ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2015

Survey: Temporal Information Retrieval

Avanti Patange, Soudamini Pawar

M. E. Student, Dept. of Computer, D. Y. Patil College of Engineering, Pune, India

Assist. Professor, Dept. of Computer Engineering, D. Y. Patil College of Engineering, Pune, India

ABSTRACT: An rising research area in the field of Information Retrieval is Temporal Information Retrieval. Because of large amount of data in the internet, and because the contents of documents are strongly time dependent, it is hard for the user to retrieve the relevant documents. Conventional Information Retrieval approach based on topic similarity alone is not enough for the search in secular document collections. The time dimension available in the documents should be integrated with document ranking for effective retrieval. This survey gives an introduction to Temporal Information Retrieval and explores the state of the art of temporal information retrieval and also challenges and application of it.

KEYWORDS: Temporal IR, Temporal Similarities, Temporal Clustering, Temporal Summarization, Temporal Queries

I. INTRODUCTION

Information retrieval is defined as "process of providing most relevant documents to the users from an existing collection". Users request for data in the form of query typically in short textual form. In recent years, time has been acquiring increasing importance within search contexts, constructing to a new research area known as temporal information retrieval (TIR) that contains a number of different challenges. In recent years many researchers has taken interest in temporal information retrieval. Its aim is to improve the effectiveness of information retrieval methods by exploiting temporal information in documents and queries. T-IR aims to fulfil search needs by merging the traditional belief of document relevance with temporal relevance. For example, users may request for documents that contains the past information (e.g., information about historical figures); documents having the most new, up-to-date information (e.g., information about weather forecasts or currency rates); or even future-related information (e.g., information about planned events to be held in a certain area).

Temporal information retrieval refers to IR tasks that analyse and exploit the time dimension embedded in documents to render alternative search features and user experience. Document search, similarity search, summarization, and clustering are examples of applications of temporal IR. Since the web is dynamic in nature, maintaining up-to-date indexes is very tedious and difficult job. It is not easy to retrieve web documents so that their temporal dimension will meet the user temporal purpose underlying the query. Basically, two types of temporal information particularly useful for temporal IR: 1) the publication or creation time of a document, and 2) temporal expressions mentioned in a document or a query.

II. STATE OF THE ART IN TEMPORAL INFORMATION RETRIEVAL

In this section, there is brief overview of related work in temporal IR: determining time for non-time stamped documents, time-aware ranking, temporal indexing, and visualization using a timeline, and searching with the awareness of terminology changes.

A. *Determining Time for Non-time stamped*

Documents determining the time of a document can be done using two methods: learning and non-learning based methods. The difference between these two methods is that the former determines time of a document by learning from a set of training documents, while the latter does not require a corpus for training. Learning-based methods use a statistical method (hypothesis testing) on a group of terms having an overlapped time period in order to determine if they are statistically related. If the resulting values from testing are above a threshold, those features are combined into a single topic, and the time of the topic is computed from a common time period associated to each term.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2015

In Non-learning methods in order to determine time of a document, temporal expressions in the document are annotated and resolved into concrete dates. A relevancy of each date is computed using the frequency of which the date appears in the document. The most pertinent date is used as a reference date for the document, however, if all dates are similar pertinent then the publication date will be used. In the end, the event time period of the document is generated by assembling all nearly dates to the reference date where their relevancy must be greater than a threshold.

Comparison result of learning and non learning based methods gives two different aspects of time. The first method gives a summary of time of events appeared in the document content, or time of topic of contents. The second method gives the most likely originated time of the document, or time of document creation.

B. *Time-aware Ranking*

Time-aware ranking techniques can be classified into two categories: techniques based on 1) link-based analysis and 2) content-based analysis. Approaches of the first category exploit the link structures of documents in a ranking process, whereas the latter approach leverages the contents of documents instead of links. Traditional link-based algorithms (i.e., PageRank and HITS) simply ignore the temporal dimension in ranking. Thus, PageRank algorithm is modified by taking into account the date of a citation in order to improve the quality of publication searches. A publication obtains a ranking score by accumulating the weights of its citations, where each citation receives a weight exponentially decreased by its age. PageRank is also extended to rank documents with respect to freshness. The difference is that freshness defines as a linear function that will give a maximum score when the date of document or link occur within the user specified period and decrease a score linearly if it occurs outside the interval.

