TAG RECOMMENDATION BASED ON USER'S BEHAVIOR IN COLLABORATIVE TAGGING SYSTEMS

Nagehan Ilhan and Şule Gündüz Öğüdücü
Istanbul Technical University, Department of Computer Engineering, Maslak, Istanbul, 34469 Turkey

Keywords: Collaborative tagging, Recommender systems, Tag suggestions, Social network analysis.

Abstract: Social bookmarking Web sites allow users submitting their resources and labeling them with arbitrary keywords, called tags, to create folksonomies. Tag recommendation is an important element of collaborative tagging systems which aims at providing relevant information to users by proposing a set of tags to each newly posted resource. In this paper, we focus on the task of tag recommendation when a user examines a document based on the user's tagging behavior. We explore the use of this semantic relationship in modeling the user tagging behavior. The experiments are performed on the data set obtained from a social bookmarking site. Our experimental result show that our method is efficient in modeling users' tagging behavior and it can be used to recommend tags for resources.

1 INTRODUCTION

Collaborative tagging systems are popular tools for creating, collecting and sharing huge amounts of social data over the Web (Golder and Huberman, 2006). Social bookmarking services allow Web users to annotate the resources with freely chosen keywords called tags. The tags given by a user to a resource reflect the interest of the user in the resource as well as the understanding of the content of the resource. Most of the social bookmarking Web sites assist users during the labeling process by recommending tags. Recommending tags can employ on various purposes such as increasing the probability of a resource's getting annotated or reminding the user what a resource is about. There are numerous social bookmarking Web sites providing these services, the most popular being Delicious[1]. Delicious is a widely used social bookmarking service devoted to tag URL's. The aim of this work is to model the tagging behavior of users in order to recommend them personalized tags related to the document they are interested in.

In this paper, we propose a method to enrich the model of tagging behavior in a folksonomy by adding some semantics based on the WordNet hierarchy of concepts (Fellbaum, 1998). We focus on modeling users' tagging behavior effectively which in turn will increase the recommendation accuracy. Our model does not only consider previously used bookmarks of the users but also takes into account the content of the document. This feature is also helpful to handle cold-start situations. Our objective is firstly to extract tagging pattern of users by analyzing the similarity between user tags and the content of the document in order to represent this relationship between folksonomy tags and the content. The content of a document is divided in this study into five different components called document sections (e.g. page title, main content, heading 1 etc.). We find out effect rates of different document sections on user tagging behavior while she/he is bookmarking a Web page. Then, we calculate score points for each user that reflects the probability of choosing tags by a user that appear in a particular section of the document. We generate our recommendation set by considering the calculated rates of the user.

The rest of the paper organized as follows. We mentioned related works in Section 2. Our proposed method is introduced in Section 3. We then present our experiments and discuss results in Section 4. Finally, Section 5 concludes the paper.

2 RELATED WORK

There exist statistical investigations about the usage dynamics and tagging patterns of tag collections (Golder and Huberman, 2005)(Kipp and Campbell, 2007).

[1] http://del.icio.us.com/

Ilhan N. and Gündüz-Öğüdücü Ş..
TAG RECOMMENDATION BASED ON USER'S BEHAVIOR IN COLLABORATIVE TAGGING SYSTEMS.
DOI: 10.5220/0003151005700573
In *Proceedings of the 3rd International Conference on Agents and Artificial Intelligence* (ICAART-2011), pages 570-573
ISBN: 978-989-8425-40-9

In (Lee and Chun, 2007) content-based tag recommendation which uses graph representation is presented. Their system recommends the tags extracted from the content of a blog using an artificial neural network which uses WordNet and word frequencies in the training step. An example of content-based tag recommendation which uses graph representation is presented in (Lee and Chun, 2007). Their system recommends the tags extracted from the content of a blog using an artificial neural network which uses WordNet and word frequencies in the training step.

The authors in (Tatu et al., 2008) utilize information from resource content and the folksonomic structure of the graph. They use the graph to create a set of tags related to the resource and a set of tags related to the user. Then the system enrich tag vocabularies of the set of tags related to resource or user by WordNet based search for words that represent the same concept in order to recommend to the user. A method which creates resource related tags with the keywords found in the resource's title and extending them with the tags that co-occur with the base tags in the system is presented in (Lipczak et al., 2009). Existing tag recommendation studies use previous tags that has been assigned to the resource by other users. Thus, they become insufficient when a new resource appears. Our recommendation model utilize content of the Web document, hence new or frequently assigned resources does not alter our recommendation success.

