Kwan Yi
University of Kentucky
School of Library and Information Science
502 King library South
Lexington, KY 40506   USA

A Pilot Study of Enhancing Subject Discovery of Textual Web Resources

Abstract: The aim of this study is to explore to what degree hyperlinked external resources contribute to the automated subject-related term indexing. Empirical evidence shows no additional enhancement of performance with the additional resources. It also implies that target Web pages are closer in subject to *siting* pages than *sited* pages.

Résumé : L'objectif de cette étude est d'explorer à quel degré les ressources hypertextes externes contribuent à l'indexation automatique par sujet. L'observation empirique ne montre aucune amélioration additionnelle de la performance avec les ressources supplémentaires. Ceci implique également que le sujet des pages web ciblées se rapproche davantage du sujet des pages web *sélectionnant* que des pages web *sélectionnées*.

1. Introduction

Because the Web has become a huge repository of information, the importance of and interest in the Web has rapidly increased during the last decade. A number of different approaches have been attempted to organize and utilize the Web. Pre-classifying the Web (Web directories), indexing the Web (search engines), and assigning metadata to the Web resources in order to integrate Web resources into library OPACs (Online Public Access Cataloging systems) or digital libraries are among the popular approaches. Classification can be seen as the *subject*-based systematic way of organizing items (Maltby, 1975) and subject has played a predominant role in organization and classification (Chan, 1994). In Web organization and retrieval, the discovery of primary subjects or topics of Web resources may be an essential aspect. Also, an IFLA (International Federation of Library Associations and Institutions) working group began to investigate the new trends of Web-based subject access to Web resources. With the proliferation of cataloging records for accessing digital resources in the networked environments, subject discovery is a crucial element of cataloging records or other metadata.

However, in spite of considerable recent efforts, full control and utilization of the rich set of disordered Web information is still far from achievement. Zeng and Chan (2004, 388) stated "Have we fully exploited technological capabilities in our efforts to improve subject access to the myriad resources now available in the networked environment?" More attention and efforts need to be devoted to the development of subject discovery and access tools of Web resources and Web pages.

Given the context, discovering index terms for a Web page is a fundamental issue in the area of Web organization and classification. An underlying challenge is the scarcity of text in Web pages (Pierre, 2001). The purpose of this pilot study is to identify alternative resources that can be supplementary to original target Web pages and to evaluate their performance in the context of automated subject indexing. This study aims to explore to what degree external Web resources (*siting* and *sited* pages) hyperlinked-related to a Web resource contribute to subject discovery of the resource. In this study, a number of popularly used statistical term selection methods will be evaluated in accordance with external alternative resources. Two research questions are proposed:

1. In automated subject term indexing, what results can be achieved when external resources are added to the original Web pages? Are they constant or varied over different subjects and term selection methods?
2. In automated subject term indexing, to what degree does the performance of today's popular statistical term selection methods vary over different external resources?

A considerable amount of research and studies have been made to deal with the same challenge of the textual scarcity of Web pages in Web classification (Golub, 2006), but these are clearly different from this study that is primarily focused on automated indexing and subject discovery.

Web documents are virtually linked to other documents through hyperlinks (called *inlinks* when used for being pointed or *outlinks* when used to point to others) on the Web. Similar to the definitions of *inlinks* and *outlinks*, *inlink*-documents (the term of 'siting pages' will be used henceforth to refer to *inlink*-documents) of Web document *A* can be defined as a set of Web documents that contain hyperlinks pointing to the document *A*, and *outlink*-documents of Web document A can be defined as a set of Web documents to which hyperlinks in the document *A* point (the term of 'sited pages' will be used henceforth to refer to *outlink*-documents). A Web graph views the Web as a graph consisting of a set of nodes and links between them, with a node denoting a Web resource or document and a link indicating a hyperlink. Siting- and sited-Web documents might be closest to a target Web resource in the perspective of Web graph.

This paper is laid out as follows: explorations of statistical term selection methods, employment of data collection, experimental evaluations, and conclusion.

