Browse with a Social Web Directory

Hao Huang[†,‡] Yunjun Gao[†] Lu Chen[†] Rui Li[§] Kevin Chiew[♯] Qinming He[†]

[†]College of Computer Science, Zhejiang University, China
[‡]School of Computing, National University of Singapore, Singapore
[§]Computer Science Department, University of Illinois at Urbana-Champaign, USA
[♯]School of Engineering, Tan Tao University, Vietnam
[†]{hh14311, gaoyj, chenl, hqm}@zju.edu.cn, [§]ruili1@uiuc.edu, [♯]kevin.chiew@ttu.edu.vn

ABSTRACT

Browse with either web directories or social bookmarks is an important complementation to search by keywords in web information retrieval. To improve users' browse experiences and facilitate the web directory construction, in this paper, we propose a novel browse system called Social Web Directory (SWD for short) by integrating web directories and social bookmarks. In SWD, (1) web pages are automatically categorized to a hierarchical structure to be retrieved efficiently, and (2) the popular web pages, hottest tags, and expert users in each category are ranked to help users find information more conveniently. Extensive experimental results demonstrate the effectiveness of our SWD system.

Categories and Subject Descriptors

H.4.3 [**Information Systems Applications**]: Communications Applications—*Information browsers*

Keywords

Browse system; web page categorization; social ranking

1. INTRODUCTION

Besides search by keywords, browse systems provide an alternative way for web information retrieval [6], especially when there is a lack of appropriate keywords or comprehension and learning processes are required during the retrieval tasks. Thus far, we have seen two generations of browse systems, i.e., (1) web directories, such as Yahoo Web Directory and Open Directory Project (www.dmoz.org), in which web pages are collected and categorized manually by skillful editors, and (2) social bookmarking and tagging services, such as Del.icio.us and citeulike.com, which enable users to annotate web pages by themselves. Nonetheless, these browse systems still have a few minor flaws in two aspects, namely (1) lack of an automatic categorization on web pages. To be specific, the maintenance of a web directory requires extensive human efforts, while the social bookmarks and tags provided by various individuals often suffer from noisy and ambiguous annotations as well as various personal categorizations [4]; and (2) lack of page ranking, although web pages ranked by their popularity often help users find information more efficiently and conveniently.

Aiming at improving users' browse experiences and facilitating the web directory construction, we propose a novel browse system known as SWD (**S**ocial **W**eb **D**irectory), which adopts the social information (e.g., web page tags) from Del.icio.us to automatically categorize web pages with the help of the well-edited hierarchical catalog structure in Open Directory Project, and ranks the web pages in each category by estimating their popularity. Furthermore, SWD can also identify expert users and the hottest tags in each category which enrich the category's description, and thus can support community based-exploratory browse [5].

Extensive experiments have been conducted on 18,304 web pages from Open Directory Project together with 604,032 annotations from Del.icio.us, and the results show that compared with a straightforward method, i.e., using the pure content of each web page for categorization, our SWD system improves 27.9% of accurary on web page categorization and 72% of efficiency. Moreover, our case study confirms that the expert users and hottest tags in each category identified by SWD are of high quality and can help users explore the category information more effectively.

2. SYSTEM OVERVIEW

The system contains three parts, i.e., a training data depository, and two models constructed with this depository.

The training data depository is automatically obtained by merging the well-edited web page catalog of Open Directory Project (ODP for short) and web page tags from Del.icio.us. Specifically, the web pages of which the URLs are within both ODP and Del.icio.us are collected into our training data depository. Then, any web page in the depository has information including users who collected it, tags used to annotate it, and the category it belongs to.

With the training data depository, there are two models constructed in SWD, i.e., (1) the web page categorization model, which classifies any newly collected web page into an existing category according to its tags, and (2) the social ranking model, which provides a ranking list of the popular web pages, hottest tags, and expert users in each category.

Fig. 1 shows two snapshots of our SWD system, in which Part A lists the top categories, Part C the sub-categories, Parts B & D the corresponding web pages, Part E the hottest tags in current category, and Part F the expert users.

SIGIR'13, July 28–August 1, 2013, Dublin, Ireland.
Copyright 2013 ACM 978-1-4503-2034-4/13/07 ...$15.00.

