Determining the User Intent Behind Web Search Queries by Learning from Past User Interactions with Search Results

Ariyam Das
Yahoo! Research and Development
Bangalore, India
ariyam@yahoo-inc.com

Chittaranjan Mandal
Department of CSE and School of IT
IIT Kharagpur, India
chitta@iitkgp.ac.in

Chris Reade
Department of Informatics and Operations Management
Kingston University, UK
Chris.Reade@kingston.ac.uk

ABSTRACT

Understanding the intent of users behind their web search queries is very useful in serving them with relevant advertisements and in ranking the search results effectively. Existing approaches broadly classify the user intent behind web queries into three categories: *navigational*, *informational* and *transactional*. In this study, we present a query classification framework that attempts to automatically determine the nature of the query. Assuming that the query has a predictable goal, our framework either classifies the goal into one of the three classes, or if the query goal is ambiguous, the framework predicts which classes have a high association with the given query. Our proposed approach deals with some of the limitations of previously reported methods by studying how users have interacted with the search results in the past. In this work, we first build and train an efficient web page classifier that categorizes a web page into *navigational*, *informational* or *transactional* classes by operating on twelve best distinguishing features selected through Correlation-based Feature Selection algorithm from a total of 152 url, html, lexical and bag of words features extracted from the click-through results of Yahoo-Bing search. We then analyzed the click-through results of Yahoo-Bing search engine with our classifier for a given set of queries and applied a set of fuzzy rules to classify the user goals behind the corresponding queries as either ambiguous or into any one of the known three classes. The goals for the same set of queries were manually classified through a questionnaire involving 50 participants. The initial results are compared and presented to show the efficiency of our proposed user goal identification technique.

Keywords

User goals, query classification, web search, user behavior

1. INTRODUCTION

Identifying the end user goal in web search can be extensively utilized in improving ad targeting and page ranking as well as in the presentation of the search results in a big way. Andrei Broder in [1] classified the query goals into following three categories based on the user intent:

1. *Navigational* - The underlying intent of the user is to reach a particular site.
2. *Informational* - The goal of the user is to gather some information from one or more web pages.
3. *Transactional* - The intent is to perform some web-mediated activity, like downloading files, purchasing items online, etc.

Based on the above taxonomy, Kang and Kim in [5] first proposed an automatic query goal identification scheme to distinguish only between navigational and informational queries. Lee et al. [2] extended this work and achieved an accuracy of 90% in classifying queries between navigational and informational classes by considering click distribution and anchor link distribution for automatic query classification. The authors in [2] primarily focused on identifying features that can strongly classify navigational queries. Baeza-Yates et al. [4] used supervised and unsupervised learning to classify queries as informational, not informational, or ambiguous. Most of these approaches have not considered all the three classes together to identify the user goals in web queries. Jansen et al. in [3] first considered all the three classes but did not look beyond the query and url for the classification purpose. The authors in their study [3] experimented with 400 queries and achieved an accuracy of 74% in their automatic classification; they found nearly 25% of the queries to be vague or multi-faceted. We build upon the latter work based on the intuition that the user goal for a given query may be learned from how users in the past have interacted with the returned results for the query.

The rest of the article is organized as follows. We present our overall approach in section 2. The questionnaire design for manual classification is discussed in section 3. The results are analyzed in section 4. Finally, section 5 concludes the article.

2. PROPOSED APPROACH

Our overall proposed approach works in the following three steps.

2.1 Building and Training a Web Page Classifier

The 19th International Conference on Management of Data (COMAD), 19th-21st Dec, 2013 at Ahmedabad, India.

We first briefly summarize the overall features that can be used to distinguish navigational, informational and transactional pages.

1. *Url Features* - These are mainly used to identify navigational pages. Navigational pages, being homepages of websites, generally have distinguishing *url* features such as smaller url depth, url length, occurrence of query keyword in the domain name, etc.

2. *HTML Features* - Different html elements such as tables, images, download buttons, etc. dominate in transactional pages. These html features along with the presence of other prominent features can help in differentiating transactional pages.

