ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 3, Issue 1, pp: (19-26), Month: January - March 2015, Available at: www.researchpublish.com

Web Based Information Retrieval Using Dynamic Classified Average Precision Crawling Approach

[1]R.Geethanjali, [2]Mr.S.Muruganantham

[1] M.Sc., Research Scholar, Department of Computer Science, Kongu Arts and Science College, Erode, India
[2] M.Sc., M.Phil, Assistant Professor & Head, Department of CT & IT, Kongu Arts and Science College, Erode, India

***Abstract*: The network is the largest collection of automatically accessible documents, which makes the richest source of information in the world. The different people may have distinct queries to search results and retrieve their results when they submit query in a search engine. The problem of clustering the feedback sessions are addressed in this research paper.**

The Research paper ""WEB BASED INFORMATION RETRIEVAL USING DYNAMIC CLASSIFIED AVERAGE PRECISION CRAWLING APPROACH" provides the best precision value and accuracy to the search results. The novel approach is inherited here to infer user goals by analyzing search engine query logs. The titles, snippets are created based on the clicked sequence of the query. Combining the titles and snippet, the feature representation can be derived. Feedback sessions are constructed from click through and can efficiently reflect the information needs of the user. They adopt a novel approach to generate the pseudo documents to better represent the feedback sessions for clustering.

Considering that if user search goals are inferred properly, the search results can also be restructured properly, since restructuring web search results is one application of inferring user search goals. In previous work the "Classified Average Precision" to evaluate the restructure results and describes the method to select the best cluster number. Previous studies have mainly focused on using manual query based logging examination to identify user web search results. In this dissertation they show whether and how they can automate this goal-identification process.

Here a new criterion "Dynamic Classified Average Precision Crawling (DCAPC)" is proposed to evaluate the performance of inferring user search goals. The dissertation proposes DCAPC (Dynamic Classified Average Precision Crawling) approach, an evaluation method based on restructuring web search results to evaluate whether user search goals are inferred properly or not and give better precision values. Experimental results are presented using user click-through logs from a commercial search engine to validate the effectiveness of the proposed methods. The quality of the extracted trait proposed DCAPC query logs as a helpful, though little explored, resource for in sequence data extraction.

1. INTRODUCTION

1.1 Overview of Data Mining

Data mining involves the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large data sets. The primary ingredient of any Data Mining is the database. A database is an organized and typically large collection of detailed facts concerning some domain in the outside world. The aim of Data Mining is to examine this database for regularities that may lead to a better understanding of the domain described by the database. In Data Mining it is generally assumed that the database consists of a collection of individuals. Depending on the domain, individuals can be anything from customers of a bank to molecular compounds or books in a library. For each individual, the database gives us detailed information concerning the different characteristics of the individual, such as the name and address of a customer of a bank, or the accounts owned.

Data mining, the extraction of hidden predictive information from large databases, is a powerful technology with great potential that helps to focus on the most important information in data warehouses. Data mining tools predict future trends and behaviors, allowing us to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations. Data mining techniques can be implemented rapidly on existing software and hardware platforms to enhance the value of existing information resources. Data mining is the practice of automatically searching large stores of data to discover patterns and trends that go beyond simple analysis. Data mining uses sophisticated mathematical algorithms to segment the data and evaluate the probability of future events. Data mining is also known as Knowledge Discovery in Data (KDD).

The key properties of data mining are:

- Automatic discovery of patterns
- Prediction of likely outcomes
- Creation of actionable information
- Focus on large data sets and databases

Data mining can answer questions that cannot be addressed through simple query and reporting techniques.

Automatic Discovery

Data mining is accomplished by building models. A model uses an algorithm to act on a set of data. The notion of automatic discovery refers to the execution of data mining models. Data mining models can be used to mine the data on which they are built, but most types of models are generalizable to new data. The process of applying a model to new data is known as scoring.

Prediction

Many forms of data mining are predictive. For example, a model might predict income based on education and other demographic factors. Predictions have an associated probability. Prediction probabilities are also known as confidence.

