Cannabis_*TREATS_*cancer: Incorporating Fine-Grained Ontological Relations in Medical Document Ranking

Yunqing Xia[1], Zhongda Xie[1], Qiuge Zhang[2], Huiyuan Wang[2], and Huan Zhao[3]

[1] Department of Computer Science, TNList,
Tsinghua University, Beijing 100084, China
{yqxia,xzd13}@tsinghua.edu.cn
[2] Information Networking Institute
Carnegie Mellon University, Pittsburgh, PA 15213, USA
{zqg0830,wanghuiyuan.anna}@gmail.com
[3] Department of Computer Science and Engineering
Hong Kong University of Science and Technology, Hong Kong
hzhaoaf@ust.hk

Abstract. The previous work has justified the assumption that document ranking can be improved by further considering the coarse-grained relations in various linguistic levels (e.g., lexical, syntactical and semantic). To the best of our knowledge, little work is reported to incorporate the fine-grained ontological relations (e.g., <*cannabis*, TREATS, *cancer*>) in document ranking. Two contributions are worth noting in this work. First, three major combination models (i.e., summation, multiplication, and amplification) are designed to re-calculate the query-document relevance score considering both the term-level Okapi BM25 relevance score and the relation-level relevance score. Second, a vector-based scoring algorithm is proposed to calculate the relation-level relevance score. A few experiments on medical document ranking with CLEF2013 eHealth Lab medical information retrieval dataset show that the proposed document ranking algorithms can be further improved by incorporating the fine-grained ontological relations.

Keywords: Medical document ranking, ontological relation, medical concept, relevance.

1 Introduction

Nowadays, powerful search engines are available and people may consult the search engines with queries like *cannabis and cancer*. The underlying information need is actually the connection between the two things rather than information of the two things. Considering the following three sentences:

(S#1): He suffers from *cancer* but he never quits *cannabis*.
(S#2): Studies prove that *cannabis* can be an effective treatment for

C. Zong et al. (Eds.): NLPCC 2014, CCIS 496, pp. 275–285, 2014.

cancer.
(S#3): The report indicates that long-term *cannabis* use may cause lung *cancer*.

Now, let's use *cannabis and cancer* as the query. Considering merely terms, we would find that the three sentences are equally relevant to the query. However, sentence S#1 is not truly relevant because no relation occurs between term *cannabis* and *cancer* though both terms are mentioned. Such a mistake occurs in medical information retrieval systems because ontological relation is not considered in document ranking. In our research in medical information retrieval, we find more than 20 percent queries usually involve fine-grained ontological relations, e.g., <*drug*,TREATS,*disease*>.

The previous work has justified the assumption that relations of various linguistic levels are helpful to improve document ranking [1–9]. A majority of research is conducted on statistical term dependency. Other work is conducted on syntactic dependency and semantic relation. Undoubtedly, the above coarse-grained relations are useful, but the discovered relations are lack of meaning. For example, <*thingX*,ISA,*thingY*> indicates a general hypernymous relation, in which *thingX* and *thingY* can be any things.

In medical domain, an early work is reported in [3] which made use of fine-grained ontological relations between medical concepts in cross-language medical information retrieval. Enlightened by the positive results, we conduct a further study which handles the fine-grained medical documents and applies them to enhance medical document ranking.

In this work, we design three combination models (i.e., summation, multiplication and amplification) to calculate the new query-document relevance score, which combines the term-level relevance score and the relation-level relevance score. We calculate the term-level relevance score with the standard Okapi BM25 algorithm[1] in Lucene[2]. For the relation-level relevance, we design a vector-based scoring algorithm which first represents query and documents with eighteen-dimension vectors, and then calculates the relevance score using cosine formula. A few experiments are conducted in medical document ranking task with dataset from CLEF2013 eHealth Lab on medical information retrieval, which show that the proposed document ranking algorithms can be further improved by incorporating the fine-grained ontological relations.

The reminder of this paper is organized as follows. In Section 2, we summarize the related work. In Section 3, we present our document ranking method. We present evaluation as well as discussion in Section 4 and conclude this paper in Section 5.

2 Related Work

The early attempts to incorporate relations in textual information retrieval (IR) started are based on concepts or semantics. A concept-based solution is

[1] `http://en.wikipedia.org/wiki/Okapi_BM25`
[2] Lucene: `http://lucene.apache.org/`

discussed in [10], in which term dependencies are first taken into account. Later on, lexical-semantic relations are used in [11] to build a structured representation of documents and queries. Due to shortage of large-scale semantic knowledge resource, the semantics based approach improves IR system slightly. Instead, Khoo et al. (2001) focused on merely cause-effect relations [12]. In [13], a relation-based search engine, i.e. OntoLook, constructs a concept-relation graph to model semantic relationships amongst concepts. In [14], semantic relationships are revisited using ontology. We notice that the semantic relations used in these IR systems are coarse-grained. That is, most relations formalize general connections of things but have little to do with real meaning.