3. *Lexical Features* - Features such as words and sentences per paragraph, amount of text per paragraph, etc. dominate in informational pages. These along with html features are particularly helpful in distinguishing between transactional and informational pages.

4. *Bag of Words Features* - Transactional and informational query classes can be tied with some specific keywords. These words are manually selected and weighted differently depending on its occurrence in meta text, title text, headings, special text, anchor text, alternate text and input text. Li et al. in [6] studied extensively on how to utilize these bag of words features to identify transactional pages. For example, frequent occurrences of keywords such as *buy*, *cart*, *online store*, etc. can indicate a transactional page. Few words like *homepage*, *welcome*, etc. can be used to distinguish navigational pages. However, one cannot rely on the bag of words features to identify informational pages as frequently occurring keywords for all domains of information are not easily predictable.

Once we have identified all the relevant features, we then developed a html parser that extracted a total of 152 distinguishing features from the click-through results of Yahoo-Bing search for classifying a web page. We next build a corpus to train our classifier by manually classifying a substantial number of pages from Yahoo-Bing search results. A total of 415 instances were manually classified for training the three-fold classifier with 132 navigational pages, 164 informational pages and 119 transactional pages.

Then we applied the Correlation-based Feature Selection algorithm and selected a subset of twelve best correlated features for our classifier. After the feature selection, we compared the performance of our classifer using different machine learning algorithms viz. *Naive Bayes*, *J48*, *Random Forest*, *SMO* and *Random Committee*. Figure 1 reports the 10 fold cross validation accuracy of our classifier for the different machine learning algorithms. We subsequently selected the *Random Committee* classification technique, achieving the highest 10 fold cross validation accuracy of 93%, for our classifier.

2.2 Processing Historical Click-Through Results

The historical click-through results of Yahoo-Bing search under consideration, comprised of the following: a transformed query term, an encrypted user ID (if the user had logged in), a browser cookie, a valid identifier assigned to

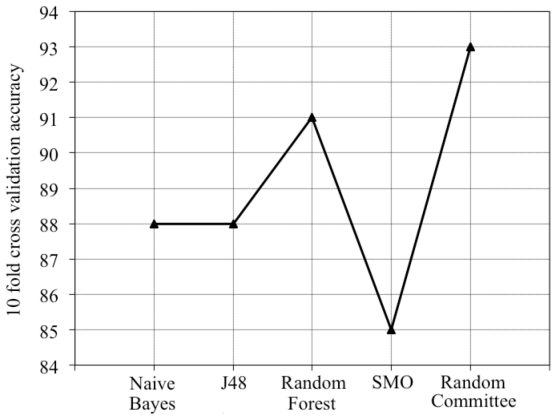

Figure 1: Measuring 10 fold cross validation accuracy with different machine learning algorithms

the user at the start of a new network session with Yahoo, the url of the result clicked at by the user who issued the query, the rank of the result that the user clicked and the timestamp of the click event.

For a given set of user queries, we processed the corresponding historical click-through results according to the following steps:

1. We ignored duplicate clicks from the same user on the same click-through url.

2. For a given query term, if the same user (as identified from the user ID or browser cookie as appropriate) clicks on more than one search result within the same network session, we assign weight on each of those clicks depending on the time spent on each of those results. We calculate the time spent on a search result by computing the difference between the timestamps of two consecutive user clicks. If the differences are more than a threshold, the weight for each click is given as the fraction of time spent on the respective search results. If the time difference does not exceed the threshold, then we consider clicks to be of equal weights. The latter is taken into consideration for scenarios where user would click and open search results in separate tabs and windows in the beginning itself. Intuitively, this overall weight assignment is appropriate because if a user clicks and spends considerable time on more than one search result, it would imply that the earlier clicked search results have not satisfied the userś goal completely. Figure 2 shows the percentage of users in one day who clicked on one or more Yahoo-Bing search results for the same query within the same network session. By adding weights to the clicks on the basis of the time spent by a user on each clicked result, we are incorporating the knowledge of how users have interacted with the search results for the same query term in the past. The time for the terminal click is assigned as the average time spent by other users for the same query term.

3. In this step, we classify the web pages corresponding to the click-through urls of the queries by our three-way classifier.