Grouping

Other forms of data mining identify natural groupings in the data. For example, a model might identify the segment of the population that has an income within a specified range, that has a good driving record, and that leases a new car on a yearly basis.

Actionable Information

Data mining can derive actionable information from large volumes of data.

Definition

Data mining is a collection of techniques for efficient automated discovery of previously unknown, valid, novel, useful and understandable patterns in large database. The patterns must be actionable so that they may be used in an enterprise's decision making. Data mining can be performed on data represented in quantitative, textual, or multimedia forms. An application, compared to other data analysis applications, such as structured queries (used in many commercial databases) or statistical analysis software, data mining represents a difference of kind rather than degree. Many simpler analytical tools utilize a verification-based approach, where the user develops a hypothesis and then tests the data to prove or disprove the hypothesis.

The effectiveness of this approach can be limited by the creativity of the user to develop various hypotheses, as well as the structure of the software being used. In contrast, data mining utilizes a discovery approach, in which algorithms can be used to examine several multidimensional data relationships simultaneously, identifying those that are unique or frequently represented. As a result of its complex capabilities, two precursors are important for a successful data mining exercise; a clear formulation of the problem to be solved, and access to the relevant data.

Reflecting this conceptualization of data mining, some observers consider data mining to be just one step in a larger process known as knowledge discovery in databases (KDD). Other steps in the KDD process, in progressive order, include data cleaning, data integration, data selection, data transformation, data mining, pattern evaluation, and knowledge presentation. Data mining has become increasingly common in both the public and private sectors. Organizations use data mining as a tool to survey customer information, reduce fraud and waste, and assist in medical research. However, the proliferation of data mining has raised some implementation and oversight issues.

These include concerns about the quality of the data being analyzed, the interoperability of the databases and software between agencies, and potential infringements on privacy. Also, there are some concerns that the limitations of data mining are being overlooked as agencies work to emphasize their homeland security initiatives .

1.2 Data Mining Methods

In general, data mining methods can be classified into two categories: predictive and descriptive. Predictive data mining methods predicts the values of data, using some already known results that have been found using a different set of data. Predictive data mining tasks include: Classification, Prediction. Descriptive mining tasks characterize the general properties of the data in database. This is done by identifying the patterns and relationships in the data. These models are not based on any underlying theory or mechanism through which the data arose rather they are simply a description of the observed data. Descriptive data mining tasks include: Clustering, Association analysis and Summarization.

Major issues in Data Mining

Mining different kinds of knowledge in databases

Since different users can be interested in different kinds of knowledge, data mining should cover a wide spectrum of data analysis and knowledge discovery tasks, including classification, prediction, and cluster and association analysis.

Interactive mining of knowledge at multiple levels of abstraction

Since it is difficult to know exactly what can be discovered within a database, the data mining process should be interactive. Interactive mining allows users to focus the search for patterns, providing and refining data mining requests based on returned results.

Performance issues

These include efficiency, scalability, and parallelization of data mining algorithms.

- Efficiency and scalability of data mining algorithms: To effectively extract information from a huge amount of data in databases, data mining algorithms must be efficient and scalable.
- Parallel, distributed, and incremental mining algorithms: The huge size of many databases, the wide distribution of data, and the computational complexity of some data mining methods are factors motivating the development of algorithms that divide data into partitions that can be processed in parallel.

Presentation and visualization of data mining results

Discovered knowledge should be expressed in high-level languages, visual representations, or other expressive forms so that the knowledge can be easily understood and directly usable by humans.

Handling outlier or incomplete data

The data stored in a database may have outliers or noise, exceptional cases, or incomplete data objects. System should be able to deal with these.

Pattern evaluation: the interestingness problem

A data mining system can uncover thousands of patterns. Many of the patterns discovered may be uninteresting to the given user, representing common knowledge or lacking novelty. Several challenges remain regarding the development of techniques to assess the interestingness of discovered patterns, particularly with regard to subjective measures.

