International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2015): 6.391

Optimized Ranking Framework for Information Retrieval

Snehal Ukarande[1], Ashish Manwatkar[2]

[1, 2]Indira college of Engineering and Management, Pune

Abstract: ***Current era is of fast information retrieval. There are lots of research methodologies which are arising to give most fast and correct result set. This paper sheds light on fast and to the point information retrieval methodology for giving user with the most relevant document in secure way. Learning to rank is being progressively more trendy research area in the machine learning. Problem of ranking aims to encourage an ordering or inclination of relation among a set of instances. Learning to rank for information retrieval has gained a lot of interest in the recent years as ranking is the main problem in many information retrieval applications, like document retrieval, multimedia retrieval, text summarizing, collaborative filtering, question answering and online advertising machine translation etc. The large amount of the web documents makes it usually impracticable for the common users for finding their desired information by surfing over net. As a result, effective information retrieval is being more vital and also search engine has turned out to be important tool for people to search their required information.***

Keywords: Ranking Model, Learning to Rank, Information Retrieval, Data Mining

1. Introduction

Learning to rank is a new research domain which is evolving since the last decade. The search engines are crucial for finding and getting information on the web and other information systems. To a better extent the ranking function are used to determine the excellence of search engines, they are also used to create the results according to user's query as ranking is the vital part of the information retrieval system. When user queries, the documents should be given rank according to the relevance factor to the query. Different types of machine learning algorithms are being used to learn the ranking function. Hence, Ranking has widespread applications such as commercial search engines and recommendation system that can find out relevance factor between the relevant documents in context of given user's query and place them in order of their significance in the rank list. Such classification explores the queries and documents are given where every query is coupled with a perfect ranking list of the documents. The model for ranking is then formed using the classification process according to the given query

Learning to rank methods in the information retrieval allows retrieval systems to incorporate hundreds or even thousands of randomly defined features. Important thing is, these approaches, without human intervention, learn the most useful combination of the features in the ranking function which depends on the availability of the data for the classification. The evaluation metrics are requisite to compute the quality of search engine that is one of the most useful metric in ranking i.e. Discounted Cumulative Gain, which is used to compute ranking quality of the search engines. The information retrieval is habitually used to evaluate usability of web search engine algorithms or other related applications. Discounted cumulative gain measures the significance, of the document based on its best position in the rank list. If the relevant document is at lower location then it is not more helpful for the user to gain knowledge. The rule is to incorporate both query level selection as well as document level selection for ranking and introducing an expected discounted cumulative gain loss optimization algorithm, for selecting most informative and relevant document associated to query.

2. Literature Survey

Learning to rank has three common approaches they are: Pair wise approaches, point wise approaches and list wise approaches. These three dissimilar approaches can be trained to rank in different ways. With the objective of ranking unlike input and output spaces may be defined, unlike hypotheses may be used and gives different loss function. The Point wise approaches are the former approaches [2]. The basic hypotheses of this approach is utilized for mapping the document's ordinal scale in the numeric values using regression and classification method, it tries to compare the relevance result of every other two documents, then comparison result is produced. Based on that result the document will be given rank. Binary classifier method is used by pair wise approach that will tell which document is better in a given two documents.

Objective of using binary classifier is minimizing average number of inversions for ranking functions. The list wise approaches is similar with the basic idea of pair wise approach, it simply compares the relevance list of documents based on the query, as a substitution of trying to get ranking score for each document independently. It uses soft Rank and Ad Rank algorithms for giving the rank. If compared with conventional active learning algorithms; there is yet work is going on in the active learning for ranking in current years. The problem of text selection based on query in ranking is demonstrated by Carbonell and Donmez[3]. The ambiguity sampling is simplest and general approach in active learning, issue in sampling is that, an algorithm which is selecting queries for which the label ambiguity samples have highest relevance score [4]. The main drawback in this type of approach is noise and variance. Noise free classification function is used by Active learning algorithm lessens the noise and reduces variance which is proposed in [5]. Query by Committee algorithm [8] another frequently used

approach for active learning is choosing a query which is once added to the training set ultimately increases the objective function value which is getting optimized [6]. Most of the other ranking algorithms such as Rank SVM and Rank Boost [7] have suggested for adding the most relevant pair of documents to the training set, the predicted relevance scores of the documents are very close under the current ranking models. In the form of binary relevance, greedy algorithm is proposed which will select the document and differentiates two different ranking systems in the terms of average precision. The comparison of relevant document selection methodologies for ranking are proposed in [8]. L. Yang, L. Wang [4] projected greedy query selection algorithm which will help to minimize query density and query diversity. Some experimental and theoretical work related to query sampling are demonstrated in [5] the results show that better having more number queries but less number of documents per query rather having more documents and less number of queries.

