Effective Information Search and Retrieval for Answering Time Sensitive Queries

V.Bharath Kumar
Department of Computer Science
M.Tech, Artificial intelligence
JNTUCEA, Anantapur
veldurthi.bharathkumar@gmail.com

Abstract

World Wide Web accumulates large volumes of documents every day. Search engines are used to search for required information from WWW. The search is generally carried out based on the similarity of the documents being searched for. The results are presented as per ranking of the items. However, for best ranking mechanism only similarity is not adequate. For some class of queries time could be an important dimension for searching in addition to the content similarity. Such queries are known as time sensitive queries that are processed and ranked based on the publication time and similarity. The existing research focused on retrieving recent queries. Recently Dakka et al. presented a general framework for handling time – sensitive queries. In this paper we propose a framework that extends their work by considering more time related dimensions such as republication date and time, review articles of the documents with later dates etc. This improves the robustness of the system with respect to answering time sensitive queries as it can make use of review articles and summarize events in temporal domain. Thus the system is made capable of analyzing contents of web documents on different dimensions in addition to their publication date. We built a prototype application to demonstrate the proof of concept. The empirical results revealed that the proposed framework for multi-dimensional time sensitive queries is effective.

Index Terms – Time sensitive queries, information retrieval, multi-dimensions

1. INTRODUCTION

Web documents are searched by Internet users through popular search engines like Google, Yahoo, Bing and so on. The search engines focus on the similarity of the search to retrieve and rank the results. However, the time is one of the important dimensions that can be used to leverage quality of results by using it ranking along with similarity. There are two observations from real world examples. The first observation is that time dimension can’t be explicitly modeled by topic-similarity ranking. It does mean that the time dimension cannot be directly incorporated. The second observation is that distribution of relevant documents is not often reflected by the results of topic similarity ranking.

Nevertheless, in [1] it is considered that time dimension can be used to rank query results in addition to similarity. Dakka et al. proposed a general framework that allows using time as dimension for presenting query results by ranking temporally. Temporal values are very useful for some class of queries. Such queries are known as time-sensitive queries. The queries with time component can produce more meaningful results for such class of queries as the ranking makes use of time domain in addition to the similarity. Their framework make use of various steps in order to take temporal queries as input and generate best ranked results that are based on temporal characteristics. However, their work considers only publication time of a document while making ranking.

In this paper we improve the work done in [1] as we also make use of other temporal dimensions such as review papers and their publication time in addition to the original document’s publication date and time. The remainder of this paper is structured as follows. Section II reviews relevant literature. Section III provides details about proposed framework which extends the work done in [1]. The section IV provides experimental results while section V concludes the paper.

2. PRIOR WORKS

Recency queries were explored in [2] by Le and Croft. A language modeling framework is used by the experiments [3], [4], [5], [6]. The prior
probability is assumed by most of the modeling approaches. Most of the queries over Internet through search engines are the general queries in which the search engines consider only the relevancy of the documents based on similarity for producing results. However, ranking is done in order to present best results that can avoid user’s wait time. Time sensitive queries can help improve the quality of the results as it can satisfy some class of queries where time is considered as a dimension for search. In our approach time sensitive queries are supported through a framework which makes use of document’s publication time for improving the rank of the document in the search results. The temporal characteristics of the document are considered for making time sensitive queries. In this paper our techniques are compared with that of Li and Croft’s techniques such as QL-Recency and RM-Recency.

Recently in [7] a temporal aware technique was improved by using blogs for event search. Here a temporal value is used before the blog posting which is presented in the form of a histogram that shows top 500 posts. These posts are related to given query and in the same fashion DAY technique combines time related value prior with relevance to the top query results. Prior probabilities change in the document is exploited by many researchers who impose prior beliefs in the retrieval process using inlink [8], Page Rank and web pages handling for various groups which contain values relevant [9], [10]. Evolution of topics over time was introduced. Such techniques were introduced by Blei and Lafferty [11] for evolving scientific topics. Locating emerging topics was explored in [12]. This study has been done in for detecting novelty [13], [14], and [15]. However, these techniques did not include time information in the query processing.

Various temporal values of given queries are analyzed in [16] and [17]. The study includes distribution of query results over a period of time and the corresponding events and query results which are precise. Queries have been categorized by them into various types such as temporally unambiguous, temporal ambiguous and temporal queries. The answering of time sensitive queries where queries that were manually identified. Manual classification of queries can be done with respect to time sensitivity. It also affects the performance of the algorithms that are used to retrieve information.

