Journal of Medical Systems (2018) 42: 27
https://doi.org/10.1007/s10916-017-0876-3

SYSTEMS-LEVEL QUALITY IMPROVEMENT

Automatic Decision Support for Clinical Diagnostic Literature Using Link Analysis in a Weighted Keyword Network

Shuqing Li[1] · Ying Sun[2] · Dagobert Soergel[2]

Received: 10 January 2016 / Accepted: 13 December 2017 / Published online: 23 December 2017

Abstract
We present a novel approach to recommending articles from the medical literature that support clinical diagnostic decision-making, giving detailed descriptions of the associated ideas and principles. The specific goal is to retrieve biomedical articles that help answer questions of a specified type about a particular case. Based on the filtered keywords, MeSH(Medical Subject Headings) lexicon and the automatically extracted acronyms, the relationship between keywords and articles was built. The paper gives a detailed description of the process of by which keywords were measured and relevant articles identified based on link analysis in a weighted keywords network. Some important challenges identified in this study include the extraction of diagnosis-related keywords and a collection of valid sentences based on the keyword co-occurrence analysis and existing descriptions of symptoms. All data were taken from medical articles provided in the TREC (Text Retrieval Conference) clinical decision support track 2015. Ten standard topics and one demonstration topic were tested. In each case, a maximum of five articles with the highest relevance were returned. The total user satisfaction of 3.98 was 33% higher than average. The results also suggested that the smaller the number of results, the higher the average satisfaction. However, a few shortcomings were also revealed since medical literature recommendation for clinical diagnostic decision support is so complex a topic that it cannot be fully addressed through the semantic information carried solely by keywords in existing descriptions of symptoms. Nevertheless, the fact that these articles are actually relevant will no doubt inspire future research.

Keywords Literature recommendation service · Clinical decision support · Link analysis · Keyword co-occurrence analysis

Introduction

In recent years, retrieval of the most relevant research articles from very large collections has become a challenging task for information services. The topic has also been attracting increasing interests within the research community, which requires effective means to reduce ambiguity in searches and return results that more precisely meet user requirements. The evaluation of this research has also been included by TREC clinical decision support track. For example, the TREC clinical decision support track 2015 and 2014 focussed on the retrieval of biomedical articles that could help answer generic questions of a specified type regarding specific case reports [1]. However, it has become apparent that these difficulties are more severe than that in other research fields and numerous new problems need to be addressed. Many existing researches still have many limitations and more improvements need to be explored. This is also the reason why 2017 TREC PM Task and 2016 TREC CDS Task were still held.

Clinical decision support involves three tasks: diagnosis, testing and treatment. Accurate test selection and suggestions for treatment are both fully reliant on effective and accurate diagnosis. It formed the focus of the current study, which also extended our previous research works. From the point of view

This article is part of the Topical Collection on *Systems-Level Quality Improvement*

✉ Shuqing Li
leeshuqing@gmail.com

Ying Sun
sun3@buffalo.edu

Dagobert Soergel
dsoergel@buffalo.edu

[1] College of Information Engineering, Nanjing University of Finance & Economics, Nanjing 210023, China

[2] Department of Library and Information Studies, Graduate School of Education, University at Buffalo, New York State University, Buffalo, NY 14260, USA

of methodology, the retrieval tasks undertaken in clinical diagnosis are more complex than those for which common retrieval systems are typically used. One problem arises simply from the terminology used, which is the basic element of information retrieval. In fact, medical science is a research area with a terminology that includes several specialised medical terms, acronyms and items from the Greek alphabet. Beside this, it is not enough to only make use of a few simple terms to classify the symptoms noted in a diagnosis because the same symptom may indicate several different diseases when combined with other symptoms [2]. Finnally, the results for diagnosis are not single links or some texts, but articles that provide detailed information on the appropriate treatment, given the description of the symptoms. Therefore, the full relationship of sentences and segments in articles should be explored and mined for the purpose of recognizing interesting and valuable knowledge.

These considerations contribute the basis of our research. The main objective of our research was to more accurately and completely mine the relationships between medical terms, and provide an efficient and effective approach to identify relevant biomedical articles based on the analysis of relationship of sentences and segments. Our study is based on two important assumptions. First is that well-formed keywords and their relations are essential for analysis and retrieval. Filtered keywords, MeSH and automatically extracted acronyms are reliable sources of terms that can be used in our research. The second assumption is that effective analysis should have understandable process and meaningful ranking order. In other words, our basic assumption is that an article containing these valid segments is more likely to discuss relevant diseases or contain other useful information. Whether segment is valid fully depends on the occurrence of key descriptor of symptom because most of description of patient symptom is so simple and short. The whole process of our work was built on these assumptions.

This study is organised as follows. In Section 2, we discuss related studies and set out the motivation for the research. In Section 3, the principles and approaches that underpin the collection of standard keywords are described. In Section 4, we introduce the process for extracting the relations between keywords and documents based on keyword normalisation. In Section 5, we discuss our approach to generate automatic recommendations from the clinical diagnosis literature based on the extraction of relevant sentences and existing diagnosis-related keywords. In Section 6, the evaluation and results of the experimental investigations are discussed. In the final section, we present our conclusions and plans for future research.

Related studies

Effective clinical diagnosis has three indispensable requirements: accurate data, adequate knowledge and access to recommendations from the medical literature [3]. Early systems were able to conduct differential diagnoses and suggest further information that could improve the precision of the judgement. These included Dxplain [4, 5] and QMR [6]. Other systems can summarise patient records and present physicians with the results in an easy-to-understand format [7].

However, the success of these systems depends mainly on a highly standardised writing format in the patient health records [8]. In the meantime, Jaspers et al. suggested that existing clinical diagnosis decision support systems provided insufficient evidence of patient outcomes [9]. Other scholars have pointed out that the effectiveness of clinical diagnosis decision support systems is greatly dependent on their acceptance by physicians [10]. Wright et al. recommended that more attention be given to the roles which management and governance play in the application and development of technology for clinical diagnosis decision support [11].

