Ranking models in Information Retrieval: A Survey

R.Suganya Devi
Research Scholar
Department of Computer Science and Engineering
College of Engineering, Guindy, Chennai, Tamilnadu, India

Dr D Manjula
Professor
Department of Computer Science and Engineering
College of Engineering, Guindy, Chennai, Tamilnadu, India

Vidhya.R
ME(Final Year)
Department of Computer Science and Engineering
College of Engineering, Guindy, Chennai, Tamilnadu, India

Abstract- Information search and retrieval is one of the most prime fields of importance in today's computing world. Information retrieval is the process of obtaining information relevant to the users need, from a huge collection of documents. Ranking the documents collection is traditionally based on Topic Similarity. To improve the effectiveness of the retrieval, time can be incorporated into the ranking models. In this paper, we will carry out a survey of Traditional Ranking Models and how time can be incorporated in those ranking models.

Index terms – Information search and retrieval, Topic Similarity, Time Based Ranking Models

I. Introduction

Information retrieval can be defined as the process of retrieving documents by using structured or unstructured keywords that satisfy the user's need. Users request their needs in the form of queries. A query is a set of questions presented to the database in a predefined format. A user, in need of an information from a specific period, must provide the time explicitly in a normal query. This may not be possible always in searches involving large amount of data like blogs and news archives. But, time plays a major role in searching and retrieving relevant documents in a large collection of documents like the ones mentioned above.

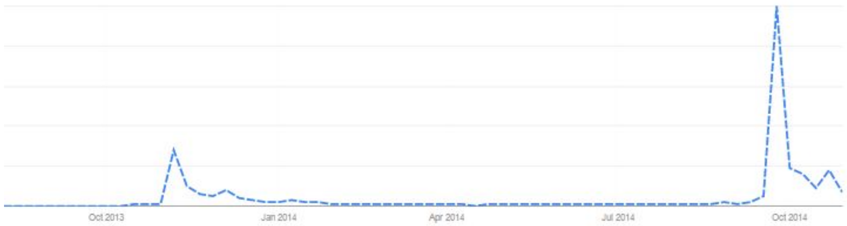

Fig. 1Interest of the keyword "Mangalyan" over the span of one and half years [Source: Google Trends]

Fig 1 shows the interest generated by the users of the keyword "Mangalyan" over the span of last one and half years. It can be seen that there are two peaks of interests corresponding to the date of launch of Mangalyan mission and the date of reach of orbit of the same. A user with the keyword Mangalyan is probably looking for information from these two periods. But, different documents with same topic Mangalyan may be generated daily over the period of time. So, in order to isolate the document that the user needs, i.e. from the two "peaks" of interest, we need to be able to find that time interval without the user needing to explicitly specify them. In such documents topic similarity alone is not enough for retrieving the relevant information effectively. So, a new type of query called time sensitive query can be implemented here. A query can be termed to be time sensitive, if the results of that query tends to be concentrated over small intervals of time rather than being spread uniformly. This paper presents a survey about the various traditional topic similarity based ranking models and how those models can be made to be time sensitive by various approaches.

II. INFORMATION RETRIEVAL

Information retrieval is the process of retrieving unstructured records i.e. records primarily in free from language text, from a huge collection of documents relevant to an information need. In information retrieval, a request to retrieve information is called a query. Information retrieval tries to find and retrieve documents that are relevant to the queries which are usually generated by the user. A document is said to be relevant if it satisfies the users need. Since the document collection is voluminous, a large number of documents may be termed to be relevant to the user. Herein lies the need to rank the documents with respect to generated query before returning the results to the user, so that the user can find “relevant” documents much faster.

Relevance is a topic of much importance in Information retrieval. It is an abstract measure of the degree of satisfaction of the user’s information need. Retrieval of a document can be based on two things:

i) Topic based Retrieval

ii) Time based retrieval

In the next section we discuss important Topic based Retrieval Ranking Models.

III. TOPIC BASED RANKING MODELS FOR IR

Traditionally Topic similarity has been used for ranking and retrieval of document. For retrieving relevant documents based on topic similarity four types of models are widely used for ranking, namely Set theoretic models, Algebraic Models, Probabilistic Models and Feature based retrieval models.