Most of the internet users depends on search engines for extraction of the information by providing a query from any walk of life. These queries are processed by search engine and a specific information retrieval or mining algorithm is applied to obtain the cluster of documents which is relevant to the query. Once the documents are retrieved they have given relevance factor. The documents with topmost relevance factor are the most useful documents. This task of collection of the most relevant document at top of the list is called ranking of the documents.

Fundamentally, there are two types of ranking models: static ranking model and dynamic ranking model. In former days ranking algorithms were based on former information about the websites. PageRank, SALSA,HITS, RankNet and fRank are example of the static algorithms. These use static features of web pages therefore known here as static Ranking algorithms. The static ranking algorithms don't take into account of the interaction with user and faces issues like query uncertainty and diversity in intent of user. There is an intrinsic trade-off amongst number of results provided for user intended object and number of objects retrieved .A way to combine or contradictory objective of high recall and result diversification is provided by dynamic ranking algorithm. These algorithms run the interaction with the user to know his intention amongst the various possible intents, otherwise they try for reordering the result of the first retrieval process and provides refined results to the user. They focus on both the relevance factor and diversity.
Ranking can be useful at different applications. For example, Jenq-Neng Hwang, H.Shihand Chung-Lin Huang has demonstrated a system which uses content based attention ranking which utilizes visual and contextual attention model for baseball videos[11]. Author has analyzed the way people are being excited towards the watched video content and projected a content-driven attention ranking strategy ,which is enabling client users to continuously browsing the video as per their preference. The Google PageRank algorithm is extended to the attention rank algorithm ,which sorts the websites based on the significance, measures the user interest efficiently and user interface level for each video frame. Embedding visual attention model based on object with context attention model derives the degree of attention. This can more reliably takes the benefit of the human perception characteristics, also effectively identifies which video contents can attract user's attention. The information of user's feedback is used in re-ranking procedure for further improvement in the retrieval accuracy. Demonstrated algorithm is particularly evaluated on broadcast baseball videos[11].

Ranking models can be useful for particular domain search. With the tremendous emergence of vertical search domains, application of the broad ranking to different domains is no longer advantageous due to domain dissimilarities, while building of a unique ranking model for each domain is difficult for labeling data as well as time consuming for training models. Bo Geng, considered these hurdles by stating a regularization-based algorithm called ranking adapting RA-SVM, via which existing ranking model can be adapted to a new domain, likewise the training cost and the quantity of labeled data is reduced at the same time the performance is guaranteed[12]. This algorithm prerequisite is the predicting the existing ranking models, than their internal representation . In addition to this, authors assumed that documents having similarity in the particular domain feature space should be having consistent rankings and add some constraints for controlling the slack variables of RA-SVM and margin adaptively. At the last, ranking adaptability measurement is stated for quantitatively estimating an existing ranking model which can be adapted for a new domain.

3. System Overview

Information retrieval is being vital research area in computer science. Information retrieval is related to the searching and retrieval of knowledge-based information from database. Information retrieval is generally measured as a branch of computer science that deals with the storage and access of information.

Main motive of information retrieval system is finding most relevant information that suffices user information needs. General representation of the ranking framework is given in the Fig. 1.

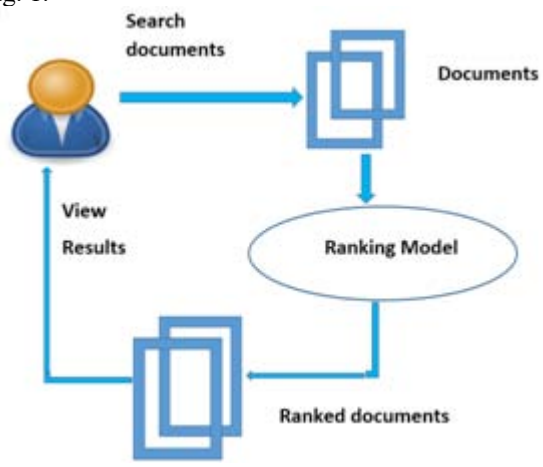


Figure 1: Architecture of document retrieval system.

The documents present in storage database will be searched by the user as shown in fig.1. As per the need and query, the earlier ranked documents will be displayed as output to the

user. If the documents are appropriately ranked then the loss of data is optimized during searching relevant data. The user always searches for a document and due to inappropriate ranking the most relevant data is also not be able to be retrieved. Hence proper ranking models should be used in order to retrieve total relevant data therefore reducing the loss.