Efficiency and scalability of data mining algorithms

To effectively extract information from a huge amount of data in databases, data mining algorithms must be efficient and scalable.

Mining information from heterogeneous databases and global information systems

Local and wide-area computer networks (such as the Internet) connect many sources of data, forming huge, distributed, and heterogeneous databases. The discovery of knowledge from different sources of structured, semi-structured, or unstructured data with diverse data semantics poses great challenges to data mining. Data mining may help disclose high-level data regularities in multiple heterogeneous databases that are unlikely to be discovered by simple query systems and may improve information exchange and interoperability in heterogeneous databases.

1.3 Clustering And Classification

Clustering and classification are both fundamental tasks in Data Mining. Classification is used mostly as a supervised learning method, clustering for unsupervised learning (some clustering models are for both). The goal of clustering is descriptive, that of classification is predictive since the goal of clustering is to discover a new set of categories, the new groups are of interest in themselves, and their assessment is intrinsic. In classification tasks, however, an important part of the assessment is extrinsic, since the groups must reflect some reference set of classes.

Clustering Technique

Clustering is a useful technique for the discovery of data distribution and patterns in underlying data. The goal of clustering is to discover both the dense and the sparse region in a data set . Cluster analysis itself is not one specific algorithm, but the general task to be solved. It can be achieved by various algorithms that differ significantly in their notion of what constitutes a cluster and how to efficiently find them.

There are two main approaches to clustering- hierarchical clustering and partitioning clustering. Besides, clustering algorithms differ among themselves in their ability to handle different types of attributes, numeric and categorical, in accuracy of clustering, and their ability to handle disk-resident data.

Hierarchical Clustering

The hierarchical clustering techniques do a sequence of partitions in which each partition is nested into the next partition in the sequence. It creates a hierarchy of clusters from small to big or vice versa. The hierarchical techniques are of two types agglomerative and divisive clustering techniques . Agglomerative clustering is a bottom up strategy starts by placing each object in its own cluster and then merge these atomic clusters into larger and larger clusters, until all of the objects are in single cluster. Divisive clustering is top down strategy that does the reverse of agglomerative clustering by starting with all objects in one cluster. It subdivides the cluster into smaller and smaller pieces, until each object form a cluster on its own.

Classification

Classification is a data mining function that assigns items in a collection to target categories or classes. The goal of classification is to accurately predict the target class for each case in the data.

For example, a classification model could be used to identify loan applicants as low, medium, or high credit risks. A classification task begins with a data set in which the class assignments are known. The classes in the classification problem are made predefined and non-overlapping before the application of any algorithm. There are three approaches to address the classification problem . The first is to divide the space defined by the data points into regions. Each region corresponds to a given class. Any instance that falls in a certain region is identifies as belonging to that particular class. The second approach is to find the probability of an instance belonging to each class. The class that gets the highest probability is assumed to contain that instance. The third approach is to find the probability of a class containing instance. The class with the highest probability captures that instance. The traditional and well accepted method of classification is the induction of decision trees, which partition the dataset to develop rules. The other popular algorithms that are used for classification tasks include Artificial Neural Networks (ANN) and the Support Vector Machines (SVM).

1.4 Web Mining

The Web is the largest collection of electronically accessible documents, which make the richest source of information in the world. The problem with the Web is that this information is not well structured and organized so that it can be easily retrieved. Search engines help in accessing web documents by keywords, but this is still far from what is needed in order to effectively use the knowledge available on the Web. Machine Learning and Data Mining approaches go further and try

to extract knowledge from the raw data available on the Web by organizing web pages in well defined structures or by looking into patterns of activities of Web users.

Web mining - is the application of data mining techniques to discover patterns from the Web. It is the mining of data related to World Wide Web (WWW). It is a technique used to crawl through various web resources to collect required information, which enables an individual or a company to promote business, understanding marketing dynamics, new promotions floating on the Internet, etc .

There is a growing trend among companies, organizations and individuals alike to gather information through web data mining to utilize that information in their best interest. The users make use of the web in several ways.

