# A Comparitive Study Of Web Search Ranking Based On User Feedback Sessions

Shani.S.Das, Rejimoan.R

Mtech Student, Dept. of CSE, SCTCE, TVM, India; Asst.Prof. Dept. of CSE, SCTCE, TVM, India shani.sdas@gmail.com, rejimoan@gmail.com

**ABSTRACT:** Internet is a place of various information and it contains huge amount of data. As the web content rises, it became difficult to organize and manage the data. These datas must be organised in such a way that the search engine must be able to retrieve it efficiently.Various methods are there which helps to improve the search results in search engines by inferring user search goals.Users all over the world have different views and requirements for searching. Most of the search goals coincide.It is the function of the search engine to satisfy the user search goals.Analyzing user search goals is a best practice to make the search results efficient.These sessions are called feedback sessions which helps to infer user search goals.

Keywords : Web content; User Queries; Query Ranking Methods

#### **1** INTRODUCTION

A Web in a collection of inter-related files which reserves on one or more servers. Web mining helps to extract knowledge from web data.Web data mining is mainly classified as Web content mining(both the extraction of useful data and information), Web Structure mining (Using graph theory to analyse the nodes and connection structure of Websites), Web usage mining(Extracting useful information from server logs) [1].In this paper we are dealing with Web usage mining. Finding and analysing the feedback sessions from server logs will helps to improve the web search results and speeds up the processing of search engines. For every users access to web page the web server records it in weblogs. These web logs contains the information about the requested URL , the IP address from where the request has originated and the duration of the session. It is based on the web log records the feedback sessions are explored. In addition to the feedback sessions the web logs are used to know the click frequency of a particular URL which helps to get how many times a particular URL has been clicked by the user and it also helps to infer the behavior of the user that is what is needed by the user for a particular query.Feedback sessions alone cannot serve to infer the user search goals.Based on the feedback sessions construct a pseudo document on that session for analyzing the accurate result. This pseudo document consists of keywords to identify the URL. This is called as enriched URL[2]. The enriched URL's are clustered together based on the frequency of URL's to form pseudo documents. The user queries are classified as navigational queries, informational queries and transactional queries. The navigational queries are those by which the user has some particular website or web page in mind. For example if a user needs to check the gmail account then the user will enter the corresponding URL into the browsers search bar.Queries which give a huge results corresponding to the users search is called informational queries. In other words the users doesnt have a particular website in mind. Transactional queries are those which are submitted by the user inorder to complete a transation. In general these queries always have words relating to transactions such as buy, purchase, order etc. The rest of the document is organised as follows :Section 2 describes the various methods for infeering user search goals. A comparison of various approaches is included in Section 3. Section 4 concludes the survey

### 2 TECHNIQUES FOR INFERRING USER SEARCH GOALS

# 2.1 Automatic Identification of User Goals in Web Search

Every query submitted by the user has a particular intention. Based on these user goals and its analysis it is very useful to identify what the user expect when he is submitting a query.Manual goal identification method [3] is used to identify the target of the user query prior to this.In manual query classification whenever a user submits a query ,q the result is i(q), the percentage of users who submits this query as informational .By this way a query is classified as either informational or navigational.The proposed system mainly focus on aut matic identification of user search goals.This methos uses two criteria to identify the user search goals.

· Past user click behaviour

The users goals are identified from how the users have interated with the past returned results of this query. If the query is navigational then the user clicks must be less on the retured results and if the query is informational then the user will click on many links in the returned results. By this it is possible to identify whether a query is informational or navigational. Click distributions are used to identify this.

Anchor Link Distribution

There is a correlation between the number the query and the number of anchor destinations from the web link with the same anchor text as the query.That is if the anchor text is pubmed then its correspoding URL is www.ncbi.nih.gov.The procedure is used to identify how many such URLs are associated with the anchor text .By counting the number of times a particular URL is clicked then the highest frequncy of the clicked URL indicate that the particular anchor text is associated with that URL.

#### 2.2 Context - Aware Query Suggestion by Mining Click-Through and Session Data

Query Suggestion [4] plays an important role in the forming the feedback sessions to infer the user search goals. The mined patterns in the existing Query Suggestion are not Context - Aware. Context-Awareness consists of mainly two steps.

