Re-ranking the Results Based on user profile.

Anuradha R. Kale, Prof. V.T. Gaikwad, Prof. H.N. Datir

Email: anuradhakale20@yahoo.com

ABSTRACT

Search engines have become indispensable gateways to the huge amount of information on the Web. Because users typically look only at the first few pages of search results, ranking can introduce a significant bias to their view of the Web and their information gain. This work addresses two common problems in search, frequently occurring with underspecified user queries: the top-ranked results for such queries may not contain relevant documents to the user's search intent, and fresh pages may not get high rank scores for an underspecified query due to their freshness and to the large number of pages that match the query, despite the fact that a large number of users have searched for parts of their content recently. Such problems can be solved if users' search goals are well understood. Here re-ranking method is proposed which employs semantic similarity to improve the quality of search results. By fetch the top N results returned by search engine, and use semantic similarities between the candidate and the query to re-rank the results. First convert the ranking position to an importance score for each candidate. Then combine the semantic similarity score with this initial importance score and finally get the new ranks. This re-ranking method work on User profile Data (USD). After getting this new rank, we re-rank the data according to the relevance of USD.

Keywords : Ranking, Web search, search engine, Imporntace.

1 INTRODUCTION

The sheer amount of Web pages and the exponential growth of the Web suggest that users are becoming more and more dependent on the search engines' ranking schemes to discover information relevant to their needs . Typically, users expect to find such information in the top-ranked results, and more often than not they only look at the document snippets in the first few result pages. Highly ranked documents have greater visibility, which translates into getting more attention and eventually leads to popularity . Thus, the ranking systems have introduced a critical bias to the users' perception of information.

As the information on the Internet explodes astonishingly, search engines play a more and more important role. However, due to the diversity of web users' search requests, it is urgent for search engines to improve their keyword-based search techniques. Currently search techniques are mainly based on keyword matching. However, this technique has the following weak-nesses. First, web users cannot express their search intention accurately using several keywords half the time. Hence the exactly-matched results do not consequentially satisfy the web users. Secondly, keyword matching cannot guarantee the selected candidates have high correlation with the user query, given the different positions and meanings of the keywords. Third, under the circumstances of keyword matching, the top- ranked search results for a given query must contain the keywords as much as possible, otherwise they will lose their ranking positions although their contents exactly discuss the same thing. This will also lead to an awkward situation: spammers try their best to pollute the web document corpus with term spamming tricks such as repetition, dumping and weaving.

Another problem about current search engines is their ranking schemes. PageRank is the most popular ranking algorithm, however, it is based on the popularity of web documents, not the quality. Therefore, a newborn web document usually cannot get highly-ranked positions due to their freshness and thus little reputation. How to promote the new documents and maximize quality of search results seen by users is becoming a more and more challenging work.

In this paper, semantic analysis method to remedy the shortcomings of the current search techniques is used. The search based on lexical semantics instead of keyword matching can better adapt to the thinking pattern of human beings, and thus search results are more relevant to users' search intention. Meanwhile, using semantic factors can conciliate the freshness and make the high-relevant new pages get moderate rank promotion.

In this work, fetch the top N results returned by search engines such as Google for user queries, and use semantic similarities between the candidate and the query to re-rank the results. We first convert the ranking position to an importance score for each candidate. The semantic similarity score combine with this initial importance score and finally we get the new ranks.

2 RELATED WORK

There is a large body of work that investigates methods to rank webpages globally or dependent on a target query. PageRank and HITS (Hypertext Induced Topic Search) are two well-established ranking metrics that use the link structure of the Web. Both of them build upon the assumption that the quality of a webpage can be inferred by the number and the quality of pages linking to it. PageRank computes a global ranking score for all the pages on the Web independent of user queries and does not take into account the particular topics in which the search engine users are interested. HITS, on the other hand, works on a query-specific subgraph of the Web, so that the ranking scores are biased by the issued query. But HITS normally requires more query-time processing and is more susceptible to localized link spam. A number of algorithms have been developed to solve the above problems. Hilltop works mostly on popular topics and depends on a set of expert pages that are identified to be authoritative in the query domain to rank the other pages. Topic-sensitive PageRank pre-computes a vector for a specific topic, and then uses these topic-sensitive vectors at query-time to bias the final ranking towards the particular topic(s) denoted by the issued query. In recent years, there are also developments that try to learn the ranking preferences of the users.