Any user interact with the web for the purpose which include finding relevant information, discovering new knowledge from the web, personalized web page synthesis and learning about individual users. Different web pages have different authoring styles, and although each page has surrounding key words, it is just too complex for traditional text mining algorithms.

According to analysis targets, web mining can be divided into three different types, which are

- Web usage mining
- Web content mining
- Web structure mining

2. LITERATURE SURVEY

A. Web Usage Mining

The amount of data continues to grow at an enormous rate even though the data stores are already vast. The primary challenge is how to make the database a competitive business advantage by converting seemingly meaningless data into useful information. How this challenge is met is critical because companies are increasingly relying on effective analysis of the information simply to remain competitive.

Implicit Feedback

T. Joachims, L. Granka, B. Pang, H. Hembrooke, and G. Gay examines the reliability of implicit feedback generated from clickthrough data in WWW search. Analyzing the users' decision process using user clicktracking and comparing implicit feedback against manual relevance judgments, they conclude that clicks are informative but biased. While this makes the interpretation of clicks as absolute relevance judgments difficult, they show that relative preferences derived from clicks are reasonably accurate on average.[10]

The paper [11] propose and evaluate a method for the automated segmentation of users' query streams into hierarchical units and the classifiers can improve on timeout segmentation, as well as other previously published approaches, bringing the accuracy up to 92% for identifying fine-grained task boundaries, and 89-97% for identifying pairs of queries from the same task when tasks are interleaved hierarchically.

Query Recommendation Using Query Logs in Search Engines

The paper [13] present the results from a human subject study that strongly indicate the feasibility of automatic query-goal identification. Then, propose two types of features for the goal-identification task: user-click behavior and anchor-link distribution. The experimental evaluation shows that by combining these features it can correctly identify the goals for 90% of the queries studied. R. Baeza-Yates, C. Hurtado, and M. Mendoza proposed a method that, given a query submitted to a search engine, suggests a list of related queries. The related queries are based in previously issued queries, and can be issued by the user to the search engine to tune or redirect the search process. The method proposed is based on a query clustering process in which groups of semantically similar queries are identified.

Varying Approaches to Topical Web Query Classification

Topical classification of web queries has drawn recent interest because of the promise it offers in improving retrieval effectiveness and efficiency. However, much of this promise depends on whether classification is performed before or after the query is used to retrieve documents. The paper [4] examines two previously unaddressed issues in query classification: pre vs. post-retrieval classification effectiveness and the effect of training explicitly from classified queries vs. bridging a classifier trained using document taxonomy.

3. PROBLEM FORMULATIONS

The World Wide Web (www or w3 commonly known as the web) is the largest database available with growth at the rate of millions of pages a day and presents a challenging task for mining web data streams. The different people may have distinct queries to search results and retrieve their results when they submit query in a search engine. The problem of clustering the feedback sessions are addressed in the dissertation.

The user interface compactly displays web pages in a hierarchical category structure. Heuristics are used to order categories and select results within categories for display. Users can further expand categories on demand. Tooltip-like overlays are used to convey additional information about individual web pages or categories on demand. They compared our category interface with a traditional list interface under exactly the same search conditions. This dissertation describes each of these components in more detail to tackle this problem, and proposes summarization of individual queries into concepts, where a concept is a small set of queries that are similar to each other. Using concepts to describe contexts, it also addresses the sparseness of queries and interprets users' search intents more accurately. To mine concepts from queries, the clicked URL for queries is taken as the features of the queries. In other words, it mines concepts by clustering queries in a click-through bipartite. In DCAPC, it proposes how to mine concepts of queries. With the help of concepts, a context can be represented by a short sequence of concepts about the queries asked by the user in the session. The next issue is how to find the queries that many users often ask in a particular context.

