Automatic Identification of user goals in web search based on classification of click-through results

Synopsis of the Thesis to be submitted for the Award of the Degree of

Masters of Technology in Computer Science and Engineering

by

Amar Kumar Dani

(03CS3014)

Under the guidance of

Prof. Chittaranjan Mandal & Prof. Pabitra Mitra

Department of Computer Science & Engineering

Indian Institute of Technology

Kharagpur-721302, India

May, 2008

Certificate

This is to certify that the report entitled '**Automatic Identification of user goals in web search based on classification of click-through results**' submitted by Mr. **Amar Kumar Dani** to the Department of Computer Science & Engineering, Indian Institute of Technology, Kharagpur in partial fulfillment of the requirement for the degree of **Master of Technology** during the academic **year 2007-2008** is a record of authentic work carried by him under my supervision and guidance.

Prof. Chittaranjan Mandal
Department of Computer Science and Engineering
Indian Institute of Technology
Kharagpur 721302, INDIA
May, 2008

Prof. Pabitra Mitra
Department of Computer Science and Engineering
Indian Institute of Technology
Kharagpur 721302, INDIA
May, 2008

Abstract

The Web is a huge resource for people who use search engines to search for specific pages related to their specific needs. As a result, search engines are continuously striving to improve their ranking algorithms to efficiently fulfill end users' search needs. While such algorithms are effective in handling large volumes of web documents and queries, an understanding of web queries remains quite primitive. In recent years, extensive study has been performed to characterize how users seek information on the web. Such studies focus on how users modify queries and what are the possible user goals in web search. This project is inspired by a study about identification of user goals in web search carried out by Broder which described how the goal behind a web query can be classified into three categories: Navigational searches are those which are intended to find a specific web site that the user has in mind; informational searches are intended to find information about a topic; transactional searches are intended to perform some web-mediated activity. The objective of this work is to identify automatically if the user query has a predictable goal and if it does have a unique goal, what it really is. The results are very promising. The identification of user goals can ultimately be used to achieve efficient and effective ranking of search engine results. The design of a Search Engine based on user goals is also presented in the work.

Introduction

1.1 Classical Information Retrieval

Classic IR (information retrieval) is inherently predicated on users searching for information, the so called "information need". But the need behind a web search is often not informational - it might be navigational (give me the url of the site I want to reach) or transactional (show me sites where I can perform a certain transaction, e.g. shop, download a file, or find a resource). A central tenet of classical information retrieval is that the user is driven by an information need. But the intent behind a web search is often not informational. In fact, informational queries constitute less than 50% of web searches.

1.2 Web Information Retrieval

In a web information retrieval environment, information need is associated with some underlying task. The web is a unique searching environment that necessitates further and independent study. This “unique search environment” represents the recent interest in complex subject of understanding the user goals when submitting a query to a search engine. Web Search users tend to make use of short queries to represent their needs, implying that a search engine must make use of other features and algorithms that enhance the relevancy of the search results.

1.3 A taxonomy of web searches

In the web context the "need behind the query" is often not informational in nature. Broder classified web queries according to their intent into 3 classes:

Navigational Queries

The purpose of such queries is to reach a particular site that the user has in mind, either because they visited it in the past or because they assume that such a site exists. Examples: google, yahoo, American airlines

Informational Queries

The purpose of such queries is to find information assumed to be available on the web in a static form. No further interaction is predicted, except reading. Informational pages are characterized by lot of textual information which is meant to be read by the user. Examples: bird flu, kidney stones, pregnancy, etc.

Transactional Queries

The purpose of such queries is to reach a site where further interaction will happen. This interaction constitutes the transaction defining these queries. We define a transactional page as one where a user can perform some transaction where a transaction is constituted by being able to place an order for some product or to be able to download a file or get to the resource indicated by the query term. Examples: myspace layouts, msn messenger, free ringtones, funny pictures, etc.

1.4 Literature Survey

Based on the taxonomy presented by Broder, Kang and Kim proposed an automatic query goal identification scheme to distinguish between Navigational and Information queries. They divided a set of web WT10g into 2 sets, DBTopic and DBHome, and based on these sets they extracted features such as the distribution of terms in a query, the mutual information between the query terms, the usage rate of query terms as anchor texts and POS information. However, the authors concluded that there is a significant inadequacy in the proposed approach for classifying queries.