Query expansion has been shown to be an effective method to bridge the gap between users' internal information needs and corresponding external expressions, and several methods for query expansion have been proposed. Thesauri-based query expansion generally relies on statistics to find certain correlations between query terms. Pseudo-relevance-feedback query expansion uses the initially retrieved documents as a source for relevance feedback. In, Cui et al. mine click-through records of search results to establish mappings from query terms to strongly correlated document terms, which are then used for query expansion with the goal to match better the lexical characteristics of the web document space. In, Kraft and Zien investigate a method to generate lists of weighted expansions for queries from the anchor texts of the retrieved documents. In , Billerbeck et al. propose and effective method of obtaining query expansion terms from the past queries that retrieved the documents in the collection associated with a target query, reporting 26%-29% relative improvements over unexpanded retrieval on the TREC-9 and TREC-10 collections.

Search results can also be refined in an interactive manner and fulfilled in an iteration of search-and-feedback cycles . Such explicit feedback methods require users to make explicit judgments of the documents' relevance. Thus, one major disadvantage of this approach is that users can be quite reluctant to offer explicit feedbacks due to the extra efforts in doing so. Also, a system based on explicit feedback is easier to spam. In contrast, implicit feedback can be collected unobtrusively by analyzing users' search history , and exploiting click-through patterns. Search results can also be refined by employing collaborative filtering algorithms to take into account similar users' preferences . Recently, there have been developments that exploit query logs for search results refinement. In, past sequence of queries are used to complement the current query in estimating document relevance. In , previous queries are selected based on the similarity between their search results and those retrieved by the target query. These queries are then used to suggest an extended document list. A formal user model is proposed in based on the immediate search context for personalized ranking. Preceding queries in the same search session are used to expand the current query, improving the relevancy of ad hoc retrieval.

The contribution is a novel re-ranking algorithm that uses distributional information about the query context, as extracted from search engine logs, which effectively and efficiently improves the ranking of search results. Although our approach shares certain common grounds with existing studies, there are significant differences between the proposed method and previous approaches. Additionally, we have evaluated the performance of our method on a dataset that contains real-world editorial judgments and is much larger than those used in most existing studies, and carried out a series of comprehensive experiments to select the best parameters.

3. Algorithm

3.1 Re-ranking Method

Semantic analysis method is use to remedy the shortcomings of the current search techniques. The search based on lexica semantics instead of keyword matching can better adapt to the thinking pattern of human beings, and thus search results are more relevant to users' search intention. Meanwhile, using semantic factors can conciliate the freshness and make the high-relevant new pages get moderate rank promotion. In our work, we fetch the top N results returned by search engines such as Google for user queries, and use semantic similarities between the candidate and the query to re-rank the results. We first convert the ranking position to an importance score for each candidate. Then we combine the semantic similarity score with this initial importance score and finally we get the new ranks.

Importance

since our result is heavily depended on the search engine's quality and result, how to grade the web pages returned from the search engine is important.

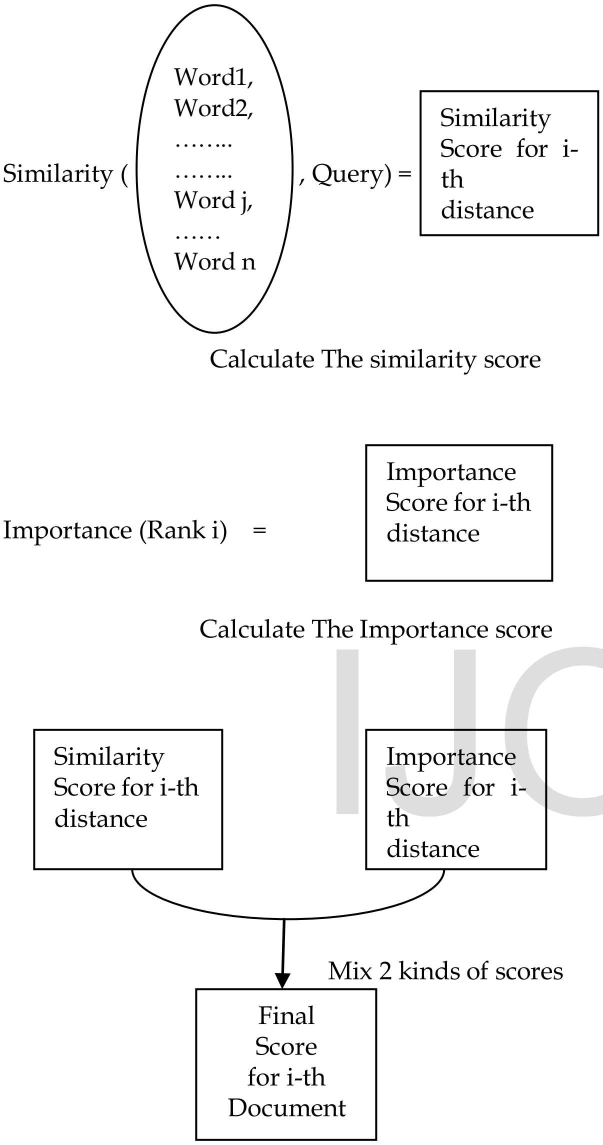
How to measure the importance of the results at different positions. As we know, the search results are returned by search engines according to their importance and relevance. The most important web pages usually are returned at the top positions, and hence attract much more attention from users. On the contrary, the unimportant pages are returned at the bottom positions. Therefore, a discount factor is needed which progressively reduces the document value as its rank decreases.

