IFLA WLIC 2013
SINGAPORE

Enviado en: 13 diciembre 2013

Clasificación de los recursos web utilizando los términos generados por los usuarios

Spanish translation of the original paper: Classification of web resources using user generated terms

Margaret E.I. Kipp
School of information studies, University of Wisconsin-Milwaukee, Milwaukee, United States.
E-mail address: kipp@uwm.edu

Soohyung Joo
School of information studies, University of Wisconsin-Milwaukee, Milwaukee, United States.
E-mail address: sjoo@uwm.edu

Inkyung Choi
School of information studies, University of Wisconsin-Milwaukee, Milwaukee, United States.
E-mail address: ichoi@uwm.edu

TRADUCTORA: **Silvia Rubio Martín**, Biblioteca Nacional de España

This is a Spanish translation of "*Classification of web resources using user generated terms*" Copyright © 2013 by **Silvia Rubio Martín**. This work is made available under the terms of the Creative Commons Attribution 3.0 Unported License:
http://creativecommons.org/licenses/by/3.0/

Resumen:

En este estudio, proponemos un método útil para clasificar recursos web basado en los marcadores sociales de información creados por los usuarios. Intentamos examinar si las etiquetas sociales pueden ser una herramienta para la clasificación de páginas web de determinados temas. En el estudio, aplicamos dos métodos estadísticos: el Análisis de Componentes Principales (en inglés, PCA) y el agrupamiento jerárquico para la clasificación de sitios web de divulgación médica. Lo primero que hicimos fue utilizar el método PCA para identificar las diferentes áreas temáticas de las páginas web seleccionadas y así se extrajeron

seis apartados: mujeres, personas mayores, niños/padres, medicamentos, hombres e investigación. En segundo lugar, se realizó un análisis de agrupamiento jerárquico para reunir sitios web similares en niveles jerárquicos diferentes. Estos dos métodos revelaron que las etiquetas sociales representan bastante bien las características de las páginas web individuales del campo de la divulgación médica. Este estudio proporciona, por tanto, un procedimiento que permite que las etiquetas sociales puedan ser utilizadas para la clasificación automática de recursos de la Web.

Palabras claves: Organización de los recursos web, clasificación, etiquetas sociales, términos generados por los usuarios, métodos de minería de datos, agrupamiento jerárquico.

---

1. INTRODUCCIÓN

Como el número de recursos web ha crecido de manera asombrosa, la organización de estos recursos ha llegado a ser un asunto importante para los investigadores y profesionales de la información. A diferencia de los materiales impresos tradicionales o de las publicaciones periódicas electrónicas, la mayor parte de los recursos web no están organizados según los métodos tradicionales. A pesar de los esfuerzos para desarrollar de manera consensuada modelos de organización, no existen metadatos estandarizados que se utilicen de forma generalizada para todos los recursos online. Además, debido a que los recursos online crecen más rápido que los recursos tradicionales, es casi imposible clasificarlos manualmente. A pesar de estas limitaciones, se necesita una organización de los recursos web para que los usuarios puedan acceder de manera eficiente a la ingente información disponible en Internet. La categorización de los recursos web permite mejorar las estrategias de búsqueda de los usuarios y les ayuda a encontrar nuevos documentos relevantes relacionados con sus investigaciones en el entorno de la Web (Xie and Joo, 2012). Muchos investigadores, dándose cuenta de la importancia de la organización de los recursos web, han intentado buscar formas eficientes de clasificación automática para subsanar la complejidad de la red y mejorar la accesibilidad a la información disponible en Internet. La categorización automática de los recursos basada en el texto completo de los documentos es el método más comúnmente utilizado, sin embargo, por lo que nosotros sabemos, los términos colaborativos creados por los usuarios apenas se han utilizado en los sistemas de clasificación de recursos web.

El etiquetado social se ha convertido en un tema popular y los investigadores estudian sus patrones y características únicas. Muchas bibliotecas y servicios de información han adoptado los servicios de marcadores sociales como complemento a los metadatos ya existentes. En este estudio, nos adentramos en otra aplicación práctica de las etiquetas sociales como herramientas para la organización de los recursos web. Hemos investigado si los términos colaborativos de los usuarios pueden ser útiles para la clasificación de este tipo de recursos y pueden convertirse en una alternativa necesaria a la organización de los recursos online basada en el texto completo. Como las etiquetas sociales contienen palabras clave para describir los recursos en la web, partimos de la base de que estas etiquetas podían servir de ayuda para clasificar esos mismos recursos. Para comprobar si era posible una clasificación basada en las etiquetas sociales, empleamos dos métodos cuantitativos, el Análisis de Componentes Principales (PCA) y el agrupamiento jerárquico y restringimos la investigación al campo de la divulgación médica. Los investigadores han utilizado muchos métodos diferentes en sus estudios sobre el etiquetado social, pero muy pocas investigaciones han empleado la reducción dimensional o la técnica del agrupamiento jerárquico. Nuestra propuesta se basa en utilizar el etiquetado social como un método para la clasificación de los recursos web por semejanza. Implementaremos dos técnicas de minería de datos, el Análisis de Componentes Principales (PCA) y el agrupamiento

