IFLA WLIC 2013
SINGAPORE

Submitted on: June 20, 2013

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

Abstract:

In this study, we suggest a useful method to classify web resources based on social tag information generated by users. We attempted to examine whether social tags could be a tool of classifying websites in a certain domain. We applied two statistical methods, including principal component analysis (PCA) and hierarchical clustering for classifying websites in the domain of consumer health information. First, PCA method was applied to identify different dimensions of the selected websites. Six dimensions were extracted from PCA: women, seniors, kids/parenting, drugs, men, and research. Second, we conducted a hierarchical clustering analysis to group similar websites in different hierarchical levels. These two methods reveal that social tags well represent the characteristics of individual websites in the domain of health information. This study yields a methodological implication that social tags can be used to automatically classify resources on the Web.

Keywords: Web source organisation, classification, social tag, user generated terms, data mining methods, principal component analysis, hierarchical clustering.

1 INTRODUCTION

As the number of web resources has increased explosively, organizing web resources has become an important topic for information professionals and researchers. Unlike traditional printed materials or electronic periodicals, most of web resources are not organized with

traditional methods. Despite concerted efforts to develop web resource organization models, there is no standard metadata widely used for all online resources. Furthermore, as online resources grow more quickly than traditional resources, it has been almost impossible to manually classify them. Despite these limitations, web resource organization is still needed and will be beneficial to many users increasing the ability to effectively access a plethora of information available on the web. Categorization of web resources better supports users' browsing strategies and helps users extend their search interests to relevant documents in the Web environment (Xie and Joo, 2012). Recognizing the importance of web resource organization, many researchers have tried to explore efficient means of automatically classifying web resources to alleviate the complexity and improve the accessibility to the Internet information. Mechanical categorization of resources via fulltext documents is the most commonly used method, however, to the best of our knowledge, users' collaborative terms are not widely used in web resource classification systems.

Social tagging has become a popular topic, and researchers have studied its unique features and patterns. Many libraries and information services have adopted social tagging services to complement existing metadata. In this study, we look into another practical implication of social tags as a tool to organize web resources. We investigated whether users' collaborative terms could be useful to classify web resources and be a compelling alternative to previous fulltext-based clustering approaches in online resources. As social tags contain key words for describing resources on the web, we assumed that tags can serve as an aid to classifying web resources. To test the possibility of social tagging based classification, we employed two quantitative methods, Principle Component Analysis (PCA) and Hierarchical Clustering, in the domain of consumer health information. Researchers have used many different methods in social tagging studies, but few studies have applied dimensional reduction or hierarchical clustering technique. Our approach is to employ social tags as a method for classifying web resources by similarity. We will implement two data mining techniques – Principal Component Analysis (PCA) and Hierarchical Clustering, to use user-generated terms in clustering web resources in a specific domain, online consumer health information.

2 PREVIOUS WORKS

Previously, efforts were made to automatically classify web resources based on keyword analysis. In particular, researchers have strived to extract subject terms or metadata from online documents to enable mechanical web information organization (Ricca et al.2004; 2007; Tonella et al. 2003). These methods helped structure randomly scattered web resources in organized patterns. However, these methods primarily rely on the content of documents themselves, without any interpretation of the resources from the users' perspective. Also, analyzing fulltext for web source classification requires significant system loads, and metadata-based classification has limited applications as coverage of metadata is usually limited to predefined fields.

Users' collaborative terms, collected in the form of social tags, can be an alternative to previous fulltext indexing or metadata-based approaches. As social tagging has received attention as a means of organizing web-based resources, researchers examined online user’s tagging practices in different domains or situations. For example, Kipp and Campbell (2007) discovered that tagging practices were congruent with traditional indexing in some ways and created an extra personal dimension regarding time and task related terms. Even though several studies compared the indexing role of tagging to that of controlled vocabulary (Yoon 2009; Kipp 2005; 2011), few studies have explored the unique role of tags as indexing tool for web sources. Cattuto et al. (2008) applied network analysis to examine the co-occurrence

of social tags from online bookmarking systems. They introduced the notion of resource distance based on the collective tagging activity of users to build a weighted network of resources and semantic relations. Kipp and Joo (2010) explored the semantic structure of web space based on tagging patterns using structural equation modeling. However, few researchers have attempted to utilize social tags to classify web resources. Previous studies focused on description of tagging practices, rather than practical application of tagging in web source classification.