We propose the following formula to calculate each web's importance score.

$$importance\ (i) = \frac{1 - (i - 1) / tot}{log_2\ (i + 1)}$$

where *i* is the original PageRank serial number (i.e., the original ranking position) and *tot* is the number of the fetched web pages for a query. The formula indicates that the top results have significant importance to the search keywords and thereby are much valuable for web users.

In order to make calculation simple, the result is normalized. For example, if *i* equals to 1, 5, 7, 3, the importance of the first web page is 7, which is the highest score. Along with the increasing of the ranking position number of a returned result, its importance score is also increased sharply.


Similarity (
Word1,
Word2,
……..
……..
Word j,
……
Word n
, Query) =
Similarity Score for i-th distance
Importance (Rank i) =
Importance Score for i-th distance
Similarity Score for i-th distance
Importance Score for i-th distance
Mix 2 kinds of scores
Final Score for i-th Document

Calculate The similarity score

Calculate The Importance score

Figure 1: re-ranking method

RELEVANCE

Relevance is the semantic similarity between keywords and a specified web document. Our algorithm is different from other semantic similarity methods. We calculate the similarities between the keywords and User profile data to get the final relevance.

Firstly, we get all the web page content and use user profile data (USD) match with this web pages result. In this if contain matched with USD then we increased original rank with 1. If title matched with USD then we increased original rank with 5. And if url matched with USD then we increased original rank with 10.

In this way we get the highest rank for the relevant data that user gives the input. We arranged this USD based rank data in decreasing order, so which data contain highest rank which would be display first, and that will be the most relevant data.

THE RE-RANKING ALGORITHM

```
1Calculate the importance (i) for each web page which are
   extracted for result.
2 Arrange this rank of i in descending order
3 Now matched the title with USD, if matched then
        Original rank i + 1;
4 If contain matched then
        Original rank i + 5;
5 If url matched then
        Original rank i + 10;
6 Finally we get result in descending order.
```

3.2 Record Matching for Data Duplication

Web databases compose the deep or hidden Web, which is estimated to contain a much larger amount of high quality, usually structured information and to have a faster growth rate than the static Web. Most Web databases are only accessible via a query interface through which users can submit queries. Once a query is received, the Web server will retrieve the corresponding results from the back-end database and return them to the user.

To build a system that helps users integrate and, more importantly, compare the query results returned from multiple Web databases, a crucial task is to match the different sources' records that refer to the same real-world entity. When multiple web databases are consider, the most common problem is happen that they may contain duplicate records. When data are extract from multiple databases for same query then it may happen two databases contain same records and both the records get extracted and display in output.

To overcome this problem we implement Exact Semantic Search Method. We are using the Comparative Method, where we are comparing each Title with all the other ones, and then checking if the title is already present or not.

Semantic search

As search engines became popular amongst Netizens, a need of semantic search has become a necessity. As the context and semantics of the information in the web pages indexed depends on multiple factor, semantic search has become a complex task. Google Research Lab has worked out many generic algorithms however they are successful in certain conditions only. A focus on personalization of semantic search was given where in one can restrict the domain set and search parameters depending on the personal information. For example, one can build the semantic search techniques based on the domain of some company.