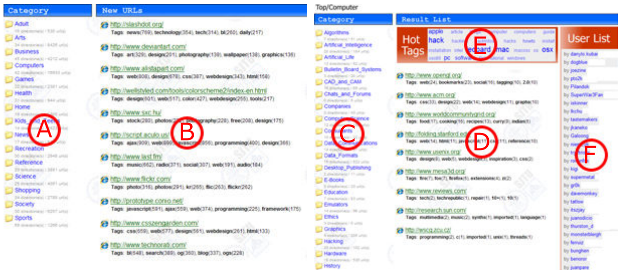

Figure 1: Interface of SWD system

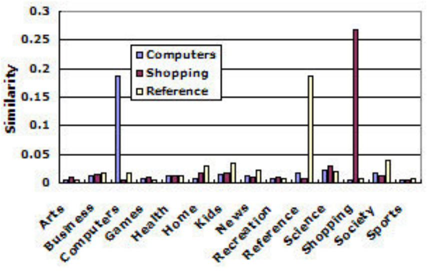

Figure 2: Average Similarity between Categories

3. WEB PAGE CATEGORIZATION MODEL

Compared with previous methods which based on web page contents and URLs [2,10] , we propose a more efficient approach to web page categorization by only investigating the web page tags annotated by users. This is motivated by our preliminary study on the tags, which reveal that these tags can usually describe the web page accurately, and meanwhile, web pages under the same category often have similar tags and vice versa. For example, Fig. 2 illustrates the average tag-based similarity between the web pages in different categories, which is defined as

$$\overline{sim}(c_x, c_y) = \frac{\sum_{p_i \in c_x, p_j \in c_y} cosine(p_i, p_j)}{|c_x| \times |c_y|}$$

where $|c_x|$ and $|c_y|$ denote the number of web pages in categories c_x and c_y, p_i and p_j are web pages in each category, $cosine(p_i, p_j) = \frac{p_i \times p_j}{|p_i| \times |p_j|} = \frac{\sum_k w_{ki} \times w_{kj}}{\sqrt{\sum_k w_{ki}^2} \times \sqrt{\sum_k w_{kj}^2}}$. Here, w_{ki} represents the tag T_k's weight in a web page p_i, and can be calculated by the *tf-idf* metric, i.e.,

$$w_{ki} = \frac{freq(T_k, p_i)}{\max_{1 \leqslant l \leqslant t} freq(T_l, p_i)} \times \lg \frac{N}{n_k}$$

in which $freq(T_k, p_i)$ is the number of times the tag T_k is used to annotate web page p_i by different users, N the total number of web pages, n_k the number of the web pages containing tag T_k, and t the total number of web page tags collected in SWD. From Fig. 2, we can learn that tags are discriminative features for categorization since the average tag-based similarity between web pages of the same category is much higher than those from different categories.

Based on the above discussion, we can utilize web page tags to represent each web page p_i as a vector below.

$$p_i = (w_{1i}, w_{2i}, \ldots, w_{ki}, \ldots, w_{ti}). \qquad (1)$$

Then, classification methods such as Naive Bayes, SVM, and KNN can be adopted to help SWD categorize web pages by using these vectors.

4. SOCIAL RANKING MODEL

In traditional web directories, web pages of each category are listed randomly or chronologically, resulting in that web pages of high quality are often hidden in the mass collections and thus are inconvenient for retrieval. In social bookmark fields, although a few ranking algorithms are proposed to provide the most popular web pages in each category based on the web page tags [1,9], they often surfer from a relatively high time complexity or just offer a static rank which is not related to the category.

To avoid the above limitations of existing browse systems, we provide a social ranking model for our SWD system, which can rank the popular web pages as well as expert users and the hottest tags of each category efficiently.

4.1 Popular Web Pages

Since each web page p_i has a set of associated users and a set of tags, it can also be represented as (1) a binary vector p_i^{user} of which each element p_{ir}^{user} ($1 \leqslant r \leqslant u$, u is the total number of users who have collected web pages for SWD) indicates whether web page p_i is collected by the rth user, and (2) a binary vector p_i^{tag} of which each element p_{ik}^{tag} ($1 \leqslant k \leqslant t$) denotes whether web page p_i has the tag T_k. With these two binary vectors, we measure web pages' popularity based on the following two features.

