Improving Web Search by Categorization, Clustering, and Personalization

Dengya Zhu, and Heinz Dreher

CBS and DEBII, Curtin University of Technology, GPO Box U1987, Perth, WA Australia
dengya.zhu@postgrad.curtin.edu.au, h.dreher@curtin.edu.au

This research combines Web snippet[1] categorization, clustering and personalization techniques to recommend relevant results to users. RIB – Recommender Intelligent Browser which categorizes Web snippets using socially constructed Web directory such as the Open Directory Project (ODP) is to be developed. By comparing the similarities between the semantics of each ODP category represented by the category-documents and the Web snippets, the Web snippets are organized into a hierarchy. Meanwhile, the Web snippets are clustered to boost the quality of the categorization. Based on an automatically formed user profile which takes into consideration desktop computer information and concept drift, the proposed search strategy recommends relevant search results to users. This research also intends to verify text categorization, clustering, and feature selection algorithms in the context where only Web snippets are available.

Keywords: text categorization, clustering, personalization, Web searching, Web snippets.

1 Introduction

The low quality of Web search [1] in terms of *recall* and *precision* stems from
1) the synonymous and polysemous characteristics of natural languages [2];
2) information overload on the Web [3, 4];
3) the imperfection of the information retrieval models so far developed; and
4) the lack of consideration of personal search interests and preferences [5, 6].

Text categorization [7] and clustering [8] are predominant approaches used to address problems of large amounts of information, and the challenges resulting from the polysemous characteristics of natural languages. Text categorization, or supervised learning, is the automatic assigning of predefined categories to free documents [9], while document clustering, or unsupervised learning, tries to discover groups in a document set such that similar documents are grouped together. For text categorization, the main issue is that it is expensive to obtain sufficient human edited training data. The main challenge for clustering algorithms is that the automatically

[1] A Web snippet, returned from search engines, contains only the title of a Web page and an optional very short (less than 30 words) description of the page.

formed cluster hierarchy may mismatch the human mental model [4, 10]. Furthermore, when only Web snippets, which are not as informative as full text document, are available, the developed algorithms for text categorization/clustering have not been sufficiently verified. This lack of 'informativeness' also makes it difficult to judge the relevance of these snippets of information, while relevance judgment is at the core of information retrieval [11].

Personalization is regarded as a promising approach to improve the relevance of Web search results because it concerns not only retrieval based on literally relevant information, but also a user's information consumption patterns, searching strategies, and applications used [12]. There are two main issues for personalized searching: concept drift [13, 14]; and privacy protection [15].

To approach the above issues, *RIB* – Recommender Intelligent Browser is proposed. The main purpose of *RIB* is to combine text categorization and clustering techniques to address synonymy, polysemy, and information overload problems by re-ranking, hierarchically organizing, and ontologically filtering returned Web snippets; to personalize Web search results by means of building a user profile based on a reference ontology - "a shared taxonomy of entities" [16] - created from a Web directory (such as the ODP); and taking search concept drift into consideration. *RIB* will recommend to users the re-ranked relevant results according to the user profile.
The contribution of this paper is twofold. First, a new approach to boost the quality of Web snippet categorization is proposed. The approach first estimates the *inter-similarities* between the Web snippets and the semantic of categories of an ontology; and then estimates the *intra-similarities* among the Web snippets to form some clusters which are used to boost the quality of categorization. Second, *RIB*, a novel Web information retrieval approach aims at recommending refined results to users based on automatically learned user profiles and ontologically filtering search results. *RIB* is to be developed and its performance in terms of *precision* is expected to comparable with or superior in some way to the results of *Windows Live Search API*.

2 Related Work

Text Categorization. Text categorization automatically assigns predefined categories to free documents [9]. Klas and Fuhr [17] use tf-idf weighting scheme [18] and probabilistic retrieval model to classify Web documents under the hierarchical structure of *Yahoo! Web Directory*. The texts of all documents belonging to a category are concatenated to form a so-called megadocument. To classify a document, the first n best terms (according to their idf values) are selected as a query vector. [19] proposes to disambiguate single-term queries by clustering and categorizing Web search results based on the meanings of WordNet for the queries.

