Graph-based Concept Weighting for Medical Information Retrieval

Bevan Koopman[1,2], Guido Zuccon[1], Peter Bruza[2], Laurianne Sitbon[2], Michael Lawley[1]

[1]Australian e-Health Research Centre, CSIRO, Brisbane, Australia

[2]Faculty of Science & Technology, Queensland University of Technology, Brisbane, Australia

{b.koopman, p.bruza, laurianne.sitbon}@qut.edu.au, {guido.zuccon, michael.lawley}@csiro.au

ABSTRACT

This paper presents a graph-based method to weight medical concepts in documents for the purposes of information retrieval. Medical concepts are extracted from free-text documents using a state-of-the-art technique that maps n-grams to concepts from the SNOMED CT medical ontology. In our graph-based concept representation, concepts are vertices in a graph built from a document, edges represent associations between concepts. This representation naturally captures dependencies between concepts, an important requirement for interpreting medical text, and a feature lacking in bag-of-words representations.

We apply existing graph-based *term* weighting methods to weight medical concepts. Using concepts rather than terms addresses vocabulary mismatch as well as encapsulates terms belonging to a single medical entity into a single concept. In addition, we further extend previous graph-based approaches by injecting domain knowledge that estimates the importance of a concept within the global medical domain.

Retrieval experiments on the TREC Medical Records collection show our method outperforms both term and concept baselines. More generally, this work provides a means of integrating background knowledge contained in medical ontologies into data-driven information retrieval approaches.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.3 Information Search and Retrieval

General Terms

Theory, Experimentation

Keywords

Medical Information Retrieval, Graph Theory

ADCS'12, December 05 – 06 2012, Dunedin, New Zealand.

1. INTRODUCTION

Most information retrieval (IR) models represent documents as bag-of-words, that is, the representation does not consider word order or term dependence. However, alternative representations, such as graph-based representations have shown that taking term dependence into account can improve retrieval performance [3]. In these approaches a document is modelled as a graph, where terms are vertices and edges represent relations between terms. The importance of a term within a document is proportional to its connectedness to other terms and can be estimated with graph-based measures such as the PageRank algorithm [11].

At the same time there is an increasing body of research within the IR community focused on systems for medical information retrieval [6]. The nature of medical natural language presents some specific challenges — vocabulary mismatch is more prevalent and there is greater interdependence between terms (e.g., between diseases and treatments or organisms and diseases) [12, 7]. This motivates the use of alternative IR models that incorporate more semantic approaches to capture the innate dependencies between terms in medical natural language.

In this paper we apply existing graph-based term weighting approaches to medical IR. Rather than applying these approaches to the original term representation of documents, we first convert the documents into medical concepts defined by the SNOMED CT medical ontology. The motivation for this conversion is that concept-based representations have a proven track record in medical IR [19, 9, 7]. Concepts (the counterpart of terms in this context) are weighted according to their connectedness within the graph using an adapted PageRank algorithm. In addition, we propose a novel background weighting method that incorporates the importance of the concept within the global medical domain (rather than just a single corpus); this is done by injecting domain knowledge from the SNOMED CT ontology into the weighting function. A consequence of this method is that a large number of query concepts are actually excluded, which proves effective as a query concept selection method.

The remainder of the paper is organised as follows: Section 2 provides the background on graph-based IR and concept-based representations for medical IR. Section 3 details our graph-based concept-weighting model, including the injection of domain knowledge from the SNOMED CT ontology in the weighting function. Section 4 describes the evaluation methodology using the TREC Medical Record Track and presents results. Section 5 discusses our findings and considers future work.

2. BACKGROUND

This section provides (i) background and related work on concept-based representations of documents for medical IR; and (ii) graph-based term weighting methods for information retrieval. The following section will present our model which combines these two approaches and extensions by the injection of domain knowledge.

2.1 A 'Bag-of-Concepts' Model for Medical IR

Broadly, concept-based IR aims to make use of external knowledge sources (such as thesauri or ontologies) to provide additional background knowledge and context that may not be explicit in a document collection and users' queries. Performance in concept-based IR is highly dependent on the specific domain model or ontology used. General applications (those that utilise WordNet or Open Directory) struggle to outperform keyword-based systems [16, 13, 5]. However, biomedical applications — which use domain specific ontologies — do demonstrate consistent improvements [19, 9, 7]. Generally, concept-based approaches fall into two categories: (i) Those that maintain the original term representation of documents and only utilise concept-based representations (typically of the query) at retrieval time. The query expansion method of Liu et al. [9] is an example of this. (ii) Approaches that translate the original terms in a document into concepts prior to indexing. Zhou et al. [19], Egozi et al. [5] and Koopman et al. [7] take this approach, thereby utilising a 'bag-of-concepts' representation of a document; they demonstrate significant improvements over a term baseline. This latter approach is the one we adopt to develop a graph-based *concept* weighting model. Therefore, we provide some additional details of the 'bag-of-concepts' model below.

The conversion of text to concepts is achieved by a natural language processing system called MetaMap [1], developed by the U.S. National Library of Medicine. MetaMap analyses biomedical free-text and identifies concepts belonging to Unified Medical Language System (UMLS). MetaMap is widely adopted in clinical NLP [10] and IR [6, 9]. Using MetaMap, both queries and documents are converted, hence the 'bag-of-concepts' representation. For example, the text '`vascular dementia`' found in a document would be replaced with the UMLS concept id `C0011269`; Koopman et al. [7] provide further details of this process.

