A Machine Learning Approach for Identifying Disease-Treatment Relations in Short Texts

Malla Rukmini Durga [1], V. V. Sivarama [2], M Srihari Varma [3]
[1] M.Tech, Department of Computer Science and Engineering, **S***.R.K.R Engineering College*
[2] Assistant Professor, Department of Computer Science and Engineering, **S***.R.K.R Engineering College*
[3] Assistant Professor, Department of Computer Science and Engineering, **S***.R.K.R Engineering College*

ABSTRACT

Medical information retrieval plays an increasingly important role to help physicians and domain experts to better access medical related knowledge and information, and support decision making. Integrating the medical knowledge bases has the potential to improve the information retrieval performance through incorporating medical domain knowledge for relevance assessment. However, this is not a trivial task because of the challenges to effectively utilize the domain knowledge in the medical knowledge bases. In this paper, we proposed a novel medical information retrieval system with a two-stage query expansion strategy, which is able to effectively model and incorporate the latent semantic associations to improve the performance. This system consists of two parts. First, we applied a heuristic approach to enhance the widely used pseudo relevance feedback method for more effective query expansion, through iteratively expanding the queries to boost the similarity score between queries and documents. Second, to improve the retrieval performance with structured knowledge bases, we presented a latent semantic relevance model based on tensor factorization to identify semantic association patterns under sparse settings. These identified patterns are then used as inference paths to trigger knowledge-based query expansion in medical information retrieval. Experimentsshowed that the performance of the proposed system is significantly better than the baseline system, and is comparable with state-of-the-art systems; (b) demonstrated the capability of tensor-based semantic enrichment methods for medical information retrieval tasks.

INTRODUCTION

WITH the exponential growth of medical related information, the retrieval of high quality results is becoming more critical. For medical applications like the Clinical Decision Support System (CDSS)[1], an effective and reliable information retrieval (IR) system is the basis to provide scientific evidences to support the clinical decision making, facilitate the translation of latest research outcomes into practice and improve the quality of health care[4]. The challenges are from (a) the inherent complexity of medical languages such as obscure medical terminologies and

ambiguous abbreviations, and (b) the associated variety of information needs from different types of users, such as patients and physicians. The complexity and ambiguity of medical languages result in vocabulary mismatch between queries and documents, which makes the conventional keyword-based IR methods often ineffective in medical IR tasks[2]. Semantic knowledge bases and concept mapping techniques are with the potential to solve this problem through annotating and analyzing the information with a common medical terminology and semantic relations including the classifications, term dependencies, hierarchies, etc.

Integrating the semantic relations is able to access and use the rich information of domain knowledge to enable accurate inferences for varying information needs[3]. Ontologies and associated semantic information are useful resources to extract structured knowledge for IR. For instance, a user submits a query to search for the information of fever treatment. There is a document introducing "aspirin", a medication used to treat fever. Human experts may judge that this document is relevant to the user's query. However, this relevance could not be automatically identified without the knowledge bases which contain the association in the form of a triple ("aspirin", "may_treat", "fever")[7].

Studies also showed that the proper incorporation of semantic information in knowledge bases is critical to the performance of medical IR . One of the most effective approaches is knowledge-based query expansion, which is a well-known method to bridge the gap between query terms and actual user information needs. With the expanded set of terms, there is a higher probability to identify and retrieve relevant documents that do not contain the terms in the original query. In addition, query expansion is flexible to be integrated with existing IR systems[6].

EXISTING SYSTEM

- Sfakianaki et al. proposed a natural language processing framework to automatically transform a clinical research question to a query that contains only terms of biomedical ontologies. Their research demonstrated the capability of biomedical ontologies and entity annotation algorithms to bridge the gap between clinical questions in natural language and biomedical literature.
- Mao et al. proposed a new medical IR system enhanced by manually assigned subject terms (Medical Subject Headings, MeSH). The proposed system constructs generative concept models to capture the associations between queries and documents Otegi et al. performed both query expansion and document expansion using a lexical database on IR tasks for question answering, and showed that their methods are complementary with pseudo-relevance feedback.

