A Survey on Temporal Information Retrieval Systems

Litty K Mathews
Post Graduate Student
Karunya University

S.Deepa Kanmani
Assistant Professor
Karunya University

ABSTRACT

Temporal Information Retrieval is an emerging research area in the field of Information Retrieval. Due to the immense amount of data in the WWW, and because the contents of documents are strongly time-dependent, it is very tough for the user to retrieve the relevant documents. Traditional Information Retrieval approaches based on topic similarity alone is not sufficient for the search in temporal document collections. The time dimension available in the documents should be incorporated with document ranking for efficient retrieval. This survey gives an introduction to Temporal Information Retrieval and explores the different time-aware retrieval methods and temporal ranking methods for specific types of time-sensitive queries.

Keywords

Temporal information retrieval, temporal ranking, Recency sensitive queries, Time-aware retrieval model. Year qualified queries.

1. INTRODUCTION

The World Wide Web is a vast repository of data. This, however, would be wasted if necessary information could not be found, analyzed, and exploited. The WWW is expanding day by day; the people are using search engines for different purpose The goal of any Information Retrieval (IR) system is to identify the documents that are relevant to the query. The problem in searching over documents is that documents are time-dependent and accumulated over time which results in a large number of irrelevant documents in a set of retrieved documents. Therefore, the users have to spend more time in finding the documents that are satisfying his/her information need. The volume of information generated in this digital world is increasing day by day, the notion of using time as an important factor becomes more important for a large number of searches.

News items mainly include recent information or new events. It is assumed that if news is old then it is not relevant for search. Consider a historian he is interested in knowing about the tsunami that occurs in past years. He searches in the news archives expecting to retrieve the details of the event- not necessarily the latest news, but a report on the latest news about that query is retrieved. Most of the relevant documents for that query would be obtained for the period 2004-2005 or associated with the time that event happened. The most of the relevant documents for that query is for the time period of 2004-2005 or associated with the time that event happened. The timeliness is one of the key aspects that determine a document's credibility besides relevance, accuracy, objectivity and coverage. Both temporal relevance and topic similarity are needed for efficient retrieval.

2. TEMPORAL INFORMATION RETRIEVAL

Temporal Information Retrieval means to satisfy temporal needs and enable the retrieval of temporal-relevant documents. The temporal information provided in the documents can be exploited by the content analysis, query analysis, retrieval and ranking models. Temporal Information Retrieval applications are document exploration, information filtering, similarity search, question querying, temporal summaries, temporal question answering, timelines and user interfaces, clustering and spatio-temporal information extraction. The retrieval effectiveness can be increased by incorporating the time dimension into the search. For example using the creation or publication time of documents if it is available or mining the query logs .

The most important research in Temporal Information Retrieval (T-IR) and its related sub-areas are Temporal Dynamics (T-Dynamics), Fresh Information Retrieval (Fresh-IR), Temporal Markup Languages (T-M Languages), Temporal Taggers (T-Taggers), Temporal Indexing (T-Index), Time-Aware Retrieval Models (T-R Model), Temporal Ranking (T-Rank), Temporal Clustering (T-Cluster), Timeline Interfaces (Timelines), Temporal Search Engines (T-S Engine), Collective Memory (C-Memory), Web Archives (W-Archives), Topic Detection and Tracking (TDT),Temporal Question Answering (T-Q Answering), Temporal Snippets (T-Snippets), Future Retrieval (F-Retrieval).

The retrieval model should rank documents by their relevance with respect to the time dimension. Much research is going on the field of temporal information retrieval to improve the retrieval results. This paper addresses the searching over temporal document collections, where documents are published and/ or edited over time, and the contents of documents are strongly time-dependent. There has been some productive research on using time for different search applications but only little work has been done on exploiting temporal information associated with documents for search in news archives or blogs.