As an uncommon view, it has been argued that rather than being used more broadly, intelligent diagnosis decision support systems should only focus on drug selection and prescription support [12]. Against a range of quality indicators, Romano and Stafford reported finding no significant difference between visits in which clinical decision support systems were and were not used [13]. However, they recognised that the support provided by these systems can improve the quality of diet counselling offered to high risk adults.Moreover, in recent years, the greatest progress has been made in the application of diagnostic support to specific diseases, many of which have standardised diagnostic and therapeutic criteria, making them easier to automate. For example, Hoeksema et al. identified clinical decision support as a promising approach for improving guideline-based care for treating paediatric asthma [14]. Such a system can provide assessments of impairment, risk, control and severity, and can generate treatment recommendations for new patients, based on the guidelines of the United States National Asthma Education and Prevention Programme. Experimental studies have reported that the implementation of an evidence-based clinical decision support system in an emergency department was associated with a 20% decrease in use of CT pulmonary angiography for evaluating acute pulmonary embolism and a 69% increase in yield [15]. Recent evidence and experiments have confirmed the feasibility of clinical decision support [16]. While so many different techniques can be used in such applications, in this introduction, we considered only three: automatic question answering, rule bank based on artificial intelligence and link-based analysis.

One increasingly popular means for clinical decision support is automatic question answering [17]. In this approach, complex problems are decomposed into fact-seeking questions or mapped to similar, simpler questions [18, 19]. Combining this method with a semantic domain model, several potential solutions such as the PICO framework have

been proposed. In this framework, all diagnoses are divided into four types: Problem, Intervention, Comparison and Outcome(PICO). All patient symptoms are reformulated using normalised expressions, allowing for automatic answer extraction [20]. From the query results, the diagnosis can be refined by applying the existing medical knowledge base and clustering methods from natural language processing [21]. However, this method is highly dependent on the correct description of symptoms [22].

In recent years, artificial intelligence-based rule banks have also attracted increasing research interests as a valid way of implementing automatic diagnosis. For example, automated diagnosis of sea cucumber diseases has been demonstrated, in which a reasoning machine using a back-propagation (BP) neural network made use of a rule bank constructed from typical cases [23]. Several other diagnosis simulation systems for respiratory diseases used the BP neural network model [24]. However, as these systems must have access to an effective training dataset, much progress has been made in the diagnosis of specific diseases rather than general diseases.

Compared with these methods, link-based analysis has a wider range of application and has been shown to be effective when dealing with big data. One common type is bipartite network analysis [25], in which all nodes are classified into two groups, allowing their relevance to be estimated using iterative algorithms. Because of good results, other specific types of link-based analysis have been applied in clinical diagnosis support. Giannis, Polykarpos, Nektarios and Michalis employed PageRank to measure the importance of each vertex (word) within the graph-of-word for in order to rank in decreasing order by relevance criteria with respect to queries in 2015 Clinical Decision Support track [26]. Jiang et al. built a co-occurrence network to mine the potential medical literatures and adopt the value of pagerank to regard as the measure of node importance in the The TREC 2015 Clinical Decision Support track [27]. Chen, Lu and Liang used this method to explore the pathogenesis of hereditary diseases, using the degree of relevance between the genetic diseases and virulent genes [28]. However, our experiments have shown that this method cannot produce satisfactory results if only keywords are used as the semantic unit. As the information carried by a single keyword is limited, the corresponding resolution is also low. But it is true that this method is useful in the analysis of gene fragments, since gene fragments carry more information and exhibit a higher degree of differentiation. A second type, which uses a hyperlink-based algorithm, has been widely used in fields such as Web page recommendation. Previous research has confirmed the use of this approach in the automatic construction of a domain ontology in library and information science [29], and in the automatic discovery of key paths from academic articles [30]. Based on the semantic information contained in keywords, it provides a feasible way of measuring the relevance of articles and is especially appropriate for the recommendation of medical literature in the clinical diagnosis support service. As link-based analysis does not require pre-existing normalised expression of symptoms or access to multiple rule banks, it is more universal in applications and of particular relevance to our own research, which is focussed on the retrieval of recommended articles based on the description of patient symptoms.

Collection of standard keywords

Keywords list

As discussed below, keywords extracted from the list given in an article perform more poorly than keywords from other sources such as standard lexicons. The reasons for this are as follows.

1) Different medical databases often apply different rules, while authors may use non-standard formats, making it difficult to extract keywords correctly using a single approach. Even when all articles are written in the XML format and most use standard HTML tags to separate keywords, we have found that colons, semicolons and even backslashes are often used. In some cases, it is impossible to resolve the different parts because of technical failures.
2) Invalid keywords are common. In addition to the use of keywords from the stop list and ones that are wrongly spelled, some keywords appear only as numbers. More surprisingly, some articles list ‘keyword’ itself as a keyword and here again, a diverse range of spellings are encountered, including ‘Keywords’, ‘Key terms’ and ‘Key indexing terms’.

To address these problems, we designed an effective and simple algorithm for keyword extraction, based on the statistical analysis of document frequency. The main steps are as follows:

1) Retrieve all keywords from all articles using a delimiter of HTML tags;
2) Delete keywords that appear in the stop list and remove invalid terms such as pure numbers, field names, or terms that are too long (such as greater than or equal to 100);
3) Aggregate the document frequencies of keywords in each field (such as title, abstract, body text) of each article;
4) Delete those keywords whose document frequency falls below a specific threshold (such as 2).

This allows all valid keywords to be acquired. Step 3 is the most important. Although we do not collect all separators or the diversity of invalid keywords in advance, the application

of a specific threshold for document frequency can filter the relevant items successfully. The hypothesis that correct spellings are more frequent than misspellings is worthy of trust in most cases. This approach has two further advantages. The first is that it expands the keyword list for each article, since it accumulates all keywords in all fields. Even if a keyword does not appear in the keyword list, it can be acquired from the title and body of the article. A second advantage is that the reformulated keyword list is more accurate. If the keyword list contains misspellings, valid keywords can be retrieved from the title or body after all the fields have been scanned for all valid keywords and those with lower document frequencies have been removed.

Medical lexicon

Many medical lexicons are available, including the Medical Subject Headings (MeSH) thesaurus and the Unified Medical Language System (UMLS). We chose MeSH as the supplementary source of keywords because of experimental evidence of the usefulness of MeSH terms in search environment [31]. Previous results have suggested that an augmented system with two additional components supporting MeSH term searching and MeSH tree browsing was more effective than the simple search system, offering improved user-perceived topic familiarity and question–answer performance [32].