4. SYSTEM METHODOLOGY

A. Proposed Methodology

Data Mining refers to “extracting” or “mining” knowledge from large amounts of data. It is also called as knowledge mining from data. Search engine is one of the most important applications in today’s internet. Users collect required information through the search engine in the internet. Analyzing user search goal is essential to provide best result for which the user looks for in the internet. Feedback sessions have been clustered to discover different user search goals for a query. Pseudo-documents are generated through feedback sessions for clustering. To understand the user search goals efficiently using Dynamic Classified Average precision Crawling (DCAPC) algorithm.In the dissertation, a Genetic approach has been proposed to infer user search goals for a query by clustering its feedback sessions represented by pseudo documents. They introduce feedback sessions to be analyzed to infer user search goals rather than search results or clicked URLs. Both the clicked URLs and the unclicked ones before the last lick are considered as user implicit feedbacks and taken into account to construct feedback sessions. Therefore, they provide feedback sessions can reflect user information needs more efficiently. They also map feedback sessions to pseudo documents to approximate goal texts in user minds. The pseudo-documents can enrich the URLs with additional textual contents including the titles and snippets and Meta data activities. Based on these pseudo-documents, user search goals can then be discovered and depicted with some keywords. Finally, a new criterion DCAPC is formulated to evaluate the high performance of best accuracy of user search goal inference.

5. EXPERIMENTAL RESULTS

A real search result clustering system is designed. The system accepts query inputs from users and pass them to one of the following search engines: Google, MSN, and AltaVista (but in this dissertation, only Google is used). This system is used for both training data collection and algorithm evaluation using DCAPC approach. Here the experimental results are given by using 100 ambiguous queries for search result. The following Table 1 represents the evaluation mean average values of proposed method by comparing CAP, VAP, RISK, and AP precision values.

TABLE 1. Evaluation Results of CAP, VAP, and RISK & AP Mean Average Values for 100 Queries

S.NO	Methods	Mean Average
1.	CAP	0.81
2.	AP	0.92
3.	RISK of AP	0.20
4.	VAP of AP	0.96

ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 3, Issue 1, pp: (19-26), Month: January - March 2015, Available at: www.researchpublish.com

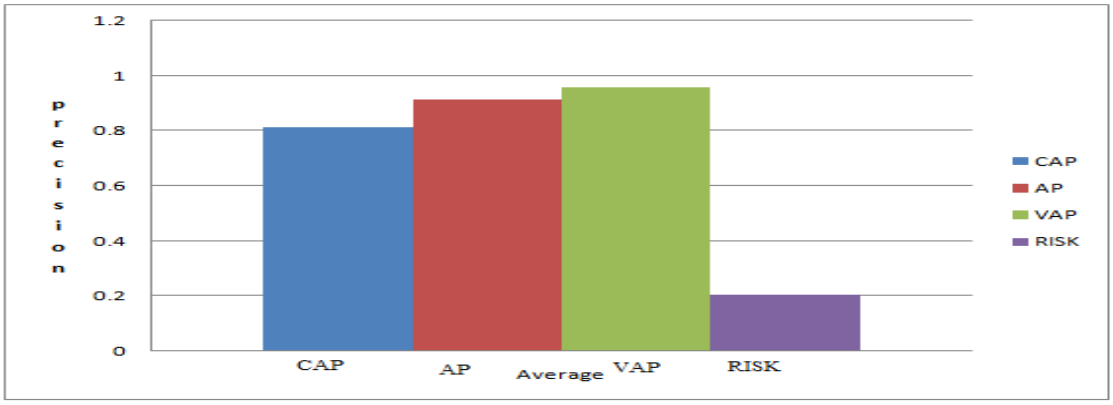

Figure 5.1 Comparison of Mean Average values of CAP, AP, VAP, RISK

The Figure 5.1 clearly depicts the values of CAP, VAP, RISK, AP in which it can be easily determined that the proposed AP values are higher than existing CAP values. The voted AP (VAP) which includes more clicks in feedback sessions. The value of VAP in existing is near to 0.755 but in proposed the VAP values reach their maximum level of 1 for most of the given queries .The Risk factors are also calculated in this experiment.