jerárquico, para utilizar los términos creados por los usuarios con la finalidad de agrupar los recursos web de un ámbito concreto, la información online sobre salud.

2. TRABAJOS PREVIOS

Las primeras investigaciones en este campo se centraron en la clasificación automática de recursos web basada en el análisis de las palabras claves. En concreto, los investigadores se esforzaron en extraer de los documentos online términos de materias o metadatos para posibilitar la organización automática de la información de la web (Ricca et al. 2004; 2007; Tonella et al. 2003). Estos métodos ayudan a estructurar en patrones los recursos web dispersos al azar, pero se encuentran con dos problemas importantes. En primer lugar, se basan principalmente en el contenido de los propios documentos, sin tener en cuenta la interpretación de los recursos desde la perspectiva de los usuarios. En segundo lugar, existe el problema de que el análisis del texto completo de recursos web para su clasificación requiere descargas importantes, y la clasificación basada en metadatos tiene aplicaciones limitadas ya que la cobertura de los metadatos está normalmente limitada a unos campos predeterminados.

Los términos colaborativos creados por los usuarios, recogidos en forma de etiquetas sociales, pueden ser una alternativa a la indexación previa basada en el texto completo o a las propuestas basadas en metadatos. Las etiquetas sociales son consideradas como un método de organización de los recursos web, por lo que los investigadores han examinado las prácticas de etiquetado social de los usuarios en diversos ámbitos y situaciones. Por ejemplo, Kipp y Campbell (2007) descubrieron que las prácticas de etiquetado eran congruentes de alguna manera con la indización tradicional y creaban una serie de términos relacionados dotados de una dimensión personal adicional. Aunque muchos estudios han comparado la indización basada en etiquetas con los vocabularios controlados (Yoon 2009; Kipp 2005; 2011), muy pocas investigaciones han examinado el rol único de las etiquetas como herramientas de indización de recursos web. Cattulo et al. (2008) aplicó el análisis de redes para examinar la coexistencia de etiquetas sociales en los sistemas de marcadores en línea. Ellos introdujeron la idea de que la actividad de marcado colectivo por parte de los usuarios puede construir una red sólida de recursos y relaciones semánticas. Kipp y Joo (2010) investigaron la estructura semántica de la Web basada en los patrones de marcado usando ecuaciones estructurales. Sin embargo, muy pocos investigadores han intentado utilizar las etiquetas sociales para la clasificación de recursos web. En definitiva, los estudios previos se han centrado fundamentalmente en la descripción de las prácticas de etiquetado más que en la posible utilización de las propias etiquetas para la clasificación del contenido de la Web.

3. METODOLOGÍA

Para examinar si los términos colaborativos de los usuarios podían ser utilizados en la clasificación de recursos web, restringimos el ámbito de investigación al campo de la divulgación médica en Internet. El conjunto de datos que manejamos para la investigación estaba formado por 34 páginas web de divulgación médica extraídas de CAPHIS (Consumer and Patient Health Information Section). Los marcadores se sacaron de Delicious.com y las páginas seleccionadas tenían, por lo menos, cincuenta marcadores. CAPHIS ofrece una lista de páginas web relacionadas con la información sobre la salud y las clasifica en diferentes grupos: mujeres, hombres, personas mayores, padres e hijos y medicamentos. El esquema de clasificación de CAPHIS, creado manualmente, se utilizó para evaluar los resultados de la clasificación basada en etiquetas. De los datos recolectados se excluyeron tanto los términos excesivamente generales como los términos demasiado específicos. En total, se extrajeron de las páginas elegidas 6416 términos únicos. De entre ellos, se seleccionaron 654 términos que

aparecían al menos en cinco páginas simultáneamente y de esos 654 se excluyeron los 79 términos generales que se encontraban con más frecuencia en las páginas. Al final quedaron 575 términos para el análisis.