The ODP categories are also used to classify Web snippets [10, 20]. The semantic aspects of the ODP categories are extracted, and category-documents are formed based on the extracted semantic characteristics of the categories. A special search browser is being developed to obtain Web snippets by utilizing *Yahoo! Search Web Service API*. Similarities between vectors represent Web snippets and the category-documents are compared. A majority voting strategy is used to assign a Web snippet

to the proper category without overlapping. One weakness of the research is while the *precision* is improved, there is a decrease in *recall*.

Web Snippet Clustering. One of the early works on Web snippet clustering is *Scatter/Gather* [21] which uses a partitional algorithm named Fractionation. It is found in the research that search results clustering can significantly improve similarity search ranking. *Grouper* [22] is another example of early work on clustering Web search results. Zeng et al. [23] propose the Web snippets clustering problem can be dealt with as a salient phrase ranking problem. The Web documents are assigned to relevant salient phrases to form candidate clusters, which are then merged to form the final clusters.

Personalization. Pitkow et al. [12] use the information space of the ODP to represent their user model. Again using the ODP, [1] creates a user profile according to a reference ontology in which each concept has a weight reflecting the user's interest in that concept. URLs visited by a user are periodically analyzed and then classified into concepts in the reference ontology. Chirita et al. [5] also suggest using the ODP metadata and combining a complex distance function with Google *PageRank* to achieve a high quality personalized Web search. Godoy and Amandi [6] propose an incremental, unsupervised Web Document Conceptual Clustering algorithm to set up user profiles. They use kNN to determine the similarities between concepts in user profiles and Web pages.

3 Conceptual Framework of RIB

RIB intends to investigate how does the use of Web snippet categorization and personalization enhance the relevance of Web search results by comparing three sets of search results:
1) the results directly obtained from meta-search engines [24, 25];
2) the results categorized without considering the clustered results; and
3) the categorized results refined with clustered results.
We also want to check to what degree the combination of categorization and clustering boosts the performance (in terms of *recall*, *precision*, and *F1*) of Web snippet categorization. We compare the categorized results of Support Vector Machines, k-Nearest Neighbors , and naïve Bayesian, with and without combining the results clustered by LSI [2], k-means [8], and Expectation Maximization clustering algorithms. The conceptual framework of *RIB* is illustrated in Fig.1. Obviously, *RIB* is not going to simply put all the algorithms together that will do nothing better except dramatically increase the computational complexity. The algorithms are mentioned here because one purpose of this research is to evaluate the effectiveness of the algorithms for Web snippets.

Meta search engine. The Meta-search engine obtains search results directly *from Yahoo! Search Web Service API* or *Windows Live Search API* after an application ID is applied. Both search APIs allow developers to retrieve from their Web databases

directly. For non-commercial licenses, the maximum number of results per query for Yahoo! is 100; and Microsoft API can return up to 1000 results. In this research, *Windows Live Search API* is employed because it provides full-size result sets the same as all the popular search engines, providing an opportunity to make a real-word comparison between *RIB* and Microsoft Live Search.

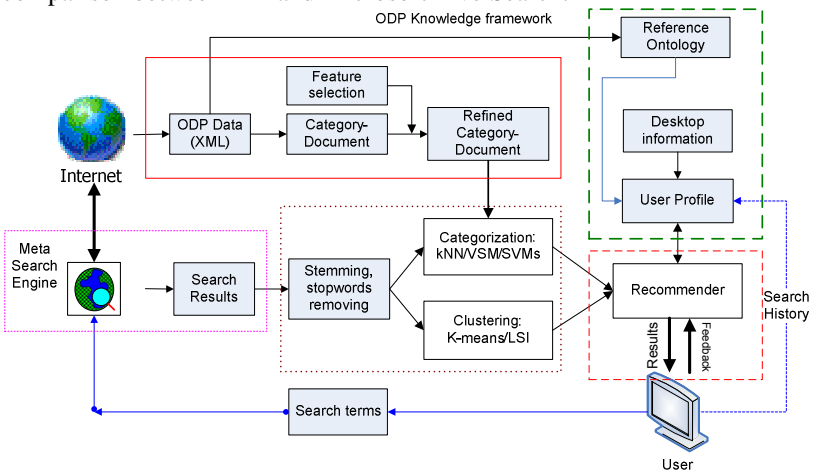


Fig. 1. The conceptual framework of Recommender Intelligent Browser

The Category-document extraction and feature selection. The ODP is selected as a predefined knowledge hierarchy because it is the largest and most comprehensive Web hierarchy edited by humans. A category-document is created based on these two files [10]. The category-document set extracted from the ODP is refined by feature selection algorithms [7, 26] such as χ^2, Mutual Information, Odds Ratio, and Information Gain [7]. Data from *structure.rdf* is used to map the ODP knowledge hierarchy to a reference ontology, which will represent users' search preferences.

