# A Survey of Query Log Processing Techniques and Evaluation of Web Query Intent Identification

Madhuri A. Potey
College of Engineering,
Pune, India
mapotey@gmail.com

Dhanashri A. Patel
DYPCOE, Akurdi,
Pune, India
dhanashripatel@gmail.com

P. K. Sinha
College of Engineering,
Pune, India
psinha@cdac.in

***Abstract*—Query log is the pouch of valuable information that records user's search queries and related actions on the internet. By mining the recorded information, it is possible to exploit the user's underlying goals, preferences, interests, search behaviors and implicit feedback. The wealth of mined information can be used in many applications such as query log analysis, query recommendation, query reformulation, query intent identification and many more to improve performance of search engine by providing more relevant results. Over the past decade, there has been tremendous work done for improving search engine results to flourish the users for searching. This paper reviews and compares some of the available methods to give an insight into the area of query log processing for information retrieval. Our approach classifies web query intent based on knowledge extraction from query log analysis.**

**Keywords— Query Log Processing; Query Reformulation; Query Intent Identification; Query Recommendation; Query Chains;**

## I. INTRODUCTION

Wide use of the web in day to day living to access information with a single click has increased popularity of research in community of web information retrieval. While surfing the web, users search activities are recorded in query log such as queries submitted by users $(q_i)$, identifier for the user who submitted the query $(u_i)$, timestamp $(t_i)$, and the clicked URL $(URL_i)$. Hence query log can be analyzed to exploit user search behavior to generate better search results [1].

While searching the web, user formulates a query to represent the information need that is usually short and ambiguous; this affects the search results by retrieving irrelevant documents. This can be reduced by reformulating the query with the help of query rewriting or recommendations [2, 3].

Also, the query log can be analyzed to organize it in well accessible manner by using different classification and clustering mechanisms. This analysis result can be used to detect spam [4] which automatically helps to improve search efficiency by filtering them.

The search engine user submits a query with some underlying goal and interest to seek required information. Thus, the user's search intentions can be determined by analyzing their search behavior from query log [5, 6]. The identified search intents can be further used to improve user satisfaction [5].

Along with all above methods, implicit feedback of the user can be utilized for efficient retrieval. The implicit feedback can be obtained from search engine log by finding sequence of subsequent queries by the same user [13].

Even after a lot of research, the growing need of information at a single click encourages researchers to work more in this area. The study and review of recent work related to all above mentioned methods has inspired us to shed light on their relative performance and attempt to improve them.

The rest of the paper is organized as follows. Section II gives literature on work done in most promising query log processing strategies; the taxonomy of query log processing techniques is discussed in section III. The comparative analysis is discussed in section IV. Section V describes proposed work. Experimental results are given in section VI. Section VII summarizes conclusions and research directions.

## II. LITERATURE REVIEW

Even though the query log is a rich source of information on user search actions, the knowledge extraction from large scale query logs in an efficient and effective manner is the crucial challenge. The literature is reviewed to identify different approaches proposed by researchers in order to extract essential features from query log.

### *A. Query Reformulation and Suggestion*

Query reformulation is the frequent insertion, deletion or modification of query terms to a seed query made by users in the hope of retrieving results efficiently.

The analysis and evaluation of query reformulations in web search log extracts the query refinement pattern of web searchers. Jeff Huang and Efthimis N. Efthimiadis proposed the taxonomy of around thirteen methods of query reformulation and listed missed reformulations, as well as the effectiveness measured by adding users click behaviors [2]. Query log helps to reformulate queries and generate suggestions [3, 11]. Clustering reformulations/refinements based on intents approach can help to improve query suggestions [9, 10]. Queries reformulated with different strategies other than query log based strategies would be a separate topic of research.

978-1-4673-4529-3/12/$31.00 

### B. Query Log Analysis

Query log analysis helps in digging additional information such as user preferences, profiling, personalization etc. along with query expansion [15].

Query log analysis provides interaction analysis tools, which help users to understand their own web search behaviors. Work has been done to help users to perform acquiring, organizing, maintaining, retrieving and using information tasks with Web pages, Web search results, emails and other types of files, users need to support their ability to analyze their personal information tasks [15]. Also, query log analysis helps in spam detection by recognizing all such queries which result top N-number of spam pages [4]. Adam Fourney, Richard Mann and Michael Terry, have proposed a system which characterizes internet usability of people from query log and educate themselves on products [16]. The user seed query can be expanded with the help of query log mining, which helps to improve search results and achieve end users satisfaction. The pre-processed queries can be mined using association rules or by finding correlation between query terms and document terms [17, 18]. Continuous sequence of query reformulation is termed as "query chain" that can be obtained by analyzing query log, and used for learning ranking or for query recommendations [13, 14].

