Centrality based Document Ranking

A K Singh C Ravindranath Chowdary

Department of Computer Science and Engineering, IIT (BHU)
Varanasi 221005, India
{nlprnd@gmail.com,rchowdary.cse@iitbhu.ac.in}

Abstract. In this paper, we address the problem of ranking clinical documents using centrality based approach. We model the documents to be ranked as nodes in a graph and place edges between documents based on their similarity. Given a query, we compute similarity of the query with respect to every document in the graph. Based on these similarity values, documents are ranked for a given query. Initially, Lucene[1] is used to retrieve top fifty documents that are relevant to the query and then our proposed approach is applied on these retrieved documents to re-rank them. Experimental results show that our approach did not perform well as the documents retrieved by Lucene are not among the top 50 documents in the Gold Standard.

1 Introduction

A huge amount of information is available on the World Wide Web (WWW) and more is being added to it frequently, making the web a great source of information, but also impossible to navigate manually. Often, information related to a topic is available in multiple web sites. Sometimes information is spread over multiple web pages or documents. Current popular retrieval systems are not specialized for a particular domain of users. For example, having an IR system for academic purpose may have to address challenges specific to that domain. The same applies to clinical documents, which were the theme of one of the tracks in TREC-2014. They have terminology and ontology that is specific to the clinical domain and very uncommon elsewhere. A regular IR system may fail to rank documents from such a domain, dealing with symptoms, diagnosis and treatment etc., appropriately. As the results for our submission show, even if we apply a regular retrieval engine just for shortlisting documents and then apply a more sophisticated technique for re-ranking the shortlisted documents, we may still get very poor results because a regular IR system like Lucene may not even be able to shortlist documents for further ranking by a different method.

1.1 Background

Information retrieval has picked up its pace from the early twenty first century and has become one of the prime areas for both research and industry. The

[1] http://lucene.apache.org/core/

seminal paper by Page et al in 1998 [4, 17] ushered academicians into IR. The notions of authority and hub pages[13] were used to build systems initially. Keyword based search taken as bag of words dominated for quite a long time [1]. Semantic search [7, 14, 15] also has made significant progress. Even though the most popular IR systems are domain-independent, considerable work has been done on domain specific IR systems [8, 12, 3]. More directly related to our work, there has been sustained interest in the area of medical or clinical information retrieval [10] and on clinical decision support systems [2, 6, 11]. Several open source IR systems are also available to researchers such as Lucene [9].

1.2 Model

In this section we discuss the proposed model for generating ranking of the documents. Documents are modeled as a weighted graph. Each document is considered as a node and an edge is present between any two nodes if the *similarity* between them is above a threshold. *Similarity* is calculated as given in Equation 1.

$$sim(\overrightarrow{n_i}, \overrightarrow{n_j}) = \frac{\overrightarrow{n_i}.\overrightarrow{n_j}}{|\overrightarrow{n_i}||\overrightarrow{n_j}|} \tag{1}$$

where $\overrightarrow{n_i}$ and $\overrightarrow{n_j}$ are term vectors for the nodes n_i and n_j respectively. The weight of each term (t_i) in $\overrightarrow{n_i}$ is calculated as $tf * isf$. Term frequency (tf) of t_i is defined as the number of occurrences of t_i in n_i and inverse sentential frequency (isf) of t_i is defined as the logarithm of the total number of nodes in the document divided by the number of nodes in which t_i is present. A term that distinguishes a node from other nodes has a higher isf value when compared to a term that occurs in many nodes.

1.3 Description of the Proposed Model

All the documents are first preprocessed as usual, i.e., all non-text/noise should be filtered. After removing stop words from the filtered documents, the remaining words are stemmed before calculating *tf*, *isf* and *similarity*. Similarities between all pairs of nodes is calculated. A low similarity between two nodes indicates that the two nodes are not related and a high similarity value indicates a strong relation between them.

We use a methodology that is similar to the one proposed in [16], for calculating the node score with respect to a query term. While calculating the node score, both the similarity (relevance) of a node to the query term and the neighbourhood (nodes that have edge scores above a threshold) of it are considered. Initially, each node is assigned a query similarity score and then these scores are propagated to their neighbours as given in Equation 2. This process is iterated till the scores of the nodes converge (weights of all the nodes in successive iterations fall below a threshold(0.0001)). A node score for each node with respect to each query term $q_i \in Q$ where $Q = \{q_1, q_2, ..., q_t\}$ is computed using Equation 2.