As with the 'bag-of-words' representation, the 'bag-of-concepts' does not incorporate the innate dependencies between concepts that exist in medical natural language. An alternative to bag-of-words representations are graph-based representations of documents, which aim to represent relations between terms in a document as edges in a document graph [3]. We now consider previous graph-based approaches with an eye for how they might be applied to our bag-of-concepts representation, thus capturing the innate relations that may exist between medical concepts.

2.2 Graph-based Term Weighting

Graph-based models have been applied in information retrieval, generally as part of connectionist approaches [4]. Shifting weights between vertices in a graph is the basis for the Inference Network model of Turtle & Croft [14], and the basis for the InQuery language used as part of the popular search engine Lemur[1]. Graphs provide a convenient means of representing information for IR applications — the propagated learning and search properties of a graph provide a powerful means of identifying relevant information items [3] (be they terms or documents). Graph-based algorithms, such as the popular PageRank algorithm [11] are examples of graph theoretic properties that can be utilised very effectively in a information retrieval scenario.

Blanco & Lioma [3] developed a graph-based term weighting model that represents each document as a graph: vertices are terms and edges are relations between terms. Relations may be simple co-occurrence relations within a context window, or more complex grammatical relations. The importance of a term within a document can then be estimated by the number of related terms and their importance, much in the same way PageRank estimates the importance of a page via the pages that link to it.

We hypothesise that Blanco's model adapted to a concept representation of documents may be a powerful tool for medical IR as it would capture the dependencies between concepts found in medical free-text. We therefore integrate Blanco & Lioma graph-based term-weighting model into previous concept-based approaches to medical IR, this is done in the next section. The remainder of this section provides an explanation of the original graph-based model and provides an example of its application on an exerpt of medical text.

In Blanco & Lioma's graph-based term weighting model, a term i in a document is represented by the vertex v_i. A vertex is connected to other vertices, $\mathcal{V}(v_i)$ denoting the set of vertices connected to v_i. The weight of v_i within a document is initially set to 1 and the following PageRank function is run for several iterations

$$S(v_i) = (1-\phi) + \phi * \sum_{v_j \in \mathcal{V}(v_i)} \frac{S(v_j)}{|\mathcal{V}(v_j)|} \quad (0 \leq \phi \leq 1) \quad (1)$$

where ϕ is the damping factor which controls "vote recycling" from the original PageRank algorithm [11]. Blanco & Lioma showed that only a small number of interations (< 50) is required to obtain convergence [3]. Edges between vertices are based on relations between terms. Term relations can be implemented as the co-occurrence between two terms within a set context window N[2].

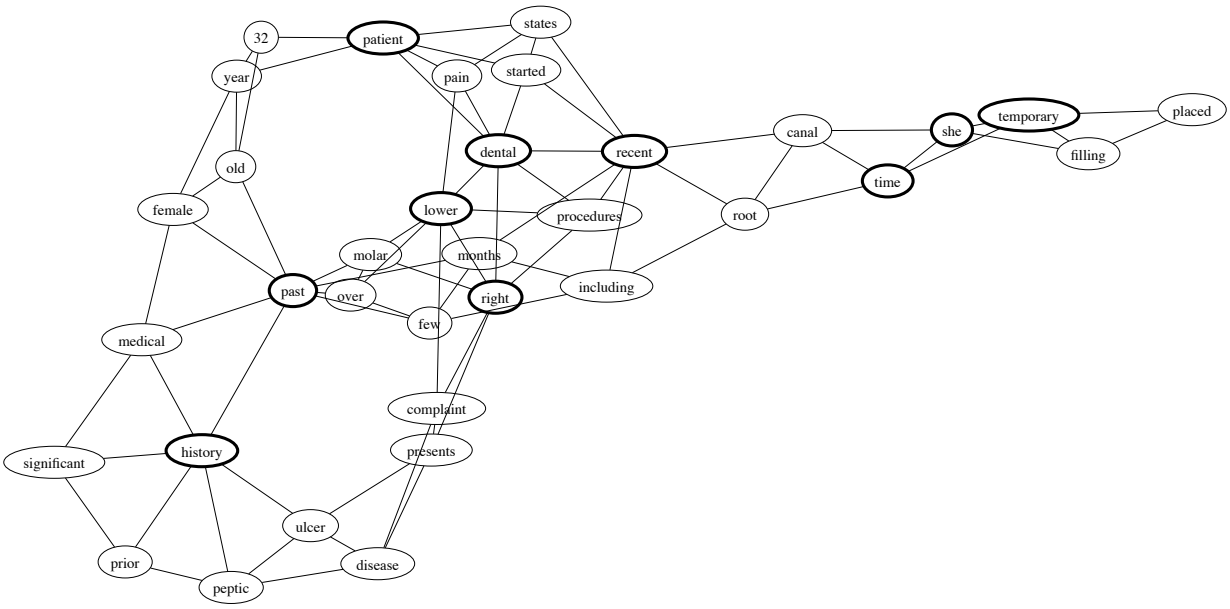
Next, we present an example of the graph produced when this method is applied to a small sample of medical text; this is done to highlight some of the characteristics of graph-based representations. Firstly, an example medical text document is shown in Figure 1(a). From this sample text Figure 1(b) shows the corresponding graph built using a context window of $N = 3$ terms in total. The vertex scoring algorithm of Equation 1 is applied to each vertex and the ten terms with the highest score are highlighted. These include the terms `dental`, `patient` and a number of temporal terms (`history`, `past`, `time`, `recent`). Those terms with higher scores provide an indication of the important terms appearing in this document. The next section shows how this information is included into the retrieval method.

