Classification-based Retrieval Methods to Enhance Information Discovery on the Web

Yogendra Kumar Jain[1] and Sandeep Wadekar[2]

[1]Head of Department, Computer Science & Engineering
Samrat Ashok Technological Institute Vidisha, M.P., India
ykjain_p@yahoo.co.in

[2]Research Scholar, Computer Science & Engineering
Samrat Ashok Technological Institute Vidisha, M.P., India
sandy12dec@gmail.com

ABSTRACT

The widespread adoption of the World-Wide Web (the Web) has created challenges both for society as a whole and for the technology used to build and maintain the Web. The ongoing struggle of information retrieval systems is to wade through this vast pile of data and satisfy users by presenting them with information that most adequately it's their needs. On a societal level, the Web is expanding faster than we can comprehend its implications or develop rules for its use. The ubiquitous use of the Web has raised important social concerns in the areas of privacy, censorship, and access to information. On a technical level, the novelty of the Web and the pace of its growth have created challenges not only in the development of new applications that realize the power of the Web, but also in the technology needed to scale applications to accommodate the resulting large data sets and heavy loads. This thesis presents searching algorithms and hierarchical classification techniques for increasing a search service's understanding of web queries. Existing search services rely solely on a query's occurrence in the document collection to locate relevant documents. They typically do not perform any task or topic-based analysis of queries using other available resources, and do not leverage changes in user query patterns over time. Provided within are a set of techniques and metrics for performing temporal analysis on query logs. Our log analyses are shown to be reasonable and informative, and can be used to detect changing trends and patterns in the query stream, thus providing valuable data to a search service.

KEYWORDS

Knowledge Discovery, Classification Method, Search Engine Technique

1. INTRODUCTION

The widespread adoption of the World-Wide Web (the Web) has created challenges both for society as a whole and for the technology used to build and maintain the Web. For many years, information retrieval research focused mainly on the problem of ad-hoc document retrieval, a topical search task that assumes all queries are meant to express a broad request for information on a topic identified by the query. This task is exemplified by the early TREC conferences, where the ad-hoc document retrieval track was prominent [2]. In recent years, particularly since the popular advent of the World Wide Web and E-commerce, IR researchers have begun to expand their efforts to understand the nature of the information need that users express in their queries. The unprecedented growth of available data coupled with the vast number of available online activities has introduced a new wrinkle to the problem of search; it is now important to attempt to determine not only what the user is looking for, but also the task they are trying to accomplish and the method by which they would prefer to accomplish it. In addition, all users

DOI : 10.5121/ijmit.2011.3103

are not created equal; different users may use different terms to describe similar information needs; the concept of what is \relevant" to a user has only become more and more unclear as the web has matured and more diverse data have become available [3, 5]. Because of this, it is of key interest to search services to discover sets of identifying features that an information retrieval system can use to associate a specific user query with a broader information need. All of these concerns fall into the general area of *query understanding*. The central idea is that there is more information present in a user query than simply the topic of focus, and that harnessing this information can lead to the development of more effective and efficient information retrieval systems. Existing search engines focus mainly on basic term-based techniques for general search, and do not attempt query understanding. This thesis addresses these shortcomings by presenting a pair of novel techniques for improving query understanding [1] and [4] and [15].

2. BACKGROUND AND RELATED WORK

A log of billions of web queries, that constitutes the total query traffic for a six-month period of a general-purpose commercial web search service. Previously, query logs were studied from a single, cumulative view. Understanding how queries change over time is critical to developing effective, efficient web search services. The web is a dynamic, uncooperative environment with several issues that make analysis of web search very difficult. These include:

1. The web is a dynamic collection: its data, users, search engines, and popular queries are constantly changing [4, 6].

2. Typical web search engine traffic consists of many hundreds of millions of queries per day and is highly diverse and heterogeneous, requiring a large sample of queries to adequately represent a population of even one day's queries [7].

It is difficult to accurately capture the user's desired task and information need from web queries, which are typically very short [4].