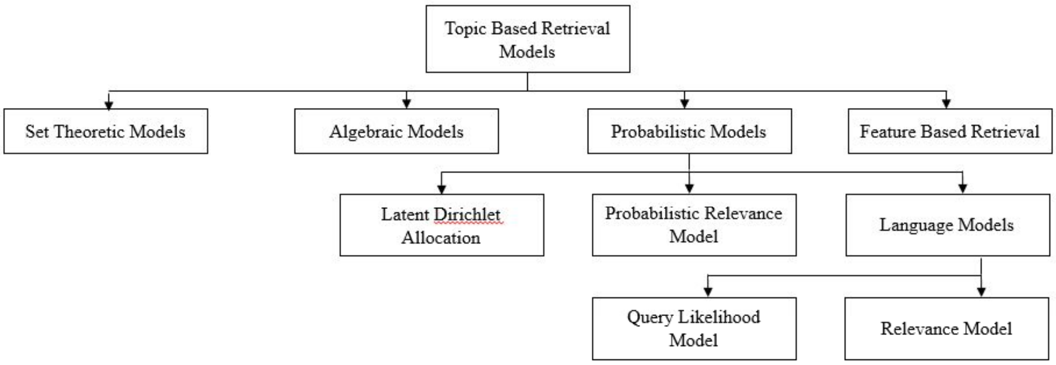

Figure 2. Topic Based Retrieval models

SET THEORETIC MODEL or Boolean model is the simplest form of Information retrieval. Here the each document in the collection is modelled as a bag of words. Here the query is given as a combination of words with a series of Boolean operators. For each word in the document a binary weighing system is used to denote the degree of relevance. The main disadvantage of this method is that the effectiveness of this retrieval model is entirely dependent upon the users’ ability to formulate a complex query.

*IN ALGEBRAIC MODELS*or Vector space model, the degree of relevance is given as the similarity between query and the document. Here, queries and documents are represented as vectors of term weights by using a term weighting scheme, e.g., *tf-idf*. The similarity of the term-weight vectors of q and d can be computed using the cosine similarity. The main disadvantage of this model is that, it makes no assumption about term dependency, which might lead to poor results [18]. In addition, the vector space model makes no explicit definition of *relevance*. In other words, there is no assumption about whether relevance is binary or mutivalued, which can impact the effectiveness of ranking models.

PROBABILISTIC MODELS exploit the property of uncertainty in information retrieval process and rank the documents based on probability of relevance. Here they assume relevance as a binary property and relevance of a documents is independent of other documents i.e it overcomes both the major disadvantages faced by the algebraic models. Probability of relevance accounts for the partial matches to be made possible. This makes probabilistic models sounder than Boolean or algebraic models. The Latent Dirchlet Allocation, Probabilistic relevance model and Language models are the most widely used models. They are explained below in detail.

LATENT DIRICHLET is based on a generative probabilistic model that models documents as mixtures over an underlying set of topic distributions. Here documents are random mixture of semantic topics and topics in turn are characterized by a distribution over words [1]. It works by trying to generate each word in a document by a multinomial distribution and then using topic to generate the word itself. Assuming this generative model for a collection of documents, LDA then tries to backtrack from the documents to find a set of topics that are likely to have generated the collection. The advantage of is that LDA is a probabilistic model with interpretable topics. The major disadvantage of LDA is that unsupervised nature of LDA makes its evaluation hard.

PROBABILISTIC RELEVANCE MODELS can be used to make an estimation of the probability that a document is relevant to a query. The underlying assumption of this model is that the probability of relevance depends on the query and document representation. It assumes that an ideal answer set that is preferred by the user as the answer set for a query maximizes the overall probability of relevance. But, this model has a major drawback that, it assumes the probability of relevance is independent of the terms and it does not have index term weighting. One prominent probabilistic relevance model incorporates a BM 25 function.

LANGUAGE MODELS are generally a probability distribution for generating a sequence of terms in a language. This underlying assumption means that all terms are mutually dependent which a major improvement is over probabilistic relevance models. Moreover, language models employ simple probabilistic approaches that are easy to evaluate. Language Modelling is used in Information Retrieval is using the idea that a document can be a match to a query only when the document is capable of generating that query i.e the document consists the keywords in the query, a number of times. Ponte and Croft [2] first introduced language modeling in information retrieval. This was followed by other scholars [3, 4, 5, 6] proposing some variations with similar framework. There are three types of language models. They are,

i) Query Likelihood Models
ii) Document Likelihood Model
iii) Comparing Query and Document likelihood Model

Generally for a language model the probability that a document belongs to the query,

$$p(d / q) \alpha\, p(q / d)\, p(d)$$

Where, p (q/d) is the probability that the query was generated by the document and p(d) is the posterior probability distribution. Traditionally, in topic similarity based searches, the posterior probability of the documents is considered to be uniform as the ranking models does not incorporate the time dimension into its search and hence is ignored. In Query likely hood model, the document is considered to be collection of words and the likelihood of query Q consisting individual words q_1, $q_{2....,}q_k$ being generated by the document,The likelihood for each document computed by this method is used for ranking.