[1]Lemur Project, `http://www.lemurproject.org`.

[2]Other relations may consider grammatical modifiers or part-of-speech information.

"The patient is a 32-year-old female with a past medical history significant for a prior history of peptic ulcer disease who presents with a complaint of right lower dental pain. The patient states that she was started on recent dental procedures, on a right lower molar, over the past few months, including a recent root canal, at which time she had a temporary filling placed."

(a) Example medical text document.



(b) Term graph of example medical text document; stop words removed.

Figure 1: Resulting term graph 1(b) built from the above medical document 1(a). Built using co-occurrence window $N = 3$. Bolded nodes indicate the 10 terms with greatest score within the document (according to Equation 1).

2.2.1 Retrieval Function

The graph-based vertex score of Equation 1 is now integrated into a retrieval function. Typical retrieval functions estimate the relevance between a document and a query as

$$R(d,q) \approx \sum_{t \in q} w(t,q) * w(t,d) \tag{2}$$

where $w(t,q)$ is the weight of the term in query, often uniform for ad-hoc queries, thus $w(t,q) = 1$. The second component, $w(t,d)$, is the weight of the term in the document. The graph-based score provides a means of estimating $w(t,d)$

$$w(t,d) = idf(t) * S(v_i) \tag{3}$$

where $S(v_i)$ is the vertex score from Equation 1 and $idf(t)$ is the inverse document frequency of the term. The retrieval function from Equation 2 can be reexpressed as

$$R(d,q) = \sum_{t \in q} w(t,d) \tag{4}$$

In the next section we apply the graph-based term weighing method to the use of concept-based representations and later show how doing so improves the performance of a medical IR system.

3. GRAPH-BASED CONCEPT WEIGHTING

Building a graph of concepts is done in the same way as building a graph of terms: a context window of fixed length is moved across a document, concepts which co-occur within the context window are connected with an edge in the graph of concepts. Although the process of creating the graph for terms and concepts is the same, the resulting graph itself can differ significantly for the concept representation. To demonstrate this we revisit the example text document and resulting graph from Figure 1. Converting the example text document to concepts and constructing the graph results in the graph illustrated in Figure 2. The concepts are identified by their concept id in both the document and the graph, but we also include their description in parentheses to make the example readable. The PageRank function from Equation 1 is applied and the 10 vertices with the highest scores are highlighted in the figure.

There are many more concepts in the concept graph than terms in the term graph. This is because a single term can map to multiple concepts, for example, the term *HIV* maps to three concepts: `C0019682` *HIV Virus*, `C0019693` *HIV (Disease)* and `C0019699` *HIV+ finding*. Alternatively, multiple terms can map to a single concept, for example, the phrase *Peptic ulcer disease* maps to the single concept `C0030920`.