2.2. Prior Work in Query Log Analysis

Examinations of search engine evaluation indicate that performance likely varies over time due to differences in query sets and collections [3]. Although the change in collections over time has been studied (e.g., the growth of the web) [4], analysis of users' queries has been primarily limited to the investigation of a small set of available query logs that provide a snapshot of their query stream over a fixed period of time. Existing query log analysis can be partitioned into large-scale log analysis, small-scale log analysis and some other applications of log analysis such as categorization and query clustering. A survey covering a great deal of relevant prior work in search studies can be found in [6]. Jansen and Pooch provide a framework for static log analysis, but do not address analysis of changes in a query stream over time [11]. Given that most search engines receive on the order of between tens and hundreds of millions of queries a day [7], current and future log analysis efforts should use increasingly larger query sets to ensure that prior assumptions still hold.

2.3. Automatic Classification of Web Queries

Accurate topical classification of user queries allows for increased effectiveness, efficiency, and revenue potential in general-purpose web search systems. Such classification becomes critical if the system is to return results not just from a general web collection but from topic-specific back-end databases as well. Successful query classification poses a challenging problem, as web queries are very short, typically providing few features. This feature sparseness, coupled with the dynamic nature of the query stream and the constantly changing vocabulary of the average user hinders traditional methods of text classification. Understanding the topical sense of user queries is a problem at the heart of web search. Successfully mapping incoming general user queries to topical categories, particularly those for which the search engine has domain-specific knowledge, can bring improvements in both the efficiency and the effectiveness of general web search Much of the potential for these improvements exists because many of today's search engines, both for the Web and for enterprises, often incorporate the use of topic specific back-end databases when performing general web search. That is, in addition to traditional document retrieval from general indices, they attempt to automatically route incoming queries to an appropriate subset of specialized back-end databases and return a merged listing of results, often giving preference to the topic-specific results. This strategy is similar to that used by meta-search engines on the web [5]. The hope is that the system will be able to identify the set of back-end databases that are most appropriate to the query, allowing for more efficient and effective query processing. Correct routing decisions can result in reduced computational and financial costs for the search service, since it is clearly impractical to send every query to every backend-database in the (possibly very large) set. If a search service has a strict operational need for low response time, erroneous routing decisions could force much larger scaling than is necessary [6]. An accurate classification of a query to topic(s) of interest can assist in making correct routing decisions, thereby reducing this potential cost.

2.4. Prior Work in Query Classification

Classifying texts and understanding queries are both fundamental concerns in information retrieval, and so we can mention only a fraction of the potentially relevant work. In particular, while work on classifying documents is relevant to classifying queries, we focus on query-oriented work as being most relevant here. Query classification makes very different demands on a classification system than does document classification. Queries are very short (the studies discussed below find average web query lengths between 2.0 and 2.6 terms), and thus have few features for a classifier to work with. Further, unlike a full text document, a query is just a stand-in for the thing we would really like to classify: the user's interest. Sebastiani has surveyed recent work on document classification [4]. More narrowly, our particular interest is in queries to web search engines. Past research can be divided (with some overlap) into work where manual or automated classification (and other techniques) are used to help understand a body of queries, versus work on automatic classification of queries as a tool to improve system effectiveness or provide new capabilities.

2.5. Related Topics and Applications of Query Classification Method

Web directories are subject taxonomies where each category is described by a set of terms that consign the concept behind the category and a list of documents related to the concept. To classify queries into directories could be useful to discover relationships among queries and the concepts behind each query. In this sense there has been substantial work. For example, in previous proposes a Bayesian approach to classify queries into subject taxonomies. Based on the construction of training data, the queries are classified following a breadth first search strategy starting from the root category and descending one level with each number of iteration. The main limitation of the method is the following: queries are always classified into leaves because leaves maximize the probability of the path *root - leaf.*

In the recent proposed competition was focused on the classification of queries into a given subject taxonomy. To do that, the competition organizers provided a small training data set composed by a list of queries and their category labels. Most of the papers were based on classifiers which learn under supervised techniques. The winning paper was written and applies a two phase framework to classify a set of queries into subject taxonomy. Using a machine learning approach, they collected data from the web for training synonym based classifiers that map a query to each related category. In the second phase, the queries were formulated to a search engine. Using the labels and the text of the retrieved pages, the queries were enriched in their descriptions.

