Content Classification by Folksonomies: Framework of Social Bookmarking System

Shih-Ming Pi
Department of Information Management, Chung Yuan Christian University, Taoyuan, Taiwan
Email: smpi@im.cycu.edu.tw

Hsiu-Li Liao
Department of Information Management, Chung Yuan Christian University, Taoyuan, Taiwan
Email: wenlly@im.cycu.edu.tw

Su-Houn Liu
Department of Information Management, Chung Yuan Christian University, Taoyuan, Taiwan
Email: vandy@im.cycu.edu.tw

Chen-Wen Lin
Department of Information Management, Chung Yuan Christian University, Taoyuan, Taiwan
Email: darkness@im.cycu.edu.tw

***Abstract*—Social bookmarking is a recent phenomenon which has the potential to give us a great deal of data about pages on the web. In this paper, we present an improved framework to web content classification based on Folksonomy. Since Folksonomy is keyword-based, it is associated with semantic problems. Various academicians have constructed ontologies to solve semantic problems. However, ontology depends on expert knowledge of the problem domain, and the process of constructing knowledge depends on the participation of knowledge engineers. This study presents an improved weighting mechanism to solve the semantic problems and the problematic effects of poor classification. An experimental prototype called FSBS (Folksonomy Social Bookmarking System) was developed. Testing indicates that the FSBS can effectively reduce the number of classification results by more than 30% significantly improved the quality of tagging, and increased user satisfaction. We believed that our works have provided a feasible framework for an intelligent social bookmarking system.**

***Index Terms*— folksonomy, social bookmarking, classification mechanism**

I. INTRODUCTION

The recent emergence and success of folksonomies and the so-called tagging with services such as del.icio.us or Flickr have shown the great potential of this simple yet powerful approach to collect metadata about resources. The classification of various documents in the field of information retrieval has been extensively studied explored. TFIDF (Frequency and Inverse Document Frequency) is the method for calculating the weight of an article, and use the method to classify documents. Other classification approaches include K-Nearest Neighborhood, Artificial Neural Network, and others. Although these classification methods are extensively applied, those based on keywords are associated with a semantic issue. Although in recent years many scholars have constructed ontology to solve the semantic problem. However, ontological construction depends on expert knowledge in the problem domain, and the process of constructing knowledge requires the participation of knowledge engineers. Therefore, the most serious problems associated with the ontological approach are how to define expert knowledge in a manner that adequately represents domain knowledge.

In dot-com bubble era, many Internet companies have closed down. However, the benefits provided by the Internet are ongoing. These websites have evolutionary from Web1.0 to Web2.0. Many enterprises have developed innovative Web2.0 applications and are eager to use them. Traditional automated document classification is enhanced by user tagging using metadata [6]. The Internet is a platform that supports tagging. We can see the example like del.icio.us and HEMIDEMI websites. Del.icio.us begin in 2003, it was the first system to provide website bookmarking. It includes its own favorite web bookmarks, and allows other users' inquiries by using keywords. HEMIDEMI was established in 2005 in Taiwan as a bookmarking site, which also allows users to set favorite bookmarks. Although it has two bookmarks tool that use Folksonomy, searching and browsing the results of classification causes information overload, reducing user-friendliness.

Unlike traditional categorization systems, the process of tagging is nothing more than annotating documents with a flat, unstructured list of keywords called tags. Although the number of peer-reviewed research on tagging is still comparatively low, several studies have already analyzed the semantic aspects of tagging and why it is so popular and successful in practice [2]. User-set labels are prone to three main problems: synonyms, semantic issues associated with keywords and classification issues [3]. These various problems have not

doi:10.4304/jsw.7.4.745-748

yet been solved. This study aims to improve the Folksonomic weighting mechanism to solve the semantic problems of Folksonomy and eliminate the effects of poor classification. This study tries to answers two research questions. (1) Can we propose a mechanism to improve the problem of synonymous words in Folksonomy and the effect of automated document classification? (2)How can the usefulness of this mechanism be verified?