3. FRAMEWORK FOR MULTI-DIMENSIONS BASED TIME-SENSITIVE QUERY

The proposed framework is meant for answering multi-dimensions based time sensitive queries. The framework is an extension to the work done by Dakka et al. [1] which considers only publication time of articles as a dimension for query processing along with similarity. However, our framework is based on multiple temporal dimensions such as republication date and time, review articles of the documents with later dates etc. This improves the robustness of the system with respect to answering time sensitive queries as it can make use of review articles and summarize events in temporal domain. Thus the system is made capable of analyzing contents of web documents on different dimensions in addition to their publication date. The overview of our framework is as show in figure 1.

As can be viewed in fig. 1, the proposed framework takes time sensitive query as input and follows a series of steps in order to produce the final results which are ranked based on multiple temporal dimensions. First of all it identifies time intervals in the search results. Then it applies scoring techniques that will integrate the temporal aspects into the ranking mechanism. Then ranking is made. Afterwards, the multiple temporal dimensions are utilized in order to apply re-ranking which will improve the results further.

Finding Temporal Relevance
Documents are typically time stamped but the search engines do not use that part for ranking query results. In our framework we use that temporal information along with other temporal dimensions as mentioned above. The temporal relevant \( p(t|q) \) is computed as follows.

\[
p(t|q) = \frac{p(q|t) \cdot p(t)}{p(q)} \cdot \sum_{d \in \text{dates}(D)} p(q|t) \cdot p(t)
\]

Where dates(D) represents time span of D.
Algorithm used to find temporal relevancy is presented below. We used this algorithm with required changes to accommodate multiple temporal dimensions. The algorithm is presented in listing 1. The algorithm estimates \( p(q|t) \) values using time based approach for each \( t \) and query \( q \). It takes query \( q \) and set of documents \( D \) as input and generates time based probability for each \( t \).

**Step 1:** Compute query frequency histogram for \( q \) and \( D \) using publication time and other dimensions

**Step 2:** Use histogram characteristics to partition the times into bins

**Step 3:** Compute the value for \( p(q|t) \) for each time \( t \) using \( t \)'s bin

**Listing 1 – Algorithm for computing time based probability**

Query Frequency Histograms

There are many binning techniques used to for identifying important time intervals. They are known as Daily Frequency (DF), Fixed Interval Frequency (FIXED), Moving Window (WIN), Running Mean (MEAN), Bump Shapes (BUMP).

4. EXPERIMENTAL RESULTS

We have done experiments to demonstrate the efficiency of our framework and the results are compared with prior techniques. Experiments are made using BMP-QL with TQ 361 dataset, BUMP-QL with TQ 401 dataset, BUMP-RM with TQ361 dataset and BUMP-RM with TQ401 dataset. We made the experiments with multiple time dimensions and compared our results with the results presented in [33].

As can be seen in fig. 2 the experiential results with various kinds of queries and datasets are presented. The horizontal and vertical Axes represent recall and precision respectively. The results reveal that time sensitive algorithms have better performance in terms of accuracy. These experiments are made with time as the dimension.

As can be seen in fig. 3 the experimental results with various kinds of queries and datasets are presented. The horizontal and vertical Axes represent recall and precision respectively. The results reveal that time sensitive algorithms have better performance in terms of accuracy. These experiments are made with multiple time dimensions. When compared with the results presented in 2, our results show better performance.

As can be seen in fig. 4 the experimental results with various kinds of queries and datasets are presented. The horizontal and vertical Axes represent recall and precision respectively. The results reveal that time sensitive algorithms have better performance in terms of accuracy. These experiments are made with time as the dimension.
As can be seen in fig. 4 the experimental results with various kinds of queries and datasets are presented. The horizontal and vertical Axes represent recall and precision respectively. The results reveal that time sensitive algorithms have better performance in terms of accuracy. These experiments are made with multiple time dimensions. When compared with the results presented in 4, our results show better performance.

5. Conclusion
In this paper we propose a framework for processing multi-dimensional time sensitive queries. The framework considers not only publication time but other dimensions like republication date and time, review articles of already published documents with later dates and so on. The multiple time related dimensions ensure that the time sensitive queries are processed with much more utility in organizing and ranking the results. Our framework is an extension to the work of Dakka et al. [1] where they focused only on publication time as the dimension for answering time sensitive queries. Our framework improves the robustness of the system with respect to answering time sensitive queries as it can make use of review articles and summarize events in temporal domain. Our prototype application demonstrates the effectiveness of the framework.

6. References