4. For each query, we then check how many users have visited navigational pages, how many have viewed transactional pages and how many have read informational pages, by counting the number of *weighted* clicks on the navigational, transactional and informational pages.

5. For computing the navigational clicks, we perform an additional adjustment. We compared the domain name of the websites and if they were found similar, we added their counts into one. Since, there exists only one correct navigational page for a query, the navigational page with maximum clicks is taken to be the total navigational clicks and the clicks for other navigational pages are added to the transactional clicks. This adjustment is apt for cases where the end goal of a user is transactional although he might have visited several navigational pages of sites offering those services.

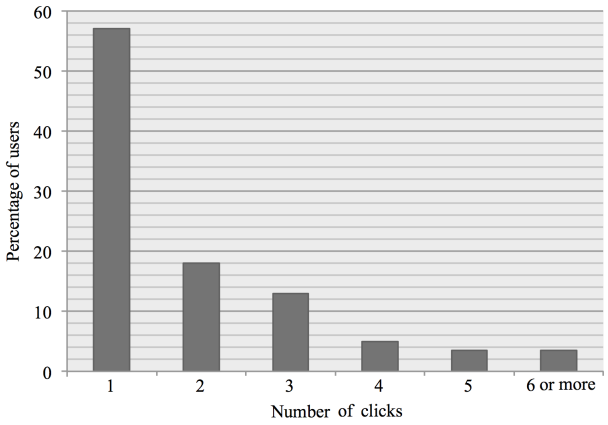

Figure 2: Percentage of users clicking on one or more search results

2.3 Applying Fuzzy Rules for Classifying Query Goals

Once the click-through results are processed, we compute the ratio of total weighted clicks distributed between the three goals for each query, to calculate the following fuzzy variables:

goals $\leftarrow$ *navig*,*transactional*,*informational*

$\forall$x $\in$ *goals*, truth(x) = $\frac{weighted\ clicks\ on\ pages\ belonging\ to\ x}{total\ weighted\ clicks}$

$\forall$x $\in$ *goals*, predict(x)
= truth(x) - max($\bigcup_{i \in goals\ and\ i \neq x}$truth($i$))

Using the above variables, we formulate the following fuzzy logic rules, $\forall$x $\in$ *goals*, where thresholds (θ_{navig}, $\theta_{transac}$, $\theta_{inform} \geq \theta_{ambig} > 0$) are assigned during the experiments.

RULE 1: *If* predict(x) $> \theta_x$ *then*
query goal is ***Unambiguous*** *and classified as* x

RULE 2: *If* predict(x) $\not> \theta_x$ *and* predict(x) $\in [-\theta_{ambig}, \theta_{ambig}]$ *then*
query goal is ***Ambiguous*** *and can belong to* $\boldsymbol{x}$

The first rule classifies the unambiguous goals into three classes, while the second rule identifies if the goal is ambiguous or not and which are the probable classes it can belong to. For ambiguous queries, the distinction between the classes can be made fuzzier by having three thresholds for each of the classes instead of just one threshold θ_{ambig}. The above rules work by looking for the most dominant class for each query. If the dominant class has significantly higher clicks (votes) than its competitors then the dominant class is unambiguously accepted as the query goal. Otherwise the query goal is predicted to be ambiguous and the dominant class and its closest competitor(s) are assigned as the probable query goals.

3. QUESTIONNAIRE DESIGN FOR MANUAL CLASSIFICATION

To measure the accuracy of our query classification framework, we conducted an experiment where we selected 100 popular queries, fired into Yahoo search over the last year and asked 50 participants through a survey to indicate the most probable goal if they were to issue such query. We also included software names and names of people that were reported to be ambiguous by Lee et al. in [2]. The participants responded through a questionnaire. This questionnaire design was also critical to collecting reliable results from the users. In the questionnaire it was more important to ask the user to classify the descriptive intention of the query rather than educate the users about the taxonomy and then ask them to classify the query into the three classes directly, since even if two participants had exactly the same descriptive intention, they might end up casting that intention into different choices [2]. We presented the participants with the following choices in the questionnaire:

1. You already have a website in your mind (one particular website only) and your intention is to reach that website with the help of the search engine.