In this paper, is a survey of scope of two research areas in the field of temporal information retrieval, Time-Aware Retrieval Models (T-R Model) and Temporal Ranking (T-Rank). Firstly, we present the work done in time-aware Information Retrieval. There are models that take into account the document timestamp, and other models which consider the temporal relevance of the document's content. Here we consider only the models that take into account document timestamp, disregarding the document's content.

2.1 Time-Aware Retrieval Models (T-R Model)

When searching a temporal document collection like news archives or blogs, the time dimension incorporated in the retrieval model improves the retrieval process. Time-aware Ranking retrieval models the documents based on the keyword score and temporal score of the query. The time aware ranking methods perform better than the topic-similarity ranking. Eg., TF-IDF and language modeling. The time dimension that is used in the time-aware retrieval models are the publication time or creation date and the temporal expressions mentioned in the documents. The time-aware ranking methods are based on the following two main approaches 1) ranking documents by a linear combination of the textual and temporal similarity 2) a probabilistic model generating

document a query from the topic and temporal part of a document independently.

A temporal query q consists of keywords q_{text} and temporal expressions q_{time}. A document d consists of the textual part d_{text} ,and the temporal part d_{time} composed of the publication date PubTime(d).Both the publication date and temporal expressions will be represented using the time model [2]. Here the query likelihood approach is used for ranking documents according to the estimated probability of the query. The textual and temporal part of the query q are generated independently from the corresponding parts of the document d as:

$$P(q \mid d) = P(q_{text} \mid d_{text}) \times P(q_{time} \mid d_{time}) \quad (1)$$

The textual similarity part $P(q_{text} \mid d_{text})$ can be determined by an existing text-based query likelihood approach, e.g., the original Ponte and Croft model [11]. To generate query temporal expression in q_{time} from d, a document temporal expression t_d is drawn at uniform random from document temporal expressions d_{time}. Second, a query temporal expression tq in q_{time} is generated from a temporal expression t_d in d.

$$P(q_{time} \mid d_{time}) = \prod_{t_{q \in q time}} P(t_q \mid d_{time}) \quad (2)$$

$$= \prod_{t_q \in q_{time}} (\frac{1}{|d_{time}|} \sum_{t_d \in q_{time}} P(t_q \mid t_d)$$

The probability of generating tq from t_d or P(tq|t_d) can be calculated using LMT and LMTU [2] method. This is an approach that integrates the temporal expressions into a language model retrieval framework.

Nattiya Kanhabua et al [5] studied implicit temporal queries where no temporal criteria are provided, and how to increase retrieval effectiveness for such queries. Three approaches have been proposed to determine the time of queries when no temporal criteria are provided.The first method performs dating queries using keywords only. The second method performs by dating queries with a technique inspired by Pseudo-Relevance Feedback (PRF) that uses the top-k retrieved documents in dating queries. The third method also uses the top-k retrieved documents by Pseudo relevance Feedback (PRF) and assumes their creation dates as the time of queries. Based on these approaches the documents can be re-ranked using the determined time of queries. Instead of using a language modeling approach as in [2], a mixture model approach of keyword score and time score has been proposed. The mixture model based approach is given as:

$$S(q,d) = (1-\alpha).S'(q_{text}, d_{text}) + \alpha.S''(q_{time}, d_{time}) \quad (3)$$

where α is a parameter indicates the importance of both similarity scores textual similarity $S'(q_{text}, d_{text})$ and temporal similarity $S''(q_{time}, d_{time})$. The textual similarity can be implemented using any of existing text-based weighting models. The d_{time} is referred to Publication Time(d). The probability of generating q_{time} from d_{time}, or $S''(q_{time}, d_{time})$ can be computed as:

$$S''(q_{time}, d_{time}) = P(q_{time} \mid d_{time}) \quad (4)$$

$$= \frac{1}{|q_{time}|} \sum_{t_q \in q_{time}} P(t_q \mid d_{time})$$

where q_{time} is a set of query temporal expressions. The approach shows improvement on retrieval effectiveness,but the quality of the actual query dating processing is a limitation when aiming at further increase in effectiveness.