3 METHODOLOGY

To examine whether users' collaborative terms can be applied to classify web resources, we selected a consumer health information domain on the web as a test set. As a dataset, we selected 34 consumer health information websites suggested by CAPHIS (Consumer and Patient Health Information Section). Tags were collected from Delicious.com and the selected sites had at least fifty tags. CAPHIS offers a list of health information sites and classifies them into different groups, such as women, men, seniors, kids & parenting, and drugs. The CAPHIS classification scheme, which is manually created, was used to assess the performance of tag-based classification. The collected data were trimmed by excluding both extremely general terms and specific terms. In total, 6416 unique terms were observed across the selected sites. Among them, we first selected 654 terms that occurred at least five sites simultaneously, and then excluded 79 general terms that most frequently occurred across the sites. Finally, 575 terms were selected in the analysis.

First, Principal Component Analysis (PCA) was applied to identify the structure of the web resources in the selected domain. PCA is widely used to explore the nature of dataset structure. It is typically used to discover the dimensionality of any dataset to represent the structure of a specific domain (Jackson 1991). It reduce a large set of variables to a small set that still contains most of the information in the large set. A reduced set of variables is much easier to analyze and interpret. In this study, we first explored dimensions of the selected web resources to ensure the possibility of tagging-based classification.

Then, hierarchical clustering was carried out to actually test the classification using user terms, which is the primary purpose of this study. Ward’s linkage method was applied, which is one of agglomerative (bottom up) hierarchical clustering algorithms. Ward's method is frequently used to determine the overall similarity of document clusters based on similarities between entities. Also, it has an advantage of forming hierarchical groups of mutually exclusive subsets.

Using these two quantitative analyses, we investigated how precisely user terms classify web resources by comparing the results with the predefined CAPHIS classification. For each website, we added a prefix as follows: w_ (women's health); s_ (seniors' health); k_ (kids' health); d_ (drug information); m_ (men's health); and r_(research related information).

4 RESULTS

Findings indicate that user terms are quite useful to accurately classify online resources on the web. Both PCA and hierarchical clustering categorize the selected sites as well as the expert-created scheme of the CAPHIS classification.

A PCA result shows that tagging information can be used to represent the structure of the web resources in a specific domain. At the eigenvalue of 1.87, six dimensions were identified from the PCA. The six dimensions accounted for 61.12% of the total variance. Those

dimensions are “women,” “seniors,” “kids/parenting,” “drugs,” “men,” and “research.” Most of the sites show moderate or high factor loadings over 0.45. This reveals that social tags can be used to elucidate the structural organization of web resources.

Table 1: Rotated component matrix (Varimax rotation)

	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

To examine whether collaborative indexing can be used to organize web resources at the practical level, hierarchical clustering was conducted. As a clustering method, Ward’s linkage and Minkowski measures were employed. The Dendrogram was used to interpret the clustering result. At the cluster level 13, six groups are identified that share commonalities with each other : men, drugs, research, kids & parenting, seniors, women. Also, the Dendrogram shows that websites with similar characteristics are adjacently located. This result reveals that user indexing can be used to automatically organize web resources.