DISADVANTAGES OF EXISTING SYSTEM:

However, the performance of knowledge-based approaches was not satisfactory.

- The organizers pointed out that the poor performance of existing approaches could be resolved with available training data to tune the parameters.

PROPOSED SYSTEM:

- To develop the knowledge-based medical IR system, we incorporate the domain-specific information extracted from UMLS, a widely used knowledge base in medical domain. In the UMLS, synonymous terms are clustered into concept, and concepts are linked to other concepts in the semantic network.
- We developed a semantically enhanced medical IR system, which has a two-stage query expansion strategy (as shown in Figure 4) to integrate the pseudo relevance feedback and the knowledge-based query expansion to improve the performance of retrieving relevant documents for queries.
- First, we proposed the incremental pseudo relevance feedback (incremental PRF) approach for query expansion to obtain the initial ranking list of retrieved documents.
- Second, we developed an enhanced knowledge-based query expansion method with a novel latent semantic relevance model. The proposed method will re-rank the documents retrieved by the incremental PRF in the first stage.

ADVANTAGES OF PROPOSED SYSTEM:

- The proposed system has the potential to be adapted in other machine learning and medical informatics applications, like recommender systems, ontology learning, bioinformatics, etc.
- The superior performance of the proposed system showed the potential of incorporating knowledge bases using tensor factorization to enhance medical IR methods.

IMPLEMENTATION

MODULES

- **Admin**
- **User**

i) Incremental Pseudo Relevance Feedback (Incremental PRF)

ii) Latent Semantic Relevance Model

iii) Tensor based Semantic Association Representation

Admin:

- In this module, the Admin has to login by using valid user name and password. After login successful he can do some operations such as view all user and their details and authorize them. Admin can add medical records from dataset and update to database. Admin can view total no of users activated.

User:

- In this module, there are n numbers of users are present. User should register before doing some operations. After registration successful he has

to wait for admin to authorize him and after admin authorized him. He can login by using authorized user name and password. Login successful he will do some operations like search for medical records based on keywords and view search results.

Incremental Pseudo Relevance Feedback (Incremental PRF):

- This method is used for query expansion to obtain the initial ranking list of retrieved documents.We use incremental PRF to expand the query and obtain the top relevant documents based on the search platform.Relevance feedback methods are widely used to reformulate the original query using expansion features from the retrieved relevant documents. The popular pseudo-relevance feedback method assumes that top retrieved documents are relevant to the query so that the terms from these documents can be used for expansion.Using this algorithm, related keywords from potential related documents are exploited to improve the recall of the medical IR system. The incremental strategy is applied to evaluate the proper set of expansion terms iteratively and control the expansion process with a threshold to reduce the risk of query drift.

Latent Semantic Relevance Model

- After we performed the incremental PRF, the system will obtain a list of top documents related to the query. In this section, we introduce the knowledge-based query expansion with a latent semantic relevance model, which is able to provide the optimal expansion paths under sparse settings, and we use the expanded query to re-rank the documents to obtain the final results.

Tensor based Semantic Association Representation:

- To incorporate the UMLS semantic network for knowledge-based query expansion, the query terms are mapped to UMLS concepts first, and then the related concepts in the semantic network could be selected as expansion concepts. However, the semantic network covers a wide range of related concepts, some of which may be useless or even harmful for the retrieval of relevant documents. Intuitively, domain-specific semantic types.