The time based language model [1] framework incorporates time into both query likelihood language models and relevance-based language model. Time-based language models are based on the publication time of the document using exponential decay. The approach is focused on recency queries by computing topic-similarity scores for each document and then boosts the score of the most recent documents.

$$p(d \mid q) \propto p(q \mid d)\, p(d \mid T_d) \quad (5)$$

where d is the document ,q is the query and probability dependent on documents date T is p(d|T_d).To estimate the probability p(d|T_d) is given in equation.

$$p(d \mid T_d) = P(T_d) = \lambda e^{-\lambda(T_c - T_d)} \quad (6)$$

Here T_c is the most recent date in the document collection and T_d is the creation date of a document.The time uncertainty is captured by the exponential decay function, such that the more recent documents obtain the higher probabilities of relevance. The main contribution of this work is that time can be incorporated into the underlying retrieval model. But this approach does not handle other types of time-sensitive queries that implicitly target one or more past time period.

Diaz and Jones [4] have measured the distribution of creation dates of retrieved documents to create the temporal profile of a query. The the time relevant to a particular query is estimated by analyzing the distribution of creation dates of the documents. Based on the analysis of document collections the four features for classifying queries are temporal Kullback Leibler divergence, autocorrelation, statistics of the rank order and burst model. There are three temporal classes of queries. First one is atemporal queries (which have no periodicities), taking place at any time. The second one is temporally unambiguous queries (which contain a single spike), taking place at a specific period in time. The third one is temporally ambiguous queries (which contain more than one spike) taking place during one of several possible episodes. Temporal profile relies heavily on the underlying retrieval model to estimate probability of the query. The approach allows users to explicitly select appropriate time intervals that demand less input from users.The approach works well on collections where documents are uniformly distributed over time.

Del Corso et al [6] address the problem of ranking news articles, taking into account publication times but also their interlinkage. A ranking framework which models the process of generation of a stream of news articles, the news articles clustered by topics, and the evolution of news stories over the time. Instead of clustering technique adopted a continuous measure of the lexical similarity between news postings. The feature of ranking algorithm is the possibility to analyze the behavior of the mean value of the ranks of all the sources, over the time and for each given category. The naive time-aware algorithm shows a bad behaviour in many cases, and then they refine them in order to have a complete control of the ranking process.

Baeza Yates [10] also considered the temporal information contained in the documents for the retrieval purpose. The extracted temporal expressions from news and index the news articles

together with temporal expressions. Then retrieved the temporal information by using a probabilistic model. The document score is calculated by multiplying a keyword score and a time confidence, i.e., a probability that the document's events will actually happen.

2.2 Temporal Ranking (T-RANK)

Relevance ranking plays a very important role in the field of information retrieval. A lot of ranking algorithms have been proposed so far, based on link analysis,online ranking model,relevance feedback model.In recent years,there is also research in time based ranking,to add temporal factor in the search. The fundamental problem with the current approaches are focused only on improving the general ranking algorithms. Many methods have been developed so far but those for improving the ranking of a particular type of temporal queries are very less . Here we focused only on the temporal ranking of a specific type of temporal queries.

Ruiqiang Zhang et al [6] introduced a new method to rank a special category of time-sensitive queries that are year qualified. The method adjusts the retrieval scores of a base ranking function according to the timestamps of web documents so that the freshest documents are ranked higher. The method, which is based on feedback control theory, uses ranking errors to adjust the search engine behaviour. The method focused only on Year Qualified Queries (YQQs) by translating the user's implicit intention as the most recent year. The method is very effective for recurring event query, ranking the search results based on the adjusting the base ranking function. Query classification and score detector is needed.