C. *Temporal Indexing*

It is an approach to manage documents and index structures in temporal document databases. Using a web warehouse containing historical web pages as a testing environment, different indexing methods proposed improve the performance of temporal text-containment queries. In [23], presented a method for text search over temporally versioned documents. They proposed the temporal coalescing technique for reducing the index size, and proposed the sub list materialization technique to improve index performance concerning space and time. Documents are retrieved according to a query and user's specified time, and are ranked based on tf-idf.

D. *Visualization using a Timeline*

Visualization of search results using temporal information to place retrieved documents on a timeline is useful for document browsing. When a user enters only keywords as a query, retrieved results are too broad without giving temporal context. To narrow down a set of documents retrieved, it is necessary to give an overview of possible time periods relevant to the query and suggest that as a hint to the user.

E. *Searching with the Awareness of Terminology Changes*

Search results can be affected by the terminology changes over time, for instance, changes of words related to their definitions, semantics, and names (people, location, etc.). It is important to note that a language change is a continuous process that can be observable also in a short term period. The variation in languages causes two problems in text retrieval; 1) spelling variation or a difference in spelling between the modern and historic language, and 2) semantics variation or terminology evolution over time.

III. TIME AWARE RETRIEVAL MODEL(T-R MODEL)

When any user search for temporal document like news archives or blogs, the time dimension is comprised in the retrieval model to improve the retrieval process. The documents are models based on the keyword score and temporal score of the query in the Time-aware Ranking retrieval. Performance of time aware ranking methods is better than the topic-similarity ranking e.g. Language modelling and TF-IDF. 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. Two main approaches for time aware ranking methods are 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.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2015

IV. 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, and relevance feedback model. There is also research in time based ranking, to add temporal factor in the search. The fundamental problem with the current approaches is 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.

A new method to rank a special category of time-sensitive queries which are yearly qualified in introduced by[2] 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. Feedback control theory based methods uses ranking errors to adjust the search engine behaviour. The method highlights 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.

In [3], proposed a method to use the micro-blogging data stream to detect fresh URLs. They also use micro-blogging data to compute new and efficient features for ranking fresh URLs. Recency Sensitive Queries refer to queries where the user required 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 engaged in this topic likely 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 maintaining 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.

V. RESEARCH TRENDS

There are various research trends of temporal information retrieval. The work by Alonso et al. presents a method for extracting temporal information and explains how it can be used for clustering search results [4]. Berberich et al. describe a model for temporal information needs [1]. These two projects rely on crowd-sourcing, mainly using Amazon Mechanical Turk, for evaluating parts of their work.

The primary focus of new sources is on number of projects on applying time information in documents. For example, the Time Frames project realizes an approach to augment news articles by extracting time information [5]. Google's news time-line is an experimental feature that allows a user to explore news by time.

There are number of applications that can benefit from leveraging more temporal information either by temporal expressions or timestamps.

A. *Exploratory Search*

Exploratory search is a specialization of information exploration which represents the activities carried out by searchers who are either unknown with the area of their goal, unsure about the ways to achieve their goals, or even unsure about their goals in the first place. Research in exploratory search systems has gained a lot of attention lately as they add a significant user interface component to help users search, navigate, and discover new facts and relationships. As the amount of information on the Web keeps growing, exploratory search interfaces are starting to surface. It is not clear how to leverage temporal information. A few problems are:

- How to expose temporal information in exploratory search systems?
- What's the best way of representing temporal information as a retrieval cue?
- For which data sources, besides news, does exploratory search make sense?
- Is e-discovery a vertical application that can benefit from temporal information?