The paper is organized as follows. Section 2 discusses the key issues in theoretical background. Section 3 introduces our research methodology. Section 4 presents an experimental prototype called FSBS (Folksonomy Social Bookmarking System) and the experimentally demonstrates the effectiveness of the FSBS. Finally, section 5 draws our conclusions.

II. THEORETICAL BACKGROUND

Tagging is the ability of users to define information, and use the keyword-based approach to describe their thought about specific web content [2]. User tagging behavior is information-describing behavior. According to the photo sharing website, Flickr, the "tagging label helps you find some commonality among photographs based on a keyword or category." However, the descriptions made by users are not limited to specific photographs. They can cover films, music, bookmarked links, and blogs. Users can set any name according the tagging label and quickly find other users to share resources of all kinds. Tagging by keywords is performed by users. It is not based on the general meaning, but only on the needs of the user. It doesn't meet strict classificatory standard.

Various websites that support tagging have increased gradually. Folksonomy refers to this phenomenon [1], [8]. The term combines the words "folk" and "taxonomy". Folksonomy therefore refers to classification by users [6], [8]. Users can mark personal information, and use tags as a basis for classification. Tagging is useful not only to the original tagger, but also to other users. Folksonomy effectively involves voting by users of a classification system. In an arbitrary use of keyword-based distributed classification systems, a group of users may establish some separate tags. Web2.0 provides a timely solution to the problems of traditional classification. Folksonomy based on user-defined keywords for classification still has several problems. They include quality of users, the quality of labeling, semantic problems associated with keywords, the lack of constraints on keywords, the classification of poor results. Also, public classification lacks accuracy and tags have multiple meanings [3], [4], [6].

The purpose of information retrieval is to eliminate information overload [5]. The earliest and most extensive use involves the calculation of the TFIDF [9]. TFIDF calculates two main frequencies. The first is term frequency across a number of documents. Frequency typically represents importance. The second is document frequency.

The WordNet lexical database is a development of modern psychological theory. It is a set vocabulary, and is used to construct automated dictionary [7]. The WordNet lexical database was developed at Princeton University for cognitive research. The construction of WordNet mainly in English nouns, verbs, adjectives and adverbs, organized into sets of synonyms. WordNet has a wide range of applications and its website offers many open API (Application Program Interface) functions. The site is WordNet.net.

III. RESEARCH METHODOLOGY

This study developed an experimental prototype called FSBS (Folksonomy Social Bookmarking System). To test the effectiveness of FSBS prototypes, this study conducted a series of tests with various system settings and parameters. The system framework for the FSBS system we proposed consists of four modules: Folksonomy module, WordNet synonym analysis module, Data storage components, and user behaviors module. Figure 1 presents this system framework.

The system architecture of data storage components is based mainly various data files. The profile database contains personal records of the main users, including personally composed information and stored bookmark tags. The parameters database records mainly classifications as required parameters. TFIDF includes classification weights, and the settings of the follow-up classification criteria.

The Folksonomic system is the most important part of this study. The details of operational process as follows:

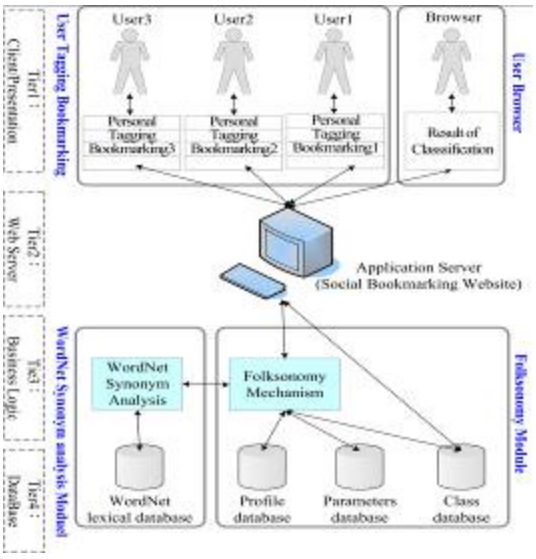

Figure 1. System framework for FSBS

(1) Read the data of user' bookmarks: before the Folksonomy process, we have to collect all the data of user' bookmarks.
(2) Read the parameters of classification: in the stage, we have to find out the TFIDF and weight from the parameter file of classification.