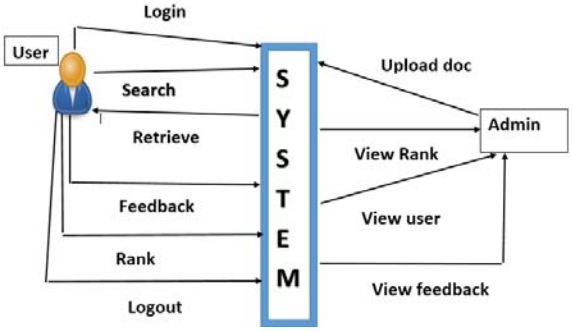


Figure 2: Architecture of proposed system.

The specific architecture of the proposed ranking framework having two interacting users mainly user and admin is shown in fig.2. User is actually using the system for getting the most relevant documents as per his requested query. User would be able to search the document by domain and the sub-domain , also would be able to give the feedback and rank to the document as per the documents usability relevant to his search. Admin would be able to upload the documents, see the ranks of the documents, see the list of users registered to the system as well as see the feedback given by the user.

4. Flow of the Proposed System

User flow is depicted in the Fig. 3.which shows user interaction flow with the system. First of all user has to register with the system. If user wants to search any data, user will get login to the system and will enter query for search ,then ranking framework will return the best result to the user which would be in the encrypted format for getting that document view user has to decrypt the document. Finally user will give rank and feedback for the document for next best search and will get logout from the system.

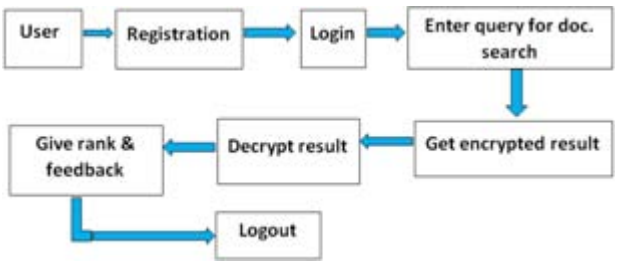


Figure 3: User interaction flow.

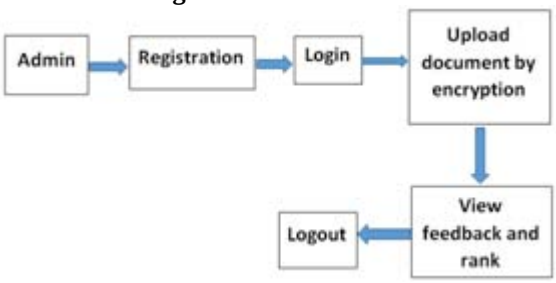


Figure 4: Admin interaction flow

The admin interaction flow with the proposed system is shown in the fig.4.For managing all the framework admin must be the registered one and has to login with the system. Then admin would be able to upload the documents ,Admin can see the feedback given by the user and ranking to the document.

5. Algorithmic steps

For getting access to the system user or admin first has to register with the system. Searching process comprises following algorithm:

1. First of all user will get login into the system.
2. Enter query for searching the document.
3. Based on query earlier ranked documents are retrieved from the database.
4. Ranking is calculated
5. System retrieves the document based on rank calculated in earlier step from db which would be the encrypted document.
6. User decrypts the document with credentials returned by admin as user is registered one and gets the most relevant document.
7. Based on the retrieved documents and the relevance factor user will give rank and feedback for the document.
8. Finally user will get logout from the system.

6. Mathematics used

Consider a set S which is having elements related to a program. The mathematical model is given as below,
Where,
S= Initial state and E= End state.
X= Input set and Y= Output set.
F = ranking function
DD= Deterministic data in this case deterministic data would be the documents which are uploaded and rank stored in the database prior to search.
NDD= Non-deterministic data would be rank and feedback given by the user after retrieving the document.
Various functions in the proposed system are given below with their input set and the output set.
F(register) = Register will register the information of User or Admin.
Input: Information
Output: Registration successful message.
F(login) = Login of User or Admin will be performed.
Input: Registration information of User or Admin.
Output: Login successful message.
F(search) = User searches a document.
Input: Document name.
Output: Searched document.
F(view) = Admin view all documents.
Input: Document list.
Output: Document.
F(upload) = Admin upload a document.
Input: Document.
Output: Document uploaded successfully message after upload operation success.
F(feedback) = User gives feedback and Admin receives it.
Input: Feedback message.

Output: A successful feedback submitted message will be displayed at user side and at admin side a feedback is received.
F(rank) : This function will calculate the rank to the document by considering the below expected loss function.
Expected loss can be calculated by the below formula.

$$EL(x) := \min \int_a^y loss(a,y)\ P(y|x, D)\ dy$$

where, loss(a,y) is loss occurred while performing the action a when true output is y. x is input set while y is output set and D is training data set.