2. Your aim is to obtain information on the *query term.*

3. Your aim is to buy or download or obtain the resource implied by the *query term.*

The users were also provided a few sample classifications as examples for their convenience.

4. RESULTS

From the questionnaire results, we calculated the fraction of candidates who indicated the goal to be navigational, informational or transactional for each query. From the survey, we observed that for queries with unambiguous goals, overwhelming majority of the respondents opted for the same choice. However for ambiguous queries, the difference in percentage of respondents for the highest and second highest query classes are around 20% or less. Based on these observations from our questionnaire results, we consider the following thresholds

$\theta_{navig} = 0.25$, $\theta_{inform} = \theta_{transac} = \theta_{ambig} = 0.2$

With these thresholds, we manually classify the query goals from the questionnaire results using the same set of generic fuzzy rules as discussed in the previous section. We also deduce the goals for each of the selected queries, using our automated query classification framework. The comparison between the manual query classifications and automated classification results derived from our classifier is presented in figure 3.

All the navigational goals were correctly identified by our classifier. Manually classified transactional goals were correctly determined, except for the query *the dark knight rises*

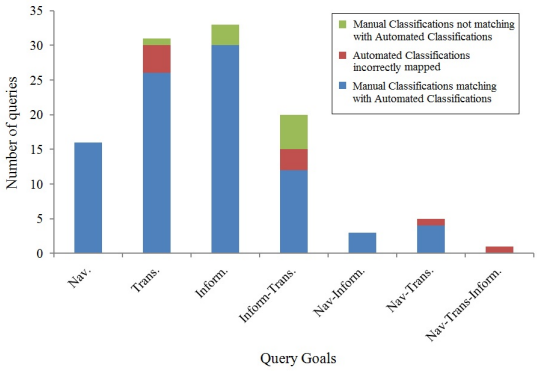

Figure 3: Manual Classifications vs Automated Classifications

which was identified as navigational-transactional goal (ambiguous) because of a large number of visits to the navigational page www.thedarkknightrises.com. Another interesting result is that for trending celebrity names such as Roger Federer, Michael Phelps, the goals were manually classifed as informational but our framework identified them as informational-transactional. On the other hand, for celebrity names like Rihanna, Justin Bieber, which were manually classified as queries having informational-transactional goals, our classifier identified the goal to be transactional . It is actually difficult to predict which class these queries should actually belong to since the user might want to get some information about these celebrities or to download their pictures and videos. The classifier is seen to work fairly well with other ambiguous queries. The query term *apple iphone 5* is the only one whose goal was detected as navigational-transactional-informational although it was manually classified as informational-transactional. This is also because of a large number of visits to the navigational page www.apple.com/iphone It is interesting to point out that, if we split the click through results for US only the goal for *apple iphone 5* becomes navigational-transactional, whereas for some parts of Asia the goal was identified as navigational-informational.

5. CONCLUSIONS

The initial results of our user goal identification technique are encouraging. We classified user goals into three classes with good precision based on the history of how users responded to prior search results. As the result shows, majority of the queries issued to a search engine have a predictable unambiguous goals which can be identified to a great extent by our classifier. Of the 76 queries with unique goals, 72 goals were correctly classified and of the 24 queries with ambiguous goals, 19 were identified correctly. In future, we intent to evaluate the accuracy of our classification technique by splitting the click-through results across demographics. Further work can also be done to hierarchically classify the transactional pages into different types of transactions like commercial transactions, download pages or resource finding pages.

6. REFERENCES

[1] Andrei Broder. A taxonomy of web search. *SIGIR* Forum, 2002.

[2] 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.

[3] Bernard J. Jansen, Danielle L. Booth and Amanda Spink. Determining the User Intent of Web Search Engine Queries. In *WWW* 07: Proceedings of the 16th international conference on World Wide Web, pages 1149−1150.

[4] Ricardo A. Baeza-Yates, Liliana Caldern-Benavides and Cristina N. Gonzlez-Caro. The Intention Behind Web Queries. In *SPIRE* 2006, pages 98−109.

[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] 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.