Finally, the queries were classified into the subject taxonomy using the classifiers through a consensus function. The main limitations of the proposed method are the dependency of the classification to the quality of the training data, the human effort involved in the training data construction and the semi-automatic nature of the approach which limits the scale of the method applications [8] and [9] and [15].

Vogel classify queries into subject taxonomies using a semi-automatic approach. First they post the query to the Google directory which scans the *Open Directory* for occurrences of the query within the Open Directory categories. Then the top-100 documents are retrieved from Google formulating the query to the search engine. Using this document collection an ordered list of those categories that the 100 retrieved documents are classified into is built. Using a semi-automatic category mapping between the web categories and a subject category the method identifies a set of the closest topics to each query. Unfortunately, the method is limited to the quality of the Google classifier that identifies the closest categories in the Open directory. Also, it is limited to the quality of the answer lists retrieved from Google. Finally, the semi-automatic nature of the approach limits the scale of the method.

Directories are hierarchies of classes which group documents covering related topics. Directories are compounded by nodes. Each node represents a category where Web resources are classified. Queries and documents are traditionally considered as Web resources. The main advantage of the use of directories is that if we find the appropriate category, the related resources will be useful in most of cases. The structure of a directory is as follows: the root category represents the *all* node. The all node means the complete corpus of human knowledge. Thus, all queries and documents are relevant to the root category. Each category shows a list of documents related to the category subject. Traditionally, documents are manually classified by human editors.

The categories are organized in a child/parent relationship. The child/parent relation can be viewed as an"IS-A" relationship. Thus, the child/parent relation represents a generalization/ specialization relation of the concepts represented by the categories.

It is possible to add links among categories without following the hierarchical structure of the directory. These kind of links represent relationships among subjects of different categories in the directory which share descriptive terms but have different meanings. For example, in the TODOCL directory, the category *sports/aerial/aviation* has a link to the category *business & economy/transport/aerial* because they share the term *aerial* that has two meanings: in one category the term is related to sports and in the other to transportation.

Other related work in the area of query classification focuses on the potential applications of query classification technology, and situations where the capabilities of an automatic query classification system would be useful. One such key area is metasearch, a web-centric search technique borne from the old problem of collection fusion [7].

A metasearch engine attempts to integrate and present data from a large number of disparate engines, which may or may not be specialized in nature [5]. Clearly, automatic query classification could be an important factor in such systems, as appropriately classifying a query could assist the metasearch engine in optimally presenting the results it gathers from its source engines, provided that sufficient metadata about their capabilities is available. Internet mediators are a form of highly specialized metasearch engine that attempt to take a single query and route it to multiple, disparate data sources based on its own analysis of the query [7], and for them, an effective query classification system could be a critical part of the decision-making process when performing effective query routing. Other areas where query classification can be effective include query disambiguation and research into predicting how well a query might perform on a given collection. Cronen-Townsend, et. al recently introduced the \Clarity" metric as a way of predicting whether or not a query will have good performance [12], [13] and [16]. The Clarity metric provides a way of measuring how closely the language of a given query fits a given collection. This information can then be used to make an estimation of that query's expected performance; if it appears low, the query can be marked for further processing. Clarity has been used by other researchers as a means of estimating and reducing the ambiguity of queries, but this is an expensive process that requires external resources [14]. Automatic query classification could assist systems in estimating the clarity of a specific query, possibly by providing similar queries from the query's category for further analysis.

At this point, it is relevant to specify which kind of applications will be considered as output of the query log mining process. We work on three kind of applications:

1. Query recommendation: focusing on the reformulation of the original query, these kind of applications aim at identifying relationships between the original queries and alternative queries, such as generalization/specialization relationships.
2. Document recommendation: These kinds of applications will identify relevant documents to the original query.
3. Query classification: These kinds of applications will identify relevant queries and documents for each node in the directory, enriching their descriptions.

