International Journal for Modern Trends in Science and Technology
Volume: 03, Issue No: 05, May 2017
ISSN: 2455-3778
http://www.ijmtst.com

Information Retrieval for Aspect Diversity using Optimal Ranking Labeling Strategy

A.Aruna[1] | K.Tamilselvan[2]

[1,2]Department of CSE, AVC College of Engineering, Mayiladuthurai, Tamilnadu, India.

To Cite this Article
A.Aruna and K.Tamilselvan, “Information Retrieval for Aspect Diversity using Optimal Ranking Labeling Strategy”, *International Journal for Modern Trends in Science and Technology*, Vol. 03, Issue 05, May 2017, pp. 257-260.

ABSTRACT

To meet biologists information need better, Information Retrieval (IR) techniques designed for biomedicine domain have been addressed, focusing on how to effectively retrieve the needed information. Given a query, an IR system can search for its relevant documents, and rank the documents based on their relevance degrees to the query. Unlike traditional IR,biomedical IRfaces some domain specific challenges, most of which are due to the abundance of the terminologies. To meet the information need more completely, biomedical IR system should cover the relevance documents from different aspects, where an aspect of relevance documents refers to a subset of relevant documents related to the same terminologies. Therefore, biomedical retrieval systems not only focus on obtaining the most relevant documents to a given query, but also emphasize the query-related aspects coverage in the document ranked list, which is mostly denoted as the diversity of the searching result.

Keyword*- Information Retrieval*

I. INTRODUCTION

In recent years, research articles in biomedicine domain have increased exponentially, which makes it difficult for biologists to manually capture all the information they need. To meet biologists information need better, Information Retrieval (IR) techniques designed for biomedicine domain [1] have been addressed, focusing on how to effectively retrieve the needed information. Given a query, an IR system can search for its relevant documents, and rank the documents based on their relevance degrees to the query. Unlike traditional IR, biomedical IR faces some domain specific challenges, most ofwhich are due to the abundance of the terminologies. Different articles may use different terminologies to represent the same concept, and as a result, two relevant documents [5] for the same query may vary a lot.In biomedical information retrieval,ranking only based on document relevance is not sufficient to meet the information need, because relevantdocuments may beredundant with each other. Aspect retrieval wasproposed toreduce the redundancy and improve result diversity. Diversity here means that when a user submits a query to a retrieval system, he (or she) is provided with the diversified results covering as many aspects of the query as possible, and then the user will find what he desires the most. To meet the information need more completely, biomedical IR [1] system should cover the relevance documents from different aspects, where an aspect of relevance documents refers to a subset of relevant documents related to the same terminologies. Therefore, biomedical retrieval systems not only focus on obtaining the most relevant documents to a given query, but also emphasize the query-related aspects coverage in the document ranked list, which is mostly denoted as the diversity of the searching result. The paper goal is to retrieve the most relevant documents covering as many aspects as possible.

II.EXISTING SYSTEM

Existing methods can be divided into two categories:

- Diversity orient retrieval methods
- Learning to rank methods

A.DIVERSITY ORIENT RETRIEVAL METHODS

Most existing diversity orient retrieval methods can be divided into two categories: Implicit approaches and explicit approaches. Implicit approaches model query-related aspects by modeling the relationship among documents. On the other hand, explicit approaches model the query-related aspects using external resources, such as the top-ranked documents.

B.LEARNING TO RANK METHODS

Learning to rank [3] is grouped into three approaches: the point wise approach, the pair wise approach and the list wise approach [11]. Different approaches model thelearning to rank process [3] indifferent ways.Point wise approach is usedtoprovide the exact relevance degree of each document is what we are going to predict.Pair wise approach does not focus on accurately predicting the relevance degree of each document; instead, it cares about the relative order of two documents. List wise approach [11] takes the document list as the object to calculate the difference between the predicted ranking list and the target ranking list of documents. Intuitively, list wise approach [11] utilizes the most ranking information to construct the ranking model.