### C. Query Log Based Intent Identification

Identifying underlying goals of users i.e. search intent, while submitting the query is an ongoing research topic. According to Broder's taxonomy, the query intent may be informational (searching information on the web), navigational (user wants to reach a particular website), or transactional (user wants to perform the web-mediated task) [6]. Further, Rose and Levinson introduced subcategories for both informational and transactional queries [19].

Study of approaches to identify user intents is carried out extensively by David, Daniel and Kilian [5]. Approaches introduced by many researchers, have shown noticeable improvement in intent identification [7, 8, 9, and 10]. Query intent classification algorithm classifies user intent based on desired contents and use of larger data sets provide more accurate percentages of user intent classification [7]. Another approach introduces a clustering algorithm that uses the user session information along with query URL entries, which identifies same intent query clusters and addresses problem of queries which are polysemic in nature [9]. Also, inferring user intent from reformulation and clicks is possible [12]. Additionally, intents can be identified by utilizing query text and web knowledge [8].

## III. TAXONOMY OF QUERY LOG PROCESSING TECHNIQUES

Figure 1 shows taxonomy of query log processing techniques discussed in section II.

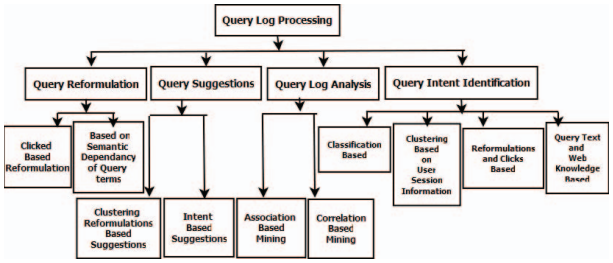


**Fig 1. Taxonomy of Query Log Processing Techniques**

## IV. COMPARITIVE ANALYSIS OF QUERY LOG PROCESSING TECHNIQUES

Table I gives comparative study of query log processing techniques discussed in section II.

## V. OUR PORPOSED WORK

The observations from table I, highlight the scope of improving search results, hence we propose an approach to identify query intent of informational queries and reduction in misclassification of a single term queries by using query chain [5, 10].

### A. Overview of Proposed Work

The proposed system $S = \{Q_L, C, QC, QI\}$ classifies queries given in the query log '$Q_L$' to a given taxonomy of query intent with main three categories $C = \{C_I, C_T, C_N\}$ i.e. informational, navigational and transactional respectively [6]. This system uses the features such as term features, bigram features and content features of query text, adopted from the work of Dayong Wu, Yu Zhang, Shiqi Zhao, and Ting Liu [8]. The query log '$Q_L$' is a set of records $\{q_i, u_i, t_i, URL_i\}$ used for classification. After getting class of informational queries '$C_I$', it is further classified into '$n$' categories $C_I = \{c_1, c_2, ...c_n\}$ using same features. Along with classification of queries from query log '$Q_L$', query chains $QC = \{QC_1, QC_2, ....., QC_n\}$ are formed. For instance, if query log contains queries such as *"Pizza hut", "Pizza Delivery", "Veg Pizza recipe", "Online order pizza", "Pizza Recipe"* for user '$u_i$' then query chains $QC$ contains $QC_1$ = {*"Pizza hut", "Pizza Delivery", "Online order pizza"*}, and $QC_2$ = {*"Recipe of Veg Pizza", "Pizza Recipe"*}. Then intent can be identified $QI = \{QI_1, QI_2, ....., QI_n\}$ based on query chains for input query '$q$' submitted by user.