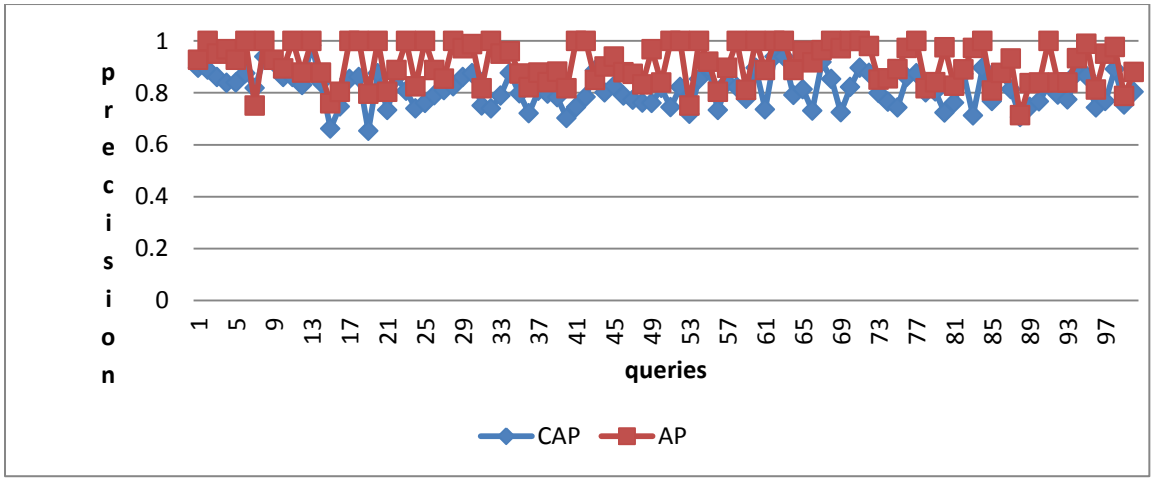

Figure 5.2: Comparative results of proposed AP and CAP for 100 Queries

The Figure 5.2 shows the comparison of CAP and AP values where the AP value Higher than the CAP values.

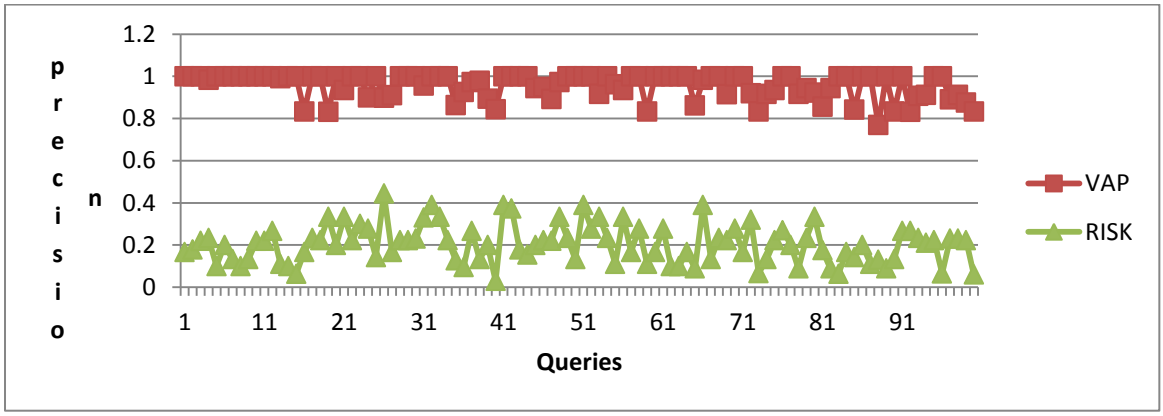

Figure 5.3: Comparative results of VAP and RISK factors of AP

The Figure 5.3 shows the comparison of VAP and RISK values where the VAP value achieve the highest value of 1 and the Risk value gradually decreased.

6. CONCLUSIONS AND FUTURE ENHANCEMENT

A. Conclusion

This DCAPC studied learning an implicit query from user click movement patterns. The learning was done with a machine learning model that uses document relevance feedback from other sessions as an indirect ground truth. In particular, user click movement patterns during reading were used to generalize feedback across sessions.