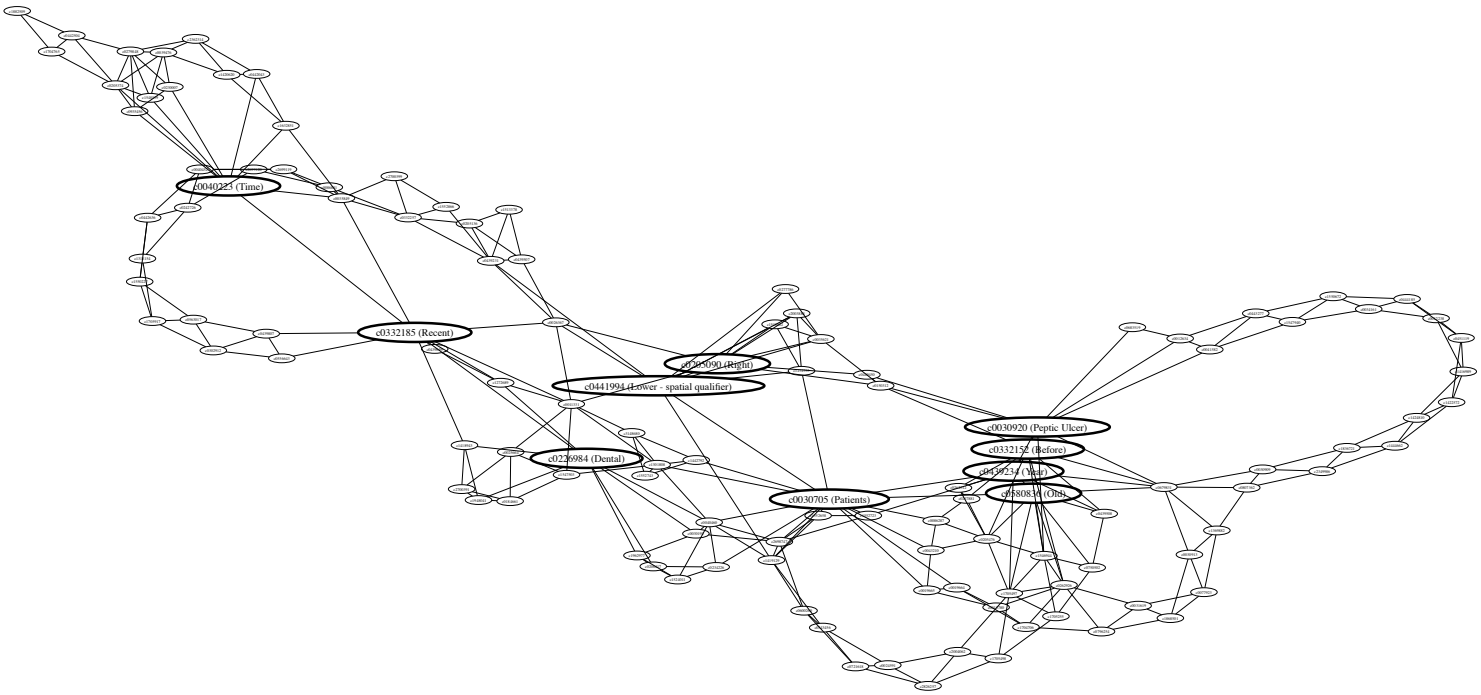


Figure 2: Resulting concept graph built from the medical document from Figure 1(a). Built using co-occurrence window $N = 3$. Bolded nodes indicate the 10 concepts with greatest score within the document (according to Equation 1).

Comparing the term graph from Figure 1 and the concept graph from Figure 2 we observe that both contain similar high score items — `dental` appears in both, as does `patient` and temporal items like `history`, `year`, `recent` and `time`. However the one major difference is the concept `Peptic Ulcer`, which appears in the concept graph, but not in the term graph. The reason for this is twofold: firstly, when converting to concepts the n-gram `peptic ulcer` from the original text maps to the single concept `c0030920`; secondly, when represented in graph form the concept is highly connected and therefore receives a high score. `Peptic Ulcer`'s high score reveals it as an important concept within the concept graph (and therefore this document), a feature not present in the term graph.

3.1 Concept Retrieval Function

Applying the weighting and retrieval functions to concepts we simply substitute terms for concepts. Thus, the original term weighting function from Equation 3 is updated to weight a concept c within document d_c as

$$w(c, d_c) = idf(c) * S(v_i) \tag{5}$$

The original retrieval function is updated to

$$R(d_c, q_c) = \sum_{c \in q_c} w(c, d_c) \tag{6}$$

where d_c is the document converted to concepts and q_c is query converted to concepts.

3.2 Injecting Domain Knowledge into the Weighting Function

The health informatics community has invested considerably in the development of medical domain knowledge resources, for example, the SNOMED CT ontology. These resources describe in great detail[3] the coverage of topics and terminology used within the medical domain. Incorporation of this large external resource into an IR system is not a trivial task. However, if effective integration can be achieved the IR system could potentially make far more informed judgements regarding relevance when presented with a user's query. Towards this goal, this section describes a method for injecting domain knowledge into the weighting function.

The concepts in our concept-based graph model are taken from the SNOMED CT medical ontology. SNOMED CT also defines explicit relationships between concepts, for example the *HIV* virus concept is related to the *AIDS* disease concept. SNOMED CT therefore can also be modelled as a graph, with concepts as vertices and relationships as edges. A concept's number of edges can be an indicator of the concept's importance within the medical domain. Consider the simple example for the concept *Asthma*, which is related to 50 different other concepts, a subset of which are shown in Figure 3.

Concepts important to the medical domain, such as diseases and treatments, are carefully modelled by the designers of SNOMED CT and contain detailed relationships to other concepts. In contrast, concepts that are peripheral to the medical domain are only broadly defined and typically contain only a small number of relationships. In contrast to the *Asthma* example, SNOMED CT defines the concept *Dog*,

[3]SNOMED CT contains approximately 311,000 concepts and 1,360,000 relationships between concepts.

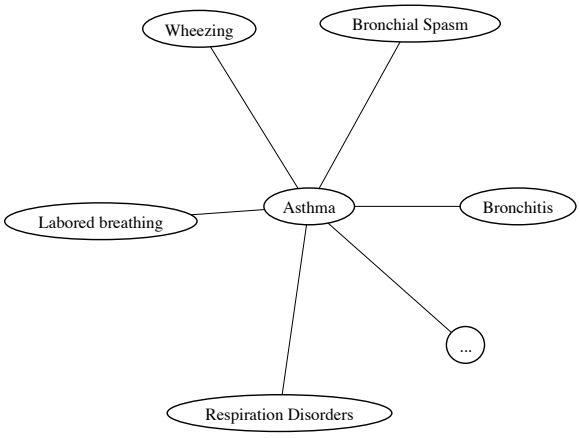


Figure 3: The concept *Asthma* is related to 50 other concepts in the SNOMED CT ontology, an indication of its importance within the medical domain.

which is related to only 5 other concepts — an indication it may be of lesser importance.