TABLE I. COMPARATIVE STUDY OF QUERY LOG PROCESSING TECHNIQUES

| Sr. No. | Method Reference | Query Log Processing Techniques | Data Source | Approach to Analyze Query Log | Feature Extraction | Strength | Limitations |
|---|---|---|---|---|---|---|---|
| *1* | Jeff Huang and Efthimis N. Efthimiadis [2]. | Query Reformulations and Suggestions | AOL query log released in 2006 | High Precision Rule Based Classifier | Reformulation Type (13) extracted | 1. Query session boundary detection<br>2. Improve query assistance<br>3. Personalized search<br>4. Interfaces supporting reformulation. | Unable to detect substitutions absence, because word substitutions are wordnet database dependent. |
| *2* | Ashok Veilumuthu and Parthasarathy Ramachandran [9]. | | AOL log. | The Beeferman and Bergers' agglomerative graph based iterative clustering has been adopted to cluster the proposed <session,query>-URL bipartite graph | -- | Solves the problem of polysemic queries in keyword based searches. | -- |
| *3* | Eldar Sadikov Jayant Madhavan Lu Wang Alon Halevy [10]. | | Six months query log of Google. | Markov-model | -- | Comparison of cluster produced using Markov-model is done with clusters using document-click and session-query approaches. | 1. Query refinements limited upto 80 queries.<br>2. No. of document states of each refinements limited to 15. |
| *4* | Shuo-En Tsai, Yi-Shin Chen, Chia-Yu Tsai and Shih-Wei Tu [11]. | | MSN search log | Novel Query Suggetion Approach-Pattern Recognition Query Suggestion. | 1.Frequency of no. of times the keyword appears in session.<br>2.Order of *last* query in session that contain the keyword.<br>3.Click frequency-represents sum of clicked URL's of all queries that contain the keywords. | Top K- suggestions generated by scoring candidates based on trigger and context scores. | -- |
| *5* | Carlos Castillo, Claudio Corsiy, Debora Donato, Paolo Ferraginay, Aristides Gionisz [4]. | Query Log Analysis. | Yahoo! UK search engine (WEBSPAM-UK2006) | View graph and anticlick graph is used to characterize spammicity of queries and documents | Syntactic and semantic features | Web Spam detection. | -- |
| *6* | Paolo Boldi, Francesco Bonchi and Carlos Castillo [14]. | | Yahoo! UK search engine log 2008 | Query log mining | 1. Textual Features<br>2. Session Features<br>3. Time-Related Features | Query flow graph which is used to find query chains and to generate query recommendation. | -- |
| *7* | Adam Fourney, Richard Mann, Michael Terry [16]. | | Query log of top-tier internet search engines | | | It is advantageous for<br>1. Harvesting<br>2. Ordering<br>3. Labelling<br>4. Filtering<br>5. Grouping of search queries related to given product. | -- |
| *8* | Patrick Ngok and Zhiguo Gong [17]. | | -- | Data mining association technique | Query feature vector | Log mining based query expansion. | -- |
| *9* | Cui, Ji-Rong Wen, Jian-Yun Nie, And Wei-Ying Ma [18]. | | Two months user log from Encarta Web Site | Novel approach based on mining techniques | Correlation between query terms and document terms. | Automatic Query Expansion. | -- |
| *10* | David J. Brenes, Daniel GayoAvello, Kilian Perez Gonzalez [5]. | Query Log Based Intent Identification | MSN Query Log for May 2006 | -- | Comparison of query intent detection methods | 1.Review of intent detection method<br>2.Relative performance gives promising lines of work for research domain.<br>3.Also gives unfeasibility of all methods by means of pooling. | 1. Only Navigational intents has been studied<br>2. All intent detection methods were not replicated<br>3. Larger collection of Web pages need to be used to extract anchor texts. |

| Sr. No. | Method Reference | Query Log Processing Techniques | Data Source | Approach to Analyze Query Log | Feature Extraction | Strength | Limitations |
|---|---|---|---|---|---|---|---|
| *11* | Bernard J. Jansen, Danielle L. Booth, Amanda Spink [7]. | Query Log Based Intent Identification | Dogpile search engine transaction log | Classification technique. | -- | 1. Intent identification of Web search queries.<br>2. Validate taxonomy by automatic classification of large query set. | 1. Each query is assigned to one and only one category. |
| *12* | Dayong Wu, Yu Zhang, Shiqi Zhao, Ting Liu [8] | | Sogou (Chinese search engine log) | Intent identification based on query text and web knowledge. | 1. Dependency Relation and Word Sense Features<br>2. Bigram Features<br>3. Term Features<br>4. Content Features of results retrieved by query. | Web Query Intent Identification. | Misclassification of queries because of following reseasons:<br>1. Ambiguous intent of queries<br>2. Some navigational queries are classified as the informational queries.<br>3. Some transactional queries are misclassified as informational queries.<br>4. Some query text and snippet text are pre-processed incorrectly |
| *13* | Filip Radlinski, Martin Szummer, Nick Craswell [12]. | | TREC queries | Approach to identify popular meaning of queries using search log and click behaviours.<br>1. Input query is expanded.<br>2. Then filtering is done.<br>3. Finally queries are Clustered based on their popularity. | Extracts popularity of queries. | Identifies meaning of queries. | -- |

### *B. Query Chain Formation*

We have used query chain concept to utilize the valuable information of query log. Query chain '$QC$' is nothing but a sequence or chain of queries with a similar information need, often performed by the user. Following steps are used to find the query [14].