In late 1990's, researchers started to study effect of syntactic term dependency relation on information retrieval. Syntactic term dependency was first used in [2] to improve Japanese information retrieval. However, the syntactic parsing tools at that moment were slow and less accurate. Recently, Park et al. (2011) proposed a quasi-synchronous dependence model based on syntactic dependency parsing for both queries and documents [7].

In the meantime, a majority of research is conducted to incorporate statistical term dependency in IR system. A general language model was proposed in [1] which presents word dependency with bi-grams. Gao et al. (2004) proposed a dependency language model in ranking documents based on statistical term dependency (i.e., linkage) [4]. This model is later revised in [6] with more general term dependency in syntactic and semantic levels. Very recently, statistical high-order word association relation was exploited by [8] in document ranking using pure high-order dependence among a number of words. Term association was further studied in [9] for probabilistic IR by introducing a new concept cross term in modeling term proximity.

It should be noted that relations have been also used in question answering. In [15], a general rank-learning framework was proposed for passage ranking within question answering systems using linguistic and semantic features. Semantic relations were also discussed in [5] to improve question analysis and answer retrieval.

Little research work is conducted to incorporate relations in medical IR since Vintar et al. (2003) achieved positive results with the fine-grained ontological relations [3], in which the relations are used to filter cross-lingual web pages in a boolean manner. In the past five years, TREC medical track [16] and CLEF eHealth Lab [17] were organized to advance the research on medical IR. However,no medical IR system uses fine-grained ontological relations in document ranking. Start from Vintar et al. (2003)[3], we design a unified ranking algorithm which combines the traditional BM25 relevance score and the proposed relation-level relevance score. The difference lies in that the relevance score is re-calculated in our work.

3 Methodology

The core of this work is assigning each document a refined relevance score that reflects relevance of a document to the query considering not only terms but also

fine-grained ontological relations. This goal is achieved in three steps. First, we calculate the term-level relevance score using BM25. Second, we calculate the relation-level relevance score using our method. At last, we adopt the following three popular combination models to calculate the refined relevance score ($r*$) by combing the term-level relevance score (r) and the proposed relation-level relevance score (l):

– Summation

$$r^* = \alpha \times r + (1 - \alpha) \times l \tag{1}$$

in which α is the normalization factor.

– Multiplication

$$r^* = r \times l \tag{2}$$

– Amplification

$$r^* = r \times \beta^l \tag{3}$$

in which β is the exponential base. We set $\beta = e$ in our study according to empirical study.

This section focuses on calculation of the relation-level relevance score, i.e. lp, which is achieved as follows: First, ontological relations are discovered within text of query and documents; Second, query and documents are represented with relation vector; Third, relational relevance score is calculated by comparing the vectors; In what follows, we elaborate the key modules in our document ranking method.

3.1 Ontological Relation Discovery

In Wikipedia, *ontology* is defined as *the nature of being, becoming, existence, or reality, as well as the basic categories of being and their relations*[3]. According to this definition, we further define the ontological relations as follows.

Definition: *Ontological relation*

An ontological relation is defined as the real-world relation between existential beings (things or events).

Compared with the general semantics relations such as synonym and polysemy, the ontological relations reflect fine-grained real-world semantic relationship such as *person_PRESIDENT_OF_nation* and *medicine_CURES_disease*.

For the three example sentences in Section 1, the corresponding ontological relations are given below:

[3] `http://en.wikipedia.org/wiki/Ontology`

(R#1): NULL
(R#2): *cannabis*_TREATS_*cancer*
(R#3): *cannabis*_CAUSES_*cancer*

Compiling the ontological relations is a tricky job, even for the specific medical domain. Fortunately, 57 types of ontological relations are defined in SemMedDB [18]. However, some relations are either overlapping with other relations or less important according to statistics in SemMedDB. To simplify the problem, we employ three medical experts to handle relations manually. Finally, an agreement on the following eighteen relations is reached: PROCESS_OF, METHOD_OF, LOCATION_OF, PART_OF, OCCURS_IN, STIMULATES, MANIFESTATION_OF, CONVERT_TO, AUGMENTS, ASSOCIATED_WITH, PREVENTS, USES, TREATS, PREDISPOSES, PRODUCES, DISRUPTS, CAUSES and INHIBITS.