All keywords in MeSH are divided into three hierarchical levels: Descriptor, the top level, Concept, the second level and Term, the lowest level. Each concept belongs to one descriptor and each term belongs to one concept. Concept is the most formal and most useful part of MeSH, since the descriptor is mainly used to group related concepts into existing classes, and each concept contains many various terms which are often synonyms with different spellings. Concepts in MeSH include all descriptors and terms also include all concepts. The detailed structure is shown in Fig. 1.

It is more helpful that MeSH provides the mappings between different concepts. This allows synonyms and related keywords to be combined with this mapping of concepts and with the hierarchy of MeSH so that the keywords in each article can be expanded and the accuracy and coverage of retrieval can be improved. Since some different concepts are semantically similar to each other in medical science, we can include more terms into existing concept. If a correlative mapping exists between two concepts, all the terms belonging to one concept can be added to the keyword list of the other. For example, if Concept2 is semantically similar to Concept3 (see bold bidirectional arrow), all the terms of Concept3 will be included into the term collection of Concept2 and vice versa. In Fig.1, the dotted arrows indicate including extra terms to concepts.

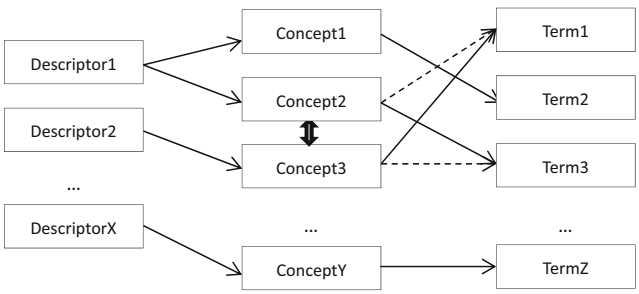

Fig. 1 The relation of three hierarchical levels in MeSH

Acronyms

Acronyms are common in the description of symptoms and diseases in the medical literature. Although they seldom appear in keyword lists, almost every medical article contains acronyms in its body text. However, few lexicons of acronyms are available on the Web, and most of these are incomplete. It was, therefore, necessary to take a novel approach to the automatic identification of acronyms.

The ideal assumption is that each acronym occurs in articles that also contain the corresponding original terms. However, this is not the case, due to the great diversity of writing styles and abbreviation rules. For example, some acronyms use full or partial capitalisation, while others use lowercase throughout; Some articles use only the acronym in the text and do not spell out the corresponding full terms. Based on these observations, we designed a two-step extraction algorithm. In the first step, all pairings of acronyms and original terms are identified. From these, valid mappings are identified between them. In the second step, a judgement is made about each acronym and its true meaning in the article.

Here, we first introduce step one and will discuss the second step in the section 4.2.

A central focus in the first step of our heuristics algorithm is the way in which pairs of acronyms and original terms always follow a basic rule with the format: A*B*(AB*). Here, A and B denote the capitalised letters of the first two keywords in the original term, since nearly all acronyms use capitalisation of the first two terms. An asterisk can represent some Latin letters, numbers, or Greek letters. This algorithm scans the full body text and retrieves all strings that have this feature. In more detail, the steps for each article are as follows:

1) Retrieve all opening curly braces;
2) Check whether at least two sequential letters after the opening brace are capitalised;
3) If yes, retrieve the position of these two capitalised letters before the opening brace. If their order is the same as that after the brace, all keywords from the first capitalised capital letter to the brace are extracted as the original term and then further checked for the inclusion of invalid characters such as colons or periods. An acronym from the brace to the first non-alphabetical and non-numeric

character after the brace is also extracted. A pairing of acronym and original term is then generated;

4) Validate and filter again after extracting all pairs of acronyms and original terms. For example, delete pairs in which the length of the acronym is equal to or longer than the original term, delete pairs in which the acronym is part of another acronym, and both have the same original terms and delete pairs in which letters in the acronym occur in a different order to those in original terms;
5) Aggregate the document frequencies of all acronyms and delete those with lower frequencies, since they may represent misspellings or make no sense, even if literally correct.

Extraction of the relationship between keywords and documents

The basic algorithm

Once a full collection of standard keywords has been constructed, the relation of keywords to documents can be acquired. Different methods are required for extracting keywords of different types. We now introduce the extraction algorithm and discuss related strategies.

Before processing, all the letters in the document are converted to lowercase. Multiple spaces are replaced with a single space. Greek letters are retained, since they are frequently used in medical articles to denote diseases and medicines.

Since the number of keywords is large, it is inefficient and time-consuming to scan each keyword from the standard keyword collection for each article. Therefore, a single traversal algorithm is proposed, in which keywords of a maximum length are matched first. Each article can then be scanned only once from beginning to end. The basic steps are as follows:

1) For one article, read the first letter of this article and locate the pointer here;
2) Get the greatest length of keywords starting with that letter pointed by current pointer, and read the character sequence with this length from the pointer;
3) Search this sequence in the keywords list;
4) If matched, move the pointer to the next letter behind this sequence in this article and jump to step 6. If the length of this sequence is equal to zero, locate the pointer to the next word from current location, and jump to step 2.
5) If not matched, delete the last word from this character sequence and do step 3 again;
6) Locate the beginning of the next word from the current pointer, and jump to step 2.

Extraction based on keyword type

Different strategies of extraction can be applied to different types of keywords.

For keywords that appear in the keyword list, the algorithm in 4.1 can be applied directly to extract all the keywords from the body text.

For the keywords that appear in the MeSH lexicon, the mapping to concepts can also be used to expand the current keyword list of each article. That is to say, if a term appears as a keyword in one article, and this term belongs to another concept, those terms of another concept will be added to the keyword list of this article with same term frequency.

To address acronyms, a two-step process is needed. In the first step mentioned before, all the original terms of all the acronyms within the full article are scanned using the same strategy of extracting standard keywords. In the second step, a case-sensitive scan of all the acronyms in the article is performed. Since a single acronym may refer to multiple original terms, it is necessary to decide which of these is most likely in the context. Once again, the algorithm based on the analysis of term frequency is used. The basic procedure is that all the original terms of each acronym in each article are scanned, and the original term with the highest term frequency is chosen as the candidate. If many candidates have the same term frequency, the one with the longest length is chosen. All acronyms are replaced with corresponding original terms after processing so that no acronym remains in the final relation of keywords and documents.