User Count. If a web page is collected by many users, it is often of a high quality and should be recommended to the other users. Hence, we identify these kinds of web pages by the user count $UC(p_i)$ of each web page p_i, where

$$UC(p_i) = \sum\nolimits_{1 \leqslant r \leqslant u} p_{ir}^{user}.$$

Topical Similarity. The topical similarity $TS(p_i|c_x)$ between a page p_i and a given category c_x also indicates whether this web page should have a high rank in the category.

$$TS(p_i|c_x) = consine(p_i^{tag}, c_x^{tag})$$

where c_x^{tag} is the category vector of c_x represented by tags, of which each element $c_{xk}^{tag} = \sum_i I(p_i \in c_x) \times p_{ik}^{tag}$, here $1 \leqslant k \leqslant t$ and $I(\cdot)$ is the indicator function.

Then the popularity $Pop(p_i|c_x)$ of a page p_i in a category c_x can be calculated as

$$Pop(p_i|c_x) = \frac{TS(p_i|c_x)}{\max_{p_j \in c_x} TS(p_j|c_x)} + \frac{UC(p_i)}{\max_{p_j \in c_x} UC(p_j)}.$$

4.2 Expert Users

Listing up expert users can help a new user find useful web pages via their collected web pages, and meanwhile stimulate users to collect new web pages of high qualities into our system. To this end, we present the following three metrics to help identify the expert users of each category.

Collection Quality. This metric checks whether the web pages collected by a user u_r are useful in a category c_x.

$$Quality(u_r|c_x) = \sum\nolimits_{p_i \in c_x, p_i \in u_r} Pop(p_i|c_x)$$

where $p_i \in u_r$ denotes that user u_r collects web page p_i.

Table 1: Categorization performance comparison

Method	NB		SVM	
	Mi-F	Ma-F	Mi-F	Ma-F
Pure Content Based	0.565	0.431	0.626	0.576
Description Based	0.662	0.505	0.672	0.636
Proposed SWD System	0.790	0.691	0.801	0.775
Combined Method	0.792	0.662	0.811	0.790

Interests of Users. The metric examines whether a user u_r is a specialist to a category c_x by estimating the relevance between his collection and the category.

$$Interest(u_r|c_x) = cosine(u_r^{tag}, c_x^{tag})$$

where u_r^{tag} is a binary vector of which each element u_{rk}^{tag} $(1 \leqslant k \leqslant t)$ indicates whether user u_r has used tag T_k.

Contributions of Users. This metric is provided to encourage users to collect new web pages of high qualities for our system. The contribution $Con(u_r|c_x)$ of a user u_r to a category c_x is defined as

$$Con(u_r|c_x) = \sum_{p_i \in c_x, p_i \in u_r} \left(\frac{UC(p_i)}{2^d} + \frac{TS(p_i|c_x)}{d} \right)$$

where d denotes that u_r is the dth user who collected p_i.

4.3 Hottest Tags

The hottest tags in each category help describe the category and can also navigate users to their target web pages. The hotness $Hot(T_k|c_x)$ of the tag T_k $(1 \leqslant k \leqslant t)$ for a category c_x can be estimated by investigating the tag's frequency of occurrences in this category and in all the categories, formally,

$$Hot(T_k|c_x) = \frac{\sum\limits_{p_i \in c_x} freq(T_k, p_i)}{\sum\limits_{T_l \in c_x} \sum\limits_{p_j \in c_x} freq(T_l, p_j)} \times \frac{\sum\limits_{p_i \in c_x} freq(T_k, p_i)}{\sum\limits_{y} \sum\limits_{p_j \in c_y} freq(T_k, p_j)}$$

where $T_l \in c_x$ means that a tag T_l appears in a category c_x.

5. EXPERIMENTAL EVALUATION

5.1 Experiments on Web Page Categorization

Experimental Setup. We collect 18,304 web pages over 60 categories and sub-categories from ODP together their 7,691 tags which are used as annotations for 604,032 times.

We represent each web page as a vector by Eq. (1) (see Section 3), and respectively utilize Naive Bayes Classifier (NB for short) and SVM as the classification methods.

To evaluate the average categorization performance of our SWD system across all categories, we adopt the micro- and macro-average F-scores (Mi-F and Ma-F for short) as the performance metrics, and perform 5-fold cross-validation to reduce the uncertainty of training data and test data.