The search engines like Google and Yahoo are so famous that they are in use now and then for searching various type of information available on web. A web has become a largest available data set in public domain to the extent that now-a-days; we are using a term "Information Explosion" as the data indexed by the search engines is so huge.

This explosion of information has brought some side effects of its own. The search using Google is easy but sorting the expected data out of the search results is very difficult. A keyword based search algorithms used in search engines adds more and more confusion in indentifying requisite data. Hence the scientists fear about the hiding of expected information in the large set of relevant and irrelevant information. This becomes worst when the keywords used for searching are unambiguous for ex "lotus" where lotus can be a name of a flower, name of hotel, or some individual etc. For carrying out the exact search, word sense disambiguates could be used. This process involves the use of other information present in a semantic analysis system.

Step of semantic search Algorithm:

1) Having the calculated importance icount from Re-ranking algorithm.
2) If current_count = icount; then
3) Check if this is present in the current list
4) If found = True; then
5) Go to step 8;
6) If fount = False; then
7) Add("{R}:" + current_count);
8) Exit

From the above algorithm we avoid the data duplication and display the unique results for the particular search.

4. Experimental Results and Discussion

After the extraction and implementation of the Re-ranking method in the following platform. The results are as follows-

Table1. Experimental Results

search item	No. of relevant sides for Google search	No. of relevant sides for Faroo search	No. of relevant sides for Re-rank search
Apple	12	10	150
Laptop	15	10	100
Car	20	10	170
Mobile	25	10	200

Table 1 shows the results, numbers of relevant data records extraction and data item alignment from multiple web database. This table shows the result for different product, using the different search engine and there number of relevant results for that search.

5 Conclusions

Alignment re-ranking method is implemented which employs semantic similarity to improve the quality of search results. Fetch the top N results returned by search engine, and use semantic similarities between the candidate and the query to re-rank the results. First convert the ranking position to an importance score for each candidate. Then combine the semantic similarity score with this initial importance score and finally get the new ranks. This re-ranking method work on User profile Data (USD). After getting this new rank, we re-rank the data according to the relevance of USD.

From above table1 we conclude that Re-ranking algorithm gives more relevant data based on USD.

6 Refernces

[1] Ruofan Wang, Shan Jiang and Yan Zhang: Re-ranking Search Results Using Semantic Similarity .

[2] J. Balinski and C. Danilowicz. Re-ranking method based on inter-document distances. Information Processing and Management, 41(2005), pages 759–775, 2005.

[3] Nambiar, U., and Kambhampati, S. Providing Ranked Relevant Results for Web Database Queries. In Proceedings of the World Wide Web Conference, pp. 314-315. 2004.

[4] W. Su, J. Wang, and F.H. Lochovsky, "Holistic Schema Matching for Web Query Interfaces," Proc. 10th Int'l. Conf. Extending Database Technology, pp. 77-94, 2006

[5] C. Tao and D.W. Embley, "Automatic Hidden-Web Table Interpretation by Sibling Page Comparison," Proc. 26th Int'l Conf. Conceptual Modeling, pp. 566-581, 2007

[6] R. Baxter, P. Christen, and T. Churches, "A Comparison of Fast Blocking Methods for Record Linkage," Proc. KDD Workshop Data Cleaning, Record Linkage, and Object Consolidation, pp. 25-27, 2003.

[7] Jansen, J., and Spink, A. An Analysis of Web Documents Retrieved and Viewed", In Proceedings of the 4th International Conference on Internet Computing, 2003.

[8] Olsen, S. Does Search Engine's Power Threaten Web's Independence? Retrieved at http://news.com.com/2009-1023- 963618.html., Oct. 2002.

[9] Page, L., Brin, S., Motwani, R., Winograd, T. The PageRank Citation Ranking: Brining Order to the Web. Technical Report,Stanford University Database Group,

1998.Retrievat http://dbpubs.stanford.edu:8090/pub/1999-66.

[10] Kleinberg, J. Authoritative Sources in a Hyperlinked Environment. In Proceedings of the 9th ACM SIAM Symposium on Discrete Algorithms, pp. 668-677, 1998.

[11] Bharat, K., and Mihaila, G. When Experts Agree: Using Non- Affiliated Experts to Rank Popular Topics. In Proceedings of the 10th International World Wide Web Conference, pp. 597-602, 2001.

[12] Haveliwala, T. Topic-Sensitive PageRank. In Proceedings of the 11th International World Wide Web Conference, pp. 517-526, 2002.