Normalisation of keywords

To improve effectiveness, all keywords are normalised before further analysis. For example, as most keywords in the medical literature are nouns, it is necessary to convert plurals into singulars to match existing dictionaries. Simple heuristic rules such as deletion of a final letter ‘s’ can be applied. We did not do any morphological analysis of words since we think it might be risky to widely apply lemmatization and stemming for professional medical terms which are often long and complex.

Finally, three extraction units are selected: sentences, segments and body texts. In our approach, sentences are always identified by periods and the segments by the section dividers of the HTML tagging. We aggregate all the term frequencies of each keyword into these three units to build the final data set.

Automatic recommendation of clinical diagnosis literature

Descriptions of patient symptoms are collected and summarised by clinicians and often include healthy

conditions, current status, existing diseases and related symptoms. These full diagnoses and case descriptions provide a rich and reliable resource for clinical decision support. In our experiment, test topics of track in TREC 2014 are used as descriptions of patient symptoms for further analysis.

This process has two key steps. The first is the retrieval of all possibly relevant articles, based on the existing descriptions of the patient. The second is the choice of articles that are most useful and relevant, allowing an ordered list of articles to be generated.

Keyword set from the description

Although the case history of a patient may include a number of keywords, few of these have a term frequency greater than one, since clinicians often prefer a concise writing style. If all the keywords are considered, many false positives will be generated, and many irrelevant articles retrieved. It is therefore necessary to weigh the importance of each keyword, so that only those with the highest weight (such as inversed document frequency) are selected as final query terms. We therefore proposed a two-step method. In the first step, all keywords are sorted in reverse order of document frequency, and the lower frequency items are discarded (such as lower than 2). The second step is the retrieval of combination of keywords occurring in the same sentence, since a combination of keywords will provide a more accurate guide to the possible disease. The combination of keywords is called keyword set here. Sentences are better units for this co-occurrence analysis. There are typically two keywords in co-occurring keyword sets, and less frequently one or three. The introduction of manual selection and selection by the clinician can enhance the selection of significant keyword set. This is the only point in our approach at which manual selection is introduced. We also feel that this is a natural way of improving the precision of recommendations because the exact description of patient symptoms was written by clinicians.

The collection of valid sentences

We have now constructed two data sets: one comprising the term frequencies of keywords in their occurrence unit, a sentence, segment, or body texts, the second comprising many keyword sets selected from the description of the patient's symptoms. It has been clearly shown that decision-making cannot be only based on the occurrence unit of keywords [33]. Sentences provide a more precise basis for occurrence analysis, but many potentially useful keywords may be excluded, since in many articles, a longer text is added to describe the symptoms of the patient. However, segments and body texts contain many insignificant co-occurrences of keywords on the other side.

We, therefore, introduced a novel method that combines the advantages of approaches based on different units. For each selected keyword set taken from the description of the patient, each sentence is chosen as a basic unit of analysis and checked to see whether it contains at least one keyword set. Those sentences containing at least one keyword set are selected as valid sentences, and segments containing at least one valid sentence are selected as valid segments. The basic idea of this process is that an article containing these valid segments is more likely to discuss relevant diseases or contain other useful information. The segments are then grouped by the articles in which they occur and articles with fewer valid segments are removed (such as fewer than 2). Finally, all sentences remaining in the collection of valid segments constitute the valid sentences.

Retrieval of diagnosis-related keywords

From the collection of valid sentences, diagnosis-related keywords can be extracted by a link-based algorithm combined with a weighting measurement.

PageRank is an efficient link-based algorithm for measuring the weight of each page in the Web. The core pages are often linked by a larger number of other pages, so that they have more in-degrees than those pages linked by fewer pages. According to the hyperlink analysis and the PageRank algorithm, each keyword in the collection of valid sentences can be seen as a node, and the links between these nodes can be identified from the co-occurrence of corresponding keywords. Since the link-based algorithm needs directed links, the direction of each runs from a keyword with the higher document frequency to one with the lower frequency. This design assumes that parts of an article that discuss diagnosis may contain relevant information and that keywords with relatively lower document frequency are more important and have the higher resolution power. The algorithm can enhance the weighting of these keywords through iterative computation.

The weight of each node of a keyword is calculated as follows:

$$weight_{n+1}(keyword_k) = (1-\alpha) + \alpha \times \sum \frac{idf(inKeyword_i) \times weight_n(inKeyword_i)}{C(inKeyword_i)}$$

Here, n is the step in the iteration, C is the out-degree of a node and inKeyword refers to the keyword at the beginning of the link. This algorithm weights a node based on document frequency so that keywords with the lower document frequency (higher inverted document frequency) and more in-degrees are therefore given higher weights.

Measuring the relevance of an article

The relevance of all related articles in the collection of valid sentences can be calculated from average weights of the

corresponding keywords. The more highly weighted keywords an article contains, the more relevant it is. The basic measure of an article's relevance is as follows:

$$weight(article_k) = avg\left(weight\left(keyword_i^{articlek}\right)\right)$$

Experiments

Data set

All the data was taken from medical articles provided in TREC clinical decision support track 2015. Its website is http://www.trec-cds.org/2015.html. A total of 733,328 articles were used. Each article was formatted using XML and had a unique ID. We adopted SQL Server to store all data and used T-SQL and Java to implement all program.

Extraction of the collection of keywords

Applying the extraction method introduced in Section 3, a total of 347999 keywords were taken from the keyword lists. The stop list used was based on the Reuters Corpus Volume 1, which lists 25 common terms. There are a total of 27149 descriptions in the MeSH lexicon, 51525 concepts and 218985 terms. There are 49412 mapping relations between concepts and terms. The total number of pairs of acronyms and corresponding original terms was 289670 before validation and 139888 after validation. The automatically constructed acronyms can be retrieved from http://www.njcie.com/medical/.

We carried a comparison of our acronym result with ADAM [34] which provided a downloadable and free dataset of medical acronyms. The number of their results is 57827 because they mainly utilized all terms in titles and abstracts, not full text of MEDLINE. We chose some acronyms for simple comparison listed in Table 1:

- Group1 (with the highest DF): CS (1st in ADAM), PA (2nd in ADAM)
- Gtoup2 (manually selected): AIDS, DNA

We can get two conclusions from the observation. One is that our acronym set is larger than ADAM so that many acronyms in ADAM have fewer entries and some acronyms don't even exist in ADAM such as SPSS, etc. The reason is that ADAM only utilized titles and abstracts of MEDLINE. Another is that our approach inevitably introduced many wrong forms. The reason is that authors are easier to write wrong words in bodies than in titles and abstracts. However, these wrongly-spelt original terms can be filtered further according to their lower document frequency.

