Exploiting Medical Hierarchies for Concept-based Information Retrieval

Guido Zuccon[1], Bevan Koopman[1,2], Anthony Nguyen[1], Deanne Vickers[1], Luke Butt[1]

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

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

{guido.zuccon, bevan.koopman, anthony.nguyen, deanne.vickers, luke.butt}@csiro.au

ABSTRACT

Search technologies are critical to enable clinical staff to rapidly and effectively access patient information contained in free-text medical records. Medical search is challenging as terms in the query are often general but those in relevant documents are very specific, leading to granularity mismatch.

In this paper we propose to tackle granularity mismatch by exploiting subsumption relationships defined in formal medical domain knowledge resources. In symbolic reasoning, a subsumption (or 'is-a') relationship is a parent-child relationship where one concept is a subset of another concept. Subsumed concepts are included in the retrieval function. In addition, we investigate a number of initial methods for combining weights of query concepts and those of subsumed concepts. Subsumption relationships were found to provide strong indication of relevant information; their inclusion in retrieval functions yields performance improvements. This result motivates the development of formal models of relationships between medical concepts for retrieval purposes.

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, Subsumption, SNOMED CT

1. INTRODUCTION

Search technologies that enable clinical staff to rapidly and effectively search patients health records may improve health outcomes as well as produce time and costs savings [3]. However, searching medical records can be challenging: keyword based approaches often fail to identify medical entities that are referred to with different terms, such as the synonymous terms 'heart attack' and 'myocardial disorder'. Concept-based retrieval approaches have been proposed to overcome keyword search problems [5]. In these approaches, the original free-text documents are converted to concepts defined in medical ontologies, such as the SNOMED CT ontology.

Mismatch in granularity between concepts in a query and those found in relevant documents may however hinder retrieval effectiveness. For example, a medical record document may contain detailed notes about the brand and dosage of drugs prescribed to a patient, whereas a query would contain only the general class of drugs or its active ingredient. Previous concept-based approaches are susceptible to granularity mismatch.

Within ontologies, concepts are organised in inheritance hierarchies, with parent-child, or *subsumption*, relationships. For example, the hierarchy for *Opiate* in the SNOMED CT ontology is shown in Figure 1. The figure shows that the different types of Opiate are subsumed by the parent Opiate. In a retrieval scenario, documents that contained these subsumed concepts would likely be relevant to a query that contains Opiate. Subsumption relationships are not accounted for in most current concept-based approaches for medical records information retrieval; successful use of subsumption has been shown in related domains, e.g. [2].

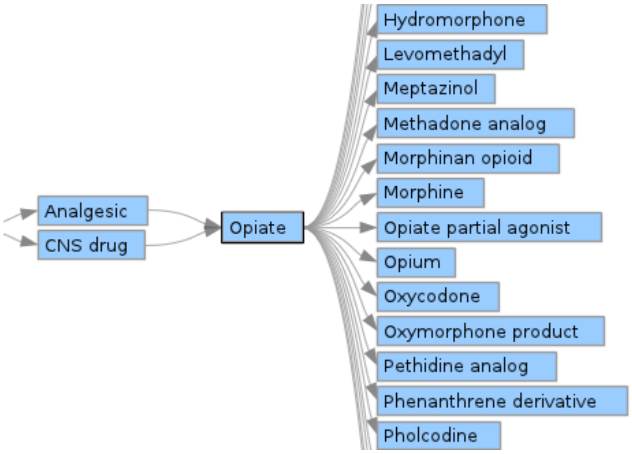

Figure 1: SNOMED CT hierarchy for the class of drug *Opiate*.

ADCS'12, December 05 – 06 2012, Dunedin, New Zealand.
Copyright 2012 ACM 978-1-4503-1411-4/12/12 ...$15.00.

We hypothesise that accounting for subsumption between concepts in retrieval methods may allow for higher effectiveness in medical search. To this end, we provide an initial empirical investigation of the use of subsumption to enhance medical information retrieval. In the experiments, we consider the scenario modelled by the TREC Medical Records Track, where a health practitioner consults a collection of electronic health records to individuate cohorts suitable for participating to a clinical study.

Subsumption information is taken from the SNOMED CT hierarchy and included in the retrieval function. In addition, a number of initial methods for combining weights of query concepts and those of subsumed concepts are evaluated. Empirical results demonstrate that subsumed concepts provide useful information that can be used to improve retrieval effectiveness.

2. METHODS

Following the work of Koopman et al. [4, 5], we implement a 'bag-of-concepts' representation of documents rather than the traditional bag-of-words. Terms are transformed to concepts using the natural language processing tool Metamap; concepts are derived from the SNOMED CT ontology.

Documents are scored according to (1) the weight of query concepts in a document, and (2) the weight of concepts in a document that have been subsumed by a query concept. For each query concept c_i we obtain the list of subsumed concepts $c_j \prec c_i$ from the SNOMED CT ontology. These subsumed concepts are included in the retrieval function, leading to the following retrieval status value (RSV):

$$RSV(d|q) = \sum_{c_i \in q} w(c_i, d) + \sum_{c_j \prec c_i; c_i \in q} \delta(w(c_j, d)) \qquad (1)$$

where $w(c_i, d)$ is the weight of concept c_i in document d, and $\delta(w(c_j, d))$ adjusts the weight of a subsumed concept c_j. That is, the score of a document for a query q is the sum of the weights associated with the query concepts and the adjusted weights of the concepts that are subsumed by the query concepts.