Identifying important concepts within the medical domain may provide an indication of what users may be interested in when searching medical documents. We would like to include this indication of importance within the medical domain into our graph-based concepts weighting model. Currently, the concept weighting scheme is based on the number of related concepts within the graph built for a single document. This method captures the importance of the concept within a document, but does not consider the importance of the concept within the wider medical domain. The original concept weight can be adjusted by the number of related concepts within the SNOMED CT ontology, representing its 'background' importance within the medical domain. The weighting function $w(c, d_c)$ of Equation 5 can then be augmented as

$$w(c, d_c) = idf(c) * S(v_i) * \log(|\mathcal{V}_s(c)|) \quad (7)$$

where $\mathcal{V}_s(c)$ is the set of edges adjacent to concept c in the SNOMED CT ontology graph. A concept's weight is therefore adjusted based on its background weight within the medical domain, similar to the way background smoothing is applied in language models based on a term's frequency within the corpus. However, the weighting using SNOMED CT is independent of the document corpus and utilises a global measure of importance for the concept within the medical domain.

4. EMPIRICAL EVALUATION

This section details our experimental setup and evaluation methodology; results are presented in the next section.

4.1 Test Collection

As the test collection we use the TREC 2011 Medical Records Track, a collection of 100,866 clinical record documents from U.S. hospitals. Documents belonging to a single patient's admission were treated as sub-documents and were concatenated together into a single document called a patient *visit* document. This was done because the unit of retrieval in TREC 2011 MedTrack was a patient visit rather than individual report. Collapsing reports to patient visits was a common practise among many TREC MedTrack participants [17]. The corpus then contained 17,198 patient visit documents.

Corpus	#Docs	Avg. doc. len.	#Vocab.
MedTrack:			
Terms	17,198*	2338 terms/doc	218,574
Concepts	17,198*	6066 concepts/doc	54,143

*100,866 original reports collapsed to 17,198 patient *visit* documents.

Table 1: Collection statistics for the TREC 2011 MedTrack corpus of clinical patient records. Statistics are provided for the original term corpus and subsequent corpus after conversion to concepts using the information extraction tool MetaMap.

The original textual documents were translated into concept identifiers using the information extraction system MetaMap, as outlined in Section 2.1[4]. Statistics for both the term and concept corpora are provided in Table 1.

4.2 Baselines for Comparison

We implement a number of baselines for comparison against our graph-based concept weighting model:

terms-tfidf: We consider a state-of-the-art bag-of-words model. In initial experiments a tf-idf implementation actually demonstrated the best performance over BM25 and a Language Model with Dirichlet smoothing. Thus, we adopt as a baseline the Lemur variant implementation of tf-idf (which uses the Okapi TF formula [18]; parameterising document length normalisation with b and term frequency weighting with $k1$). This baseline was tuned by selecting the best performing (oracle) pair of parameters values for b and $k1$ from a complete sweep of the parameter space in the ranges $b = [0, \ldots, 1]$ (with increments of 0.1) and $k1 = [0, \ldots, 40]$ (with increments of 1). The best values were $b = 0.45$ and $k1 = 3.7$. This strong tf-idf tuned baseline is denoted terms-tfidf.

terms-graph: We implemented Blanco & Lioma's graph-based weighting method and apply it to terms. The damping factor parameter ϕ from Equation 1 is set to 0.85 according to the findings of Blanco & Lioma [3]. Similarly, the number of iterations and the context window size were set at 20 and 10 respectively, in line with Blanco & Lioma. This baseline is denoted terms-graph.

concepts-tfidf: We implement a bag-of-concepts model as the same tf-idf model as for terms-tfidf, but on the concepts corpus (as opposed to the term corpus). Parameters for this baseline were tuned in the same manner as terms-tfidf; $b = 0.35$, $k1 = 5.0$. This tuned baseline is denoted concepts-tfidf.

[4]Koopman et al. found that mapping to the SNOMED CT subset of UMLS provided the best representation, we also adopt this approach [7].

4.3 Graph-based Concept Weighting Models

concepts-graph: We apply the graph-based weighting method to concepts, as described in Section 3.1. We use the same parameter settings as terms-graph for ϕ, iterations and context window. This model is denoted concepts-graph.

concepts-graph-snomed: Background information, derived from the SNOMED CT ontology, is injected into the concepts-graph weighting as described in Section 3.2, maintaining the same parameter settings. This model is denoted concepts-graph-snomed.

4.4 Evaluation Topics & Metrics

Evaluation was performed using the 34 topics from the TREC MedTrack'11 collection. Retrieval results were evaluated using Bpref and Precision @ 10 in accordance with the measures from TREC MedTrack'11. Bpref is regarded as the primary metric by MedTrack'11 and was used as the objective measure to tune the baselines terms-tfidf and concepts-tfidf.

4.5 Results

Retrieval results of the three baselines and the two graph-based concept methods are reported in Table 2.

Run	Bpref	Prec@10
terms-tfidf	0.4722	0.4882
concepts-tfidf	0.4993	0.5176
terms-graph	0.4393	0.4882
concepts-graph	0.5050 (+15%)	0.5441 (+11%)
concepts-graph-snomed	**0.5245** (+19%)	**0.5559** (+14%)

Table 2: Retrieval results on TREC MedTrack'11 using both term and concept representations, and after applying graph-based weighting and injection of domain knowledge. Percentage improvement shown over terms-graph.