There is a similar goal behind each application: to identify relevant documents to the users. This is the main objective of a search engine and finally all the search engine applications are oriented to accomplish this goal. But they work in different manners:

1. A query recommendation application searches for a better way to formulate the original query.
2. Document recommendation applications focus on the identification of documents that are relevant to the users who have formulated similar queries in the past.
3. A query classification application searches for a better description of a node in a taxonomy, adding new queries and documents to them. When the original query is classified into a directory node, a document list is recommended to the user. Also it is possible to navigate into the directory, specializing /generalizing the original query. A global view of the relationships between data mining techniques used in this thesis and the applications generated from the use of these techniques over the query log data is given in Figure 1 [11] and [13] and [17]. Of course, the applications will follow a set of design criteria. The desirable properties of the applications will be the following:

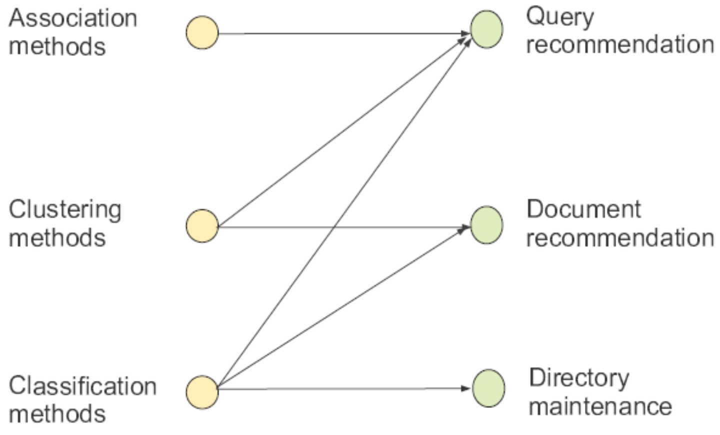

Figure 1: Relations between the data mining techniques used in this thesis and the applications generated.

Maintainability: The costs associated to the system maintenance must be low.

Variability: The user's preferences change in the course of time. This user click feature is due to two main factors: changes in the user's interests and elaboration of new documents with contents that can be interesting to the users. The applications must reflect the changes in the user's interests in order to produce quality answer lists.

Scalability: The applications designed must be able to manage huge document collections, efficiently.

User's click data analysis: The analysis of the user's click data is known as Web usage mining. Web usage mining literature focuses on techniques that could predict user behavior patterns in order to improve the success of Web sites in general, modifying for example, their interfaces. Some studies are also focused on applying soft computing techniques to discover interesting patterns in users click data, working with vagueness and imprecision in the data. Other kind of works is focused on the identification of hidden relations in the data. Typical problems related are user sessions and robot sessions identification. Web usage mining studies could be classified into two categories: those based on the server side and those based on the client side. Studies based on the client side retrieve data from the user using cookies or other methods, such as ad-hoc logging browser plugins. For example, in recent proposes mining techniques to process data retrieved from the client side using panel logs. Panel logs, such as TV audience ratings, cover a broad URL space from the client side. As a main contribution, they could study global behavior in Web communities. Web usage mining studies based on server log data present statistical analysis in order to discover rules and non trivial patterns in the data. The main problem is the discovery of navigation patterns. For example, assumes that users choose the following document determined by the last few documents visited, concluding how well Markov models predict user's clicks. Similar studies consider longer sequences of requests to predict user behavior or the use of hidden Markov models introducing complex models of user click prediction. An analysis of the Alta Vista search engine log, focused on individual queries, duplicate queries and the correlation between query terms. As a main result, authors show that users type short queries and select very few documents in their sessions. Similar studies analyze other commercial search engine logs. These works address the analysis from an aggregated point of view, i.e., present global distributions over the data. Other approaches include novel points of view for the analysis of the data. As a main result, author shows that query traffic differs from one topic to another, considering hourly analysis, these results being useful for query disambiguation and routing. Finally the analysis is using geographic information about users. The main results of this study show that queries are short, search topics are broadening and approximately 50% of the Web documents viewed are topically relevant. The literature also shows some works related with the following purpose: to determine the need behind the query. To do that, analyzed a query log data determining three types of needs: informational, navigational and transactional. Informational queries are related to specific contents in documents. Navigational queries are used to find a specific site in order to browse their pages. Transactional queries allow users to perform a procedure using web resources such as commercial operations or download. In the last two categories were refined into ten more classes. In recent years, a few papers have tried to determine the intention behind the query, following the categorization introduced above. Simple attributes were used to predict the need behind the query. To introduce a framework to identify the intention behind the query using user's click data. Following a combination between supervised and unsupervised methods, the authors show that it is possible to identify concepts related to the query and to classify the query in a given user's needs categorization. As we can see in the works described above, the user's click data has been extensively studied in scientific literature. Despite this fact, we have little understanding about how search engines affect their own searching process and how users interact with search engine results. Our contribution in this topic will be focused in this manner: to illustrate the user-search engine interaction as a bidirectional feedback relation. To do this we should understand the searching process as a complex phenomenon constituted by multiple factors such as the relative position of the pages in answer lists, the semantic connections between query words and the document reading time, among other variables. Finally, we propose user search behavior models that involve the factors enumerated. The main advantage over this work is the use of several variables involved in the searching process to produce models about how users search and how users use the search engine results [5], [16], [18] and [20].

