Otro trabajo futuro sería la incorporación de recomendaciones colaborativas, que aprovechen, además de las etiquetas del historial del usuario actual y del contenido de los recursos, las etiquetas asignadas por otros usuarios, ya sea al mismo recurso o a recursos similares. Esto permitiría agregar heterogeneidad a la lista de recomendaciones y utilizar en el etiquetado el lenguaje predominante de la comunidad.

Finalmente, otro trabajo futuro sería un modelo de recomendación multilenguaje, que se pueda utilizar en múltiples idiomas, no sólo en inglés.

## Referencias


Albarran, G. & Eduard, G. (2009). *Data Mining and SQL*. Lima, Perú: Megabyte

Baeza-Yates, R. & Ribeiro-Neto, B. (1999). *Modern Information Retrieval.* Boston, Massachusetts, USA: Addison-Wesley

Belem, F. M.; Almeida, J. M. & Goncalves, M. A. (2017). A Survey on Tag Recommendation Methods. *J. Assoc. Inf. Sci. Technol 68*(4), 830-844. doi: 10.1002/asi.23736.

Godoy, D. & Amandi, A. Hybrid (2008). Content and Tag-based Profiles for Recommendation in Collaborative Tagging Systems (pp. 58-65). *Proceedings of the 2008 Latin American Web Conference*. Espírito Santo, Brasil.

Godoy, D. & Corbellini, A. (2016). A State-of-the-art Review. *International Journal of Intelligent Systems 31*(4), 314–346. doi: 10.1002/int.21753

Gou, Z.; Han, L.; Zhu, J.; Yang, Y. & Duan, B. (2018). Personalized Search by a Multi-Type and Multi-Level User Profile in Folksonomy. *Arab. J. Sci. Eng 43*(12), 7563–7572. doi: 10.1007/s13369-018-3133-2

Hong, L.; Chi, E. H.; Budiu, R.; Pirolli, P. & Nelson, L. (2008). SparTag.us: A Low Cost Tagging System for Foraging of Web Content. *Proceedings of the Working Conference on Advanced Visual Interfaces,* 65-72. Napoli, Italy.

Hotho, A.; Jäschke, R.; Schmitz, C. & Stumme, G. (2006). Information Retrieval in Folksonomies: Search and Ranking. *European Semantic Web Conference*, 411-426. Budva, Montenegro: Springer.

Ju, S., & Hwang, K.-B. (2009). A weighting scheme for tag recommendation in social bookmarking systems. *Proceedings of the 2009th International Conference on ECML PKDD Discovery Challenge - Volume 497*, 109–118. https://dl.acm.org/doi/abs/10.5555/3056147.3056156

Levandoski, J. J.; Ekstrand, M. D.; Ludwig, M. J. et al. (2011). *Benchmarks for Evaluating Performance of Recommender System Architectures*. *PVLDB 4*, 911-920. doi: 10.14778/3402707.3402729.

Lipczak, M. (2008). Tag Recommendation for Folksonomies Oriented towards Individual Users. *Proceedings of ECML PKDD*, 84-95. Amberes, Bélgica.

Liu, H. (2018). A Tag-Based Recommender System Framework for Social Bookmarking Websites. *Inderscience Publishers 14*(3): 303-322. doi: 10.1504/IJWBC.2018.094916

Lo, R. T.; He, B. & Ounis, I. (2005). Automatically Building a Stopword List for an Information Retrieval System. *5th Dutch-Belgium Information Retrieval Workshop* (DIR) '05, 1-8. Países Bajos. https://bit.ly/2Fbz8v5

Mishne, G. (2006). AutoTag: A Collaborative Approach to Automated Tag Assignment for Weblog Posts. *Proceedings of the 15th International Conference on World Wide Web*, 953-954. Edimburgo, Escocia: ACM.

Musto, C.; Narducci, F.; Gemmis, M. d.; Lops, P., & Semeraro, G. (2009). STaR: A Social Tag Recommender System. Proceedings of the ECML/PKDD Discovery Challenge, 215-227. Bled, Eslovenia.

Peng, J. & Zeng, D. (2010). Making Item Predictions through Tag Recommendations. En: *ICEBI2010,* 483-490. Kunming, Yunnan, P. R.China: Atlantis Press.

Porter, M. (1980). An Algorithm for Suffix Stripping. *Program: electronic library and information systems 14*(3): 130–137. doi: 10.1108/eb046814

Portilla Olvera, E. & Godoy, D. (2012). Evaluating Term Weighting Schemes for Content-based Tag Recommendation in *Social Tagging Systems*. *IEEE Latin America Transactions 10*(4): 1973-1980. doi: 10.1109/TLA.2012.6272482

Qassimi, S. & Abdelwahed, E. (2019). The Role Of Collaborative Tagging and Ontologies in Emerging Semantic of Web Resources. *Computing 101*(10): 1489-1511. doi: 10.1007/s00607-019-00704-9

Salton, G., Wong., A. & Yang, C. (1975). A Vector Space Model for Automatic Indexing. *Comun. ACM 18*(11): 613–620. doi: 10.1145/361219.361220.

Singh, A. K., Nagwani, N. K., & Pandey, S. (2017). TAGme: A Topical Folksonomy Based Collaborative Filtering for Tag Recommendation in Community Sites. *Proceedings of the 4th Multidisciplinary International Social Networks Conference*, 1–7. https://doi.org/10.1145/3092090.3092095

Sood, S., Owsley, S., Hammond, K. & Birnbaum, L. (2007). Tagassist: Automatic Tag Suggestion for Blog Posts (pp.1-7*). International Conference on Weblogs and Social Media*,. USA: Boulder, Colorado.

Veerman, E.; Lachev, T. & Sarka, D. (2009). *Microsoft SQL Server 2008 - Business Intelligence Development and Maintenance.* Washington, EEUU: Microsoft Press.

Wang, K.; Jin, Y.; Wang, H.; Pengi, H. & Wang, X. (2018). Personalized Time-aware Tag Recommendation. *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, (pp. 459–466). New Orleans, LA, USA.

Wetzker, R.; Zimmermann, C. & Bauckhage, C. (2008). Analyzing Social Bookmarking Systems: A del.icio.us Cookbook (pp. 26-30). *Proceedings of the ECAI 2008 Mining Social Data Workshop*. Patras, Grecia.

Wu, Y.; Yao, Y.; Xu, F.;Tong, H. & Lu, J. (2016). Tag2Word: Using tags to generate words for content based tag recommendation. *Proceedings of the 2016 ACM Conference on Information and Knowledge Management,* 2287–2292. Indianapolis, USA: ACM.

Yu-Ta, L.; Shoou-I, Y.; Tsung-Chieh, C. & Jane Yung-jen, H. (2009). A Content-Based Method to Enhance Tag Recommendation. *Proceedings of the 21st international jont conference on Artifical Intelligence*, 2064 - 2069. Pasadena, California: Morgan Kaufmann.

Zhang, N.; Zhang, Y. & Tang, J. (2009). A Tag Recommendation System for Folksonomy. *Proceedings of the 2nd ACM Workshop on Social Web Search and Mining,* 9-16. Hong Kong, China: ACM.

Zhang, Z.-K.; Tao, Z. & Yi-Cheng, Z. (2011). Tag-aware Recommender Systems: A State-of-the-art Survey. *J. Comput. Sci. Technol 26*(5): 767--777. doi: 10.1007/s11390-011-0176-1