(3) Analysis of synonyms: in the stage, we have to send user' bookmarks to synonyms analysis module in WordNet. And we have to send back the result to Folksonomy classification module.
(4) Classification: First, we use CKIP (Chinese Knowledge and Information Processing) to proceed word segmentation. Second, we use WordNet to get the information of synonyms. Third, we use TFIDF to calculate the weight. Finally, the adjustment of the details for classification.
(5) Access the result of classification.

IV. IMPLEMENTATION AND EVALUATION OF SYSTEM

A. System Implementation

The FSBS system development environment and platform is based on convenience, effectiveness of implementation of internal references and bookmarks. JSP (Java Server Page) is adopted to establish these modules. The data storage components use the MySQL database.

The system analysis and implementation stage utilizes a method for developing prototype systems. Since prototyping is effective, fast and low-risk, it can be easily applied with advantages. The prototyping approach is the most commonly used and easiest method for developing in a Web-based system.

In order to test the feasibility of our FSBS prototype, a series of tests with various system settings and parameters was conducted. Research implications and issues of future works are discussed in the last section. The scope of the new bookmarks, their modification, deletion, and other functions, is tested. Also, many other design labels and special labeling input are tested to verify the fact that the developed classification mechanism has been fully implemented. Table 1 presents information concerning testing. When the system has been tested, the system and the experimental rules designed.

Table I. TEST DATA

Test target	Test Labels	Result of Test
Single word labels	"a";"好";"用"	Default values are not displayed, but if more than ten users are involved, the test labels are shown.
Duplication of single word labels	"aa";"aaa"; "bb"; "bbbb"	Default values are not displayed, but if more than ten users are involved, the test labels are shown.
Capitalizations conversion	"Blog"; "blog"	The two categories will involve the same category.
The classification of Chinese synonyms	"影片"; "影集"	Two labels belong to synonyms, they will involve the same category.
Inquiries about words segmentation	"好吃"; "好吃的"; "溫泉"; "溫泉之旅"	After the Chinese word segmentation processing, "好吃的" should be classified into "好吃" category; "溫泉之旅" belongs to the "溫泉" category.
The classification of exceptional words	"部落格"; "部落格介紹"	"部落格介紹" belongs to "部落格" category.

B. System Evaluation

In this study, the experiments arrange the actual users to use the FSBS system,. When users browse the FSBS prototype, the system adopts experimental design to differentiate between the experimental group and controls group. Control group adopt traditional Folksonomy; Experimental group adopt the method we proposed. After that, we compare the result of control group and experimental group.

User identities and backgrounds are beyond the scope of this study. Therefore, to reduce the number of users increase the convenience of sampling, background information was obtained from many several students and college students who participated in the experiments. A questionnaire survey was utilized to measure user satisfaction.

V. CONCLUSION

A. The effectiveness of classification mechanism

In this paper, we presented a new approach to the classification of website content by social bookmarking. We have shown how to design and implement such a system in practice and a FSBS prototype was developed to investigate its feasibility and usefulness. Our evaluation results are encouraging and the subject is worth further research. As a whole, this exploratory investigation try to solve the problems as follow:

(1) Synonyms problem: we use WordNet lexical database to support inquiries. It is useful to solve synonyms problem.
(2) number of words in the design label: We have no this kind of limitation.

(3) Improving the results of classification: To improve classification, this study develops a Folksnonomic weighting mechanism that includes the following rules - single word labels, duplication of single word labels, capitalizations conversion, the classification of Chinese synonyms, inquiries about words segmentation, and the classification of exceptional words. A scheduling mechanism is constructed, every six hours; operation rules are applied to classify information on labels.