Categorization/clustering algorithms. Lucene [27] is used to calculate similarities between documents and queries. A modified *k* majority voting algorithm has already been developed by Zhu [10] and can be used in this research. Naïve Bayesian, and k-means clustering algorithms are developed using the C# programming language.

Categorization creates some groups with distinct topics. The number of groups is to be used as *k* for the k-means algorithms because how to decide *k* is always a nontrivial problem for k-means algorithms [8].

User profile Creation. Desktop computer information, indexed by *Google Desktop Search SDK*, is used to initialize the user profile. For each of the indexed documents, the similarities sim ($\mathbf{d}_j$, $\mathbf{c}_i$) between a document $\mathbf{d}_j$ in a personal computer and a category-document representing a category $\mathbf{c}_i$ in the ODP are estimated by Lucene. When a Web page is visited, the time factor is considered [28]. The impact of concept drift will be a weighting factor which represents user search preferences [28]. Let w_i be the weight of concept $\mathbf{c}_i$ in the profile, and the width of slide time window is 400, then,

$$w_i = u(t) \times w_i, \quad u(t) = \begin{cases} 0.95 & current - most\ current\ 200\ searches \\ 0.75 & recent - past\ 201 - 400\ searches \\ 0.30 & historical - searches\ earlier\ than\ 400 \end{cases}$$

Recommender. Search results returned from the meta-search engine are categorized into the ODP knowledge hierarchy. Suppose the Web snippets are categorized into category c_i, and its corresponding category weight in the user profile is w_i (i = 1, 2, …N). According to the descending order of w_i, the corresponding category is recommended to users in the same order. Users can adjust the number of categories to be recommended.

4 Combination of Inter- and Intra-similarities

Fig. 2. illustrates how the *inter-similarity* and *intra-similarity* are combined to boost the effectiveness of Web snippet categorization. In Figure 2 (a), there are five categories labeled as C1 to C5 and five Web snippets labeled from S1 to S5. The five snippets are to be categorized under these five categories. According to the cosine similarities between the category-document and the Web snippets, and suppose one snippet can only be classified into one category, S1 and S2 are categorized under category C3; S3 is categorized under category C4; and S4 and S5 are categorized under category C1. Suppose the topic of interest is C3, when that category is selected, Web snippets S1 and S2 will be presented to the user.

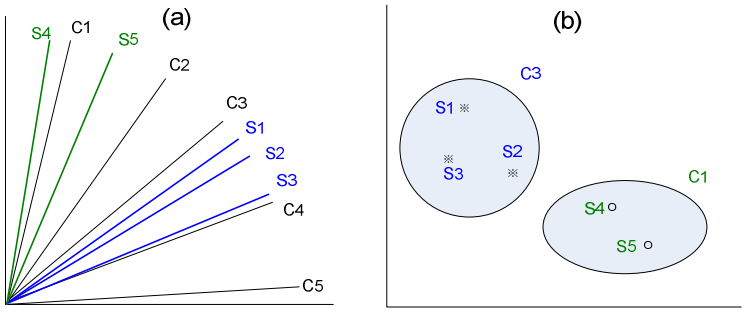
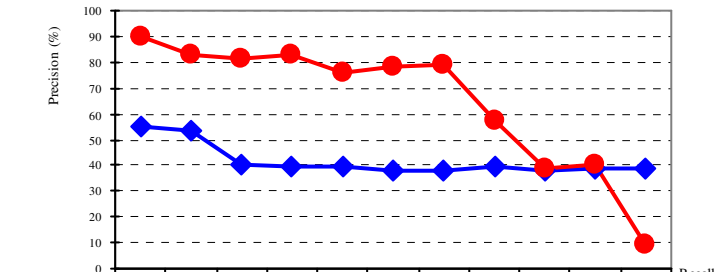


Fig. 2. Illustration of inter- and intra-similarities of Web snippets [29]

However, as can be seen from (b) in Fig. 2, the snippets S1, S2 and S3 are also similar to each other and will thus form a cluster. It is reasonable to assign category C3 to S3 as well. Therefore, to increase *recall*, one snippet should be allowed to assign more than one categories. That is, when category C3 is selected, snippets S1, S2 and S3 should all be presented; not only S1 and S2. When C4 is selected, because S3 and S2 are not in a cluster, only snippet S3 is to be presented.