Offline Model Learning Step

To adderss the data spareseness, queries are summarized into concepts by clustering click-through bipartite. From session data a concept sequence suffix tree is constructed query suggession model.

Online Query Suggestion Step

A user's search context is captured by mapping the query sequence submitted by a user to a sequence of concepts .By looking up the context in the concept sequence suffix tree, it is possible to suggest queries to the user in a context-aware manner. The advantage of this method is that the context awareness [5] donot take into account the immediately preceeding queries as context in query suggession. In offline process the learning step is used to address the data, queries are then converted into concepts by a technique called clustering and a click through bipartite.

# 2.3 Learning Query Intent from Regularized Click Graphs

This develops a query intent classifier using the click graphs .The main objective is to identify the user goals by increasing the training data using the semi supervised learning .The effectiveness of the algorithm is demonstrated by product intent and job intent classification.In both of these cases the training data is improved by using labelled queries leading to increase in classification performance.The main goal of this method is to develop a content based classifier.A query with product intent is one that refers to any tangible product that can be purchased online from the store.A query is said to be job intent [7] if the user is interested to find the job listings as the result of the query submission. There exits two preprocessing steps for constructing click graphs.

#### Removing navigational queries

When analysing the web logs based on the query classification the user clicks can be either navigational or informational.The number of clicks for the navigational queries may be sometimes less.So it results in a skewed click count.Inorder to avoid this such type of navigational queries are eleminated during the construction of click graphs.

• Clustering URLs Inorder to compensate the sparsity of data the URLs are clustered. If the URLs have similar domain names they are clustered to a single node.

The click graphs are consturcted by applying seed queries. The steps for constructing the click graphs are

- Initialize the query set (ie the seed queries) and the URL set
- Update the URLs based on the query set
- Corresponding to the URLs update the query sets
- · Repeat the steps until the query set reach the desired size

### 2.4 A New Algorithm for Inferring User Search Goals with Feedback Sessions

A session is series of queries submitted by the user to satisfy user needs.A feedback session is a combination of both

clicked and unclicked URLs. The framework consists of two steps[8]

- All the feedback sessions of a query are extracted from user clickthrough logs and mapped to pseudodocuments. These pseudodocuments are clustered and user search goals are infered from these clusters
- The original search results are restructured based on the user search goals inferred .The performance evaluation is done by using classified average precision (CAP)[8] method.

The CAP Evaluation method is an improvement of Average precision method. The average precisions are computed by,

$$AP=((1/N) \sum rel (r) R_{r/r})$$

Where, N - the number of relevant clicks

r - rank in total number of retrieved documents.

rel ()- binary function on the relevance of a given rank.

R<sub>r</sub>. No of relevant retrieved documents of rank r or less

After the average precision calculation , Voted Average Precision (VAP) is calculated .This is the AP classes based on the clicks. If the number of clicks in the two classes are same then AP with higher value is selected as VAP .CAP involves the computation of risk , based on which the web results are restructured.

Risk =  $((\sum d_{ij})C_m^2)$ , where  $d_{ij}$  is 1 then the i<sup>th</sup> and j<sup>th</sup> urls are categorized, otherwise it is zero.

And

 $CAP = VAP x (1 - Risk)^{\mu}, \mu$  is used to reduce the influence of Risk.

The feedback sessions can be mapped either by binary vector representation or by using goals texts as keywords. Among these goal texts reflect the user needs. The feadback sessions inferred is converted to a TF-IDF vector of titles and snippets. The goals are inferred based on clustering. K - means clustering method is used for this. Based on the CAP evaluation the web results are restructured

#### 2.5 Query-Sets Using impicit feedback and Query Patterns to Organize Web documents

The feedback form the weblog is known as the implicit user feedback.Based on this implicit feedback a Document representation model (DRM) [9] is made. The base of this DRM is the queries submitted by the user to the search engine. DRM is used to achieve better query results using clustering and labeling. User clicks are given preference and better choice of feature is made .This shows a frequency of query patterns and it is called as query set model.