B. *Micro-blogging and Real time search*

Micro-blogging sites like Twitter have gained a lot of attention lately as the ultimate mechanism to broadcast what's going on. Due to its nature, a typical message is very short and its lifespan is basically the crowd interest about that particular event be a football final game or an earthquake. [6] In the case of Twitter, it is very difficult to beat the timely broadcasting of an important event if one compares this to a news article. Each tweet has a time-stamp but the organization of such information is still not clear. In [7], in the news context, the reporter has to write an article that contains a few paragraphs and submit the final version through some content management version that would push it to an external website so a search engine can hopefully crawl and index it in time. In parallel, if a tweet is so important by

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2015

the time the reporter is finishing with the article, the main idea would be trending in Twitter, therefore highlighting its importance at a world scale. Some problems are:

- What is the best way to specify a time-line of events in micro-blogging?
- What is the lifespan of the main event?
- How fast and precise can one detect trending events?
- What is the fraction of new content on the topic stream?

C. *Temporal Summaries*

Temporal information is very important. One extension is to generate time sensitive summaries that can be used as temporal caption. In [8], the main goal of a snippet (or caption) is to present a document alternate that the user can fastly run down in the search results page without the need to click and read the full content of a document. In [9], there is a limit to the number of lines of text that the snippet should present. There are some problems also:

- When to show a time-stamp or temporal expressions?
- Should the snippet present the matching lines in a timeline?
- Is a temporal summary a good alternative for a document?
- For which kind of queries is a temporal summary appropriate?
- Should temporal summaries be query independent?

D. *Temporal Querying*

The temporal information extracted from documents can directly be used to allow the user of a search engine to constrain his/her query in a temporal manner. That is, in addition to a textual part, a query contains a temporal part. For example, in addition to world war a temporal constraint like 1944-1945 could be specified. The user would obviously expect documents about World War II as results for his query.

The objective when using a combination of a text and a temporal query can thus be formulated in the following way: The more both parts of the query are satisfied, i.e. the more the textual and the temporal parts fit to a document, the higher should be the rank of this document. The main problems for
such a combination of constraints is the following:

- How can a combined score for the textual part and the temporal part of a query be calculated in a reasonable way?
- Should a document in which the textual match and the temporal match are far away from each other be penalized?
- What about documents satisfying one of the constraints but slightly fail to satisfy the other constraint?

E. *Temporal Similarities*

A related research question to temporal querying is temporal document similarity. Instead of comparing a temporal query with the temporal information of a document, two documents can be compared with respect to their temporal similarity. The main problem arising here is what makes two documents temporally similar? This leads to the following questions:

- Should two documents be considered similar if they cover the same temporal interval?
- Should the temporal focus of the documents be important for their temporal similarity?
- Can two documents be regarded as temporally similar if one contains a small temporal interval of the other document in a detailed way?

F. *Temporal Clustering*

Clustering search results indicate that users are interested in dissecting a document collection by time. In [10], at the same time it is not clear for which kind of scenarios besides research like questions this approach would work. Key issues are:

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2015

- Can we find documents that are contemporary and therefore related?
- Which chronons can be more useful for clustering?
- Is a time-line the best way to cluster search results?

VI. APPLICATION OF TEMPORAL INFORMATION RETRIEVAL

There are various applications of temporal information retrieval system:

A. *Existing Temporal search engines*

An enormous amount of information is stored in web archives including web pages harvested and added into the archive repository when recrawling, as well as focused web warehouses like news archives. Information in such document repositories are helpful for both skilled users, e.g., historians, librarians, and journalists, as well as a general user searching for information needs in old versions of web pages. To date, there are existing search systems that provide accessibility to web archives. For example Google News Archive Search. This tool allows a user to search a news archive using a keyword query and a date range. In addition, the tool provides the ability to rank search results by relevance or date. However, there is a problem that has not been addressed by this tool yet, e.g., the effect of terminology evolution. Consider the example; a user wants to search for news about Earthquake in Maharashtra that are written before 2000. So, the user issues the query Earthquake in Maharashtra and specifies the temporal criteria 1990/01/01 to 2000/31/12. So only a small number of documents are returned by the tool where most of them are not relevant to the Earthquake in Maharashtra. In other words, this problem can be viewed as vocabulary mismatch caused by the fact that the term Earthquake in Maharashtra was not widely used.