After deduplication and aggregation, 311,379 keywords remained in the final collection. There were a total of 881 million occurrence pairs of keywords and sentences, 148 million sentences and 24 million segments.

The recommendation of medical articles

We tested ten standard topics and one demo topic. Each result displayed at most five articles with the highest relevance. All typical test topics and their results are presented below:

1) Topic description: A woman in her mid-30s presented with dyspnoea and haemoptysis. CT scan revealed a cystic mass in the right lower lobe. Before she received treatment, she developed right arm weakness and aphasia. She was treated; however, four years later, she suffered another stroke. Follow-up CT scan showed multiple new cystic lesions. Possible relevant articles are shown in Table 2.

The pairs of keywords we chose included dyspnoea and haemoptysis, mass and lobe, weakness and aphasia and stroke. After scanning, 88,512 valid segments were obtained, with 30,622 corresponding articles. After filtering, 58 valid segments were confirmed. These included 1728 keywords with 38,456 directed links, contained within 625 sentences and 52 articles. Table 2 suggests that the first two articles mainly concerned pulmonary arteriovenous malformation and were highly relevant to the diagnosis. Although the third discussed migraine and the forth mainly discussed infarction in the posterior cerebral artery, they also provided useful points of comparison for the diagnosis. The last article discussed rehabilitation following strokes, and its information was less useful and had lower relevance. Compared with the existing suggested three results, our result only produced the first of them (docid: 3,148,967). This was because our algorithm did not choose the other two articles as candidates for valid articles. For example, the second article (docid: 2,987,927) contained no sentence with more than two valid keywords and the discussion of symptoms was distributed thinly across a long text. The third (docid: 3,082,226) gave no description of symptoms at all. This result also reveals that a possible weakness of our approach is that it only focuses on symptom keywords as the query terms.

2) Topic description: A 58-year-old African-American woman presents to the ER with episodic pressing/burning anterior chest pain that began two days earlier for the first time in her life. The pain started while she was walking, radiates to the back and is accompanied by nausea, diaphoresis and mild dyspnoea but is not

Table 1 The comparison of ADAM and our acronym set

ADAM		Our acronym set	
Original Terms	Count	Original Terms	DF
Group 1: CS			
conditioned stimulus	702	conditioned stimulus	356
chondroitin sulfate |chondroitin sulfates	361	chondroitin sulfate	174
coronary sinus	340	coronary sinus	158
citrate synthase	334	chitosan	104
cesarean section|cesarean sections	186	caesarean section	102
corticosteroids|corticosteroid	171	cockayne syndrome	88
contact sensitivity	133	calf serum	69
contrast sensitivity	126	cesarean section	61
Caesarean Section|caesarean sections	97	compressed sensing	58
cigarette smoke	95	conditioned stimuli	45
Group 1: PA			
phosphatidic acid|phosphatidic acids	833	phosphatidic acid	429
plasminogen activator|plasminogen activators	826	protective antigen	231
pulmonary artery|pulmonary arteries	716	palmitic acid	104
physical activity	268	plasminogen activator	82
protective antigen	218	positive affect	69
posteroanterior	143	posterior anterior	54
Pseudomonas aeruginosa	140	posteroanterior	53
primary aldosteronism	118	photoacoustic	50
arterial pressure	98	pleomorphic adenoma	38
procainamide	89	primary aldosteronism	35
Group 2: AIDS			
acquired immunodeficiency syndrome|acquired immunodeficiency syndromes	6297	acquired immunodeficiency syndrome	975
acquired immune deficiency syndrome	2376	acquired immune deficiency syndrome	466
acquired immunodeficiency disease	18	acquired immuno deficiency syndrome	35
acquired immune deficiency	13	acquired immunodeficiency disease syndrome	8
acquired immunodeficiency	12	autoimmune deficiency syndrome	8
		autoimmune diseases	6
		autoinflammatory diseases	4
		autoinflammatory disorders	3
		acquiredimmunodeficiency syndrome	3
		acquired immunodeficiencysyndrome	3
Group 2: DNA			
deoxyribonucleic acid|deoxyribonucleic acids	1761	deoxyribonucleic acid	671
		deoxyribose nucleic acid	11
		desoxyribonucleic acid	7
		deoxyribo nucleic acid	6
		did not attend	5
		determined by transcript number normalized to total rna quantity and difference	4
		deoxy ribo nucleic acid	4
		double stranded nucleic acid	4
		difco nutrient agar	3
		dna fragment	3

increased on inspiration. The latest episode of pain ended half an hour prior to her arrival. She is known to have hypertension and obesity. She denies smoking, diabetes, hypercholesterolaemia or a family history of heart

Table 2 The results of demo topic

DocID	Title
3025345	A Case of a Pulmonary Arteriovenous Malformation With Ebstein's Anomaly
3148967	Stroke in hereditary hemorrhagic telangiectasia patients. New evidence for repeated screening and early treatment of pulmonary vascular malformations: two case reports
3180463	Infarctions in the vascular territory of the posterior cerebral artery: clinical features in 232 patients
3287708	Maximising adherence to study protocol within pharmaco-rehabilitation clinical trials
3420796	Sporadic Hemiplegic Migraine with Seizures and Transient MRI Abnormalities

disease. She currently takes no medications. Physical examination is normal. The EKG shows non-specific changes. Possible relevant articles are shown in Table 3.

The pairs of keywords we chose included episodic and chest pain, nausea and dyspnea, hypertension and obesity. The first article is highly relevant since its focus is on primary pericardial malignant mesothelioma, and the described symptoms are the same as that in the topic case. The second one is a literature review regarding in-flight medical emergencies. It is not relevant but as it is a review text, it contains many valid keywords and thus gains a higher weighting. The same shortcoming, of using only keywords of symptoms as the query term, was revealed. The third article was relevant as it discussed epipericardial fat necrosis. The last mainly discussed chest pain and coronary heart disease, which provided a number of useful comparison cases. The final column of Table 3 shows the score for the standard measures specified in the 2015 Clinical Decision Support Track (http://www.trec-cds.org/qrels2014.txt). These values represent the similarity degree of documents, and it is assumed that documents with higher scores should be retrieved first. The highest value is two, and the lowest is zero. Only 13,138 articles were relevant to all ten diagnosis topics, representing 1.79% of articles in the collection. The label N means that no such record was found, so that the relevance could not be determined.