Lee et al. built upon this work and substantiated the idea that the process of automatic query-goal identification is a feasible objective in Web IR. In an initial analysis following a human survey they demonstrate how more than half the queries have a predictable goal (the intention is not ambiguous) and that around 80% of those with an unpredictable goal are either software or person names. Their work also introduced two new features for automatic classification: click distribution and anchor link distribution which yielded an accuracy of 90% for query classification between navigational and informational query classes. Both features are modeled using statistical distributions from past user interaction based on the intuition that if a particular hyperlink shows authoritativeness in terms of a given query, the most probable intention is navigational.

Both Broder and Rose and Levinson observe that the "need" behind considerable amount of queries is transactional. Kang proposes a scheme that serves transactional queries postulating that hyperlinks are a good indicator in classifying queries and collecting relevant pages for transactional queries. The author suggests that by observing the actions related to a hyperlink, cue expressions related to transactional queries can be extracted from tagged anchor texts and titles. These actions are determined by observing the link types of the hyperlinks extracted from relevant web documents.

A frequent occurrence of music, text, application and service link types suggest that the intention of the query is transactional. In a separate study, Li et al. propose a mechanism for identifying transactional queries by building a transactional annotator from a corpus collected from the web that is capable of highly specific labeling of many distinct transaction types. The authors suggest that transactional features engineering, hand crafted regular expressions and an index of terms are suitable and robust for identifying transactional terms within a web document. The process relies on regular expressions that identify the existence of transactional patterns and a dictionary of negative patterns that evaluates the presence of any negative terms collected by the object identifier.

1.5 Problem Statement: Automatic query-goal identification

All the approaches to identification of user goals in web search mentioned above have not taken all the three classes of goals into consideration. Lee et al classifies the queries into navigational and informational whereas others provide features useful for navigational and transactional query classification. But, we have seen that most researchers agree on the existence of three fold user intent in web searches as proposed by Broader: navigational, informational and transactional. To be able to utilize any information regarding user intent, any search engine must be able to detect and distinguish between the three classes of user intention. Further, the query intention identification system must be able to clearly distinguish between the ambiguous queries for which the intention is not clearly identified and the unambiguous queries where the intention is clearly identified. This work focuses on automatically identifying whether the query has a predictable goal and if so, detect the goal of the query.

Automatic query-goal identification: Experimental Setup & Approach

Our query intention classifier takes the past user click behavior into account to classify the intention of a query entered into a search engine. The click through data of a search engine consists of the query and the url of the result clicked at by the user who issued the query. It is based on the intuition that user's goal for a given query may be learned from how users in the past have interacted with the returned results for this query. Figure 1 shows the various steps involved in the query classification process.

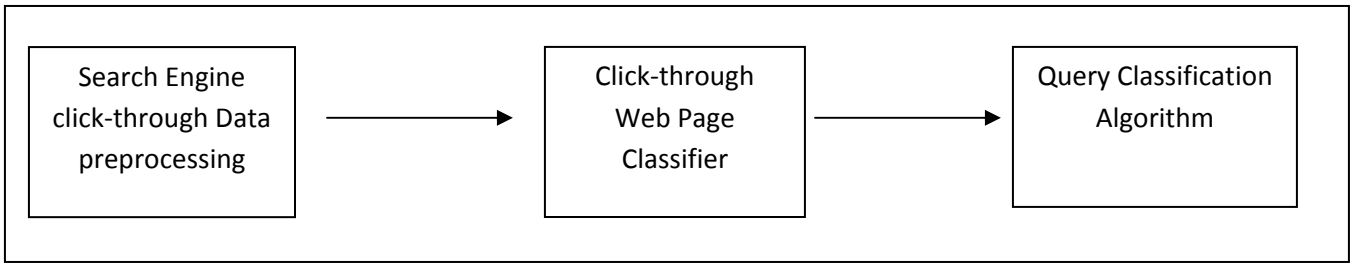

Fig 1: Steps in Query Classification

2.1 Search Engine Click through data preprocessing

In order to perform our experiments and classification algorithm, we use the AOL Search Engine click-through data.