Performance Study. Besides our SWD system, we also report the web page categorization performance of the following three methods, i.e., (1) the pure content based method, which straightforwardly uses the text content of each web page for the categorization, (2) a state-of-the-art description based method [8], and (3) a combined method, which combines the categorization results of our proposed method and the pure content based method as follows.

$$Target(p_i) = \arg\max_{c_x} \Big(\alpha \cdot P_{cont}(c_x|p_i) + (1-\alpha) \cdot P_{SWD}(c_x|p_i) \Big)$$

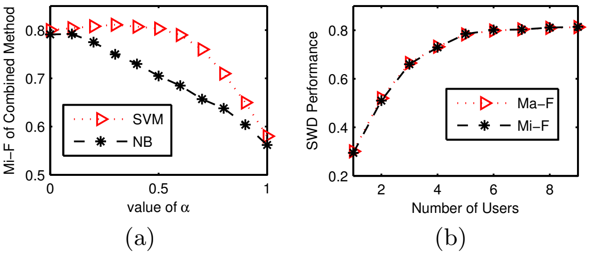

Figure 3: (a) Mi-F of combined method vs. α. (b) SWD categorization performance vs. user number

where $P_{cont}(c_x|p_i)$ and $P_{SWD}(c_x|p_i)$ denote the probability of a web page p_i belonging to a category c_x by using the pure content based method and our proposed method.

Table 1 shows that our propose method achieves a 27.9% improvement over the pure content based method and 19.1% over the description based method. In addition, the combined method gets a slight improvement over our method since it can benefits from both the two methods combined.

To investigate which method provides more profit in this combination, we vary the parameter α from 0 to 1. When $\alpha = 0$, the combined method is equivalent to ours, and when $\alpha = 1$, it is equivalent to the description based method. Fig. 3(a) shows that the combined method often performs better when the weight of our method is higher, indicating that our method contributes more in the combination.

Moreover, previous studies have shown that most of classification algorithms including NB and SVM are often sensitive to the document length when using the content of each web page for categorization [7]. However, since the number of each web page's tags is much less than that of the words within each web page's content, our proposed approach to web page categorization is more efficient on both time and space. Taking the NB algorithm as an example, experimental results demonstrate that it saves 90% in terms of space and 72% in terms of running time.

Effect of Annotation Amount. The above performance study verifies the effectiveness and efficiency of our SWD system on web page categorization. In what follows, we address another concern, i.e., whether the categorization are stable and at least how many annotations should be collected by SWD to guarantee its categorization performance.

To this end, we examine how the Ma-F and Mi-F of our categorization results are affected by the increasing number of tags when more and more users are involved, and depict the result in Fig. 3(b), in which the x-axis indicates that the first r users' annotations are used for web page categorization. The figure shows that (1) the categorization accuracy improves much on the arrival of each new user when the number of users involved is small, and (2) improves not very

Table 2: The hottest tags in 4 categories

System	Security	Programming	Internet
tabletpc	firewall	java	webdesign
tablet	network	python	blog
amiga	vpn	develop	css
sinclair	Linux	lisp	flash
arm	sysadmin	opensource	design

Table 3: Popular web pages in 3 categories and collected by expert users

Category: Java	Category: Artificial Intelligence	Category: Opensource
http://www.eclipse.org	http://www.agentlink.org/	http://sourceforge.net
http://java.sun.com	http://www.aiwisdom.com/	http://www.mozilla.org
http://pmd.sourceforge.net	http://bnj.sourceforge.net/	http://www.mozillazine.org
http://www.lowagie.com/iText	http://www.abelard.org/turpap/turpap.htm	http://www.gnu.org
http://www.onjava.com	http://www.imagination-engines.com/	http://www.opensource.org
Expert User of "Java": t****v	Expert User of "Java": a***o	Expert User of "Java": s*****3
http://java.sun.com	http://java.sun.com	http://java.sun.com
http://ant.apache.org	http://www.eclipse.org	http://ant.apache.org
http://www.eclipse.org	http://pmd.sourceforge.net	http://www.eclipse.org
http://pmd.sourceforge.net	http://jakarta.apache.org/cactus	http://jakarta.apache.org
http://argouml.tigris.org	http://argouml.tigris.org	http://java.sun.com/products/jsp

much and tends to be stable as the number of users increases. Similar as the conclusion in [3], this observation can be explained by the fact that the distribution of tags of a specific web page tends to be stable during the running of the social bookmarking systems, e.g., Del.icio.us.