SCREENS

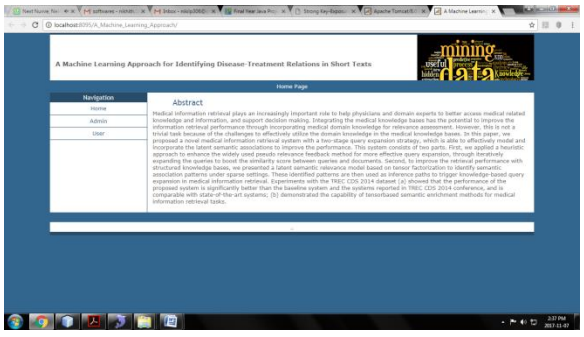

Fig: Home Page

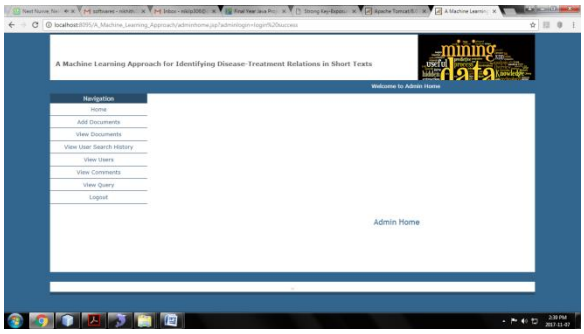

Fig: Admin Home

Fig: View Documents

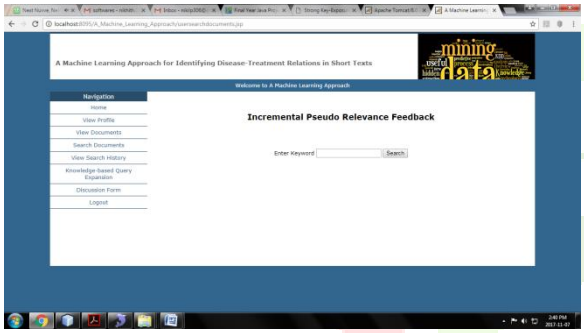

Fig: Search Documents

Fig: Aggrigate Calculation

CONCLUSION

The existing MSNs environment increasingly requiressituation awareness. Users' environment and behaviorare dynamic, and an individual's intention is also tochange. In order to adapt to the dynamic changes of useridentities in the social domain, this paper extends andenriches the Situ theory, and builds a *SocialSitu*framework for the social media networks. We design andachieve the intention serialization algorithm inmultimedia social networks. The user's frequent intentionsequence mode is obtained through the intentionserialization algorithm. When the user's identify changes,we conclude his behavior pattern with different ID, andprove that different *SocislSitu(t)* sequences are acquired inthe same *Min_Support*with the same intention when hisrole and group change. In the future works, the existingintention sequence patterns of the user could be adoptedto predict the user's more and deeper intentions. Besides,we will employ the *SocialSitu*and the proposed algorithmto improve multimedia recommendation system andsome killer applications in MSNs.

REFERENCES

[1]M. A. Musen, B. Middleton, and R. A. Greenes, "Clinical decision-support systems," in *Biomedical informatics*, ed: Springer, 2014, pp. 643-674.

[2]L. Goeuriot, L. Kelly, G. J. Jones, H. Müller, and J. Zobel, "Report on the SIGIR 2014 workshop on medical information retrieval (MedIR)," in *ACM SIGIR Forum*, 2014, pp. 78-82.

[3]G. Zuccon, B. Koopman, and P. Bruza, "Exploiting inference from semantic annotations for information retrieval: Reflections from medical ir," in *Proceedings of the 7th International Workshop on Exploiting Semantic Annotations in Information Retrieval*, 2014, pp. 43-45.

[4]C. Carpineto and G. Romano, "A survey of automatic query expansion in information retrieval," *ACM Computing Surveys (CSUR)*, vol. 44, p. 1, 2012.

[5]R. Socher, D. Chen, C. D. Manning, and A. Ng, "Reasoning with neural tensor networks for knowledge base completion," in *Advances in Neural Information Processing Systems*, 2013, pp. 926-934.

[6]W. Shen and J.-Y. Nie, "Is Concept Mapping Useful for Biomedical Information Retrieval?," in *International Conference of the Cross-Language Evaluation Forum for European Languages*, 2015, pp. 281-286.

[7]B. Ermiş, E. Acar, and A. T. Cemgil, "Link prediction in heterogeneous data via generalized coupled tensor factorization," *Data Mining and Knowledge Discovery*, vol. 29, pp. 203-236, 2015.

[8]E. Acar and B. Yener, "Unsupervised multiway data analysis: A literature survey," *IEEE transactions on knowledge and data engineering*, vol. 21, pp. 6-20, 2009.