In Document Relevance models or document likelihood models, relevance feedback is incorporated and the joint probability of observing a word w occurring along with the query words q_1, $q_{2....,}$ q_k is estimated. Lavrenko and Croft [6] have given two methods for estimating the joint probability, both of which differ in a minute assumption made i.e. in first method w is sampled in same way as query words whereas in second method two different mechanisms are used for sampling words and queries. In this paper the first method is adopted as they reported first method to be slightly more efficient. If w and q_1, $q_{2….,}$ q_k are independent, we get

$$P(Q/M_D)=\prod_{w\in Q}P(w/M_D)\prod_{w\notin Q}(1-P(w/M_D)$$

These are the extensively used traditional ranking models that use topic similarity for ranking models. But, these ranking models cannot produce an effective retrieval in a collection of documents whose distribution vary in time such as blogs, news archives, personal emails etc. because in such document collections, the relevance of the documents is closely related to its temporal characteristics such as publication date, date of creation etc. Hence it is necessary to introduce time dimension into these ranking models to improve the effectiveness of the retrieval.

Table 1 - Comparison of various Topic based retrieval models

MODELS	DEGREE OF RELEVANCE	MERITS	DEMERITS
Set Theoretic Model	Binary Weighing System	- Simple Concept	- Dependent upon users' ability to formulate a complex query
Algebraic Models	Similarity between Query and Document	- Not completely dependent on users abilities.	- No explicit definition of relevance
Latent Dirichlet Allocation	Multinomial Distribution	- Probabilistic model with interpretable topics.	- Unsupervised nature of LDA - Evaluation is hard.
Probabilistic Relevance Models	Identical Query and Document Representation	- Probability of relevance depends on the query and document representation	- It assumes the probability of relevance is independent of the terms
Language Models	Simple Probablistic Approach	- It assumes the probability of relevance is dependent of the terms. - It employs simple probabilistic approaches that are easy to evaluate.	

IV. TIMEBASED RANKING MODELS FOR IR

In this section, a brief overview of existing how time can be exploited into ranking and time based ranking models is presented. There are two methods of time based ranking techniques: Link based analysis and Content based analysis. Since data about links may not be available always, this survey will focus on content based analysis.

Existing works on Time stamped Documents:

One of the most important available temporal data in a document is its publication time. Using the publication time, the ranking for time sensitive document collections can be improved. Li and Croft [7] introduced time into existing language models like query likelihood model and proposed a time based language model. They considered that the posterior probability of time sensitive document collections is not uniform unlike previous approaches where they were considered to be uniform and they used an exponential distribution dependent on publication for posterior probability and thereby assigning documents with recent creation dates with higher probability of relevance. Here only the meta data of publication time was used.

Time can also be incorporated into probabilistic models. Robertson et all. [14] [15] stated that to produce optimal ranking of a set of documents, they must be ranked by the posterior probability of the documents belonging to the relevance class of the query. As the original BM25 model doesn't have the ability to handle the non-textual features of a document such as time, Craswell et all [16] introduced c_d and t_d which are content and time dependence. Here c_d is ignored as the temporal relevance of day t_d does not depend on content but rather on the density of relevant documents.

Jones and Diaz [8] classified the temporal profile of the documents into atemporal, temporally unambiguous and temporally ambiguous based on the publication dates of these documents. The Features they described were Kl Divergence, Autocorrelation, Statistics of the Rank Order P(t/Q) and Burst model. They proposed a method to automatically identify and classify the documents based on temporal profile.

Existing works on Non - Time stamped Documents:

But for documents for which date of publication is unknown, determination of time is very important. Generally two methods are presented for identification of non-time stamped documents: Learning methods and Non Learning Methods. Learning Methods use statistical analysis to give a distribution of events mentioned in the documents [9, 10]. Gilad Mishe [11] improved the retrieval of information from blogs by introducing a temporal prior using the top 500 posts of topically similar contents. Non learning methods find the most probable relevance date by using frequency of the date appearing on the document [12, 13]. A hybrid method was proposed by Chen et. al. to find the document date by machine learning techniques [17].