2. Statistical Term Selection Methods

Term selection methods are techniques of selecting a subset of terms from an original set of terms used in automatic text applications such as document indexing and document classification. Two popular applications of term selection methods might be text classification and word association. In text classification, term selection methods have been a useful tool in reducing the size of vocabulary to be dealt with as well as increasing classification accuracy (Koller & Sahami, 1997). In word association, they were widely used to measure statistical association among terms (Chung & Lee, 2001). The following methods are selected due to their popularities (Manning & Schutze, 1999).

2.1 Document frequency (DF)

Document frequency is the number of documents where a term occurs among a collection of documents. The assumption of DF is that terms in low document frequency are non-informative in selecting terms relative to category prediction (Yang & Pedersen, 1997), a situation which is supported by the Zipf's law. Any ideal threshold of document frequency (2 in our case) is not known in theory, but a number in a range of 2-4 is commonly used in text categorization applications (Sebastiani, 2002). Considering the relatively small number of documents used in this study, the value of 2 was employed. In this study, document frequencies of all the terms that appeared in our collection were calculated and terms whose document frequency was less than or equal to 2 were removed. After the removal, low-DF terms are presented to be relatively more informative than high-DF terms.

2.2 Information gain

Information gain is an information entropy-based method of measuring the effectiveness of a feature in a collection of documents (Mitchell, 1997). Information entropy measures the level of impurity in a collection of documents. Information gain calculates the reduction of entropy of a collection of documents after a feature is introduced and considered into the collection. Let C = {C_1, ... ,C_n} be a set of classes, where there are n different number of classes to be considered. The information gain for a class *k* and a term *t* used in this study is defined to be:

$$IG(k,t) = -\sum_{i=1}^{n} P(c_i)\log_2 P(c_i) + P(c_k,t)\sum_{i=1}^{n} P(c_i \mid c_k,t)\log_2 P(c_i \mid c_k,t)$$
$$+ P(t)\sum_{i=1}^{n} P(c_i \mid \bar{c}_k,\bar{t})\log_2 P(c_i \mid \bar{c}_k,\bar{t})$$

This definition is extended from the one used for the m-ary category (Yang & Pedersen, 1997).

2.3 Log-likelihood ratio (LR)

The likelihood ratio for a specific hypothesis is the ratio of the maximum value of the likelihood function over the subspace represented by the hypothesis to the maximum value of the likelihood function over the entire parameter space, and a log-likelihood version is favored due to quickly asymptotical qui-square distribution in the case of binomial and multinomial distriubtions (Dunning, 1993). A version of log-liklihood ratio was examined (Manning & Schutze, 1999) and applied for the co-occurrence of two terms (Chung & Lee, 2001) and for the application of finding relevant Dewey Decimal Classifications for Library Classification Subject Headings (Vizine-Goetz, 2001). The same version was modified in this study for the discovery of dependence between term and class.

2.4 Mutual information (MI)

Mutual information is a method of measuring the amount of shared information between two random variables (one variable is class and the other is term in our case) based on entropy. As two involved variables are more independent of each other, the value of mutual information is close to 0. It further increases as the degree of dependence and the entropy of the variables become larger (Manning & Schutze, 1999). The mutual information MI(*c*, *t*) between a class *c* and a term *t* is defined as (Dumais, Platt, Heckerman, & Sahami, 1998):

$$MI(c,t) = \sum_{c\in\{0,1\}} \sum_{t\in\{0,1\}} P(c,t)\log\frac{P(c,t)}{P(c)P(t)}$$

For each of all possible pairs of classes and terms, the mutual information is calculated. A ranked list of terms is presented for each class.

3. Web Data Set

The data set used in this study was constructed from the Yahoo! Directory where Web pages are manually classified by professionals into the Yahoo!-made classification scheme. The following issues were taken into consideration in selecting categories to be considered: (1) subject coverage – A wide range of diverse subjects each of which is devoted towards a single subject rather than multiple subjects is sought so that 'Arts' and 'Science' are more favorable than 'Reference' and 'News and Media'; (2) the number of Web pages available – In each class, a number of Web pages are listed under the heading of 'POPULAR SITES' along with other links. Our interest is limited to Web pages only, but not links that often refer to other Yahoo! categories. As it is reported that a certain percentage of Web pages do not contain any text in either the title or body or even both (Perrre, 1997), Yahoo! categories containing more than 10 Web pages were considered as a least requirement; (3) subject depth - To avoid the discrepancy in subject depth among different categories, the category levels among different categories were balanced.