Table 3 The results of topic 1

DocID	Title	Score
2994533	Primary pericardial malignant mesothelioma and response to radiation therapy	N
3258729	Epipericardial fat necrosis – a rare cause of pleuritic chest pain: case report and review of the literature	2
3490454	Does the patient with chest pain have a coronary heart disease? Diagnostic value of single symptoms and signs – a meta-analysis	2
3681367	Medical emergencies on board commercial airlines: is documentation as expected?	N

Table 4 The results of evaluation

Article order in result	1st		2nd		3rd		4th		5th	
Query	q1	q2	q1	q2	q1	q2	q1	q2	q1	q2
User 1	5	5	4	5	3	3	3	2	1	3
User 2	5	5	2	1	4	5	5	4		
User 3	5	4	5	5	4	3	3	5	5	5
User 4	5	5	5	4	4	5	4	5	5	5
User 5	5	5	2	3	1	1				
User 6	4	4	5	5						
User 7	4	3	3	4	3	3	4	5	5	5
User 8	3	4	3	2	5	5	5	5		
User 9	5	5	5	5	3	4	3	3	4	3
User 10	5	5	5	5	4	3	3	3	3	4
User 11	1	2	3	4	5	5	5	5	5	5

q1 and q2 respectively express two queries from each user

User evaluation

We then invited 11 users to evaluate the results. Five were clinicians with specific knowledge of medical science from 454 Hospital in Nanjing, and six majored in Information Management Department at Nanjing University of Chinese Medicine and therefore had a rich comprehension of information retrieval. Each user was asked to evaluate two queries and record his/her satisfaction with the five highest weighted recommendations. We used a Likert-type scale with five steps, in which five was maximum satisfaction, and one meant 'not satisfied'. The results are shown in Table 4.

As a part of queries produced fewer than five results, there are some blanks in Table 4. The total satisfaction level of all users was 3.98, or 33% above average satisfaction. The fewer the number of results, the higher the average satisfaction turned out to be. The detailed results are given in Table 5.

We validated the effectiveness of these evaluation results by checking the consistency of evaluation of the participants. Consistency is better when the first recommendation has the

Table 5 Average satisfaction of results with different number

The total number of articles	Average satisfactory
1	4.27
2	4.07
3	3.94
4	3.95
5	3.98

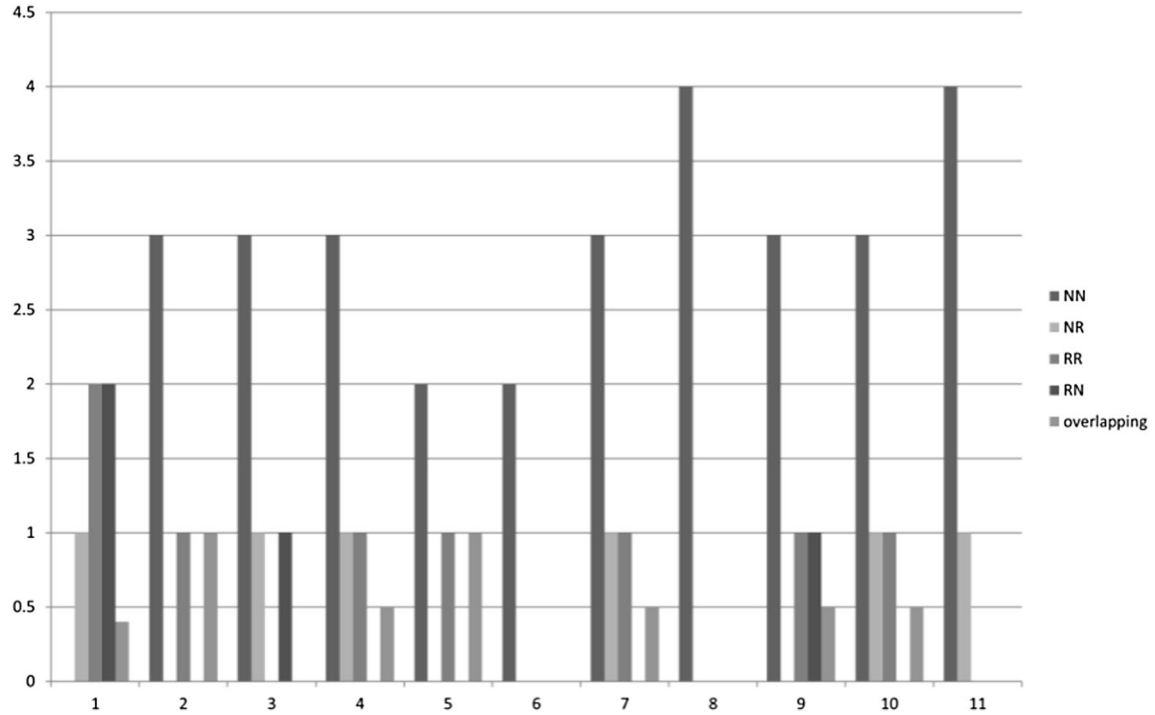

Fig. 2 The detailed information of consistence validation

highest score and the last has the lowest score. Since there were two different evaluation results for each topic, the two results for one topic were used to calibrate the consistency of user satisfaction. We assigned R if the evaluation result fitted the ideal value, and N otherwise. The overlapping index was calculated as RR/(NR + RR + RN). The mean overlap was 0.489. A detailed validation is shown in Fig. 2.

We also validated these data against the NDCG(Normalized Discounted Cumulative Gain) to measure how effectively this method ranked the results. The mean for all NDCG results was 0.962. Detailed information is shown in Fig. 3.