$$w_{q_i}(s) = d\frac{sim(s,q_i)}{\sum_{m\in N} sim(m,q_i)} + (1-d)\sum_{v\in adj(s)} \frac{sim(s,v)}{\sum_{u\in adj(v)} sim(u,v)} w_{q_i}(v) \quad (2)$$

where $w_{q_i}(s)$ is the node score of node s with respect to the query term q_i, d is the bias factor, N is the set of all nodes in the document and $sim(i,j)$ is computed as given in Equation 1. First part of Equation 2 computes relevance of s to the query term and the second part computes the relevance of s with its neighbours. Also, the second part captures the amount of relevant information that neighbours of s have with respect to the query term. The bias factor d gives trade-off between these two parts and it is determined empirically. For higher values of d, more importance is given to the relevance of s with respect to the query term when compared to the relevance between s and its neighbours. The denominators are for normalization. This method was proposed in [18, 5] for text summarization.

2 Discussion

We used the above approach to rank the documents as part of the shared task. Due to shortage of computational resources, before applying the above method, we used a two-step process to shortlist documents. In the first step, Lucene was used to return the top 1000 documents. In the second step, Lucene was again used to select only the top 50 relevant documents for a given query. These 50 shortlisted documents were then re-ranked using the method described in the preceding section.

We did use a naive query expansion heuristic for both the runs (summary and description). We prepared a hand-crafted list of synonyms for each of the query types, viz. diagnosis, test and treatment. This list was used to expand the queries given to Lucene.

Clearly, the results in this setup were going to be heavily dependent on the performance of Lucene for the clinical decision support task. As it turned out, the results were very poor, which suggests that using a general purpose IR sytem in this way is not a good idea.

References

1. R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1999.
2. E. S. Berner. *Clinical Decision Support Systems: Theory and Practice*. Springer, New York, 1998.
3. J. Bing. *Handbook of Legal Information Retrieval*. Elsevier Science Inc., New York, NY, USA, 1984.

4. S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *Seventh International World-Wide Web Conference (WWW 1998)*, 1998.
5. C. R. Chowdary and P. S. Kumar. ESUM: An efficient system for query-specific multi-document summarization. In *ECIR '09: Proceedings of the 31th European Conference on IR Research*, Lecture Notes in Computer Science, pages 724–728, Toulouse, France, April 2009. Springer.
6. E. Coiera. *Clinical decision support systems*, pages 331–344. Hodder Arnold Publishers, 2003.
7. R. Guha, R. Mccool, and E. Miller. Semantic search. In *INTERNATIONAL WORLD WIDE WEB CONFERENCE*, pages 700–709. ACM, 2003.
8. A. Hanbury and M. Lupu. Toward a Model of Domain-Specific Search. In *Proc. of OAIR*, 2013.
9. E. Hatcher and O. Gospodnetic. *Lucene in Action*. Manning Publications, 2004.
10. W. Hersh. *Information Retrieval: A Health and Biomedical Perspective: A Health and Biomedical Perspective*. Health Informatics. Springer, 2008.
11. J. Holstiege, T. Mathes, and D. Pieper. Effects of computer-aided clinical decision support systems in improving antibiotic prescribing by primary care providers: a systematic review. *Journal of the American Medical Informatics Association*, 22(1):236–242, 2015.
12. C. B. Jones and R. Purves, editors. *Proceedings of the 7th Workshop on Geographic Information Retrieval, GIR 2013, 5th November, 2013, Orlando, Florida, USA*. ACM, 2013.
13. J. M. Kleinberg. Hubs, authorities, and communities. *ACM Computing Surveys*, 31(4es), 1999.
14. E. Makela. Survey of semantic search research. In *Proceedings of the Seminar on Knowledge Management on the Semantic Web*, 2005.
15. P. Mika, T. Tudorache, A. Bernstein, C. Welty, C. A. Knoblock, D. Vrandecic, P. T. Groth, N. F. Noy, K. Janowicz, and C. A. Goble, editors. *The Semantic Web - ISWC 2014 - 13th International Semantic Web Conference, Riva del Garda, Italy, October 19-23, 2014. Proceedings, Part I*, volume 8796 of *Lecture Notes in Computer Science*. Springer, 2014.
16. J. Otterbacher, G. Erkan, and D. R. Radev. Using random walks for question-focused sentence retrieval. In *HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 915–922. Association for Computational Linguistics, 2005.
17. L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, November 1999.
18. M. Sravanthi, C. R. Chowdary, and P. S. Kumar. QueSTS: A query specific text summarization system. In *Proceedings of the 21st International FLAIRS Conference*, pages 219–224, Florida, USA, may 2008. AAAI Press.