To be formal, the ontological relation is represented by a three-tuple: $< C\#1, r, C\#2 >$, where $C\#1$ and $C\#2$ represent two medical concepts and r a relation. The medical concepts are obtained with MetaMap [4].

In this work, ontological relation is discovered based on keywords which are mentioned in SemMedDB annotations of predicate instances. To reduce complexity, we only use the high-frequency ones (i.e., 8,015 unigram keywords and 114,839 bigram keywords). We find some keywords indicate different relation in different context. Thus the discovered ontological relations can be modeling in a probabilistic manner with a priori distribution within texts. We thus extend the above four-tuple to include relation probabilities to 20-tuple: $< C\#1, \{r_1; p_1\}, ..., \{r_{18}; p_{18}\}, C\#2 >$, where r_i represents the i-th relation and p_i its probability, which is estimated in SemMedDB using the simple MLE (maximum likelihood estimation) technique.

3.2 Representation of Query and Document Using Ontological Relations

Considering an 18-dimension vector that entails the aforementioned 18 relations, we now create a relation vector $V = \{r_1 : w_1, ..., r_{18} : w_{18}\}$ for a piece of medical text. We use relation keywords mentioned in Section 3.1 in detecting ontological relations in text. We map query and document to relation vectors with different approaches.

(1) Query
Query is usually too short to indicate a deterministic medical relation, especially when no keyword is mentioned. We choose to assign equal probability to each possible relation. We first consult the UMLS [5] and extract all possible relations that may occur between concepts via the keyword (if any) in the query. Then

[4] MetaMap is a medical concept annotation tool available at `http://mmtx.nlm.nih.gov/`
[5] UMLS is a medical knowledge base that can be downloaded via `http://www.nlm.nih.gov/research/umls/`.

equal probability is assigned to the relations and we obtain a relation vector for the query. For the query *cannabis and cancer*, we obtain a relation vector below:

(RV#0) (0,0,0,0,0.5,0,0,0,0.5,0,0,0,0,0,0,0,0,0)

where the 5-th relation is TREATS and 9-th relation is CAUSES.

(2) Document
Document may mention the query for more than one times. Thus we need to resolve medical relation for each mention. For presentation convenience, we need first to fix the window for relation detection. Here, we use sentence as an example window in the following description. We map each query-mentioning sentence to a relation vector.

Difference between sentence and query lies in that sentence gives a much larger context thus can indicate a deterministic relation of a keyword. Thus the relation vector for a sentence contains only one non-zero value dimension. For the three example sentences in Section 1, we obtain three sentence-level relation vectors (RV) below:

(RV#1) (0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0)
(RV#2) (0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0)
(RV#3) (0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0)

After relation vector for each window is obtained, we merge all these vectors thus obtain an overall vector for the document.

3.3 Calculating the Relation-Level Relevance Score

Once query and documents are represented with 18-dimension vectors, we are able to adopt vector-based distance measures in relation scoring. Given two vectors $v_i = \{w_{i1}, w_{i2}, ..., w_{i18}\}$ and $v_j = \{w_{j1}, w_{j2}, ..., w_{j18}\}$, we adopt the Cosine distance measure in distance calculation:

$$Cos(v_i, v_j) = \frac{\sum_{k=1}^{18} w_{ik} w_{jk}}{\sqrt{\sum_{k=1}^{18} w_{ik}^2 + \sum_{k=1}^{18} w_{jk}^2}} \tag{4}$$

4 Evaluation

Data
We used the dataset in CLEF2013 eHealth Lab Medical IR task [17] in our experiments, which covers a broad range of health topics, targeted at both the general public and healthcare professionals.

The test queries are extracted from the 50 queries in CLEF2013 Medical IR task. As we intend to prove contribution of medical relations to medical document ranking, we select queries that involves more than one medical concepts.

Finally, we obtain fourteen queries: #6, #7, #8, #11, #12, #16, #17, #18, #23, #24, #25, #39, #40 and #49.

The CLEF2013 Medical IR dataset, denoted with CLEF, contains 1,878 relevant documents judged by nurses from the pool of 6,391 documents. As NA (non-annotation) documents are retrieved by our method, we employed three medical students to assess relevance of these documents. We calculate Kappa coefficient value between every two assessors and obtain the average Kappa coefficient value 0.82. In this way, we obtained an extended dataset, denoted with CLEF+.

Evaluation Metrics
Two metrics are used in our evaluation: (1) p@10: precision at top 10 web pages. (2) nDCG@10: normalized Discounted Cumulative Gain at the top 10 returned web pages. (3) MAP: Mean average precision at top 10 returned web pages.