5 Experimental Results

Our early stage experimental data [10] reveal that Web snippet categorization under the ODP can improve the relevance of Web search. The experiment uses five ambiguous search-terms to obtain search results from *Yahoo! Search Web Service API*, the similarity between the Web search results and the ODP category-documents are calculated by Lucene. A majority voting algorithm is used to make a final categorization decision. For each search-term, 50 search results are taken into consideration. One unique information need is specified for each of these search-terms and one search result is classified to one ODP category. The relevance of Web search results and the supposed information needs are judged by five human experts. Their judgments are summarized to make the final binary relevance judgment decisions. Because the Web search results are often categorized into more than one of the ODP categories, when estimating *precision* and *recall*, two categories with most relevant results are selected. The standard 11 points *precision-recall* curves of the results of Yahoo! API, and our categorized results are shown in Fig. 3.



	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Yahoo!	55.0	53.6	40.6	39.2	39.5	37.8	38.2	39.4	38.1	38.6	38.8
RIB	90.0	83.3	81.5	83.1	76.3	78.0	78.8	57.7	38.8	40.0	9.2

Fig. 3. Average recall-precision curve of Yahoo! search results and the categorized results of RIB over the five search-terms [10]

This early stage experimental result demonstrates that according to the standard 11 points *precision-recall* curve, an average 23.5% *precision* improvement is achieved.

The limitations of this early stage experiment are:

1) the ODP categories are not merged, there are 59,000 category-documents corresponding to the huge ODP categories;

2) document terms are only stemmed; no feature selection algorithms are applied. The computational efficiency therefore has scope for improvement.

6 Future Work

The next goal is to implement *RIB* which is expected to address the problems discussed in the introduction (section 1). Allowance to assign more than one categories to one search result can also improve *recall* of categorized results.

RIB will obtain 100 Web search results for each of 50 selected search-terms to get 5000 search results. Around 50 human experts will be employed, they will be divided into five groups, and each group will have 500 Web results to judge. In addition to relevance judgment, human experts this time will also decide which ODP category a result is to be assigned, and consequently give sufficient training and test data for our experiments to verify and evaluate the developed categorization, clustering, and feature selection algorithms in the context where only Web snippets are available. The effectiveness of personalization and search concept drift process will also be verified.

7 Conclusion

The purpose of this research is to improve the relevance of Web searching by recommend to users with personalized results. A new Web search system, *RIB*, which combines Web snippet categorization, clustering, and personalization was proposed. *RIB* intended to boost the Web snippet categorization by exploring not only inter-similarities between Web snippets and category-documents formed by extracting semantic characteristics of ODP categories; but also the intra-similarities among the returned Web snippets by grouping similar documents into clusters. Users search concept drift problem was addressed by adjusting the weighting factor which represents the users' search preferences in user profiles. Experimental results so far were inspiring; a 23.5% *precision* improvement was achieved when Web search results were categorized under the ODP categorization scheme; and a further boost of Web searching is expected with the implementation of *RIB*.

References

1. Gauch, S., Chaffee, J., Pretschner, A.: Ontology-based personalized search and browsing. Web intelligence and Agent System 1, 219-234 (2003)
2. Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R.: Indexing by latent semantic indexing. J. Am. Soc. Inf. Sci. 41, 391-407 (1990)
3. Montebello, M.: Information Overload--An IR Problem? In: Proceedings of String Processing and Information Retrieval: A South American Symposium, pp. 65-74. IEEE Computer Society (1998)
4. Zhu, D., Dreher, H.: IR Issues for Digital Ecosystems Users. In: Proceedings of the Second IEEE Digital Ecosystems and Technologies Conference, pp. 586-591. IEEE (2008)
5. Chirita, P.-A., Nejdl, W., Paiu, R., Kohlschütter, C.: Using ODP Metadata to Personalize Search. In: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 178-185. ACM Press (2005)
6. Godoy, D., Amandi, A.: Modeling user interests by conceptual clustering. Inform. Syst. 31, 247-265 (2006)
7. Sebastiani, F.: Machine Learning in Automated Text Categorization. ACM Comput. Surv. 34, 1-47 (2002)
8. Jain, A.K., Murty, M.N., Flynn, P.J.: Data Clustering: A Review. ACM Comput. Surv. 31, 264-323 (1999)