Considering the issues aforementioned, the five sub-categories below were selected and their corresponding abbreviations are listed in the parenthesis: (1) Yahoo! Directory > Arts > Crafts (AR); (2) Directory > Business and Economy > Trade (BE); (3) Yahoo! Directory > Computers and Internet > Communications and Networking (CI); (4) Yahoo! Directory > Health > General Health (GH); and (5) Yahoo! Directory > Science > Information Technology (SC).

Because our concern is to compare and evaluate the effectiveness of three Web sources (*target page*, *siting page*, *sited page*) for the automatic selection of subject terms, given each category, Web pages from three different sources are collected: (1) Web pages classified for the category; (2) siting pages of the pages collected in (1); and (3) sited pages of the pages collected in (1). We will call these target pages, siting pages, and sited pages for a set of Web pages collected from (1), (2), and (3), respectively.

First, a collection of target pages was obtained from the Yahoo! categories directly. The target categories contained a mix of Web pages and links to other categories and related Web pages. Only Web pages classified for the categories were manually collected for the target pages. Second, we used the Google advanced command of the *link:url* to get a set of siting pages, where *url* can be replaced by a specific Web address. Although there are other search engines available supporting the *link* command, Google is chosen because Google is the most commonly used search engine and contains the largest number of indexed Web pages among search engines (Gulli & Signorini, 2005). The result of the *link* Google command often produced a large number of returns and sometimes returned with no result. Only the first available and valid Web page was utilized as the siting page corresponding to a target Web page, *valid* in that the selected siting page does not come from the same site as the target page used. Third, a collection of sited pages were collected by the help of a computer program. A computer program was written to read HTML codes of the target Web pages and to collect external links (equivalent to URLs) defined in each of the target pages. Only the first available and valid URL was used as the Web address of the sited page, *valid* in that the URL does not root at the same site as the target page and the page for the URL is accessible. Regardless of Web sources, the following Web pages were not considered and used in our study: (1) Web pages with no text in title or body; (2) Non-English Web pages; (3) document types other than pure HTML-based Web pages. After Web pages were collected from the three sources, HTML tags were removed and only textual parts were preserved for further process.

We ended up with different numbers of collected Web pages in the distinct categories. For example, for the category of BE (Yahoo! Directory > Business and Economy > Trade), there were eleven Web pages (target pages) found in the category. In only four cases out of the eleven, both corresponding siting and sited pages that satisfied the conditions above exist. Refer to Table 1 for the other categories. Thus, the minimum number of Web pages existing from all the sources is 4, which is corresponding to the last row of the Table 1. To balance the variation in the size of Web pages according to distinct categories, only the first four sets of Web pages from the three sources were employed as the final set of the data for this study. In other words, given each category, three separate sets of Web resources (target page, siting page, sited page) are prepared with four distinct documents in each set.

Table 1. Number of Web pages collected in the Yahoo! Directory

	Category				
	AR	BE	CI	GH	SC
# of Web pages in the source of target pages	20	11	15	20	14
# of cases that all three Web pages from the three sources exist	17	4	7	13	5

Figure 1. Comparison of the four term selection methods over different categories