C. DEMERITS

- In existing system, the problem for biologist to capture all the needed information manually.
- Information Retrieval technologies can deal with the problem automatically providing users with the needed information.
- There is a great challenge to apply IR technologies directly for biomedical retrieval, because of the abundance of domain terminologies.
- The existing system only provides most relevant documents.

III. PROPOSED SYSTEM

IR for diversity [8] is a retrieval task that diversifies the search results to meet the multiple information needs of different user. Proposed system consist of two labeling strategy.

- Optimal Ranking Labeling Strategy
- Group-wise Labeling Strategy

A. OPTIMAL RANKING LABELING STRATEGY

In the optimal ranking list, more diversified relevant documents are ranked higher than less diversified ones. Optimal ranking labeling strategy is based on this idea by taking the number of aspects for one document and the frequency of aspects among all the documents into account, where the aspects for a relevant document reflect its diversity [6] degree. The Optimal ranking strategy provides the target ranking to train the ranking models, which may be more suitable for list wise learning [11] to rank approach.

B. GROUP-WISE LABELING STRATEGY

Documents with different labels are treated as a group, and the ranking task is then reduced from ranking the whole set of documents to ranking a group of documents with different labels. The division of groups is based on the diversity degrees of the documents, the group-wise framework canbe more focused on the diversified documents [8], and the final ranking model may improvethe performance [10] in terms of both relevance and diversity [5].

IV. SYSTEM ARCHITECTURE

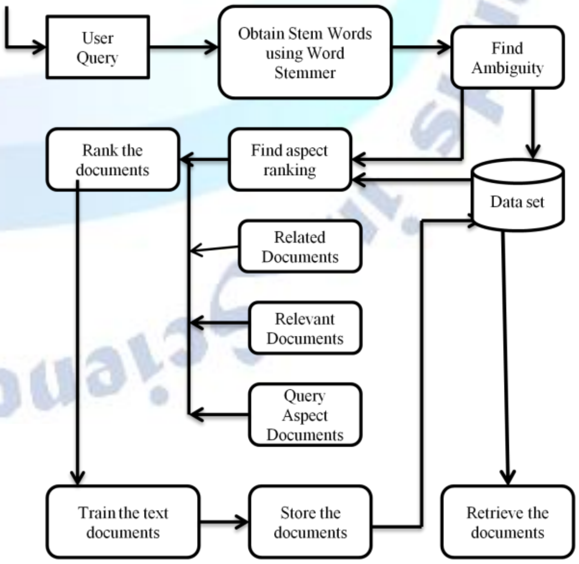

Figure: System Architecture

V.SYSTEM DESIGN

A. Preprocessing Text Data
B. Feature Extraction
C. Document Retrieval
D. Ranking and Learning Dataset

A. PREPROCESSING TEXT DATA

In this module, the user can enter the text query. This text query may be a sentence and it can be considered as the input to the paper. The Sentence Splitter is used to split the given user input. The Word Stemmer is used to identify the stop words in the given text query. Then, the identified stop words are eliminated.

B. FEATURE EXTRACTION

In this module, the Word Sense Disambiguation is used to find the ambiguity in the given user query. The UMLS (Unified Medical Language System) is a compendium of many controlled vocabularies in the biomedical sciences. The UMLS is used to provide ambiguous words in the given user text query. The Word Stemmer is also used in this module and it is used to find the stem word in the given input. The stem word is nothing but it is the specialwords or the important words in the input sentence.

C. DOCUMENT RETRIEVAL

In this module, the stem words are obtained and then document retrieval isprocessed. Thismodule is used to find the document label URLs. Document ranking is categorized into Related Documents and Relevant Documents. The Obtained document URLs consist of the ranking based on the relevance degree [5] to the given query.

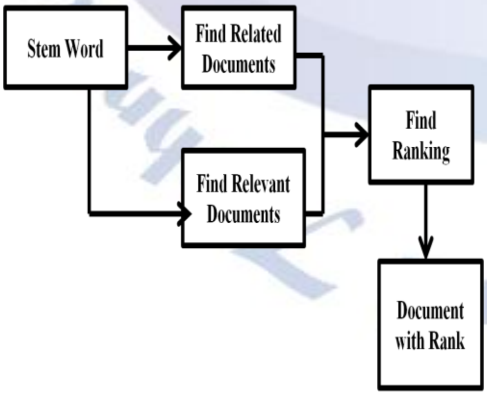

Figure: Document Retrieval

D. RANKING AND LEARNING DATASET

In this module, the obtained Document Label URLs are reranked with different aspect by using Optimal Ranking Labeling Strategy. Then the reranked documents are trained and stored in the dataset [10]. This module is used to view the document with different aspect to given user query.