5.2 Case Study on Social Ranking Model

In this experiment, we conduct some case studies to confirm the effectiveness of our social ranking model. Towards this, we collect a sample of Del.icio.us data consisting of 2,879,614 tags made by 10,109 different users on 690,482 different web pages which are automatically categorized to the hierarchical catalog of ODP by our SWD system.

In SWD, when users are browsing information of a particular category, we will show them the URLs of the most popular web pages, the hottest tags, and the expert users, all of which help enrich the description of this category and can navigate users to find the target information more efficiently and conveniently. For example, Table 2 has listed the top 5 hottest tags of 4 sub-categories in the "Computer" category, while Table 3 has listed the URLs of the most popular web pages in of another 3 sub-categories, and the top URLs collected by 3 expert users of the "Java" sub-category. From Tables 2 and 3, we can observe that (1) the hottest tags seem reasonable, these tags (e.g., "tabletpc" for "System" and "css" for "internet") summarize the hottest topics in each domain; (2) the URLs really correspond to authority web pages in each domain; and (3) the expert users always bookmark web pages of high qualities in each domain, which are very helpful for those rookie users who are interested in this domain. The case study has been also conducted on the other categories and sub-categories, the conclusion is similar to the above observation.

6. CONCLUSION

We proposed a framework of novel browsing system called Social Web Directory (SWD), which can automatically categorize web pages collected and tagged by every user and enable them to browse these web pages in an efficient way. By taking advantages of the well-edited hierarchial catalog structure in ODP and the social bookmarks in Del.icio.us, our SWD system can make up the flaws of previous browsing systems, and has the following three characteristics. (1) The system constructs a web directory automatically with the help of social bookmarks. In other words, the maintenance of a web directory does not require skilled editors to pay extensive efforts any more because our SWD system enables every user to contribute to the web directory and locate the desired resources easily. (2) The system provides a ranking list of popular web page in each category. With this ranking list, users can easily find the most useful web pages in a certain category with much less clicks and content reviews. (3) The system also supports community-based exploratory browsing since it offers the hottest tags and expert users in each category, which enrich the category description and make the browse of community information easier.

7. ACKNOWLEDGMENTS

This work was partly supported by the National Key Technologies R&D Program (Grant No.: 2011BAD21B02), NSFC 61003049, the Fundamental Research Funds for the Central Universities (Grant No.: 2012QNA5018, 2013QNA5020), and the Key Project of ZJU Excellent Young Teacher Fund.

8. REFERENCES

[1] S. Bao, G. Xue, X. Wu, Y. Yu, B. Fei, and Z. Su. Optimizing web search using social annotations. In *WWW*, pages 501–510, 2007.

[2] L. Blanco, N. Dalvi, and A. Machanavajjhala. Highly efficient algorithms for structural clustering of large websites. In *WWW*, pages 437–446, 2011.

[3] H. Halpin, V. Robu, and H. Shepherd. The complex dynamics of collaborative tagging. In *WWW*, pages 211–220, 2007.

[4] R. Li, S. Bao, Y. Yu, B. Fei, and Z. Su. Towards effective browsing of large scale social annotations. In *WWW*, pages 943–952, 2007.

[5] G. Marchionini. Exploratory search: From finding to understanding. *Communications of the ACM*, 49(4):41–46, 2006.

[6] C. Olston and E. H. Chi. Scenttrails: Integrating browsing and searching on the web. *ACM Transactions on Computer-Human Interaction*, 10(3):177–197, 2003.

[7] X. Qi and B. D. Davison. Web page classification: Features and algorithms. *ACM Computing Surveys*, 41(2):12, 2009.

[8] D. Shen, Z. Chen, Q. Yang, H.-J. Zeng, B. Zhang, Y. Lu, and W.-Y. Ma. Web-page classification through summarization. In *SIGIR*, pages 242–249, 2004.

[9] Y. Song, L. Zhang, and C. L. Giles. Automatic tag recommendation algorithms for social recommender systems. *ACM Transactions on the Web*, 5(1):4, 2011.

[10] G.-R. Xue, Y. Yu, D. Shen, Q. Yang, H.-J. Zeng, and Z. Chen. Reinforcing web-object categorization through interrelationships. *Data Mining and Knowledge Discovery*, 12(2-3):229–248, 2006.