B. The implementation and verification of classification system

(1) Actual system observation: in the study of the classification mechanism, the experiment ends when the number of labels reaches 164. Traditional public classification requires 253. The effectively reduction in the number of labels is 30%.
(2) User satisfaction survey: after completion of the experiment, the users were required to complete a satisfaction questionnaire. Statistical analysis of the results indicates that the proposed classification system exhibited significantly. The Folksonomic mechanism provides significant improvements over traditional Folksonomy, and helps users make information inquiries.

C. Future work

Overall, the results show that FSBS may help to reduce the necessary effort of content classification. Furthermore, the utility of the proposed approach can be extended and/or elaborated far beyond the scope of website content. It can also be used to overcome the drawbacks of using text mining technique on searching other semi-structured documents, such as journal papers. However, future work is still needed to improve its effectiveness and extend it into more practical utilization.

(1) Semantic issue: Future work should entail further semantic analysis, or establish labels that combine the automatic construction of ontology, to improve public classification.
(2) Integration of Chinese and English synonym: We hope to integrate the Chinese and English synonym to improve the integrity of the classification system.
(3) Integration of recommendation mechanism: If we can involve the various recommendation mechanisms in the future, such as users click flow analysis, labeling flow analysis, experts recommend mechanism or cooperative mechanisms. It will bring greater effectiveness.

ACKNOWLEDGMENTS

The authors would like to thank the financial support in Taiwan by the Taiwan National Science Council (NSC96-2416-H-033-004). And we will appreciate Ted Knoy for his editorial assistance.

REFERENCES

[1] D. Fichter, "Intranet Applications for Tagging and Folksonomies," *Online*, vol. 30(3), pp. 43-45, 2006.
[2] S. Golder, B.A. Huberman, "Usage Patterns of Collaborative Tagging Systems," *J. of Information Sci.*, vol. 32(2), pp. 198-208, 2006.
[3] L. Gordon-Murnane, "Social Bookmarking, Folksonomies, and Web 2.0 Tools," *Searcher-The Magazine for Database Professionals*, vol. 14(6), pp. 26-38, 2006.
[4] P. Heymann, G. Koutrika, H. Garcia-Molina, "Can Social Bookmarking Improve Web Search?" *Proceedings of the International Conference on Web Search and Web Data Mining*, pp. 195-206, 2008.
[5] M. Kobayashi, K. Takeda, "Information Retrieval on the Web," *ACM Computing Surveys*, vol. 32(2), pp. 144-173, 2000.
[6] A. Mathes, "Folksonomies - Cooperative Classification and Communication through Shared Metadata," http://www.adammathes.com/academic/computer-mediatedcommunication/folksonomies.html, 2004.
[7] G.A. Miller, "WordNet: A Lexical Database for English," *Communications of the ACM*, vol. 38(11), pp. 39-41, 1995.
[8] I. Ohmukai, M. Hamasaki, H. Takeda, "A Proposal of Community-based Folksonomy with RDF Metadata," *Proceedings of the 4th International Semantic Web Conference (ISWC2005)*, 2005.
[9] G. Salton, M. McGill, "Introduction to Modern Information Retrieval," McGraw-Hill, 1983.

Shih-Ming Pi is an associated professor of Information Management at Chung Yuan Christian University in Taiwan. His recent articles can be found in Communications of the ACM, European Journal of Information Systems, Journal of the Association for Information Systems, and other Chinese information management journals.

Hsiu-Li Liao is an Assistant Professor in the Department of Information Management at CYCU. She has published refereed papers in Computers & Education, Social Behavior and Personality, Lecture Notes in Computer Science, Issues in Information Systems, and other Chinese management journals. She is also a reviewer of six IS international journals.

Su-Houn Liu is a Professor of the Department of Information Management at CYCU in Taiwan. His recent publications can be found in Computers & Education, Social Behavior and Personality, International Jounal of Technolgy Management, Issues in Information Systems, and other Chinese management journals.

Chen-Wen Lin is a graduate student of the Department of Information Management at CYCU in Taiwan. His current research interests focus on information systems implementation and evaluation, management information system, and spam filtering.