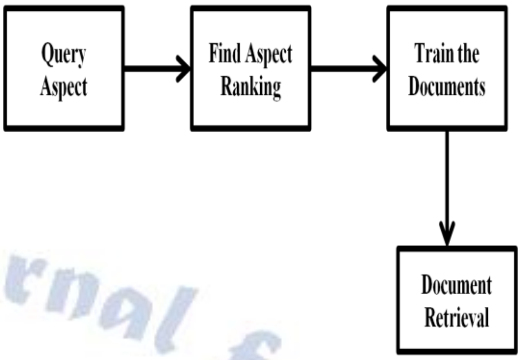

Figure: Learning and Ranking Dataset

VI. CONCLUSION AND FUTURE ENHANCEMENT

In this paper, a learning to rank based frameworkfor biomedical information retrieval, focusing on improving the retrieval performance in term of both relevance and diversity. The proposed methods are respectively based on optimal ranking strategy and group-wise learning to rank, seeking to boost the diversity of retrieved relevantdocuments. The future work attempt to explore an approach an automatic aspect mining when the dataset contains no such annotations and attempt toincorporate an aspect-related item into the loss function forlearning to rank methods, which mayproduce a moreeffective model. The future work will also develop and examine the performance of other features, especially some domain specific features to make the framework more applicable for biomedical document retrieval.

REFERENCES

[1] A Survival Modeling Approach to Biomedical Search Result Diversification Using Wikipedia" , Xiaoshi Yin, Jimmy Xiangji Huang, Senior Member, IEEE, Zhoujun Li, Member, IEEE, and Xiaofeng Zhou, 2013.

[2] Optimizing Search Engines using Clickthrough Data, Thorsten Joachims, Cornell University", Department of Computer Science, Ithaca, NY 14853 USA, 2002.

[3] Learning to Rank for Information Retrieval", Tie-Yan Liu, 2009.

[4] AdaRank: A Boosting Algorithm for Information Retrieval", Jun Xu, Microsoft Research Asia, China, Hang Li, Microsoft Research Asia, China, 2007.

[5] "Ranking Biomedical Passages for Relevance and Diversity", Andrew B. Goldberg , David Andrzejewski , Jurgen Van Gael, Burr Settles, Xiaojin Zhu Department of Computer Sciences, University of Wisconsin, Madison,2006.

[6] "A LDA-based approach to promoting ranking diversity for genomics information retrieval",Yan

Chen, Xiaoshi Yin, Zhoujun Li , Xiaohua Hu, Jimmy Xiangji Huang, 2011.

[7] "Learning to Rank for Hybrid Recommendation", Jiankai Sun, Shuaiqiang Wang, Byron J. Gao, Jun Ma,2012.

[8] Diversifying Search Results", Rakesh Agrawal, SreenivasGollapudi, Alan Halverson, Samuel Ieong, 2009.

[9] "LETOR: A Benchmark Collection for Research on Learning to Rank for Information Retrieval",Tao Qin, Tie-Yan Liu,Jun Xu, Hang Li, 2008.

[10] "Combining resources to find answers to biomedical questions", Dina Demner-Fushman,Susanne M. Humphrey,Nicholas C. Ide, 2007.

[11] "ListwiseAp-proach to Learning to Rank: Theory and Algorithm",F. Xia, T.Y.Liu, J.Wang, W.Zhang, and H.Li, 2008.

[12] Learning to Rank with Groups",Y. Lin, H.Lin, Z.Ye, S.Jin, and X.Sun, 2010.

[13] Learning to Rank with Nonsmooth Cost Functions",C.J. Burges, R.Ragno and Q.V. Le, 2006.