Comparing the term and concepts runs (terms-tfidf vs. concepts-tfidf), the concept based representation demonstrates improved performance. Comparing the effect of the graph-based weighting on terms (terms-tfidf and terms-graph) we actually observed degraded performance. However, when concepts are used to construct the graph (concepts-tfidf and concepts-graph), performance improved. The injection of domain knowledge using SNOMED CT (concepts-graph-snomed) provided additional improvements over concepts-graph in both bpref and precision. Analysis of results is presented in the next section.

Statistical significance using paired t-test was not found for any of the above results. The test collection contained only 34 query topics; van Rijsbergen comments that paired t-test may not reliably indicate statistical significance with small query sets [15]. Ideally, a larger query set or additional test collections would have been used; however, the medical domain does not currently have the diversity of evaluation resources available to other domains.

5. DISCUSSION

First, we consider the effect of using a bag-of-concepts rather than a bag-of-words representation — comparing the concepts-tfidf and terms-tfidf baselines. The use of a concept-based representation provides a 5% increase in bpref and 6% increase in P@10. This result is inline with previous concept-based approaches [7] and is encouraging for applying graph-based weighting to concept-based representations.

The effect of Blanco & Lioma's graph-based *term* weighting is now considered. When comparing the terms-tfidf and terms-graph baselines we observe that the use of graph weighting actually degraded retrieval performance by 6%. This result is contrary to the findings of Blanco & Lioma [3], who report improvements using the graph model on a number of test collections (over both tf-idf and BM25 baselines). Their corpora were newswire articles, web and blog crawls. The graph-based term weighting method may not be as suited to the peculiarities of medical IR; further analysis would be required to fully understand the reason for this.

In contrast to using terms, applying graph-based weighting to concepts *does* improve performance. Our concepts-graph model shows improvements over both the terms-tfidf and concepts-tfidf baselines, especially in precision, which exhibits an 11% improvement over the tuned terms-tfidf baseline and a 5% improvement over the tuned concept-tfidf baseline. Graph-based weighting is effective when using concepts, but not so when using terms. We hypothesise that this may be due to the fact that the concept representation encapsulates important medical n-grams as a single vertex in the graph (such as the Peptic Ulcer example from the concept graph of Figure 2). In contrast, the term-based graph does not encode these n-grams: instead, the two terms are split as separate vertices, both receiving a lower weight.

Overall, both the graph-based concept weighting methods (concepts-graph and concepts-graph-snomed), outperform the other three baselines in both bpref and precision @ 10. Although the small topic set makes statistical significance judgements difficult, we can provide some insights by considering how many queries were improved (and by how much) when using our concept graph method. Figure 5 show the change in bpref for each query using the concept-graph-snomed model when compare against the terms-graph baseline; topics ordered in decreasing change in bpref. The figure shows what

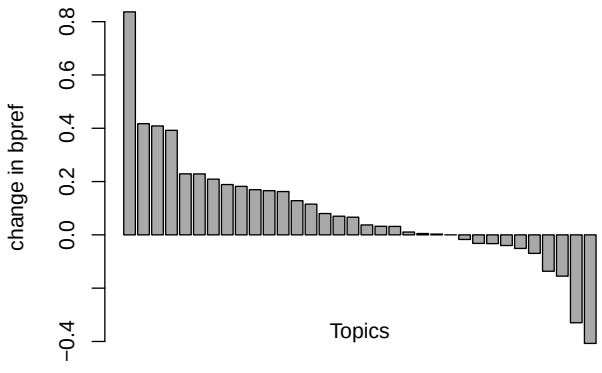


Figure 4: Per-query change in bpref for concept-graph-snomed against terms-graph baseline.

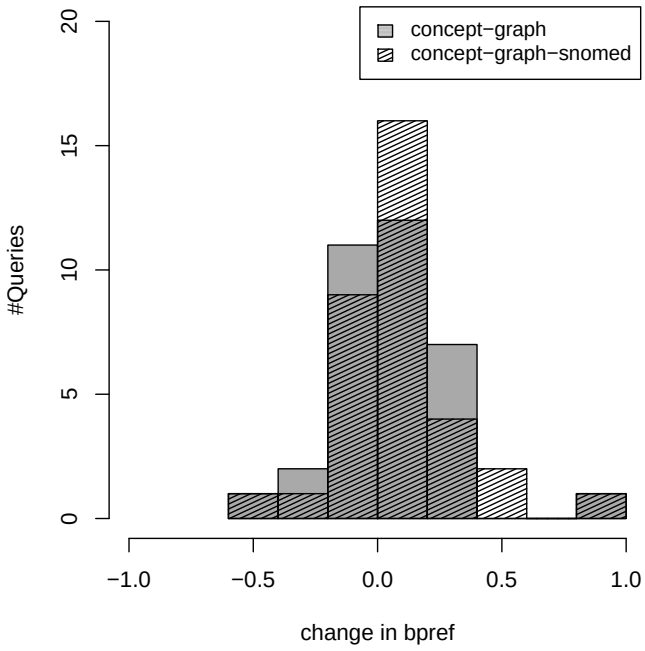


Figure 5: Histogram showing #queries exhibiting change in bpref over term-graph for both concept graph models. Results show concepts-graph-snomed tends to make more small improvement to many queries — an indicator of increased robustness.