9. Yang, Y.: An Evaluation of Statistical Approaches to Text Categorization. Inform. Retrieval 1, 69-90 (1999)
10. Zhu, D.: Improving the Relevance of Search Results via Search-term Disambiguation and Ontological Filtering. School of Information Systems, Curtin Business School, Master. Curtin University of Technology, Perth (2007) 235
11. Mizzaro, S.: Relevance: The Whole History. J. Am. Soc. Inf. Sci. 48, 810-832 (1997)
12. Pitkow, J., Schütze, H., Cass, T., Cooley, R., Turnbull, D., Edmonds, A., Adar, E., Breuel, T.: Personalized Search: A contextual computing approach may prove a breakthrough in personalized search efficiency. Commun. ACM 45, 50-55 (2002)
13. Tsymbal, A.: The problem of concept drift: definitions and related work. Technical report, Trinity College Dublin, (2004)
14. Webb, G.I., Pazzani, M.J., Billsus, D.: Machine Learning for User Modeling. User Model User-Adap. 11, 19-29 (2001)
15. Shen, X., Tan, B., Zhai, C.: Privacy Protection in Personalization Search. ACM SIGIR Forum 41, 4-17 (2007)
16. Smith, B.: Ontology. In: Floridi, L. (ed.): Blackwell Guide to the Philosophy of Computing and Informationpp. 155-166. Blackwell, Oxford (2004)
17. Klas, C.-P., Fuhr, N.: A New Effective Approach for Categorizing Web Documents. In: Proceedings of the 22nd Annual Colloquium of the British Computer Society Information Retrieval Specialist Group (BCSIGSG-00), pp. Unknown (2000)
18. Salton, G., Buckley, C.: Term-Weighting Approaches in Automatic Text Retrieval. Inform. Process. Manag. 24, 513-523 (1988)
19. Hemayati, R., Meng, W., Yu, C.: Semantic-based Grouping of Search Engine Results Using WordNet. In: Dong, G., Lin, X., Wang, W., Yang, Y., Yu, J.X. (eds.): APWeb/WAIM 2007, vol. 4045 pp. 678-686. Springer (2007)
20. Zhu, D., Dreher, H.: An Integrating Text Retrieval Framework for Digital Ecosystems Paradigm. In: Proceedings of the Inaugural IEEE Digital Ecosystems and Technologies Conference, pp. 367-372. IEEE (2007)
21. Hearst, M.A., Pedersen, J.O.: Reexamining the Cluster Hypothesis: Scatter/Gather on Retrieval Results. In: Proceedings of the 19th annual international ACM/SIGIR conference on Research and development in information retrieval, pp. 76-84. ACM Press (1996)
22. Zamir, O., Etzioni, O.: Grouper: A Dynamic Clustering Interface to Web Search Results. In: Proceedings of the Eighth International World Wide Web Conference (WWW8), pp. 283-296. Elsevier (1999)
23. Zeng, H.-J., He, Q.-C., Chen, Z., Ma, W.-Y., Ma, J.: Learning to Cluster Web Search Results. In: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 210-217. ACM Press (2004)
24. Arasu, A., Cho, J., Garcia-Molina, H., Paepcke, A., Raghavan, S.: Searching the Web. ACM Trans. Inter. Tech. 1, 2-43 (2001)
25. Meng, W., Yu, C., Liu, K.-L.: Building Efficient and Effective Metasearch Engines. ACM Comput. Surv. 34, 48-89 (2000)
26. Mladenic, D., Grobelnik, M.: Feature selection on hierarchy of web documents. Decis. Support Syst. 35, 45-87 (2003)
27. Gospodnetić, O., Hatcher, E.: Lucene In Action. Manning Publications, Greenwich (2005)
28. Zhu, D., Dreher, H.: Personalized Information Retrieval in Digital Ecosystems. In: Proceedings of the Second IEEE Digital Ecosystems and Technologies Conference, pp. 580-585. IEEE (2008)
29. Zhu, D.: RIB: A Personalized Ontology-based Categorization/Clustering Approach to Improve the Relevance of Web Search Results. In: Proceedings of Curtin Business School Doctorial Colloquium, pp. Curtin University of Technology, Perth (2007)