3. PROPOSED CLASSIFICATION APPROACHES

Prior efforts in classifying general web queries have included both manual and automatic techniques. In this section, we describe the manual and automatic classification approaches that collectively form our framework and explain our motivations for using each approach. We also introduce a new rule-based automatic classification technique based on an application of computational linguistics for identifying selectional preferences by mining a very large unlabeled query log. We demonstrate that a combination of these approaches allows us to develop an automatic web query classification system that covers a large portion of the query stream with a reasonable degree of precision for using each approach. We also introduce a new rule-based automatic classification technique based on an application of computational linguistics for identifying selectional preferences by mining a very large unlabeled query log. We demonstrate that a combination of these approaches allows us to develop an automatic web query classification system that covers a large portion of the query stream with a reasonable degree of precision.

We explored this technique by using 18 lists of categorized queries produced in the above fashion by a team of AOL^{TM}editors. We examined the coverage of these categories and found that even after considerable time and development effort; they only represented 12% of the general query stream. In addition to poor coverage, a key weakness in this approach is that the query stream is in a constant state of flux. Any manual classifications based on popular elements in a constantly changing query stream will become ineffective over time. Additionally, manual classification is very expensive. It is infeasible to label enough queries, and to repeat this process often enough, to power an exact match classification system that covers a sufficient amount of the query stream. Finally, one does not necessarily achieve high precision even on queries that are covered. A natural step is to leverage the labeled queries from the exact match system through supervised learning. The idea is to train a classifier on the manual classifications with the goal of uncovering features that enable novel queries to be classified with respect to the categories.

Each query is classified following a top-down approach. First, we determine the closest centroid to the query considering all the centroids at the first level of depth in the concept taxonomy. Then we repeat the process for each level of the taxonomy while the distance between the query and the closest centroid will be less than the distance at the previous level. The top-down approach is used to avoid the noise effects introduced by document and query terms. From our point of view, the term relevance decreases from general topics to sub-topics. In the figure 2 we illustrate the classification schema.

For our experiments we use the Perception with Margins algorithm, which has been shown to be competitive with state-of-the-art algorithms such as support vector machines in text categorization, and are very computationally efficient [7].

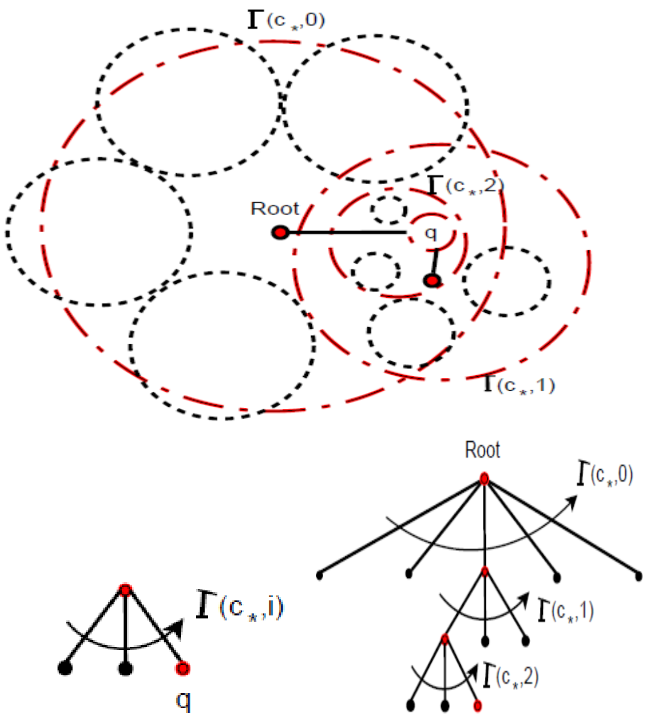

Figure 2: The hierarchical classification schema

Our implementation normalizes the document feature vectors to unit length (Euclidean normalization), which we have found to increase classification accuracy. To examine the effectiveness of the supervised learning approach we trained a perception for each category from the exact match system, using all queries in a given category as positive examples and all queries not in that category as negative examples.