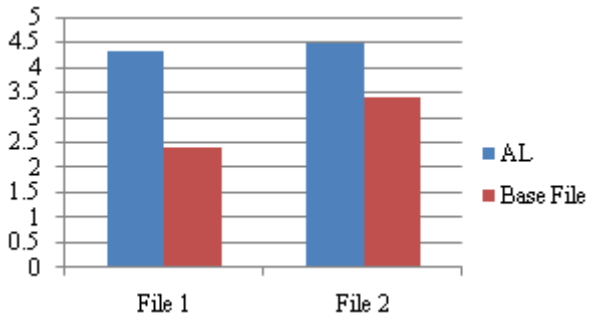
The ranking methodology used in the proposed system gives the required document in the polynomial time hence this system is P – problem system. Solutions to the query can be verified in the polynomial time with the help of proposed system hence this system is also NP – complete problem.

7. Result Analysis

Table. 1. shows that the data set of file type base set is less optimized than the Active Learning(AL) file type. The file 1 of base type data set is given rank as 8 and it is optimized 70% while file 2 of AL type is given rank as 8 and it is optimized 81%. In other words proper ranking leads to great loss optimization

Table 1: Result Analysis

Data Set	Type	Rank	Optimization
File 1	Base	8	2.4 (70%)
File 1	AL	8	4.3 (81%)
File 2	AL	2	4.5 (85%)
File 2	Base	2	3.4 (75%)



8. Conclusions

As technology improves each day new developments are continuously infiltrating our lives. Research in learning to rank is a friendly process and the must of ranking change every day depending on the requirements from the user. Active learning for ranking differs from Active learning for classification and regression including learning for ranking that has some unique features. There are many ranking algorithm which are all time consuming and also cost much in obtaining labeled data compared with those algorithm Expected loss optimization for query and document level ranking by active learning performs efficiently by providing the user the most informative documents for their references [10].

References

[1] Bo Long, Jiang Bian Olivier Chapelle, Ya Zhang, Yoshiyuki Inagaki, and Yi Chang, “Active Learning for Ranking through Expected Loss Optimization”,IEEE transactions on knowledge and data engineering, VOL. 27, NO. 5, MAY 2015.
[2] B. Qian, H. Li, J. Wang, X. Wang, and I. Davidson, “Active Learning to Rank using Pairwise Supervision,” In Proc. 13th SIAM Int.Conf. Data Mining, 2013, pp. 297–305.
[3] P. Donmez and J. G. Carbonell,“Optimizing estimated loss reduction for active sampling in rank learning”, In ICML '08: Proceedings of the 25th international conference on Machine learning, pages 248- 255, New York, NY, USA, 2008. ACM.
[4] D. Lewis and W. Gale, “Training text classifiers by uncertainty sampling”, In Proceedings of the 17thAnnual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 3-12, 1994.
[5] D. A. Cohn, Z. Ghahramani, and M. I. Jordan, “Active learning with statistical models”, In G. Tesauro, D. Touretzky, and T. Leen, editors, Advances in Neural Information Processing Systems, volume 7,pages 705-712. the MIT Press, 1995.
[6] C. Campbell, N. Cristianini, and A. Smola, “Query learning with large margin classifiers”, In Proceedings of the Seventeenth International Conference on Machine Learning, pages 111-118. Morgan Kaufmann, 2000.
[7] Y. Freund, R. Iyer, R. E. Schapire, and Y. Singer,”An efficient boosting algorithm for combining preferences”, Journal of Machine Learning Research, 4:933-969,2003.
[8] Y. Freund, H. S. Seung, E. Shamir, and N, ”Tishby Selective sampling using the query by committee algorithm”,Machine Learning, 28(2- 3):133-168, 1997
[9] A. Aslam, E. Kanoulas, V. Pavlu, S. Savev, and E. Yilmaz, “Document selection methodologies for efficient and effective learning-to-rank,” In Proc. 32nd Int. ACM SIGIR Conf. Res. Develop.Inform. Retrieval, 2009, pp. 468–475.
[10] Aditi Sharma, Nishtha Adhav, Anju Mishra ,”A Survey : Static and Dynamic Ranking ” , International Journal of Computer Applications (0975 -8887) Volume 70 No-14 ,May 2013.
[11] Huang-Chia Shih, Jenq-Neng Hwang, Fellow and Chung-Lin Huang, “content based attention ranking Using Visual and Contextual Attention Model for Baseball Videos”, IEEE transactions on multimedia, vol. 11, NO. 2, feb 2009.
[12] Bo Geng, Linjun Yang, Chao Xu, and Xian-Sheng Hua, “Ranking Model Adaptation for Domain-Specific Search”, IEEE transactions on knowledge and data engineering,vol. 24, no. 4, april 2012.