In order to further evaluate the consistence and accurate of searching results, we also used infNDCG as the metric [35]. In

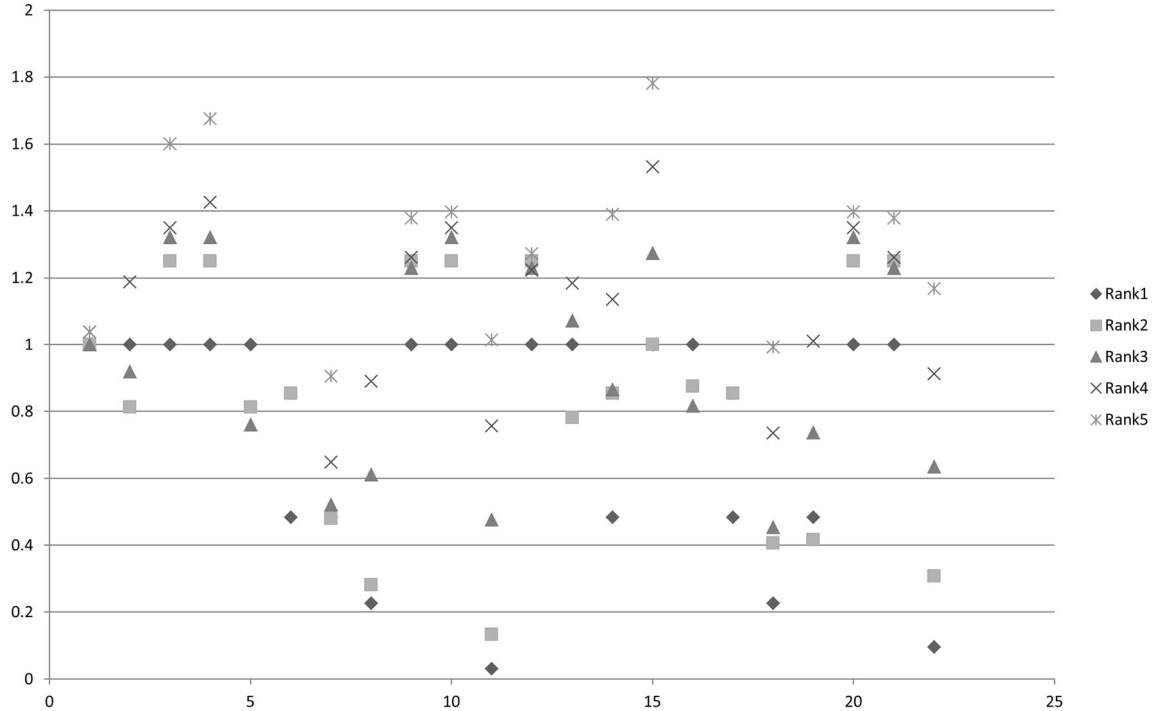

Fig. 3 The detailed evaluation results of NDCG

Table 6 The top five infNDCG of all manual tasks in Task A of TREC 2015

Participant	infNDCG
Wayne State Univ. [wsu ir]	0.3109
Northwestern/Utah/UNC [NU UU UNC]	0.3019
Univ. of Michigan [Foreseer]	0.2954
Fudan Univ. [FDUDMIIP]	0.2689
Our Method	0.2602
Demo. Univ. of Thrace [DUTH]	0.2318

our experiments, we conducted a comparison of infNDCG with TREC 2014 and TREC 2015 [1, 36]. The condition and tasks of TREC 2015 are different from those in TREC 2014. TRECT 2015 includes two tasks: Task A and Task B. Task A includes two different sub-tasks: an automatic task and manual task. Our method should be classified as a manual task since initial pairs of important keywords are manually selected in the analysis of each topic. Our infNDCG ranked fifth among all manual tasks in Task A of TREC 2015. The full results are shown in Table 6.

Discussion

We introduced an approach to medical literature recommendation for clinical diagnostic decision support, and experiments demonstrated that the combination of semantic analysis using a link-based algorithm with a keyword co-occurrence analysis was able to effectively retrieve relevant articles based on a description of patient symptoms. In a future study, we will continue to evaluate the stability of the approach. We will expand our data collection into other areas and conduct a more wide-ranging user evaluation. As noted, the utilisation of semantic information only from keywords created difficulties when the articles were too long or their contents differed too widely from the required information. Nevertheless, these articles were really relevant. Our future study will also involve the design of experiments that allow us to combine useful information from other fields, suggesting alternative approaches that may enhance the performance of the algorithm.

Acknowledgements This work was supported by the Chinese National Social Science Foundation 16BTQ030 (2016).

References

1. Simpson, M. S., Voorhees, E., and Hersh, W., Overview of the TREC 2014 clinical decision support track. In: *Proceedings of the23rd Text Retrieval Conference (TREC 2014)*. National Institute of Standards and Technology (NIST), 2014.
2. Abrahamsson, E., Forni, T., Skeppstedt, M., and Kvist, M., Medical text simplification using synonym replacement: adapting assessment of word difficulty to a compounding language. In: *Proceedings of the Workshop on Predicting & Improving Text Readability for Target Reader Populations* (pp. 57–65). Association for Computational Linguistics, 2014.
3. Safran, C., Bloomrosen, M., Hammond, W.E., Labkoff, S., Markel-Fox, S., Tang, P.C., and Detmer, D.E., Toward a national framework for the secondary use of health data: An American medical informatics association white paper. *Journal of the American Medical Informatics Association*. 14(1):1–9, 2007.
4. Elkin, P.L., Liebow, M., Bauer, B.A., Chaliki, S., Wahner-Roedler, D., Bundrick, J., et al., The introduction of a diagnostic decision support system (DXplain™) into the workflow of a teaching hospital service can decrease the cost of service for diagnostically challenging diagnostic related groups (DRGs). *International Journal of Medical Informatics*. 79(11):772–777, 2010.
5. Barnett, G.O., Cimino, J.J., Hupp, J.A., and Hoffer, E.P., DXplain: An evolving diagnostic decision-support system. *The Journal of the American Medical Association*. 258(1):67–74, 1987.
6. Shwe, M.A., Middleton, B., Heckerman, D.E., Henrion, M., Horvitz, E.J., Lehmann, H.P., and Cooper, G.F., Probabilistic diagnosis using a reformulation of the INTERNIST-1/QMR knowledge base. *Methods of Information in Medicine*. 30(4):241–255, 1991.
7. Klimov, D., and Shahar, Y., iALARM: An intelligent alert language for activation, response, and monitoring of medical alerts. In: *Proceedings of Process Support and Knowledge Representation in Health Care* (pp. 128–142). Springer International Publishing, 2013.
8. Elhadad, N., Kan, M.Y., Klavans, J.L., and McKeown, K.R., Customization in a unified framework for summarizing medical literature. *Artificial Intelligence in Medicine*. 33(2):179–198, 2005.
9. Jaspers, M.W., Smeulers, M., Vermeulen, H., and Peute, L.W., Effects of clinical decision-support systems on practitioner performance and patient outcomes: A synthesis of high-quality systematic review findings. *Journal of the American Medical Informatics Association*. 18(3):327–334, 2011.
10. Seidling, H.M., Phansalkar, S., Seger, D.L., Paterno, M.D., Shaykevich, S., Haefeli, W.E., and Bates, D.W., Factors influencing alert acceptance: A novel approach for predicting the success of clinical decision support. *Journal of the American Medical Informatics Association*. 18(4):479–484, 2011.
11. Wright, A., Sittig, D.F., Ash, J.S., Bates, D.W., Feblowitz, J., Fraser, G., et al., Governance for clinical decision support: Case studies and recommended practices from leading institutions. *Journal of the American Medical Informatics Association*. 18(2):187–194, 2011.
12. Wright, A., Sittig, D.F., Ash, J.S., Feblowitz, J., Meltzer, S., McMullen, C., et al., Development and evaluation of a comprehensive clinical decision support taxonomy: Comparison of front-end tools in commercial and internally developed electronic health record systems. *Journal of the American Medical Informatics Association*. 18(3):232–242, 2011.
13. Romano, M.J., and Stafford, R.S., Electronic health records and clinical decision support systems: Impact on national ambulatory care quality. *Archives of Internal Medicine*. 171(10):897–903, 2011.
14. Hoeksema, L.J., Bazzy-Asaad, A., Lomotan, E.A., Edmonds, D.E., Ramírez-Garnica, G., Shiffman, R.N., and Horwitz, L.I., Accuracy of a computerized clinical decision-support system for asthma assessment and management. *Journal of the American Medical Informatics Association*. 18(3):243–250, 2011.
15. Raja, A.S., Ip, I.K., Prevedello, L.M., Sodickson, A.D., Farkas, C., Zane, R.D., et al., Effect of computerized clinical decision support on the use and yield of CT pulmonary angiography in the emergency department. *Radiology*. 262(2):468–474, 2012.