*1)* First, query log records are sorted as per user id '$u_i$' and then by timestamp '$t_i$'.

*2)* After sorting, the record of each user is divided into sessions '$S_i$' of the same user whenever time difference of two queries exceeds timeout threshold '$t_\theta$'. We use $t_\theta$=30 minutes, which is typical timeout often used in web log analysis [14].

*3)* Again the sessions '$S_i$' of user '$u_i$' in query chains are divided using term features and bigram of query text [8]. Finally, for each query chain the frequency count '$f$', is computed and attached with it which represents the frequency of number of times same query is fired by user.

### *C. Query Intent Identification*

When the user '$u_i$' submits query '$q$' to search engine, we searched the query chains of same user and returned matching query chains. If query chains are representing multiple intents, then intent relevance degree '$IRD_i$' is computed using formula (1).

$$IRD_i = \max\left(QC_i(f)\right) \qquad 1 < i < n$$

Where $QC_i(f)$ represents the number of times the query in query chain $QC_i$ is fired and the maximum of that frequency represents intent relevance degree, i.e. the intent of user behind submitting the query.

## VI. EXPERIMENTAL RESULTS

### *A. Data Set*

To evaluate our approach experimentally, we randomly selected 25,000 queries from AOL query log, released on August 3, 2006 [20], simultaneously removing the repetitive queries, and finally obtained a raw query set of total 23425 queries. These queries are tagged with the help of manual human annotation into their intent categories. After annotation, a set of 23253 labeled queries is used for our work by removing 172 unknown queries. Table II shows summary of annotated queries with their intent categories and percentage. This tagged query set was experimented with for training and testing the classifier.

TABLE II. QUERIES LABELED WITH THEIR INTENT CATEGORIES AND PERCENTAGE

| Sr. No. | Intent Categories | Actual queries | Percentage |
|---|---|---|---|
| 1. | Informational | 17389 | 74.2% |
| 2. | Navigational | 5381 | 23% |
| 3. | Transactional | 483 | 2.1% |

### B. Evaluation Measure

We have used F-value to evaluate performance of intent identification as shown in formula (2).

$$F = \frac{2 \times P \times R}{P + R}$$

Where ' $F$ ' refers to F-value for each intent identification category, ' $P$ ' is the precision rate of a correct intent identification and ' $R$ ' is the recall rate of a correct intent identification.

### C. Experimental Results

In this work, we have conducted two sets of experiments to achieve required results. In the first experiment, we classified AOL query log queries in informational, navigational and transactional categories using the classification approach proposed by Dayong Wu, Yu Zhang, Shiqi Zhao, Ting Liu for Chinese query log [8]. Second experiment gives key contribution, which uses query chains generated from query log for web query intent identification. The table III shows results of these experiments.

TABLE III. CLASSIFICATION RESULTS USING EACH TYPE OF FEATURES

| Types of Features | Evaluation Measures | | |
|---|---|---|---|
| | F(%) | F(%) | F(%) |
| | Informational | Navigational | Transactional |
| TF[8] | 87 % | 92 % | 94 % |
| BF[8] | 85.63 % | 84.59 % | 52 % |
| CF[8] | 78 % | 70 % | 72 % |
| Query Chain | 95 % | 85.15 % | 95.48 % |

## VII. CONCLUSIONS

The user satisfaction is very important in information retrieval, and it can be improved significantly by providing more accurate results based on query log processing. This paper introduces taxonomy which helps to understand query log processing techniques and categories them. Further, analysis of query log processing techniques given in Table I show that there are limitations in existing techniques which necessitates further research in the area of query log processing. Hence we proposed an approach to identify intent of web queries based on query chains, which was obtained from query log. The experimental results show that our approach is effective to identify the user intention behind submitting a query to search engine. Our experiment gives overall accuracy of 80%, which demonstrates that our approach is of practical significance.

In future, we plan to exploit correctly identified intent for query rewriting, which we expect to further improve performance of searching on the web.

## REFERENCES


[1] K. Hofmann, M. de Rijke, B. Huurnink, and E. Meij. A semantic perspective on query log analysis. In *Working Notes for the CLEF 2009 Workshop*, 2009.

[2] Je Huang and Efthimis N. Efthimiadis. Analyzing and evaluating query reformulation strategies in web search logs. In *Proceedings of the 18th ACM conference on Information and knowledge management*, CIKM '09, pages 77-86, New York, NY, USA, 2009. ACM.

[3] Amac Herdagdelen, Massimiliano Ciaramita, Daniel Mahler, Maria Holmqvist, Keith Hall, Stefan Riezler, and Enrique Alfonseca. Generalized syntactic and semantic models of query reformulation. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '10, pages 283-290, New York, NY, USA, 2010. ACM.