3 PROPOSED METHOD

3.1 Analysis of Tagging Behavior

It can be assumed that Web pages can be represented by their text. In this study, this text is separated into five different sections: (1) main content for long texts in the body part of the document (C); (2) page title (P), (3) heading 1 (H1); (4) heading 2 (H2); and (5) the anchor text in the links (A). There are 6 heading tags available in HTML coding and H1 is the largest being at the top of the heading structure hierarchy. In the remaining part of this paper, dx_i denotes one of this five sections of a document d_i. A preprocessing step is performed which includes stop word removal and stemming of terms. The main content of a Web page is then represented by top-k terms that have the highest frequency among the other terms in the body part of the document. The terms in a section of the document are combined into a single vector:

$$\overrightarrow{dx_i} = (wx_1, f_{i1}), (wx_2, f_{i2}), \ldots, (wx_n, f_{in}) \quad (1)$$

where $wx_1, wx_2, \ldots, wx_n$ are terms that appear in the corresponding section dx_i and $f_{i1}, f_{i2}, \ldots, f_{in}$ are the frequencies of the terms. Thus, a Web document can be represented by 5 term vectors. Instead of commonly used TF-IDF (Term Frequency/Inverse Document Frequency) weighting scheme we used TF weighting in vector representations.

The tags assigned to a Web document are combined into a single tag vector:

$$\overrightarrow{tt_i} = (t_1, f_{i1}), (t_2, f_{i2}), \ldots, (t_l, f_{il}) \quad (2)$$

where $t_1, t_2, \ldots, t_l$ are tags assigned by users to document d_i and $f_{i1}, f_{i2}, \ldots, f_{il}$ are the frequencies of the corresponding tags in that document.

As stated earlier, the aim of this step is to find a relationship between terms appeared in the document and the tags assigned to it. For this reason, the similarity between each term vector and tag vector of the document is computed using the cosine similarity measure:

$$sim(\overrightarrow{dx_i}, \overrightarrow{tt_i}) = \frac{\overrightarrow{dx_i} \bullet \overrightarrow{tt_i}}{\| \overrightarrow{dx_i} \| \| \overrightarrow{tt_i} \|} \quad (3)$$

The second step of tag analysis comprises of determining the semantic relationship between the scope of a document and tags of this document using WordNet. Each term in each term vector of a document is converted into its hypernym and hyponym versions using WordNet. A term's hypernym is a general term whereas a hyponym is specific. The frequency f_{ij} of a term t_j in a term vector of d_i is mapped to its hypernyms/hyponyms $\{h_1, \ldots, h_j, \ldots, h_r\}$. The frequencies of synonym terms are determined in a similar way of hypernym/hyponym case. The similarity between each term vector and synonym tag vector is computed based on the cosine measure.

3.2 Personalized Tag Recommendations

We are given a set of users $\mathcal{U} = \{u_1, u_2, \ldots, u_N\}$, a set of Web pages $\mathcal{R} = \{d_1, d_2, \ldots, d_K\}$ and a set of tags $\mathcal{T} = \{t_1, t_2, \ldots, t_M\}$. In this paper, we will use the following notations:

- $tags(u_i) \subseteq \mathcal{T}$ is the set of tags used by user u_i.
- $tags(u_i, d_j) \subseteq tags(u_i)$ is the set of tags given by user u_i to a Web page d_j.
- $tags(d_j) \subseteq \mathcal{T}$ is the set of tags given to Web page d_j.
- $tags(dx_j) \subseteq tags(d_j)$ is the set of tags of Web page d_j that appear in the dx_j part of that page. Note that dx can be one of the five different sections of the document, such as main content, page title, h1, h2 or anchor text.