Anlei Dong et al [7] they propose a method to use the micro-blogging data stream to detect fresh URLs. They also use micro-blogging data to compute novel and effective features for ranking fresh URLs. Recency Sensitive Queries refer to queries where the user expects documents which are both topically relevant as well as fresh. For example, consider the occurrence of some natural disaster such as an earthquake or tsunami . A user interested in this topic probably wants to find documents which are both relevant and timely. Data gathered from twitter is useful to address recency sensitive queries. The approach is based on preserving the quality of data presented to the general web searcher by using only micro-blog data as evidence for discovering and ranking URL. Filtering of URLs from twitter posts is needed and incorporating those URLs into the larger web ranking system is also an overhead.

Metzler [9] proposed an efficient algorithm for mining /implicitly year qualified queries.An implicitly year qualified query is a query that does not contain a year , but the user may have implicitly formulated the query with a specific year in mind. Mining the query logs and analyze query frequencies over the time in order to identify the strongly time-related queries. The algorithm relies only on access to a query log with frequency information. The approach does not rely on user, clickstream, or session data.

Chang et al [8], presents a query classifier for recency and a ranking model for recent results. The query classifier builds two models representing the content of the document and the query data at time t, respectively. The two models are then compared on different instants of time and a query is considered recent if it increases its probability of being generated in two different instants. The ranking model aims at learning a ranking function based on four categories of recency-related features: timestamp features, link time features, web buzz features and page classification features. Limitations are focused only on breaking news queries. The query is first classified according to its temporal profile, and then is sent to the appropriate ranker that has been optimized for either relevance or freshness. The main disadvantage of classification-based techniques is that selecting a wrong ranker due to misclassification can significantly degrade the performance

3. COMPARISONS OF DIFFERENT METHODS

Each and every method has its own advantages and disadvantages. Most of the methods incorporated the temporal relevance in the retrieval process. But in many methods user has to provide the relevant time period for the query. The good temporal information retrieval system should have the ability to automatically identify the relevant time periods and rank them based on time of the documents.

Table I shows the characteristics of different Time-Aware Retrieval models (T-R model) discussed above. Time is an important dimension can be explored from the documents. Each document contains temporal information in the both explicit and implicit forms.Explicit temporal data can be represented as publication date or updation date and the implicit data are embedded in the content of the document itself. Many temporal information needs have a temporal dimension passed by a temporal expression in the user's query.Consider the query "FIFA World Cup in 1998", the temporal expression is "1998's".That means the query has an explicit temporal intent. Most of the papers focus on implicit temporal queries,i.e, temporal queries that contain only keywords and the relevant documents are associated to particular time periods that are not mentioned in the query.The retrieval effectiveness can be improved by identifying the relevant time periods and re-ranking the search results based on the time periods.

The table 2 shows the summary of various temporal ranking methods specific types of time-sensitive queries. The main criteria used for the comparison are the temporal information and handling which type of queries.

Table 1. Summary of Various Time-Aware Retrieval Models

AUTHOR	TEMPORAL FACTOR	TEMPORAL INTENT	MERITS	DEMERITS
Berberich et al [2]	Temporal Expressions and Publication date	Explicit	Integrated temporal expressions into query-likelihood language modeling, which considers uncertainty.	The temporal criteria are explicitly provided in a query by the user. Inconvenience
Del Corso et al [4]	Publication Date	Implicit	Time aware algorithm for ranking news articles	More complicated algorithm
Diaz and Jones [3]	Publication Date	Explicit	Efficient Query categorization. Temporal information is successfully integrated	The user explicitly selects time intervals, so demands less input from the user.

			for several retrieval tasks.	
Li and Croft [1]	Publication Date	Implicit	Time incorporated in both query-likelihood and relevance models	Focused on recency queries.
Nattiya Kanhabua [5]	Publication or Updated Date	Implicit	Efficient technique for determining the time of queries and re-ranking the search results using the time of queries.	A further improvement on the query dating is needed.