AOL Search Engine click-through data

The click-through data is taken from an AOL log of search data released to the public in August 2006. This includes around 36 million search queries from circa 658,000 of AOL's users taken from the period between March 01 2006 and May 31 2006. Each line of data includes an anonymous ID, the actual query, the date and time the query was submitted, the page rank and the domain portion of the URL as the click-through results. The query issued by the user is case shifted with most punctuation removed.

Data Processing

For experimentation purposes, the AOL data is processed to extract the queries that have been issued by many different users. Duplicate entries are eliminated from the data that is issued by the same user at different points of time. Then the data is sorted based on the number of times different users issue the same query. The queries to be used for testing are then selected across all the alphabets. Further, while selecting the queries for testing purposes, it had to be borne in mind that queries are represented from all the classes and that all forms of ambiguities are taken care of.

The AOL data click-thorough url consists of only the domain name of the site. But for classification purposes, the complete url is required. To get the complete url, a virtual user is simulated whereby the query term and the domain name together are used to search on the Yahoo search engine using the Yahoo API. The first matching result is taken to be the probable click-through. Many of the sites mentioned in the AOL data have become obsolete and hence do not match with any of the Yahoo results. The queries having sufficient clicks after eliminating the unmatched urls are taken for testing.

Questionnaire Design for Manual classification

In order to test the automatic classification results, a user survey is conducted whereby **30 users** are asked to manually indicate the goal of **65 queries**. The users are not directly asked to indicate whether the queries are navigational or informational or transactional. Instead, three options are given defining the goals and the users

only have to mention the choice number beside the query. The queries are then classified based on how many users classified the query into which class.

2.2 Click-through web page classifier

The click-through page is classified into three classes: navigational, informational and transactional. Navigational pages are those which constitute the home page of web sites. They are characterized by having a small url depth, occurrence of query keywords in the domain name of the url and having a high ratio of clicks to the total number of clicks issued for the query. Informational pages are characterized by having lots of textual material which can be read up. Transactional pages are characterized by having lots of other HTML elements including tables, images, divs, etc. They also can be classified based on occurrence of common commercial terminology like occurrence of words including 'product', 'hot product', 'download', etc. For transactional and informational queries, there does not exist one "correct" answer as the user can do the transaction from different sites as well as get information from multiple sites.

Corpus construction

A total of **322 instances** were manually classified for training the three-fold classifier with **127 navigational** pages, **92 informational** pages and **103 transactional** pages. The pages for classification are selected from the click-through results so that a representation of the click-through pages could be made to some extent.

Feature Engineering and Classification Algorithm

A total of **152 features** are extracted from the HTML pages by writing a parser of the HTML page and extracting features including HTML, url based features and bag of words features. Then, feature selection algorithm was run to extract the important features. For each set of features, the classification algorithm was run to check if the classification accuracy is increased.

Experimentation was done with several classification algorithms including SVM, Naïve Bayes, Random Forest, J48, etc. to see which results in best classification accuracy. Finally, the meta classifier RandomCommittee is used along with Attribute Selection algorithm which resulted in 17 best features. Classification accuracy (10 fold cross validation) of 91.3043 was achieved.

Web Page Classification Results

10 fold cross validation accuracy of **91.3043** was achieved using the **RandomCommittee** classification algorithm and **17 best features** selected using Attribute selection algorithm. The following tables show the accuracy achieved across the three different classes.

Class	Recall	Precision	F-Measure
navigational	0.992	0.984	0.988
informational	0.924	0.817	0.867
transactional	0.806	0.922	0.860

Table showing accuracy of classification across the 3 classes

2.3 Query Classification Algorithm

Following are the steps of the algorithm used to classify a query into navigational, informational, transactional or ambiguous query:

1. For each query, classify each click-through result into three classes: navigational, informational or transactional

2. Count the number of informational and transactional clicks for the query
3. For the navigational results, compare the domain name of the website to compare the similarity. If they are similar, add their counts into one
4. For the navigational results, the navigational result with the maximum clicks is taken to be the navigational representative. Other navigational clicks are added to transactional clicks for the query
5. The belongingness value for each class is calculated by dividing the number of clicks for each class with the total number of clicks for the query
6. The class with maximum belongingness value and the one with 2^{nd} maximum belongingness value are chosen and the difference d between them calculated. If d is greater than a threshold value t, the query is classified to belong to the class with maximum belongingness value else it is termed ambiguous with belonging to both the maximum and 2^{nd} maximum classes. Various values of threshold are experimented with and the value chosen for t is finally .2

Query Classification Results

The results of automatic classification of the 65 queries manually classified by 30 users are compared with the manual classification results. The salient features of the results are as follows:

1. Out of the 65 queries, 15 were manually classified as navigational, 19 as informational, 19 as transactional and 12 ambiguous (10 of type informational + transactional and 1 each of the other types)
2. Out of the **15 navigational queries**, **all** were detected to be navigational by the query classification algorithm
3. Of the **19 transactional queries**, **18** were classified as transactional
4. Of **19 informational queries**, **11** were detected to be informational
5. The ambiguity of type informational + navigational and navigational + transactional was detected by our algorithm, but for the type informational + transactional, out of **10 such queries, only 2 were detected to be so**
6. By analyzing the data, it was found that, the queries which represent **names of celebrities** including Anna Benson, Jessica Alba, etc. have been misclassified. Whereas users identified these queries to be of type informational, the pages corresponding to these queries were more transactional in nature (having more of downloadable images than textual material).
7. Out of the **8 Informational pages** wrongly classified, **4 are the names of famous people** and out of the **10 misclassified informational + transactional queries**, **5** were the names of famous people and these were classified as transactional pages.

Conclusion

From the results, we can conclude that most queries issued to a search engine have a predictable goal which can be identified automatically. Of the 53 queries with a unique goal, 44 were correctly classified and of the 12 ambiguous queries, 5 queries were wrongly classified as having a predictable goal whereas others were correctly detected as not having a predictable goal though not all gave the same class of unpredictability. Most queries that were misclassified were the names of people or places in which case, the users indicated an informational goal whereas the classifier detected a more transactional goal. This is due to the definition of the web page classifier which classifies the web pages corresponding to the click-through urls of these queries as transactional rather than informational. Such a distinction between informational / transactional pages is very subjective and should be done taking the final goal of optimizing search engine results presentation into consideration.

References

1. B. Schneiderman, D. Byrd, and W. B. Croft. Clarifying search: A user-interface framework for text searches. *D-Lib Magazine*, January 1997.
2. Andrei Broder. A taxonomy of web search. SIGIR Forum, 36(2):3–10, 2002.
3. Bernard J. Jansen and Udo Pooch. A review of web searching studies and a framework for future research. In J. Am. Soc. Inf. Sci. Technol., volume 52, pages 235–246, New York, NY, USA, 2001. John Wiley & Sons, Inc.
4. In-Ho Kang. Transactional query identification in web search. In AIRS, pages 221–232, 2005.
5. In-Ho Kang and GilChang Kim. Query type classification for web document retrieval. In SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, pages 64–71, New York, NY, USA, 2003. ACM Press.
6. Uichin Lee, Zhenyu Liu, and Junghoo Cho. Automatic identification of user goals in web search. In WWW '05: Proceedings of the 14th international conference on World Wide Web, pages 391–400, New York, NY, USA, 2005. ACM Press.
7. Daniel E. Rose and Danny Levinson. Understanding user goals in web search. In WWW '04: Proceedings of the 13th international conference on World Wide Web, pages 13–19, New York, NY, USA, 2004. ACM Press.
8. Yunyao Li, Rajasekar Krishnamurthy, Shivakumar Vaithyanathan, and H. V. Jagadish. Getting work done on the web: supporting transactional queries. In SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, pages 557–564, New York, NY, USA, 2006. ACM Press.
9. C. Silverstein, M. Henzinger, H. Marais, and M. Moricz. Analysis of a very large Web search engine query log. SIGIR Forum, 33(1):6 12, 1999.
10. Sullivan. Searches per day. http://searchenginewatch.com/reports/article.php/2156461, 2003.
11. AOL 500k User Session Collection creator. AOL 500k User Session Collection (collection).