B. *Analysis and Exploration over Time*

Some applications have taken time-based exploration of textual archives beyond just searching over time. Filtering and displaying information might benefit from presenting time information conveniently in some domains. In [5], time Explorer combines a number of interesting features present in other time-based systems, although extended in several important ways. First, users are enticed to discover how entities such as people and locations associated with a query change over time. Second, by searching on time expressions extracted automatically from text, the application allows the user to research not only how topics developed in the past, but also how they will continue to evolve in the future. All these features are combined in an intuitive easy-to-use interface, which is always a great challenge when designing search engines that allow for extended capabilities. In [12], other exploration-based systems have turned into other sources of textual information, for instance word evolution over time. This is naturally promising research direction, since there are available digitalized collections that span centuries. These systems need to employ a combination of several time-aware algorithms. For instance, they will require extracting time-related information from a given underlying textual collection, indexing this information along with the standard collection's contents and performing a combination of temporal query analysis and ranking after a user query is posed.

C. *Temporal Summarization*

Another stream of applications has focused into exploiting the use of time for enhanced story telling. In [13], the task of news summarization has been studied previously ranging from multi-document summarization to generate a time-line summary for a specific news story. A user enters a topic into a news search engine and obtains a list of relevant results, ordered by time. In [14], the user subscribes to this query so in the future she will continue to receive the latest news on this query. The time dimension comes into play when the user is observing a current document, and one may want to show the most relevant entities of the document for her query taking into account features extracted from previous documents. In [15], the most widespread summarization technology is the focused summaries produced by search engines, or search results snippets. Those are useful to assist users in deciding whether a document is relevant for a query or not.

D. *Temporal Clustering of Search Result*

Another popular application that makes use of time-based IR techniques is search results clustering, which is an important feature for some information retrieval applications, in particular, enterprise search systems. In [16], there is a

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2015

prototype that is able to display date and time attributes per cluster. Those attributes are extracted from the textual content of documents that belong to a particular cluster.

Furthermore, [17] extend the idea of reusing temporal information embedded in documents to enhance results presentation by introducing a time-line-based display of results. The timelines span different time granularities and display temporal information extracted automatically from documents and made explicit. They also explore how search results can be clustered according to time and how to produce temporal snippets to navigate through documents returned.

[18] Describe a method to group search result documents at a year level using a similarity measure that identifies the most relevant dates. Each group is then displayed differently in the results page, allowing for an easier exploration of the search results, as demonstrated by a user survey.

VII. CONCLUSION

The purpose of this survey paper is to provide an overview of temporal information retrieval systems. Temporal information enclosed in documents in the form of temporal expressions offer an interesting means to further enhance the functionality of current information retrieval applications. This paper covers the introduction of temporal information retrieval, some time aware retrieval models, research trends, various applications of temporal information retrieval.