Lo primero que hicimos fue aplicar el Análisis de Componentes Principales (PCA) para identificar la estructura de los recursos web del tema seleccionado. El PCA se utiliza a menudo en la investigación de la estructura de conjuntos de datos. Normalmente se usa para descubrir la dimensionalidad de cualquier conjunto de datos con la finalidad de definir la estructura de un dominio específico (Jackson 1991). Lo que hace el PCA es reducir un amplio grupo de variables y convertirlo en un pequeño grupo que todavía contiene la mayor parte de la información del grupo original. Este grupo reducido de variables es mucho más fácil de analizar e interpretar. En este trabajo, hemos estudiado en primer lugar las dimensiones de los recursos web seleccionados para consolidar la clasificación basada en etiquetas.
En segundo lugar, llevamos a cabo el agrupamiento jerárquico para poner a prueba realmente la clasificación basada en los términos creados por los usuarios, que era el objetivo inicial de esta investigación. Aplicamos el método de encadenamiento de Ward, que es uno de los algoritmos de agrupamiento jerárquico de aglomeración (de abajo arriba). El método de Ward es usado frecuentemente para determinar las semejanzas globales de los clusters de documentos utilizando las similitudes entre las entidades. Además, este método tiene la ventaja de formar grupos jerárquicos de subconjuntos que se excluyen mutuamente.
Mediante la aplicación de estos dos análisis cuantitativos, comprobamos cómo de precisa era la clasificación de los recursos web basada en los términos de los usuarios comparándola con la clasificación predefinida de CAPHIS. Para cada página web, añadimos un prefijo, esto es: w_ (salud de la mujer); s_ (salud de las personas mayores); k_ (salud infantil); d_ (información sobre medicamentos); m_ (salud masculina); y r_ (información relacionada con la investigación).