To realistically evaluate this approach we developed a test collection that is representative of the general query stream. To determine the number of queries required to achieve a representative sample of our query stream, we calculated the necessary sample size in queries:

Representative Sample Size Formula

$$ss = (Z^2)(P)(1\text{-}p)/c^2 \qquad \ldots\ldots\ldots\ldots\ldots(1)$$

Where:

Z = Z value (e.g. 1.96 for 95% confidence)

p = percentage picking a choice, expressed as decimal (0.5 used here)

c = confidence interval, expressed as decimal (e.g., .04 = +/- 4).
By setting our confidence level to 99% and error rate to 1%, we require a sample of at least 16,638 queries.
To build the test set we had a team of human editors perform a manual classification of 20,000 queries randomly sampled from general query stream. These 20,000 queries were then used for testing our perception learner trained on the classifications from the exact match system.

3.1 Selectional Preferences in Computational Linguistics

A selectional preference study begins by parsing a corpus to produce pairs, ($x; y$), of words that occur in particular syntactic relationships. The verb-object relationship is most widely studied, but others such as adjective-noun have been used [5]. The interest is in finding how x, the context, constrains the meaning of y, its argument.
The selectional preference strength $S(x)$ of a word x, where u ranges over a set U of semantic classes. $P(U)$ is the probability distribution of semantic classes taken on by words in a particular syntactic position, while $P(U|x)$ is that distributions for cases where a contextual syntactic position is occupied by x. $S(x)$ is simply the KL-divergence of $P(U|x)$ from $P(U)$.

Selectional Preference Strength:$S(x) = D(P(u|x)||P(u)) = \sum P(u|x) lg(\mathrm{P(u|x)/p(u)})$

The ideal approach to estimating the necessary probabilities would be to use a set of ($x; y$) pairs where each y has been tagged with its semantic class, perhaps produced by parsing a semantically tagged corpus. The maximum likelihood estimates (MLEs) are:

$P\ (u) = \mathrm{n_u/N}$
$P(u|x) = (\mathrm{n_{xu}/N})\ /\ \mathrm{n_x/N} = \mathrm{n_{xu}/\ n_x}$

Where N is the total number of pairs, nu is number of pairs with a word of class u in the syntactic position of interest, n_x is the number of pairs with x providing context for that position, and n_xu is the number of pairs where both are true.

4. Results

Our experiment consisted of evaluating the quality of the answer lists retrieved by the search engine ranking method, the method based on the hierarchical classifier, and the method based on the flat classifier. In order to do this, we have considered the first 10 documents recommended by the search engine for each one of the 30 queries, the first ten documents recommended by the web directory when the closest category is determined using the flat classifier, and the first 10 documents recommended by the web directory when the query is classified using the hierarchical method. The document quality has been evaluated by a group of experts using the same relevance criteria as in the previous experiment (0-4, from lower to higher relevance). The precision for every ranking and every query is obtained, according to the position. Finally, the average precision is calculated over the total of documents recommended by position. Results are shown in Figure below 4.1. In the figure we can observe that all the evaluated methods are good quality rankings, especially for the first 5 recommended documents. The recommendation methods based on hierarchical classification and flat classification perform in a better way for the first 5 positions than the TodoCL ranking. This means that the classification in the taxonomy is a good quality one, the same as shown in the previous experiment. However, the ranking loses precision compared to the one of TodoCL, if we consider the last 5 positions. This is due to the fact that many of the evaluated queries are classified in taxonomy nodes where less than 10 documents are recommended.