[4] Carlos Castillo, Claudio Corsi, Debora Donato, Paolo Ferragina, and Aristides Gionis. Query-log mining for detecting spam. In *Proceedings of the 4th international workshop on Adversarial information retrieval on the web*, AIRWeb '08, pages 17-20, New York, NY, USA, 2008. ACM.

[5] David J. Brenes, Daniel Gayo-Avello, and Kilian Perez-Gonzalez. Survey and evaluation of query intent detection methods. In *Proceedings of the 2009 workshop on Web Search Click Data*, WSCD '09, pages 1-7, New York, NY, USA, 2009. ACM.

[6] Andrei Broder. A taxonomy of web search. *SIGIR Forum*, 36(2):3-10, September 2002.

[7] Bernard J. Jansen, Danielle L. Booth, and Amanda Spink. Determining the user intent of web search engine queries. In *Proceedings of the 16th international conference on World Wide Web*, WWW '07, pages 1149-1150, New York, NY, USA, 2007. ACM.

[8] Dayong Wu, Yu Zhang, Shiqi Zhao, and Ting Liu. Identification of web query intent based on query text and web knowledge. In *Proceedings of the 2010 First International Conference on Pervasive Computing, Signal Processing and Applications*, PCSPA '10, pages 128-131, Washington, DC, USA, 2010. IEEE Computer Society.

[9] Ashok Veilumuthu and Parthasarathy Ramachandran. Intent based clustering of search engine query log. In *Proceedings of the fth annual IEEE international conference on Automation science and engineering*, CASE'09, pages 647-652, Piscataway, NJ, USA, 2009. IEEE Press.

[10] Eldar Sadikov, Jayant Madhavan, Lu Wang, and Alon Halevy. Clustering query refinements by user intent. In *Proceedings of the 19th international conference on World Wide Web*, WWW '10, pages 841-850, New York, NY, USA, 2010. ACM.

[11] Shuo-En Tsai, Yi-Shin Chen, Chia-Yu Tsai, and Shih-Wei Tu. Improving query suggestion by utilizing user intent. In *IRI*, pages 25-30. IEEE Systems, Man, and Cybernetics Society, 2010.

[12] Filip Radlinski, Martin Szummer, and Nick Craswell. Inferring query intent from reformulations and clicks. In *Proceedings of the 19th international conference on World Wide Web*, WWW '10, pages 1171-1172, New York, NY, USA, 2010. ACM.

[13] Filip Radlinski and Thorsten Joachims. Query chains: learning to rank from implicit feedback. In *Proceedings of the eleventh ACM SIGKDD in-ternational conference on Knowledge discovery in data mining*, KDD '05, pages 239-248, New York, NY, USA, 2005. ACM.

[14] Paolo Boldi, Francesco Bonchi, Carlos Castillo, Debora Donato, Aristides Gionis, and Sebastiano Vigna. The query-flow graph: model and applications. In *Proceedings of the 17th ACM conference on Information and knowledge management*, CIKM '08, pages 609-618, New York, NY, USA, 2008. ACM.

[15] Amanda Spink and Bernard J. Jansen. People's Query Logs: Personal Information Management. In Einat Amitay, Craig G. Murray, and Jaime Teevan, editors, Query Log Analysis: Social And Technological Challenges. *A workshop at the 16th International World Wide Web Conference (WWW 2007)*, May 2007.

[16] Adam Fourney, Richard Mann, and Michael Terry. Characterizing the usability of interactive applications through query log analysis. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 1817-1826, New York, NY, USA, 2011. ACM.

[17] Patrick Ngok and Zhiguo Gong. Log Mining to Support Web Query Expansions. In: *Proceedings of the 2009 IEEE International Conference on Information and Automation*. June 22 -25, 2009, Zhuhai/Macau, China. Pages 375-379.

[18] Hang Cui, Ji-Rong Wen, Jian-Yun Nie, and Wei-Ying Ma. Query expansion by mining user logs. *IEEE Transactions on Knowledge and Data Engineering*, 15:829-839, 2003.

[19] Daniel E. Rose and Danny Levinson. Understanding user goals in web search. In *Proceedings of the 13th international conference on World Wide Web*, WWW '04, pages 13-19, New York, NY, USA, 2004. ACM.

[20] G. Pass, A. Chowdhury, C. Torgeson, "A Picture of Search" Internet: http://www.gregsadetsky.com/aol-data, *The First International Conference on Scalable Information Systems*, Hong Kong, June, 2006 [ May 17, 2012].