For each user u_i, a score is calculated to determine whether the user selects tags related to the content of the document and if so from which part of the document or (s)he assigns tags from her/his own vocabulary independent from the content of the document. First, a score value is computed for each document section-user pair which is the probability of choosing tags by that user that appear in dx section in a document:

$$score_{dx_j,u_i} = \frac{|tags(u_i,d_j) \cap tags(dx_j)|}{|tags(u_i,d_j)|} \quad (4)$$

Each document section dx_j contributes to the final set of tag recommendations with $n_{x,j}$ tags which is proportional to the score value of this section. Let the final set of recommendations consists of k tags. The number of tags in the final recommendations set that are part of dx_j is:

$$n_{x,j} = \frac{score_{dx_j,u_i}}{\sum_x score_{dx_j,u_i}} \times k \quad (5)$$

A recommendation set $R(u_i,dx_j)$ is formed for user u_i with $n_{x,j}$ tags that have the highest frequency in term vector dx_j. Finally, user u_i is provided with a set of k recommended tags $R(u_i,d_j)$ for a particular Web document d_j:

$$R(u_i,d_j) = \bigcup_x R(u_i,dx_j) \quad (6)$$

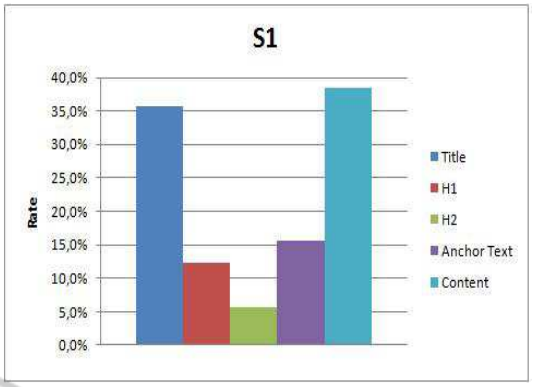

Figure 1: Similarity values between term and tag vectors of documents.

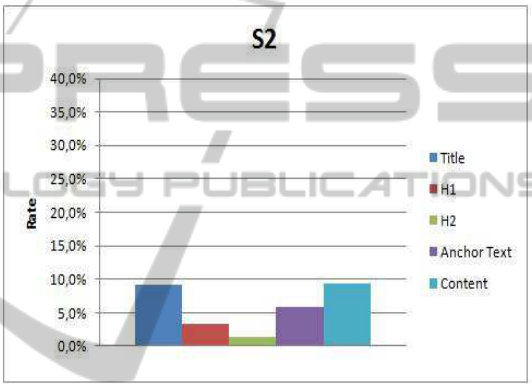

Figure 2: Similarity values between hypernym term and tag vectors of documents.

4 EXPERIMENTAL RESULTS

4.1 Data Preperation

The experiments are performed on two different datasets which are collected from the Delicious Web site. The details of the datasets are given in Table 1.

Table 1: Dataset Information.

	Urls	Users	Tags
Dataset1	1013	45654	42169
Dataset2(train)	25122	1020	82626
Dataset2 (test)	25880	1020	85321

Each Web document in each dataset is parsed to remove HTML tagging. The same preprocessing step is performed on each Web document and the set of user tags by applying a stop word removal and Porter's stemming algorithm (Jones and Willet, 1997). Each Web document is divided into 5 sections by representing each section by a term vector as explained in Section 3.1. Hypernym, hyponym and synonym vectors of each term vector of each document are constructed using WordNet (Fellbaum, 1998). Then the cosine similarity between each (hypernym/hyponym/synonym) term vector and tag vector of documents is calculated.

For simplification, we present the following experimental settings, S1-S3. In S1, the cosine similarity between each term vector dx_j and the tag vector tt_j of d_j is calculated using Eq. 3. The cosine similarity between hypernyms of term vectors and tag vectors of documents is calculated in S2. The synonym of term vector is constructed for S3 and the cosine similarity is calculated between synonym term vectors and tag vectors. In each setup, the similarity values are averaged over the entire set of documents in Dataset1.

Fig. 1, 2 and 3 show the similarity results for S1, S2 and S3, respectively. The similarity between term vector obtained from the content and the tag vector is higher than the similarities between the remaining term vectors and tag vector. The similarity value obtained by using page title is close to the similarity value of using content term vector.