4.1 Experiment 1: Methods

Three methods are compared in this evaluation, three of which are different implementations of our method:

- **BM25**: Okapi BM25 is used in relevance scoring. Default settings are adopted in the BM25 algorithm.
- **BMB**: The method described in [3] is implemented, in which the discovered ontological relations are used to filter the web pages in a *boolean* manner.
- **BMR**: Our method is implemented to combine BM25 term-level relevance score and relation-level relevance score. In this experiment, we adopt the the amplification formula as the combination model (Eq.3 and HTML tag pair as relation detection window. Such a setting is proved most effective in our experiments.

Experimental results are given in Table 1.

Table 1. Results of the document ranking methods

Method	p@10		nDCG@10		MAP@10	
	CLEF	CLEF+	CLEF	CLEF+	CLEF	CLEF+
BM25	0.450	0.516	0.448	0.504	0.112	0.129
BMB	0.437	0.521	0.435	0.514	0.106	0.117
BMR	0.456	0.534	0.452	0.519	0.124	0.144

It can be seen in Table 1 that when ontological relations are incorporated, the BMB method performs slightly worse than the BM25 baseline. This indicates that the boolean combination method does not bring performance gain. As a comparison, the proposed BMR method outperforms BM25 by 0.018 on p@10, by 0.015 on MAP@10 and by 0.015 on nDCG@10. Looking into the fourteen

queries which involve ontological relations, we find our proposed method obtains improvement on 9 queries while loss on 2 queries. The major reason for the loss is that some discovered relations in web pages are incorrect. This reminds us to plan future work on a better relation detection method.

Note that on CLEF+ dataset, BMR method improves more than that on CLEF dataset. This is because a few out-of-pool web pages are judged relevant by annotators in our work.

4.2 Experiment 2: Combination Models

In this experiment, we seek to compare the combination models described in Section 3. Accordingly, the following implementations of our method is evaluated:

- **SUMM**: Our method is implemented to adopt Eq.1 in combining the BM25 term-level relevance score and relation-level relevance score. We set $\alpha = 0.7$ in this implementation according to empirical study.
- **MULT**: We adopt Eq.2 in this implementation of our Our method.
- **AMPL**: We adopt Eq.3 in this implementation of our Our method.

Note in all the three BMR implementations, the relation detection window is set HTML tag pair, which is proved effective in our experiments. Results are presented in Table 2.

Table 2. Results of the our methods using different combination models

Method	p@10		nDCG@10		MAP@10	
	CLEF	CLEF+	CLEF	CLEF+	CLEF	CLEF+
SUMM	0.451	0.527	0.447	0.511	0.121	0.140
MULT	0.447	0.521	0.443	0.501	0.117	0.139
AMPL	0.456	0.534	0.452	0.519	0.124	0.144

We can see in Table 2 that the AMPL implementation performs best on both datasets across the three evaluation metrics. This justifies that the amplification model is advantageous over the two models. Looking into the output results, we find the MULT implementation improves BM25 method on 7 queries while loss on 3 queries.

4.3 Experiment 3: Relation Detection Window

This experiment aims to compare different relation detection windows in the proposed BMR method. The following six implementations are developed:

- **CURS**: The current sentence is used as relation detection window.
- **CURSP**: The current and the preceding sentence are used as relation detection window.

- **CURSPF**: The current and the preceding and following sentences are used as relation detection window.
- **CURP**: The current paragraph is used as relation detection window.
- **CURD**: The current web document is used as relation detection window.
- **HTML**: Text in the current HTML tag pair (e.g., TD and UL) is used as relation detection window.

Table 3 presents the experimental results of our method with the six different windows.

Table 3. Results of the our method with different relation detection window

Method	p@10		nDCG@10		MAP@10	
	CLEF	CLEF+	CLEF	CLEF+	CLEF	CLEF+
CURS	0.450	0.522	0.445	0.511	0.110	0.131
CURSP	0.451	0.524	0.448	0.513	0.111	0.134
CURSPF	0.453	0.528	0.449	0.516	0.112	0.137
CURP	0.442	0.476	0.431	0.502	0.106	0.127
CURD	0.427	0.458	0.416	0.489	0.097	0.112
HTML	0.456	0.534	0.452	0.519	0.124	0.143

Seen in Table 3 that the best window for relation detection is HTML. When the window is extended to the whole document, our method becomes worse than the BM25 baseline. This is because more errors occur in detecting relations in the whole document. Using paragraph as window also brings some errors. On the other hand, when the window is reduced to current sentence, quality of our method drops most. This is because in a sentence, many relations cannot be detected. However, the elements of these ontological relation can be found in the preceding or following sentences. This is why CURSPF outperforms CURSP and CURS. Meanwhile, a bigger context may bring noise. Thus the appropriate window is found HTML tag. This ascribes the writing style in HTML web pages.