DCAPC algorithm improves the search performance and the user experience. The experiments show that it is possible to infer the interest from user click movements at least in a controlled laboratory experimental setting. If user click tracking is cheaply available, it can be combined with explicit relevance judgments or keyword-based search to increase retrieval performance.

B. Future Enhancement

This dissertation the keywords deal only with meta data in single page. In future it may search the keywords in multiple pages .The queries used here are ambiguous one. In future it may be extended to general query. One limitation of the experimental results is a small sample size.A larger study would be needed to better evaluate the retrieval performance. The models presented here are built of standard machine learning components, and the user click movements features were harvested from various psychological studies.

REFERENCES

[1] R. Baeza-Yates and B. Ribeiro-Neto, Modern Information Retrieval. ACM Press, 1999

[2] R. Baeza-Yates, C. Hurtado, and M. Mendoza, “Query Recommendation

[3] Using Query Logs in Search Engines,” Proc. Int’l Conf. Current Trends in Database Technology (EDBT ’04), pp. 588-596, and 2004.

[4] D. Beeferman and A. Berger, “Agglomerative Clustering of a Search Engine Query Log,” Proc. Sixth ACM SIGKDD Int’l Conf. Knowledge Discovery and Data Mining (SIGKDD ’00), pp. 407-416, 2000.

[5] S. Beitzel, E. Jensen, A. Chowdhury, and O. Frieder, “Varying Approaches to Topical Web Query Classification,” Proc. 30th Ann. Int’l ACM SIGIR Conf. Research and Development (SIGIR ’07), pp. 783-784, 2007.

[6] H. Cao, D. Jiang, J. Pei, Q. He, Z. Liao, E. Chen, and H. Li, “Context-Aware Query Suggestion by Mining Click-Through,” Proc. 14th ACM SIGKDD Int’l Conf. Knowledge Discovery and Data Mining (SIGKDD ’08), pp. 875-883, 2008.

[7] H. Chen and S. Dumais, “Bringing Order to the Web: Automatically Categorizing Search Results,” Proc. SIGCHI Conf. Human Factors in Computing Systems (SIGCHI ’00), pp. 145-152, 2000.

[8] C.-K Huang, L.-F Chien, and Y.-J Oyang, “Relevant Term Suggestion in Interactive Web Search Based on Contextual Information in Query Session Logs,” J. Am. Soc. for Information Science and Technology, vol. 54, no. 7, pp. 638-649, 2003.

[9] T. Joachims, “Evaluating Retrieval Performance Using Clickthrough Data,” Text Mining, J. Franke, G. Nakhaeizadeh, and I. Renz, eds., pp. 79-96, Physica/Springer Verlag, 2003.

[10] T. Joachims, “Optimizing Search Engines Using Clickthrough Data,” Proc. Eighth ACM SIGKDD Int’l Conf. Knowledge Discovery and Data Mining (SIGKDD ’02), pp. 133-142, 2002.

[11] T. Joachims, L. Granka, B. Pang, H. Hembrooke, and G. Gay, “Accurately Interpreting Clickthrough Data as Implicit Feedback,” Proc. 28th Ann. Int’l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR ’05), pp. 154-161, 2005.

[12] R. Jones and K.L. Klinkner, “Beyond the Session Timeout: Automatic Hierarchical Segmentation of Search Topics in Query Logs,” Proc. 17th ACM Conf. Information and Knowledge Management (CIKM ’08), pp. 699-708, 2008.

[13] R. Jones, B. Rey, O. Madani, and W. Greiner, “Generating Query Substitutions,” Proc. 15th Int’l Conf. World Wide Web (WWW ’06), pp. 387-396, 2006.

[14] U. Lee, Z. Liu, and J. Cho, “Automatic Identification of User Goals in Web Search,” Proc. 14th Int’l Conf. World Wide Web (WWW ’05), pp. 391-400, 2005.