In these cases, since there is no recommendation, the associated precision equals 0, which severely disqualifies the methods based on classifiers. Fortunately, none of the queries is classified in a node with less than 5 recommended documents. Therefore, a fair comparison of the methods should be limited to the first 5 positions where, as we have seen, the hierarchical method is favorably compared with the original ranking and with the flat classifier.

Due to the fact that in general the coverage of directories is low, i.e. there are few documents recommended in each node of the directory, it is necessary to design a maintenance method in order to classify documents into nodes enriching their descriptions and improving the coverage of the directory.

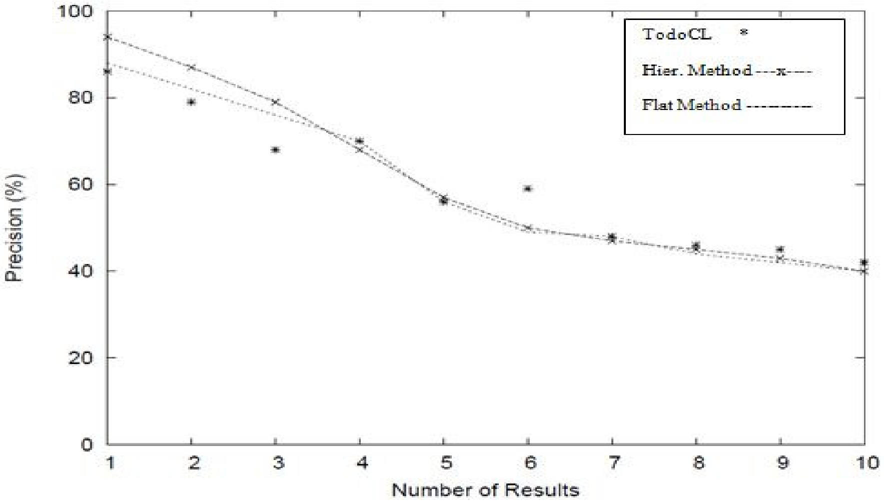

Figure 3: Average precision of the retrieved documents for the methods based on classifiers and for the search engine ranking.

5. CONCLUSIONS AND FUTURE WORK

We can conclude that our query classification method allows us to identify concepts associated to the queries. The taxonomy would permit us to specialize the query in order to provide final accurate recommendations. The proposed method also allows us to improve the precision of the retrieved documents over the first 5 positions of the answer lists. Compared with the search engine ranking method and compared with the method based on a flat classifier, the hierarchical classifier method provides better results regarding the precision of the answer lists. One of the biggest constraints of the proposed method lies in the fact that it depends strongly on the taxonomy quality. In the third experiment, we can notice that an important proportion of the nodes contain an insufficient quantity of recommended documents, which prevents a favorable comparison to the TodoCL ranking beyond the first 5 recommendations. The classification method proposed in this chapter allows the identification of well defined concepts for each query classified. The maintenance method proposed allows the Web directory to be automatically maintained without losing precision and recall. By considering the logs in the maintenance process we take into account information based on user's click data. The preferences are bound to query sessions, and thus we may obtain preferences sensible to topics users are looking for. The importance and quality of a Web page is reflected by the preferences registered in the Web log. This allows us to find many new relevant documents not considered in the directory built by human editors. Another constraint is that the taxonomy does not always allow us to identify the need behind the query. Both constraints are related to the fact that since the directory is manually maintained, it is limited in enlargement and freshness because of the frequency of the editor's evaluations and updates.

In future, we will present a method based on the proposed classification schema which permits us to maintain automatically the directories, by adding new queries and documents to each directory node.

REFERENCES

[1] Ali Sajedi Badashian, Seyyed Hamidreza Afzali, Morteza Ashurzad Delcheh, Mehregan Mahdavi and Mahdi Alipour, (2010) "Conceptual File Management: Revising the Structure of Classificationbased Information Retrieval", 5th IEEE International Conference on Digital Information Management (ICDIM), pp. 108-113.

[2] Ming Zhao and Shutao Li, (2009) "Sparse Representation Classification for Image Text Detection", 2nd IEEE International Symposium on Computational Intelligence and Design, Vol. 1, pp. 76-79.

[3] Radhouane Guermazi, Mohamed Hammami, and Abdelmajid Ben Hamadou, (2009) "Violent Web images classification based on MPEG 7 color descriptors", IEEE International Conference on Systems, Man, and Cybernetics, pp. 3106 – 3111.