5 Conclusion and Future Work

In this work, we propose a novel medical document ranking method which incorporates the fine-grained ontological relations in relevance scoring. The relation-level relevance score is measured by comparing relation vectors for query and documents. In our experiments, we evaluate not only the outperformance of our method over the state-of-the-art baseline methods, but also the influence of combination model and relation detection window on our method. Experimental results confirm that the ontological relations indeed bring performance gain in medical document ranking.

However, this work is still preliminary. For example, the eighteen types of ontological medical relations are compiled by human experts. We will explore the possibility to extend these relations to cover all possible medical relations.

The future work includes a finer algorithm for medical relation detection, a probabilistic model for relation-level relevance scoring and further attempts in applying ontological relations in general domain information retrieval. Meanwhile, some substantial evaluation is planned to compare more baselines and more parameters.

Acknowledgement. This work is supported by National Science Foundation of China (NSFC: 61272233). We thank the anonymous reviewers for the valuable comments.

References

1. Song, F., Croft, W.B.: A general language model for information retrieval. In: Proc. of CIKM 1999, pp. 316–321. ACM, New York (1999)
2. Matsumura, A., Takasu, A.: Adachi: The effect of information retrieval method using dependency relationship between words. In: Proceedings of RIAO 2000, pp. 1043–1058 (2000)
3. Vintar, S., Buitelaar, P., Volk, M.: Semantic relations in concept-based cross-language medical information retrieval. In: Proceedings of ECML/PKDD workshop on Adaptive Text Eextraction and Mining (ATEM) (2003)
4. Gao, J., Nie, J.Y., Wu, G., Cao, G.: Dependence language model for information retrieval. In: Proc. of SIGIR 2004, pp. 170–177. ACM, New York (2004)
5. Morton, T.: Using semantic relations to improve information retrieval. PhD thesis, University of Pennsylvania (2004)
6. Maisonnasse, L., Gaussier, E., Chevallet, J.P.: Revisiting the dependence language model for information retrieval. In: Proc. of SIGIR 2007, pp. 695–696. ACM, New York (2007)
7. Park, J.H., Croft, W.B., Smith, D.A.: A quasi-synchronous dependence model for information retrieval. In: Proc. of CIKM 2011, pp. 17–26. ACM, New York (2011)
8. Hou, Y., Zhao, X., Song, D., Li, W.: Mining pure high-order word associations via information geometry for information retrieval. ACM Trans. Inf. Syst. 31(3), 12:1–12:32 (2013)
9. Zhao, J., Huang, J.X., Ye, Z.: Modeling term associations for probabilistic information retrieval. ACM Trans. Inf. Syst. 32(2), 7:1–7:47 (2014)
10. Giger, H.P.: Concept based retrieval in classical ir systems. In: Proc. of SIGIR 1988, pp. 275–289. ACM, New York (1988)
11. Lu, X.: Document retrieval: A structural approach. Inf. Process. Manage. 26(2), 209–218 (1990)
12. Khoo, C.S.G., Myaeng, S.H., Oddy, R.N.: Using cause-effect relations in text to improve information retrieval precision. Inf. Process. Manage. 37(1), 119–145 (2001)
13. Li, Y., Wang, Y., Huang, X.: A relation-based search engine in semantic web. IEEE Trans. on Knowl. and Data Eng. 19(2), 273–282 (2007)
14. Lee, J., Min, J.K., Oh, A., Chung, C.W.: Effective ranking and search techniques for web resources considering semantic relationships. Inf. Process. Manage. 50(1), 132–155 (2014)
15. Bilotti, M.W., Elsas, J., Carbonell, J., Nyberg, E.: Rank learning for factoid question answering with linguistic and semantic constraints. In: Proc. of CIKM 2010, pp. 459–468. ACM, New York (2010)

16. Voorhees, E.M., Hersh, W.: Overview of the trec 2012 medical records track. In: Proc. of TREC 2012 (2012)
17. Goeuriot, L., Jones, G.J.F., Kelly, L., Leveling, J., Hanbury, A., Müller, H., Salanterä, S., Suominen, H., Zuccon, G.: Share/clef ehealth evaluation lab 2013, task 3: Information retrieval to address patients' questions when reading clinical reports. In: CLEF Online Working Notes (2013)
18. Kilicoglu, H., Shin, D., Fiszman, M., Rosemblat, G., Rindflesch, T.C.: Semmeddb: a pubmed-scale repository of biomedical semantic predications. Bioinformatics 28(23), 3158–3160 (2012)