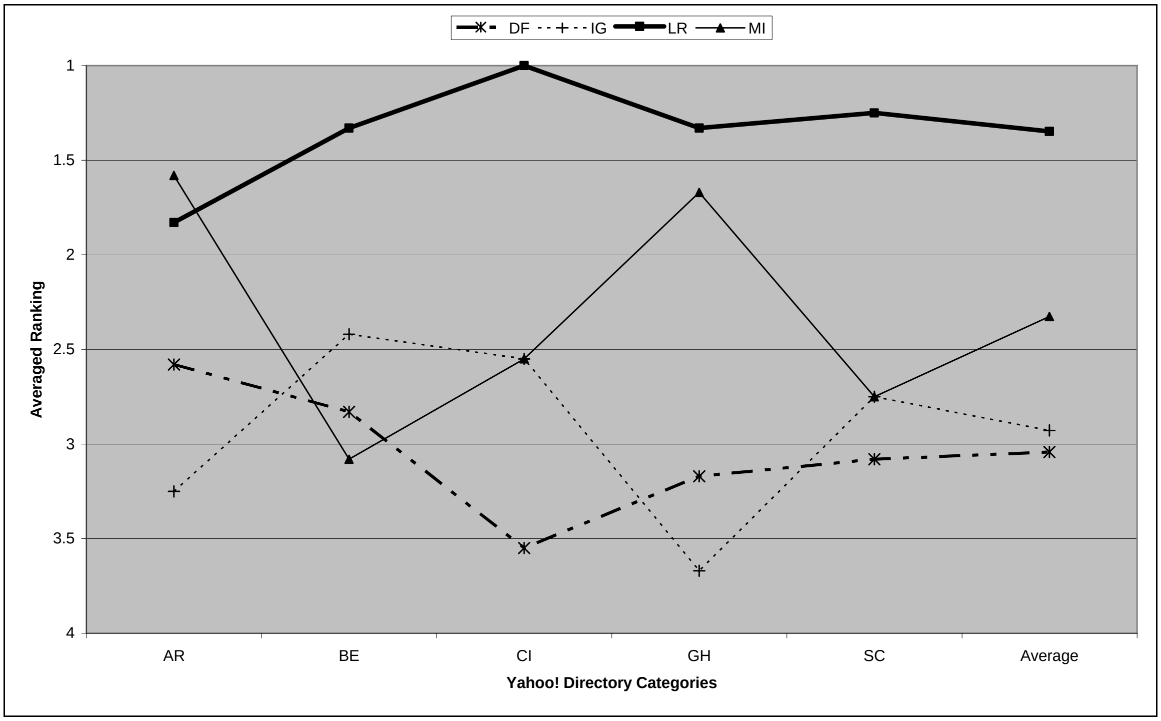
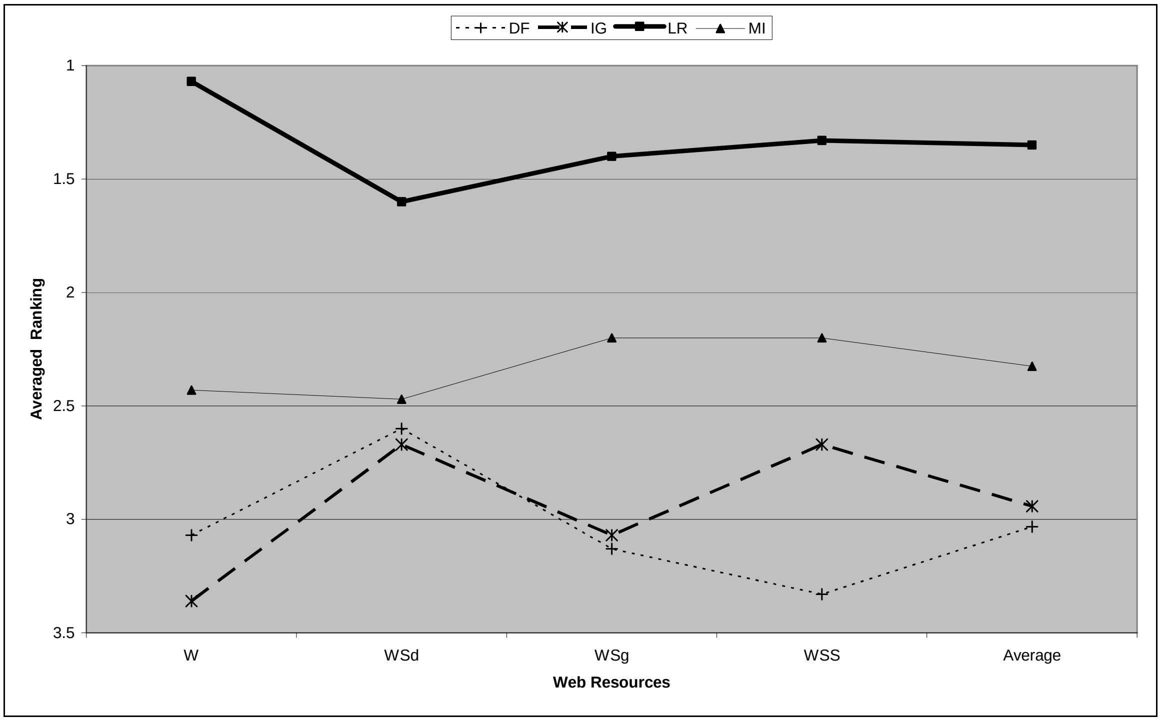