#### **3 COMPARITIVE STUDY**

Automatic Identification of User Goals in Web Search uses User Click Behaviour and Anchor Link Distribustion [3] for inferring user search goals. The advantage with this method is that it has 90% accuracy but it posses a disadvantage that it is based on potentially - biased dataset. The Context aware Query Suggesion by Mining Click-through and Session Data has both Online and Offiline methods [5]. It uses a cocept sequence to rank the results and study its relevance. The main advantage is that it covers all the user behaviours and maintains the quality of suggessions. As the coverage area increases the complexity of the method also increases. The third method, Learning from Query Intent from Regularized Click Graphs [6] uses a semi supervised [7] Click graphs for ranking. This method improves the classification performance. Eventhough this method improves the performance, it posses an impact of seed queries. A New Algorithm for Inferring User Search Goals with Feedback Sessions [8] has both online and offline learning. The advantage of this method is that it posses a CAP evaluation method which provides a high performance and ranks the results most accurately compared to other methods previously discussed. The final method, Query sets [9] uses non supervised method to infer the user goals. It posses atmost 90% accuracy and the main drawback of this method is that in needs a broader comparison with the online dictionary

Works	Methods	Advantage	Disadvantage
Automatic Identification of User Goals in Web Search	User Click Behaviour and Anchor Link Distri- bustion	90% accuracy	Based on a typical Dataset
Context aware Query Suggesion by Mining Click- through and Session Data	Online and Offiline me- thods Concept sequence suffix tree	Covers all the user beha- viours	Complexity of the method also increases as the coverage area increases.
Learning from Query Intent from Regula- rized Click Graphs	Click Graphs	Improves the classification performance	Presence of seed queries
Algorithm for Inferring User Search Goals with Feed- back Ses- sions	Online and Offiline me- thods , CAP Evalua- tion	High pefor- mance due to minimizing the influence of Risk.	Complex Com- putation, which invloves the calculation of AP ,VAP , Risk factor and CAP
Query sets Using impicit feedback and Query Pat- terns to Or- ganize Web documents	Non super- vised	Appoximately 90% accuracy	Needs compari- son with online dictionary

### 4 CONCLUSION

The web usage increases day by day. Every internet user searches the web contents for acquiring knowledge. As the usage increases the content in the web also increases and it is trivial to analyse its content. Each user has some preferences about the query results and the behave exactly for some queries.Hence it is necessary to gain the user search goals so that the queries are ranked exactly as per the users behavior. It is trivial to gain the users feedback sessions and based on that the query results can be ranked properly.

#### References

- Beitzel. S, E. Jensen, A. Chowdhury, and O. Frieder, "Varying Approaches to Topical Web Query Classification," Proc. 30th Ann.Int'I ACM SIGIR Conf. Research and Development (SIGIR '07), pp. 783-784, 2007.
- [2] Baeza-Yates. R, C. Hurtado, and M. Mendoza, "Query Recommendation Using Query Logs in Search Engines," Proc. Int'l Conf. Current Trends in Database Technology (EDBT'04), pp. 588-596, 2004.
- [3] Lee. U, 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.
- [4] Joachims. T, 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
- [5] Cao. H, 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.
- [6] Joachims. T, "Evaluating Retrieval Performance Using Clickthrough Data," Text Mining, J. Franke, G. Nakhaeizadeh, and I. Renz, eds., pp. 79-96, Physica/Springer Verlag, 2003.
- [7] Li. X, Y.-Y Wang, and A. Acero, "Learning Query Intent from Regularized Click Graphs," Proc. 31st Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '08), pp. 339-346, 2008.
- [8] Zheng Lu, Hongyuan Zha, Xiaokang Yang, , Weiyao Lin, Member, and Zhaohui Zheng A New Algorithm for Inferring User Search Goals with Feedback Sessions, IEEE Transaction on Knowledge and Data Engineering, vol no 25, March 2013
- [9] Poblete. B and B.-Y Ricardo, "Query-Sets: Using Implicit Feedback and Query Patterns to Organize Web Documents," Proc. 17th Int'I Conf. World Wide Web (WWW '08), pp. 41-50, 2008.
- [10] Zeng. H, Q.-C He, Z. Chen, W.-Y Ma, and J. Ma, "Learning toCluster Web Search Results," Proc. 27th Ann. Int'l ACM SIGIRConf. Research and Development in Information Retrieval (SIGIR '04), pp. 210-217, 2004.
- [11] Jones. R 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.
- [12] Beeferman. D 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.