[4] Esin Guldogan, and Moncef Gabbouj, (2010) "Adaptive Image Classification Based on Folksonomy", 11th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS), pp. 1-4.

[5] Shuang Deng, and Hong Peng, (2006) "Document Classification Based on Support Vector Machine Using A Concept Vector Model", IEEE/WIC/ACM International Conference on Web Intelligence, pp. 473-476.

[6] Choochart Haruechaiyasak, Mei-Ling Shyu, Shu-Ching Chen, and Xiuqi Li, (2002) "Web Document Classification Based on Fuzzy Association", 26th IEEE Annual International Conference on Computer Software and Applications COMPSAC 2002, pp. 487-492.

[7] Shou-Bindong, Yi-Myianngg, (2002) "Hierarchical Web Image Classification by Multi-Level Features", 1st IEEE International Conference on Machine Learning and Cybernetics, Vol. 21, pp. 663-668.

[8] S. M. Beitzel, E. C. Jensen, D. D. Lewis, A. Chowdhury, A. Kolcz, and O. Frieder, (2005) "Improving automatic query classification via semi-supervised learning", 5th IEEE International Conference on Data Mining, pp. 42-49.

[9] YANG Xiao-qin, JU Shiguang and CAO Qinghuang, (2009) "A Deep Web Complex Matching Method based on Association Mining and Semantic Clustering", 6th IEEE Web Information Systems and Applications Conference, WISA 2009, pp. 169-172.

[10] Nizar R. Mabroukeh, and C. I. Ezeife, (2009) “Semantic-rich Markov Models for Web Prefetching”, IEEE International Conference on Data Mining Workshops, ICDMW’09, pp. 465-470.

[11] Suleyman Salin, and Pinar Senkul, (2009) “Using Semantic Information for Web Usage Mining Based Recommendation”, 24th IEEE International Symposium on Computer and Information Sciences, ISCIS 2009, pp. 236-241.

[12] Yi Feng, (2010) “Towards Knowledge Discovery in Semantic Era”, 7th IEEE International Conference on Fuzzy Systems and Knowledge Discovery, FSKD 2010, Vol. 5, pp. 2071-2075.

[13] Elaheh Momeni, (2010) “Towards (Semi-) Automatic Moderation of Social Web Annotations”, 2nd IEEE International Conference on Social Computing, pp.1123-1128.

[14] Julia Hoxha and Sudhir Agarwal, (2010) “Semi-automatic Acquisition of Semantic Descriptions of Processes in the Web”, IEEE International Conference on Web Intelligence and Intelligent Agent Technology, Vol. 1, pp. 256-263.

[15] Taksa, Isak, Zelikovitz, Sarah, Spink Amanda, (2007) “Using Web Search Logs to Identify Query Classification Terms”, 4th IEEE International Conference on Information Technology, ITNG '07, pp. 469-474.

[16] C. N. Ziegler, M. Skubacz, (2007) “Content Extraction from News Pages Using Particle Swarm Optimization on Linguistic and Structural Features”, IEEE/WIC/ACM International Conference on Web Intelligence, pp. 242-249.

[17] Weiming Hu, Ou Wu, Zhouyao Chen, Zhouyu Fu, S. Maybank, (2007) “Recognition of Pornographic Web Pages by Classifying Texts and Images” IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 29, No. 6, pp. 1019-1034.

[18] Jun Fang, Lei Guo, XiaoDong Wang, Ning Yang, (2007) “Ontology-Based Automatic Classification and Ranking for Web Documents”, 4th International Conference on Fuzzy Systems and Knowledge Discovery, FSKD’2007, Vol. 3, pp. 627-631.

[19] Shiqun Yin, Yuhui Qiu, Chengwen Zhong, Jifu Zhou, (2007) “Study of Web Information Extraction and Classification Method”, IEEE International Conference on Wireless Communications, Networking, and Mobile Computing, WiCom’2007, pp. 5548-5552.

[20] F. Schroff, A. Criminisi, A. Zisserman, (2007) “Harvesting Image Databases from the Web”, ICCV’2007, 11th IEEE International Conference on Computer Vision, pp. 1-8.