Table 2. Summary of Various Temporal Ranking Methods

AUTHOR	**METHOD**	**TEMPORAL INFORMATION**	**MERITS**	**DEMERITS**
Anlei Dong et al [7]	Method to use the micro-blogging data stream to detect fresh URLs.	Micro blog data	Micro blogging data can be used to improve web ranking for recency sensitive queries. Data gathered from twitter useful to address recency sensitive queries.	Addressing only on recency sensitive querie Filtering of URLs from twitter posts is needed
Chang et al [8]	A retrieval system which contains a query classification algorithm that can automatically detect recency queries	Time stamp/Link time /web buzz/ page features.	Recency ranking improved under learning-to rank framework.	Further improvement needed.
Metzler [9]	Automatically mine implicitly year qualified queries	Mining query logs	Handled implicitly year qualified queries based on query logs. The approach does not rely on click,user or session data.	Only focused on implicitly YQQs
Ruiqiang et al [6]	Effective method to extract year qualified queries by mining query logs.	Mining Query logs and timestamps	Most effective for recurrent event query. No need of query expansion	Can be applied to any category of queries but proper query classification is needed.

4. CONCLUSION

The purpose of this survey is to provide an overview of temporal information retrieval systems.This paper covers the introduction of temporal information retrieval and some time-aware retrieval models and the scope of temporal ranking in specific types of time-sensitive queries. The methods that are described above are effective in retrieving the relevant document from the temporal document collections. Time-aware ranking methods show better performance compared to methods based on keyword score only. In most of the time-aware models, a mixed approach of keyword score and a time score is used. Recognizing temporal information embedded in documents in the form of temporal expressions and exploiting it for information retrieval can significantly improve the current functionality of search applications.

5. REFERENCES

[1] X. Li and W.B.Croft,"Time-Based Language Models,"Proceedings.12th ACM Conference Information and Knowledge Management (CIKM '03), 2003.

[2] K.Berberich, S. J.Bedathur, O. Alonso, and G.Weikum. "A language modeling approach for temporal information needs", In Proceedings of the 32nd European Conference on IR Research on Advances in Information Retrieval, ECIR '10, page 13-25,2010.

[3] R.Jones and F.Diaz,"Temporal Profiles of Queries", ACM Transaction on InformationSystems, vol. 25, Issue 3, article 14, 2007.

[4] Gianna M. Del Corso, Antonio Gulli, and Francesco Romani, "Ranking a stream of news". In WWW Proceedings of the 14th International Conference on World Wide Web pages 97– 106, 2005.

[5] N. Kanhabua and K. Norvaag," Determining time of of queries for re-ranking search results", In Proceedings of the 14th European conference on Research and advanced technology for digital libraries, ECDL'10, pages 261–272, 2010.

[6] Ruiqiang Zhang,Yi Chang,Zhaohui Zheng,Donald Metzler and Jian-yun Nie, "Search Result Re-ranking by Feedback Control Adjustment for Time-sensitive Query", In Proceeding NAACL-Short '09 Proceedings of Human Language Technologies, Pages 165-168,2009.

[7] Anlei Dong,Ruiqiang Zhang,Pranam Kolari,Jing Bai,Fernando Diaz,Yi Chang,Zhaohui Zheng and Hongyuan Zha,"Time is of the essence:

improving recency ranking using Twitter data", In Proceeding WWW '10 Proceedings of the 19th international conference on World wide web, Pages 331-340,2010.

[8] A. Dong, Y. Chang, Z. Zheng, G. Mishne, J. Bai, R. Zhang,K. Buchner, C. Liao, and F.Diaz." Towards recency ranking in web search". In Proceedings of the third ACM international conference on Web search and data mining,Pages 11–20, 2010.

[9] D. Metzler, R. Jones, F. Peng, and R. Zhang. Improving search relevance for implicitly temporal queries. In Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, SIGIR '09,pages 700–701, 2009.

[10] R.A.Baeza-Yates. Searching the future. In Proceedings of SIGIR workshop on mathematical/formal methods in information retrieval MF/IR, SIGIR '05, 2005.

[11] J.M Ponte and W. B. Croft. A language modeling approach to information retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR 98 ,pages 275-281,1998.