Based on these result, a hybrid recommendation

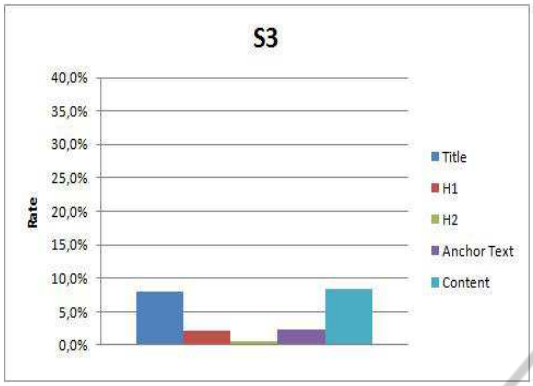

Figure 3: Similarity values between synonym term and tag vectors of documents.

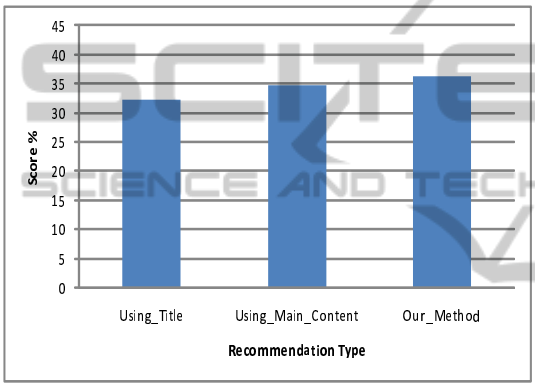

Figure 4: Recommendation Results.

set for users by only calculating the users' tagging scores on page title and content of Web documents. The recommendation set consists of 10 tags (k) which is empirically determined. Recommendation results given in Figure 4 support our prior review on similarities between tags and document content. Recommendation set generated by just using most frequent content terms outperforms the set generated by using most frequent page title terms. However, the recommendation rate of our hybrid recommendation set performs better than both sets.

5 CONCLUSIONS

In this paper, we considered the content of a resource as tag source in creating the recommendation set. We investigated the similarity between different parts of the content of the resource with the tags assigned to the resources. Our main aim was to determine which part of the document has valuable tags and can be a potential tag source. It is also examined that if the semantically related terms of the content can be used as tag source or not. Results indicate that users tend to choose terms that appear in the content of the document rather than selecting terms that are semantically similar to the terms in the document.

Afterwards, we proposed a recommendation model which rates users' tag selection to assign resources. These rates measure the likeness of user tags with different parts of the document and represents which part of the document's text is selected by the user. Then, our recommendation set is generated by considering those rates. Results show that users are more likely to select tags from main content when compared to titles and our proposed recommendation technique outperforms the recommendation methods in which tags are created using only the main content and the title terms.

ACKNOWLEDGEMENTS

The authors are supported by the Scientific and Technological Research Council of Turkey (TUBITAK) EEEAG project 110E027.

REFERENCES

Fellbaum, C., editor (1998). *WordNet: an electronic lexical database*. MIT Press.

Golder, S. and Huberman, B. A. (2005). The structure of collaborative tagging systems. *Journal of Information Science*, 32(2):198–208.

Golder, S. A. and Huberman, B. A. (2006). Usage patterns of collaborative tagging systems. *J. Inf. Sci.*, 32(2):198–208.

Jones, K. S. and Willet, P., editors (1997). *Readings in Information Retrieval*. Morgan Kaufmann Publishersand Inc., San Francisco, California.

Kipp, M. E. and Campbell, G. D. (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.

Lee, S. O. K. and Chun, A. H. W. (2007). Automatic tag recommendation for the web 2.0 blogosphere using collaborative tagging and hybrid ann semantic structures. In *ACOS'07*, pages 88–93.

Lipczak, M., Hu, Y., Kollet, Y., and Milios, E. (2009). Tag sources for recommendation in collaborative tagging systems. In Eisterlehner, F., Hotho, A., and Jschke, R., editors, *ECML PKDD Discovery Challenge 2009 (DC09)*, volume 497 of *CEUR-WS.org*, pages 157–172.

Tatu, M., Srikanth, M., and D'Silva, T. (2008). Rsdc'08: Tag recommendations using bookmark content. In *Proceedings of the ECML/PKDD 2008*, pages 96–107.