Evaluation Parameters:

Two commonly used effectiveness evaluation parameters are precision and recall. Precision is the ratio of retrieved relevant documents to relevant documents. Recall is the ratio of retrieved relevant documents to retrieved documents. Let R be set of all relevant documents, and A be set of all retrieved documents. Then

$$\text{Precision} = \frac{|R \cap A|}{|R|} \qquad \text{Recall} = \frac{|R \cap A|}{|A|}$$

F-measure is a single parameter that incorporates both recall and precision into one parameter by taking weighted harmonic mean of both precision and recall. Generally, precision at top k documents is denoted as P@k. *Mean Average Precision* (MAP) provides a summarization of rankings from multiple queries by averaging the precision values from the rank positions where a relevant document was retrieved, or *average precision.*

Table 2 - Summary of various Time based retrieval models

Author	Description	Comment
Li and Croft	Introduced Time into Language Models to retrieve recent documents.	Gives importance only to recent documents even when prior documents may be relevant.
Craswell et all	Proposed a method to introduce time into BM 25 ranking model.	Tries to evaluate non-textual features using BM 25
Jones and Diaz	Proposed a method to identify various temporal profiles and various features to measure temporal profiles.	User has to explicitly select temporal data
Gilad Mishe	Improved the retrieval of information from blogs by introducing temporal prior	Uses temporal data of blog to improve effectiveness of retrieval.

V. Conclusion

Thus a study on various approaches to information retrieval has been carried out. To study the usage of time dimension in information retrieval, it is very important to first completely understand the features and setbacks of the various approaches which has been carried above. Also extensive studies are carried out on each approach to incorporate time dimension into it. Of the various methods it was found that language modelling method had the most suitable characteristics to incorporate the time dimension into it. The reasons for these has also been worked on and presented in this paper. The various temporal classes and features of temporal profiles were also highlighted. A brief method of how time has been integrated into search and retrieval was also presented above. Using this, we can investigate further how a general framework can be developed for answering time sensitive general queries.

REFERENCES

[1] David M. Blei, Andrew Y. Ng and Michael I. Jordan "Latent Dirichlet Allocation" *Journal of Machine Learning Research 3* (2003) 993-1022

[2] J. Ponte and W. B. Croft, "A Language Modeling Approach to information retrieval". *Proceedings of the 21st annualinternational ACM SIGIR conference*, 275-281, 1998.

[3] F. Song and W. B. Croft. "A general language model for information retrieva". *Proceedings of the 22nd annualinternational ACM SIGIR conference*, 279-280, 1999

[4] D. Hiemstra. *Using language* models *for information retrieval*. PhD thesis, University of Twente, 2001.

[5] J. Lafferty and C. Zhai. "Document language models, query models, and risk minimization for information retrieval". *Proceedings of the 24th annual international ACM SIGIRconference*, 111-119, 2001.

[6] V. Lavrenko and W. B. Croft. "Relevance-based language models". *Proceedings of the 24th annual international ACMSIGIR conference*, 120-127, 2001.

[7] X. Li and W.B. Croft, "Time-Based Language Models," Proc. 12th ACMConf. Information and Knowledge Management (CIKM '03), 2003.

[8] R. Jones and F. Diaz, "Temporal Profiles of Queries," ACM Trans. Information Systems, vol. 25, no. 3, article 14, 2007.

[9] R. Swan and D. Jensen. Timemines: Constructing timelines with statistical models of word usage. In *Proceedings of KDD Workshop on Text Mining*, KDD '00, 2000.

[10] R. C. Swan and J. Allan. Extracting significant time varying features from text. In *Proceedings of the 1999 ACM CIKM International Conference on Information and Knowledge Management*, CIKM '99, pages 38–45, 1999.

[11] G. Mishne, "Using Blog Properties to Improve Retrieval," Proc. First Int'l Conf. Weblogs and Social Media (ICWSM '07), 2007.

[12] D. Llidó, R. B. Llavori, and M. J. A. Cabo. Extracting temporal references to assign document event-time periods. In *Proceedings of the 12th International Conferenc*e

[13] I. Mani and G. Wilson. Robust temporal processing of news. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, ACL '00, pages 69–76, 2000.

[14] K.S. Jones, S. Walker, and S.E. Robertson, "A Probabilistic Model of Information Retrieval: Development and Comparative Experiments - Part 1," Information Processing and Management, vol. 36, no. 6, pp. 779-808, 2000.

[15] K.S. Jones, S. Walker, and S.E. Robertson, "A Probabilistic Model of Information Retrieval: Development and Comparative Experiments - Part 2," Information Processing and Management, vol. 36, no. 6, pp. 809-840, 2000.

[16] N. Craswell, S.E. Robertson, H. Zaragoza, and M. Taylor, "Relevance Weighting for Query Independent Evidence," Proc. 28th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '05), 2005.

[17] Z. Chen, J.Ma, C. Cui, H. Rui, and S. Huang. Web page publication time detection and its application for page rank. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '10, pages 859–860, 2010.

[18] R. A. Baeza-Yates and B. A. Ribeiro-Neto. *Modern Information Retrieval – the concepts and technology behind search, Second edition*. Pearson Education Ltd., Harlow, England, 2011.