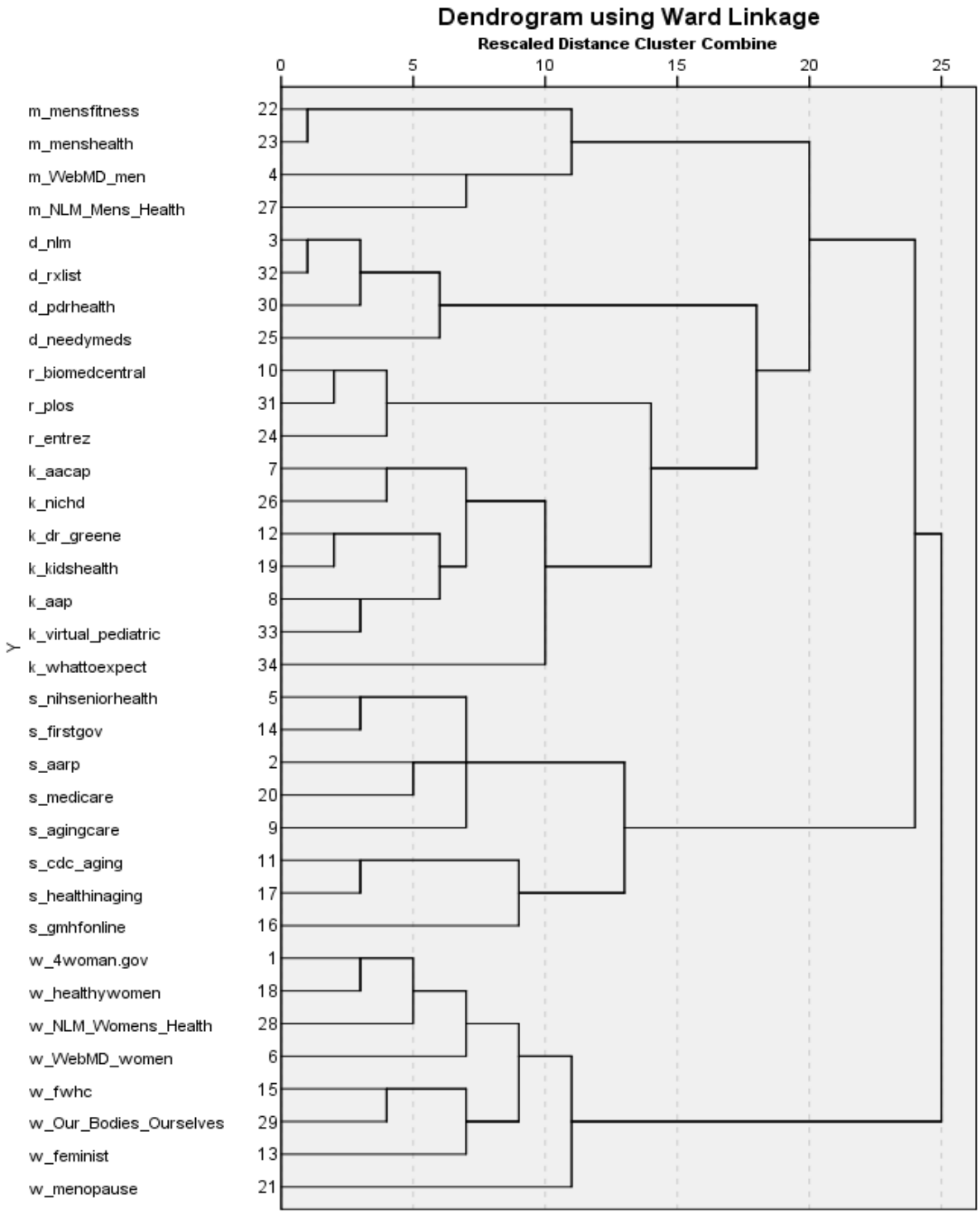
4. RESULTADOS

Los resultados indican que los términos de los usuarios son bastante útiles a la hora de clasificar de manera precisa los recursos online de la Web. Los dos métodos utilizados, el PCA y el agrupamiento jerárquico, son capaces de categorizar las páginas seleccionadas tan bien como el esquema creado por expertos de la clasificación de CAPHIS.
El resultado del PCA muestra que la información etiquetada puede ser utilizada para representar la estructura de los recursos web de determinados temas. En el valor propio de 1.87, se identificaron seis dimensiones mediante el PCA. Las seis dimensiones supusieron el 61.12% de la varianza total. Estas dimensiones son “mujeres”, “personas mayores”, “niños/padres”, “medicamentos”, “hombres” e “investigación”. La mayor parte de las páginas muestran cargas factoriales moderadas o altas, por encima del 0.45. Esto revela que las etiquetas sociales pueden ser utilizadas para determinar la organización estructural de los recursos web.

Tabla 1: Matriz de componentes rotados (Rotación varimax)

	Dimension					
	1	2	3	4	5	6
w_Our_Bodies_Ourselves	.819					
w_healthywomen	.814					
w_NLM_Womens_Health	.798					
w_4woman.gov	.735					
w_fwhc	.706					
w_feminist	.679					
w_WebMD_women	.641					
w_menopause	.451					
s_healthinaging		.889				
s_cdc_aging		.887				
s_nihseniorhealth		.872				
s_firstgov		.716				
s_agingcare		.592				
s_aarp		.513		.445		
s_gmhfonline		.506				
s_medicare		.403		.339		
k_aap			.906			
k_dr_greene			.864			
k_kidshealth			.795			
k_virtual_pediatric			.667			
k_nichd			.616			.358
k_aacap			.528			
k_whattoexpect			.483			
d_nlm				.890		
d_rxlist				.887		
d_pdrhealth				.866		
d_needymeds				.706		
m_menshealth					.923	
m_mensfitness					.904	
m_WebMD_men					.795	
m_NLM_Mens_Health					.486	
r_plos						.894
r_biomedcentral						.843
r_entrez						.803

Por otro lado, el agrupamiento jerárquico nos ha servido para averiguar si la indización colaborativa puede utilizarse a nivel práctico para la organización de los recursos web. Los métodos de agrupamiento empleados fueron el método de encadenamiento de Ward y las medidas de Minkowski. El Dendrograma se utilizó para interpretar el resultado del agrupamiento. En el nivel 13 del cluster, se identificaron 6 grupos que compartían entre sí elementos comunes: hombres, medicamentos, investigación, niños y padres, personas mayores, mujeres. El Dendrograma también mostró que las páginas web con características similares se encuentran adyacentes. Este resultado revela que la indización realizada por los usuarios puede servir de base para la organización automática de los recursos web.



5. DISCUSIÓN/CONCLUSIÓN

Esta investigación ha examinado si los términos generados por los usuarios pueden utilizarse para la clasificación de recursos web. Nuestro trabajo ha consistido en la recopilación de términos de usuarios provenientes de páginas de divulgación médica, y en su clasificación mediante agrupamiento jerárquico. Comparando los resultados del agrupamiento jerárquico con el esquema de clasificación creado manualmente, el resultado fue que el agrupamiento basado en etiquetas coincidía en un 100% la clasificación manual. Aunque este estudio se centra en un

grupo de datos de un campo muy específico, los resultados sugieren que los términos generados por los usuarios pueden ser útiles a la hora de clasificar automáticamente recursos web que, por otro lado, serían imposibles de clasificar manualmente debido a su crecimiento exponencial.
El análisis PCA revela que la información de las etiquetas muestra de manera bastante clara y precisa la estructura encubierta de los recursos web. En este análisis, la información extraída de las etiquetas puede agrupar acertadamente las páginas web mediante métodos cuantitativos apropiados. El agrupamiento jerárquico confirmó que los términos de los usuarios pueden clasificar correctamente recursos web en los seis grupos predefinidos en el esquema de clasificación manual ("mujeres", "personas mayores", "niños/padres", "medicamentos", "hombres" e "investigación"). Los resultados obtenidos están muy próximos a la clasificación de las páginas web propuesta por CAPHIS. Por lo tanto, se ha demostrado de manera práctica que la contribución de los usuarios en la descripción de recursos web online mediante etiquetas es una buena herramienta para su clasificación.