Figure 2. Comparison of the four term selection methods over Web resources



4. Comparative Evaluation

4.1 Comparison of the Term Selection Measures

The purpose of this section is to evaluate the term selection measures by comparing the rankings of four keyword lists generated by the four measures. Given a specific Yahoo! category and a Web resource, four different keyword lists are produced, each of which is based on each of the four feature methods, respectively. The four lists, each of which contain the top fifteen keywords only, are shown to three different graduate students in the program of Library and Information Science, all of whom are in the last or the second to last semester of study. Each of them is asked to rank the lists given a category and is allowed to give a tie when two or more lists are more likely to be equivalent. Figure 1 and 2 displays the performance lines of the four metrics over Yahoo! categories and the Web resources, respectively. The ranking for each mark in Figure 1 and 2 is an average of 12 (3 evaluators X 4 Web resources) and 15 (3 evaluators X 5 categories) different cases, respectively.

In both cases, the LR metric performs superior to any other metrics on average and the MI metric performs second to the LR metric. The performances of IG and DF are the third and fourth but very close to each other. The close performance of IG and DF as term selection methods were reported in the context of document classification (Yang & Pederson, 1997). The performance lines of the LR method show strong steadiness over different categories and resources.

Figure 3. Comparison of the three Web resources over Yahoo! categories

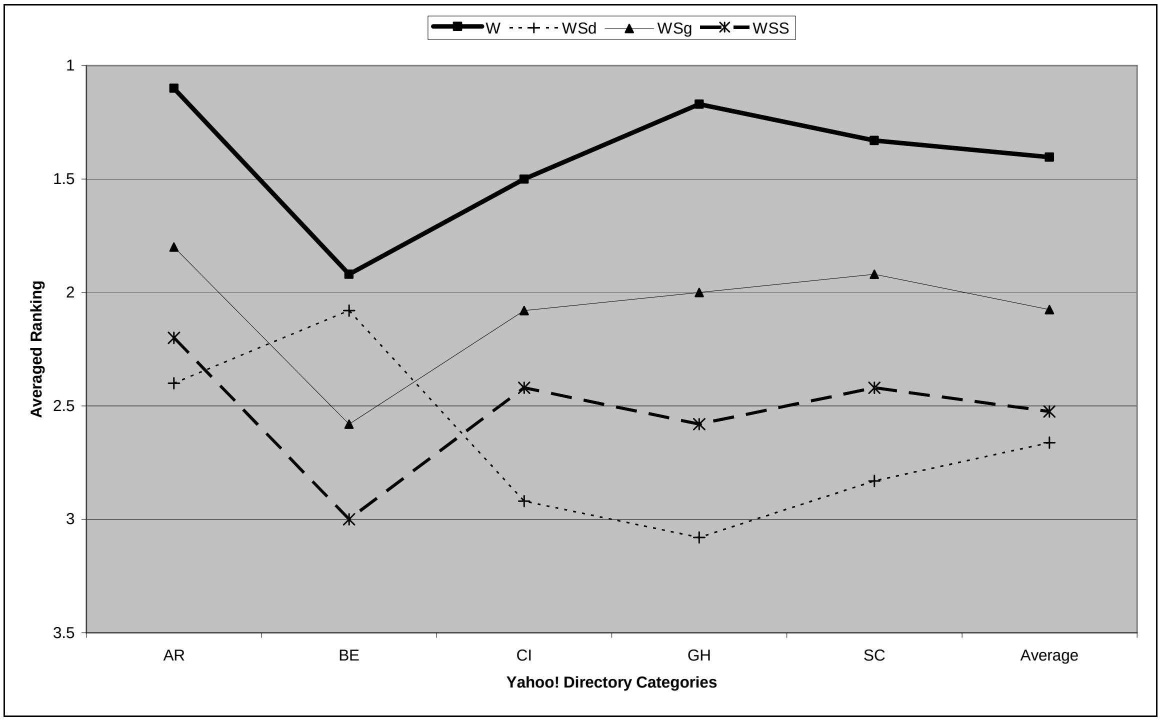
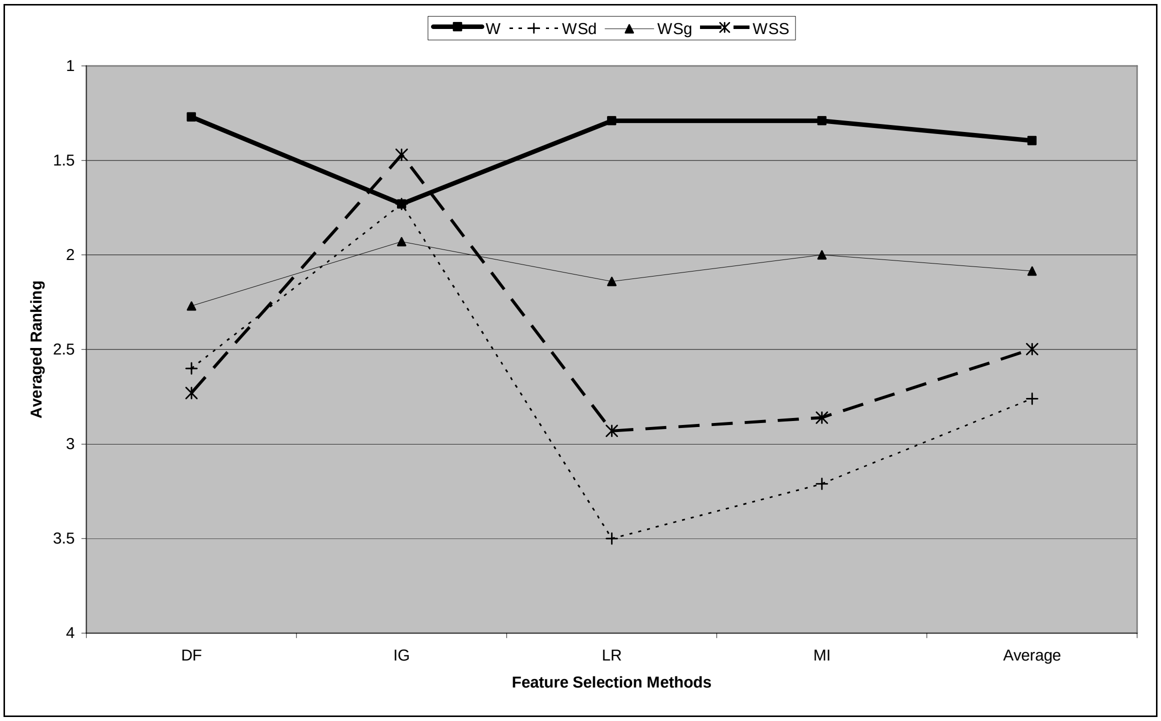


Figure 4. Comparison of the three Web resources over term selection methods



4.2 Comparison of the Web Resources

The purpose of this section is to evaluate different Web resources' performance for automatic discovery of subject terms in Web documents. Four different collections (‘W’, ‘WSd’, ‘WSg’, and ‘WSS’) are tested, each of which is made of different combination of three Web resources (target pages, siting pages, and sited pages): (1) the collection ‘W’ consists of target Web pages only; (2) the collection ‘WSd’ consists of the combined set of target pages and sited pages; (3) the collection ‘WSg’ consists of the combined set of target pages and siting pages; (4) the collection ‘WSS’ consists of the combination of all three resources available.

In each scenario, four different keyword lists based on the four collections (‘W’, ‘WSd’, ‘WSg’, and ‘WSS’) are given to the evaluators. The performances of the four collections are summarized over the Yahoo! categories and term selection methods plotted in Figure 3 and 4, respectively. Similarly to the previous section, each mark in Figure 3 and 4 refers to an average of 12 (3 evaluators X 4 term selection methods) and 15 (3 evaluators X 5 categories) different cases, respectively.