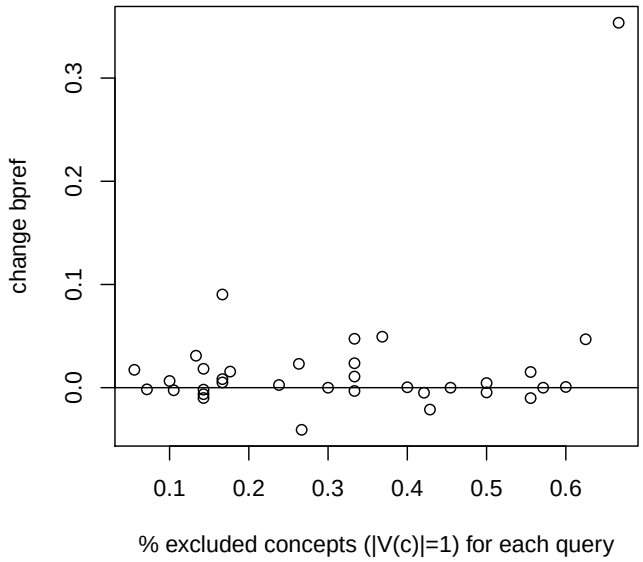


Figure 6: The change in bpref when excluding query concepts with only one edge in the SNOMED CT graph. x-axis indicates the percentage of concepts for a given query where $|\mathcal{V}_s(c)| = 1$ (and are therefore excluded).

a significance test should show — that change is seen in most queries, and change is for the better in most cases.

When comparing concept-graph-snomed to concept-graph, the injection of domain knowledge using SNOMED CT into the weighting provides an improvement in both bpref (4%) and precision (2%). Although the overall performance after injecting domain knowledge is not considerably higher, the injection method does provide some additional robustness across the query set. To illustrate this, Figure 5 shows the number of queries exhibiting change in bpref over the terms-graph baseline for both concept graph models. The histogram shows that concept-graph-snomed tends to make small variations (gains and losses) to a larger number of queries, whereas the concepts-graph has larger variations on a smaller number of queries. The former (small gains on many queries) indicates increased robustness and is more desirable for the model's general applicability. Both graph concept models do have the promising potential to benefit some queries substantially. Further study is need to enhance this aspect.

We now consider some interesting characteristics of the injection of domain knowledge. From Equation 7, the weighting of concept c is dependent on the logarithm of the number of edges adjacent to c in the SNOMED CT graph. Note, that when a concept has only one adjacent edge in the SNOMED CT graph, then the weight w_b of query concept c for document d is zero ($\log|\mathcal{V}(c)| = \log 1 = 0$). In practice, this means that query concepts that contain only one edge in SNOMED CT are essentially ignored (their weight always being 0). Intuitively, this seems an undesirable characteristic that could lead to significant degradation in performance. To understand the extend of this characteristic and how it actually affects performance, we first consider how many queries contain concepts with only one edge in SNOMED CT (and therefore had scores of 0). The 34 test queries contained 448 concepts in total; of these a total of 127 (28%) had only one edge in the SNOMED CT graph, and were therefore ignored. Intuitively, ignoring so many concepts in the topic set would have a drastic effect on retrieval performance; however empirical results show the contrary. This is confirmed by Figure 6, which compares the change in bpref after applying the SNOMED CT weighting against the percentage of concepts within a given query where $|\mathcal{V}(c)| = 1$. Points on the far right of the x-axis indicate queries where many concepts were excluded from the weighting function. Note, that every query has at least one query concept excluded after applying SNOMED CT weighting. In addition, even when large portions of concepts are excluded from the query (far right of the x-axis) there are still positive changes in bpref. These queries contained a large number of concepts which were deemed as peripheral to the medical domain and, when excluded, aided performance.

Rather than completely exclude concepts with $|\mathcal{V}_s(c)| = 1$ we did perform experiments with alternative approaches that instead of excluding the concept, simply assigned a logarithmic scaled weight (e.g., $1 + log(|\mathcal{V}_s(c)|)$ or $log(1 + |\mathcal{V}_s(c)|)$). However, the best results in bpref were obtained when query concepts with only one adjacent edge in SNOMED CT were completely excluded. We conclude that a concept's lack of connectedness to other concepts in the domain ontology indicates they provide no additional information for the query and, in fact, may be misleading.

The exclusion of certain concepts based on the SNOMED CT connectedness is in effect a form of *query reduction*. Query reduction has been considered by researchers in information retrieval; finding an ideal subset of query terms can result in substantial performance gains [8, 2]. Kumaran

& Carvalho adopted a learning to rank approach that used statistical predictors (such as IDF, tf, Mutual Information and Query Clarity) to find an optimal query subset — they found an upper bound of 30% increase in performance, but their predictors only provided an 8% increase [8]. Bendersky & Croft [2] made use of corpus based statistics (such as IDF) and corpus independent indicators, such as Google n-grams, to identify and weight 'key concepts' within the query. They show improvements in retrieval, but found no robust feature across different test collections. We have shown that the use of a concept's connectedness in the SNOMED CT ontology provides an indicator of importance; in practice, providing a good feature for the implementation of an implicit query reduction method. Unlike previous approaches, our method used only one feature and avoided the use of heavy-weight machine learning to find an optimum feature combination; we also introduce no additional parameters. An interesting avenue of future work from this study is to consider query reduction specific to medical information retrieval, especially given the rich amount of domain knowledge available in resources such as SNOMED CT.