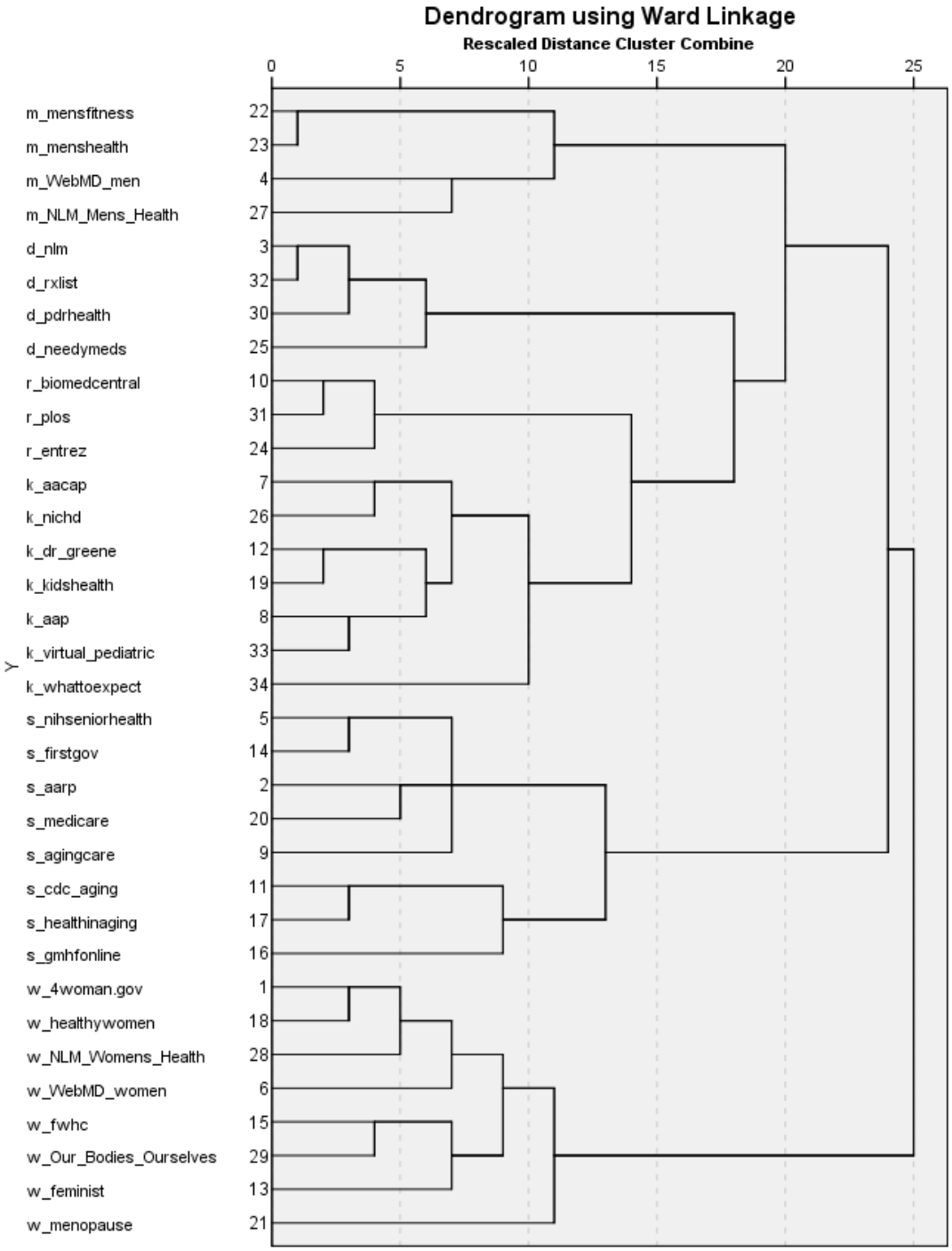

Figure 1: Dendrogram of hierarchical clustering analysis result

5 DISCUSSION/CONCLUSION

This study examined whether user generated terms can be used to classify web resources. We collected user terms from consumer health information sites, and classified them using hierarchical clustering. By comparing the clustering result with the human-created classification scheme, we found that the tagging-based clustering showed 100% precise classification performance. Although this study conducted with a domain-specific dataset, the

findings suggest that user-generated terms can be useful to automatically organize web resources that are impossible to do manually due to size.

The PCA analysis reveals that the tag information clearly and quite precisely account for the hidden structure of web resources. In this analysis, the collected tag information can generate a reasonable clustering of the websites with appropriate quantitative methods. Hierarchical clustering confirmed that user terms can classify web resources precisely into the six groups predefined by human-created classification scheme ("women," "seniors," "kids/parenting," "drugs," "men," and "research."). The results obtained by the two analyses are closely akin to the classification of websites suggested by CAPHIS. Therefore, it can be suggested that user's contribution to describing online web resources using tags has appropriate credibility in classification.

This study suggests that user collaborative terms can be used to organize web resources, which has been a challenging task in the field of library and information science. Although this study used websites in a specific domain, we can extend this method to other types of web resources, such as books, web documents, images, and other multimedia sources. For example, documents can be classified by user terms collected from Librarything (www.librarything.com/). Image files can be grouped by their similarities based on tags collected from Flickr (www.flickr.com/). This study attempts to explore usefulness of user tags on organizing web sources within a specific domain by applying two data mining methods. To expand its usefulness and examine which extend tagging is applied as an organizing tool, further studies will be conducted for exploring more possibilities of user tag's organizing web sources with a broader range of interest and terminology . Social tagging as an organizing tool provide a simpler and easier way to classify web resources than machine learning methods that require more data traffic, training process, and cost.

This study introduced two data mining methods that can be applicable to classify web resources in a specific domain. Both PCA and hierarchical clustering methods validated that social tags can be used as a tool to organize resources on the Web. The preliminary findings showed that we can apply user-generated terms to automatically classify resources on the Web. Our study demonstrates that a number of different statistical methods can be applied to generate a reasonable clustering of web resources based on tags, requiring less analysis or system training than machine learning techniques.

REFERENCES

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.