16. Tamine, L., and Chouquet, C., On the impact of domain expertise on query formulation, relevance assessment and retrieval performance in clinical settings. *Information Processing & Management*. 53(2):332–350, 2017.
17. Abacha, A.B., and Zweigenbaum, P., MEANS: A medical question-answering system combining NLP techniques and semantic web technologies. *Information Processing & Management*. 51(5):570–594, 2015.
18. Ryu, P.M., Jang, M.G., and Kim, H.K., Open domain question answering using Wikipedia-based knowledge model. *Information Processing & Management*. 50(5):683–692, 2014.
19. Ou, S.Y., An entailment-based question answering method in a restricted domain. *Journal of the China Society for Scientific and Technical Information*. 30(5):540–547, 2011.
20. Amini, I., Martinez, D., Li, X., and Sanderson, M., Improving patient record search: A meta-data based approach. *Information Processing & Management*. 52(2):258–272, 2016.
21. Demner-Fushman, D., *Complex question answering based on a semantic domain model of clinical medicine*. University of Maryland (United States), OCLC's Experimental Thesis Catalog. College Park, 2006.
22. Huang, X., Lin, J., and Demner-Fushman, D., Evaluation of PICO as a knowledge representation for clinical questions. In: *Proceedings of AMIA Annual Symposium* (pp. 359). American Medical Informatics Association, 2006.
23. Li, F., Han, S.J., and Zhang, D., The construction of sea cucumber disease diagnosis inference engine. *Computer Applications and Software*. 29(12):211–213, 2012.
24. Huang, Z.X., Zhong, C., and Li, X.R., Simulation study of respiratory disease diagnosis based on BP neural network. *Journal of Hefei University of Technology (Natural Science)*. 35(3):347–349, 2012.
25. Li, S.Q., Xu, X., and Xu, M.J., The measures of books' recommending quality and personalized book recommendation service based on bipartite network of readers and books' lending relationship. *Journal of Library Science in China*. 39(3):83–95, 2013.
26. Giannis, N., Polykarpos, M., Nektarios, L., and Michalis, V., AUEB at TREC 2015: clinical decision support track. In: *Proceedings of 24rd Text Retrieval Conference (TREC 2015)*. National Institute of Standards and Technology (NIST), 2015.
27. Jiang, J., Guan, Y., Su, J., Zhao, C., and Yang, J., HIT-WI at TREC 2015 Clinical Decision Support Track. In: *Proceedings of 24rd Text Retrieval Conference (TREC 2015)*. National Institute of Standards and Technology (NIST), 2015.
28. Chen, W.Q., Lu, J.A., and Liang, J., Research in disease-gene network based on bipartite network projection. *Complex Systems and Complexity Science*. 6(1):13–19, 2009.
29. Li, S.Q., Research on automatic construction of domain ontology in library and information science based on weighted co-occurrence of citation keywords. *Journal of the China Society for Scientific and Technical Information*. 31(4):371–380, 2012.
30. Li, S.Q., Xu, X., Qian, G., and Han, W., A method for automatic recognition and visualization of main-paths in academic documents based on vibration algorithm and domain ontology. *Journal of the China Society for Scientific and Technical Information*. 31(7):676–685, 2012.
31. Liu, Y.H., and Wacholder, N., Evaluating the impact of MeSH (medical subject headings) terms on different types of searchers. *Information Processing & Management*. 53(4):851–870, 2017.
32. Mu, X., Lu, K., and Ryu, H., Explicitly integrating MeSH thesaurus help into health information retrieval systems: An empirical user study. *Information Processing & Management*. 50(1):24–40, 2014.
33. Kaur, J., and Gupta, V., Effective approaches for extraction of keywords. *International Journal of Computer Science Issues*. 7(6): 144–148, 2010.
34. Zhou, W., Torvik, V.I., and Smalheiser, N.R., ADAM: Another database of abbreviations in MEDLINE. *Bioinformatics*. 22(22): 2813–2818, 2006.
35. Yilmaz, E., Kanoulas, E., and Aslam, J. A., A simple and efficient sampling method for estimating AP and NDCG. In: *Proceedings of Engineering in International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp.603–610). ACM, 2008.
36. Roberts, K., Simpson, M. S., Voorhees, E. M., and Hersh, W. R., Overview of the TREC 2015 clinical decision support track. In: *Proceedings of 24rd Text Retrieval Conference (TREC 2015)*. National Institute of Standards and Technology (NIST), 2015.