In both cases, the overall best performance occurs with the collection ‘W’. It is rather surprising because this is opposed to our general expectation that supplementing target pages with related resources may contribute to the enhancement of performance. In any case, the ranking of the resources in performance is in the decreasing order of ‘W’, ‘WSg’, ‘WSS’, and ‘WSd’. It appears that adding other types of resources to the target Web pages makes the performance worse. It is interesting to report that the addition of sited pages deteriorates the performance much more than the addition of siting pages does.

5. Conclusion

This is an evaluation of external resources hyperlinked to target Web pages for automated subject term indexing, associated with a number of statistical term selection metrics. The findings of this study are summarized as follows.

First, we found LR most effective in selecting subject related terms for various Web collections. IG and DF are found not comparable to the performance of LR. Mutual information has inferior performance compared to LR but superior to IG and DF.

Second, we discovered that the addition of hyperlinked external resources deteriorates the performance significantly over all the Yahoo! categories attempted. It is worth reporting that the add-on of sited pages led to the worst performance overall, compared to siting pages. This may imply that siting pages are closer in subject to the original Web pages than sited pages are.

In conclusion, it seems that the incorporation of hyperlinked resources does not work in favor of the additional terms when they are combined with the original Web pages. In this study, the pure effects of external resources are not considered. Even though they are expected to lead to lower performance, they would be valuable resources when there is no text found in original target Web pages.

Reference

Chan, Lois Mai. 1994. *Cataloging and Classification: an Introduction*. New York, NY: McGraw-Hill.

Chung, Young Mee, and Jae Yun Lee. 2001. A Corpus-Based Approach to Comparative Evaluation of Statistical Term Association Measures. Journal of the American Society for Information Science and Technology 52 (4): 283-296.

Dumais, S. T., J. Platt, D. Heckerman, & M. Sahami. 1998. Inductive learning algorithms and representations for text categorization. In Proceedings of CIKM-98, 7th ACM International Conference on Information and Knowledge Management, in Bethesda, MD, 148-155.

Dunning, Ted. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics* 19(1): 61-76.

Golub, Koraljka. 2006. Automated subject classification of textual web documents. *Journal of Documentation* 62 (3): 350-371.

Gulli, A. and Signorini, A. 2005. The indexable web is more than 11.5 billion pages. In proceedings of Special interest Tracks and Posters of the 14th international Conference on World Wide Web, May 10 - 14, in Chiba, Japna, 902-903.

Pierre, John M. 2000. Practical Issues for Automated Categorization of Web Sites. In proceeding of ECDL 2000 Workshop on the Semantic Web, in September, in Lisbon, Portugal.

Koller, Daphne, & Mehran Sahami. 1997. Hierarchically Classifying Documents Using Very Few Words. In proceedings of the 14th International Conference on Machine Learning (ICML-97), 12 July, in Nashville, TN, 170-178

Manning, C. D. and Schutze Hinrich. 1999. *Foundations of Statistical Natural Language Processing* Cambridge, MA: MIT Press.

Maltby, Arthur. 1975. *Sayers' Manual of Classification for Librarians*. 5th ed. London: Andre Deutsch.

Schutze, Hinrich, David Hull, and Jan Pedersen. 1995. A comparison of document representations and classifiers for the routing problem. In proceedings of the 18th Annual ACM SIGIR Conference, in July 9-13, in Seattle, Washington, 229-237.

Sebastiani, Fabrizio. March 2002. Machine Learning in Automated Text Categorization. *ACM Computing Surveys* 34(1): 1-47.

Vizine-Goetz, Diane. 2001. Popular LCSH with Dewey numbers. *Jouarnal of Library Administration* 34(¾): 293-300.

Yang, Y. & J. O. Pedersen. 1997. A comparative study on feature selection in text categorization. In proceedings of the 14th International Conference on Machine Learning (ICML-97), 12 July, in Nashville, TN, 412-420.

Zeng, Marcia Lei, and Lois Mai Chan. 2004. Trends and Issues in Establishing Interoperability Among Knowledge Organization Systems. *Journal of the American Society for Information Science and Technology* 55(5):377-395.