REFERENCES

1. Berberich, Klaus, Srikanta Bedathur, Omar Alonso, and Gerhard Weikum. A language modeling approach for temporal information needs. Springer Berlin Heidelberg, 2010.
2. Zhang, Ruiqiang, Yi Chang, Zhaohui Zheng, Donald Metzler, and Jian-yun Nie. "Search result re-ranking by feedback control adjustment for time-sensitive query." In Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers, pp. 165-168. Association for Computational Linguistics, 2009.
3. Dong, Anlei, 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 Proceedings of the 19th international conference on World wide web, pp. 331-340. ACM, 2010.
4. Alonso, Omar, Michael Gertz, and Ricardo Baeza-Yates. "Clustering and exploring search results using timeline constructions." In Proceedings of the 18th ACM conference on Information and knowledge management, pp. 97-106. ACM, 2009.
5. Koen, Douglas B., and Walter Bender. "Time frames: Temporal augmentation of the news." IBM systems journal 39, no. 3.4 (2000): 597-616.
6. Allan, James, ed. Topic detection and tracking: event-based information organization. Vol. 12. Springer Science \& Business Media, 2012.
7. Makkonen, Juha, Helena Ahonen-Myka, and Marko Salmenkivi. Topic detection and tracking with spatio-temporal evidence. Springer Berlin Heidelberg, 2003.
8. Allan, James, Rahul Gupta, and Vikas Khandelwal. "Temporal summaries of new topics." In Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 10-18. ACM, 2001.
9. Alonso, Omar, Ricardo Baeza-Yates, and Michael Gertz. "Effectiveness of temporal snippets." In WSSP Workshop at the World Wide Web Conference—WWW, vol. 9. 2009.
10. Alonso, Omar, Michael Gertz, and Ricardo Baeza-Yates. "Clustering and exploring search results using timeline constructions." In Proceedings of the 18th ACM conference on Information and knowledge management, pp. 97-106. ACM, 2009.
11. Matthews, Michael, Pancho Tolchinsky, Roi Blanco, Jordi Atserias, Peter Mika, and Hugo Zaragoza. "Searching through time in the New York Times." In Proc. of the 4th Workshop on Human-Computer Interaction and Information Retrieval, pp. 41-44. 2010.
12. Erkan, Günes, and Dragomir R. Radev. "LexRank: Graph-based lexical centrality as salience in text summarization." Journal of Artificial Intelligence Research (2004): 457-479.
13. Yan, Rui, Xiaojun Wan, Jahna Otterbacher, Liang Kong, Xiaoming Li, and Yan Zhang. "Evolutionary timeline summarization: a balanced optimization framework via iterative substitution." In Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval, pp. 745-754. ACM, 2011.
14. Sipos, Ruben, Adith Swaminathan, Pannaga Shivaswamy, and Thorsten Joachims. "Temporal corpus summarization using submodular word coverage." In Proceedings of the 21st ACM international conference on Information and knowledge management, pp. 754-763. ACM, 2012.
15. Alonso, Omar, and Michael Gertz. "Clustering of search results using temporal attributes." In Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 597-598. ACM, 2006.
16. Alonso, Omar, Michael Gertz, and Ricardo Baeza-Yates. "Clustering and exploring search results using timeline constructions." In Proceedings of the 18th ACM conference on Information and knowledge management, pp. 97-106. ACM, 2009.
17. Campos, Rui, Alipio Mario Jorge, Guilherme Dias, and Celia Nunes. "Disambiguating implicit temporal queries by clustering top relevant dates in web snippets." In Web Intelligence and Intelligent Agent Technology (WI-IAT), 2012 IEEE/WIC/ACM International Conferences on, vol. 1, pp. 1-8. IEEE, 2012.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2015

18. Alonso, Omar, Jannik Strötgen, Ricardo A. Baeza-Yates, and Michael Gertz. "Temporal Information Retrieval: Challenges and Opportunities." TWAW 11 (2011): 1-8.
19. Kanhabua, Nattiya, Roi Blanco, and Kjetil Nørvåg. "Temporal Information Retrieval." FOUNDATIONS AND TRENDS IN INFORMATION RETRIEVAL 9, no. 2 (2015): 92-+.
20. Mathews, Litty K., and S. Deepa Kanmani. "A Survey on Temporal Information Retrieval Systems." International Journal of Computer Applications 58, no. 4 (2012).
21. Campos, Ricardo, Gaël Dias, Alípio M. Jorge, and Adam Jatowt. "Survey of temporal information retrieval and related applications." ACM Computing Surveys (CSUR) 47, no. 2 (2014): 15.
22. Barbosa, Luciano, Ana Carolina Salgado, Francisco de Carvalho, Jacques Robin, and Juliana Freire. "Looking at both the present and the past to efficiently update replicas of web content." In Proceedings of the 7th annual ACM international workshop on Web information and data management, pp. 75-80. ACM, 2005.