Finally, the findings of this study are applicable outside of the medical domain, specifically the injection of domain knowledge representing the importance of a term outside of the corpus being indexed. We currently use connectedness in SNOMED CT as the indicator of importance. Alternative weighting could be applied based on connectedness within any other resource represented as a graph, including domain specific resources, or general resources such as WordNet.

6. CONCLUSION

This paper presents a graph-based method to weight medical concepts found in documents for the purpose of medical IR. Graph-based representations are chosen over bag-of-words representations because they capture the relationships that exist between concepts, a feature important for capturing the innate dependencies in medical natural language. Additionally, concept-based representations are used to overcome vocabulary mismatch and to encapsulate important n-grams into a single concept.

We adapt previous graph-based term weighting method and apply them to concepts; a concept's weight is based on its PageRank score within the document. In addition, we present a novel method for the injection of domain knowledge regarding the concept's importance within the wider medical domain (not just the corpus itself). This method has an interesting characteristic of excluding a large number of query terms, resulting in a form of query reduction, and surprisingly leads to improvements in performance.

Evaluation was done on the TREC Medical Records track and a number of strong baselines were provided for comparison. Results showed that our graph-based concept weighting method outperforms each of the baselines.

The graph-based concept weighting method offers a framework for integrating formal background knowledge, often locked in medical domain ontologies, into data-driven approaches typical of information retrieval.

7. REFERENCES

[1] A. R. Aronson and F.-M. Lang. An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236, 2010.

[2] M. Bendersky and W. B. Croft. Discovering key concepts in verbose queries. In *Proceedings of the 31st annual International ACM SIGIR conference on research and development in information retrieval (SIGIR)*, pages 491–498, New York, NY, USA, 2008. ACM.

[3] R. Blanco and C. Lioma. Graph-based term weighting for information retrieval. *Information Retrieval*, 15(1):1–39, 2012.

[4] T. E. Doszkocs, J. Reggia, and X. Lin. Connectionist models and information retrieval. *Annual review of information science and technology*, 25:209–262, 1990.

[5] O. Egozi, S. Markovitch, and E. Gabrilovich. Concept-Based Information Retrieval using Explicit Semantic Analysis. *ACM Transactions on Information Systems*, 29(2):1–38, 2011.

[6] W. Hersh. *Information retrieval: a health and biomedical perspective*. Springer Verlag, New York, 3rd edition, 2009.

[7] B. Koopman, P. Bruza, L. Sitbon, and M. Lawley. Towards Semantic Search and Inference in Electronic Medical Records: an approach using Concept-based Information Retrieval. *Australasian Medical Journal: Special Issue on Artificial Intelligence in Health*, 5(9):482–488, 2012.

[8] G. Kumaran and V. R. Carvalho. Reducing long queries using query quality predictors. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 564–571, NY, USA, July 2009. ACM.

[9] Z. Liu and W. W. Chu. Knowledge-based query expansion to support scenario-specific retrieval of medical free text. *Information Retrieval*, 10(2):173–202, Jan. 2007.

[10] A. N. Nguyen, M. J. Lawley, D. P. Hansen, R. V. Bowman, B. E. Clarke, E. E. Duhig, and S. Colquist. Symbolic rule-based classification of lung cancer stages from free-text pathology reports. *Journal of the American Medical Informatics Association*, 17(4):440–445, 2010.

[11] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: bringing order to the web. *Technical Report, Stanford Digital Library Technologies*, 1999.

[12] C. Patel, J. Cimino, J. Dolby, A. Fokoue, A. Kalyanpur, A. Kershenbaum, L. Ma, E. Schonberg, and K. Srinivasclass. Matching patient records to clinical trials using ontologies. *The Semantic Web*, 4825:816–829, 2007.

[13] D. Ravindran and S. Gauch. Exploiting hierarchical relationships in conceptual search. In *Proceedings of the 13th annual international ACM CIKM conference on information and knowledge management*, pages 238–239. ACM, 2004.

[14] H. Turtle and W. B. Croft. Evaluation of an inference network-based retrieval model. *ACM Transactions on Information Systems*, 9(3):187–222, July 1991.

[15] K. van Rijsbergen. *Information Retrieval*. Butterworth & Co, London, 2 edition, 1979.

[16] E. M. Voorhees. Query expansion using lexical-semantic relations. In *Proceedings of the 17th annual international ACM SIGIR conference on research and development in information retrieval*, pages 61–69, Dublin, Ireland, 1994. ACM.

[17] E. M. Voorhees and R. M. Tong. Overview of the TREC 2011 Medical Records Track. In *Proceedings of the Twentieth Text REtrieval Conference (TREC 2011)*, Gaithersburg, Maryland, USA, Nov. 2011.

[18] C. Zhai. Notes on the Lemur TFIDF model. Technical report, School of Computer Science, Carnegie Mellon University, 2001.

[19] W. Zhou, C. Yu, N. Smalheiser, V. Torvik, and J. Hong. Knowledge-intensive conceptual retrieval and passage extraction of biomedical literature. In *Proceedings of the 30th annual international ACM SIGIR conference on research and development in information retrieval*, pages 655–662, New York, USA, 2007. ACM.