Este estudio sugiere que los términos colaborativos de los usuarios se pueden utilizar para organizar recursos web, lo que ha sido en los últimos tiempos una tarea desafiante en el campo de la Biblioteconomía y las Ciencias de la Información. Aunque nos hemos centrado en las páginas web de un campo específico, podríamos extender este método a otros tipos de recursos web, tales como libros, documentos electrónicos, imágenes, y otros recursos multimedia. Por ejemplo, se pueden clasificar los documentos utilizando los términos de los usuarios recogidos en Librarything (www.librarything.com/). Las imágenes se pueden agrupar según sus similitudes basándose en las etiquetas de Flickr (www.flickr.com/). Nuestro trabajo intenta explorar la utilidad de las etiquetas de los usuarios en la organización de recursos web de un determinado campo aplicando dos métodos de minería de datos. Para ampliar su utilidad y examinar si el etiquetado social se puede emplear de manera generalizada como herramienta de organización, se deberán llevar a cabo más estudios que exploren nuevas posibilidades del marcado social en la organización de recursos web de varios campos temáticos y con nuevas terminologías. El etiquetado social como herramienta de organización nos proporciona un sistema mucho más simple y fácil para clasificar recursos web que los métodos de aprendizaje automáticos que requieren más tráfico de datos, más formación y son más costosos.

En este estudio hemos aplicado dos métodos de minería de datos que se pueden emplear para clasificar recursos web de campos específicos. Ambos, el PCA y el método de agrupamiento jerárquico, han demostrado que las etiquetas sociales se pueden utilizar como herramientas de organización de recursos de la Web. Los hallazgos preliminares muestran que podemos utilizar los términos generados por los usuarios para una clasificación automática de los recursos de Internet. Nuestro trabajo pone de manifiesto que diversos métodos estadísticos pueden servir para agrupar correctamente los recursos mediante las etiquetas que estos contienen, sin necesidad de recurrir a los métodos de aprendizaje automáticos que requieren de más análisis y más formación en el sistema.

BIBLIOGRAFÍA

Jackson, J. E. (1991).*A user's guide to principal components* (Vol. 244). Wiley-Interscience.

Kipp, M. E. I. (2005). Complementary or Discrete Contexts in Online Indexing: A Comparison of User, Creator, and Intermediary Keywords. *Canadian Journal of Information and Library Science, 29*(4):419–436.

Kipp, M. E. I. (2011). Tagging of Biomedical Articles on CiteULike: A Comparison of User, Author and Professional Indexing. *Knowledge Organization 38*(3): 245-261.

Kipp, M. E.I., & Campbell, D. G. (2007). Patterns and inconsistencies in collaborative tagging systems: An examination of tagging practices. *Proceedings of the American Society*

for Information Science and Technology,43(1): 1-18.

Kipp, M. E. I. and Joo, S. (2010). Application of structural equation modelling in exploring tag patterns: A pilot study. *Proceedings of the American Society for Information Science and Technology, 47*: 1–2. doi: 10.1002/meet.14504701325.

Ricca, F., Tonella, P., Girardi, C., & Pianta, E. (2004). An empirical study on keyword-based web site clustering. In *Program Comprehension, 2004. Proceedings. 12th IEEE International Workshop on,* 204-213.

Ricca, F., Pianta, E., Tonella, P., & Girardi, C. (2008). Improving Web site understanding with keyword-based clustering. *Journal of Software Maintenance and Evolution: Research and Practice, 20*(1): 1-29.

Tonella, P., Ricca, F., Pianta, E., & Girardi, C. (2003, September). Using keyword extraction for web site clustering. In *Web Site Evolution, 2003. Theme: Architecture. Proceedings. Fifth IEEE International Workshop on*, 41-48.

Xie, I. and Joo, S. (2012). Factors affecting the selection of search tactics: Tasks, knowledge, process, and systems. *Information Processing & Management*, 48(2): 254-270.

Yoon, J. (2009). Towards a user-oriented thesaurus for non-domain-specific image collections. *Information Processing & Management 45*(4): 452-468.