Equation 1 is a general method to integrate subsumed concepts into the retrieval function. A number of instantiations of both $w(c_i, d)$ and $\delta(w(c_j, d))$ are possible.

In the following we outline some possible variations of both type of functions; these are then empirically evaluated in Section 3.

2.1 $\mathbf{w(c_i, d)}$: Weighting Concepts

Next, we consider a number of possible instantiations of the weighting function $w(c_i, d)$. As overarching weighting schema, we used variations of tf-idf as Koopman et al. found that in the medical domain this often yields higher retrieval performance than alternative approaches, such as BM25 and language models [5].

cfidf: this corresponds to the normalised tf-idf weighting schema[1], where concepts are used instead of terms, i.e.:

$$w(c_i, d)_{\mathsf{cfidf}} = \frac{count(c_i, d)}{l_d} \cdot \log \frac{|D|}{|d(c_i)|} \qquad (2)$$

and $count(c_i, d)$ is the frequency of concept c_i in document d, l_d is the length of document d, $|D|$ is the total number of documents in the collection and $|d(c_i)|$ is the number of documents that contain concept c_i.

ncfidf: in this instantiation a concept frequency is normalised by its frequency in the collection (i.e. the maximum likelihood estimation is used for the concept frequency component), i.e.:

$$w(c_i, d)_{\mathsf{ncfidf}} = \frac{count(c_i, d)}{count(c_i)} \cdot \log \frac{|D|}{|d(c_i)|} \qquad (3)$$

and $count(c_i)$ is the frequency of concept c_i in the collection.

ecfidf: this corresponds to the enhanced tf-idf described by Zhai [8] in which the Okapi formula is used for weighting term frequencies and where concepts are used instead of terms, i.e.:

$$w(c_i, d)_{\mathsf{ecfidf}} = \frac{k_1 count(c_i, d)}{count(c_i, d) + k_1(1 - b + b\frac{l_d}{l_{avg}})} \cdot \log \frac{|D|}{|d(c_i)|} \qquad (4)$$

and l_{avg} is the average document length, and k_1, b are the Okapi parameters.

2.2 $\delta(\mathbf{w(c_j, d)})$: Integrating Subsumption

Next, we consider how the weight of a concept should be adjusted if it was subsumed by the query.

A straightforward approach would to treat subsumed concepts in the same way as query concepts, i.e. $\delta(w(c_j, d)) = w(c_j, d)$. We call this approach linear.

However, the presence of a subsumed concept in a document may offer a different indication of relevance than an actual query concept. A subsumed concept indicates a specialisation of the parent concept, and thus treated differently to an actual query concept. Intuitively a subsumed concept would be a weaker indication of relevance than a query concept. Alternatively, a subsumed concept may be a stronger indication of relevance because it is an actual specialisation of the more general concept used in the query as it is more focused and less ambiguous. To this end, we consider a number of instantiations of $\delta(w(c_j, d))$ that encompass the two alternative rationales.

sqrt($\mathbf{w(c_j, d)}$): the weight for the subsumed concept c_j in the document is adjusted according to the square root of the weight $w(c_j, d)$, i.e.:

$$\delta(w(c_j, d)) = \sqrt{w(c_j, d)} \qquad (5)$$

In this case a subsumed concept contributes less evidence towards the score of a document than a query concept.

log($\mathbf{w(c_j, d)}$): the weight for the subsumed concept c_j in the document is the logarithm of the weight $w(c_j, d)$, i.e.:

$$\delta(w(c_j, d)) = \log[w(c_j, d)] \qquad (6)$$

If $w(c_j, d)$ is less than one, then the subsumed concept receives a negative weight[2]. In this case, the weight of a subsumed concept but be considerably higher than that of a query concept to influence the score.

[1] Note, the standard tf-idf weighting (no document length normalisation) performed significantly worse.

[2] We excluded the case $w(c_j, d) = 0$ to avoid $\log[w(c_j, d)] = -\infty$; in this case a zero weight is assigned to $\mathsf{log}(\mathbf{w(c_j, d)})$.

pow($\mathbf{w}(\mathbf{c_j}, \mathbf{d})$): the weight for the subsumed concept c_j in the document is the square of the weight $w(c_j, d)$, i.e.:

$$\delta(w(c_j, d)) = [w(c_j, d)]^2 \tag{7}$$

In this instantiation, more weight (and thus importance) is given to subsumed concepts rather than query concepts.

exp($\mathbf{w}(\mathbf{c_j}, \mathbf{d})$): the weight for the subsumed concept c_j in the document is the (natural) exponential function of the weight $w(c_j, d)$, i.e.:

$$\delta(w(c_j, d)) = e^{w(c_j,d)} \tag{8}$$

Here, subsumed concepts become the main influence of the document's score.

3. EMPIRICAL EVALUATION

3.1 Baselines

To understand the empirical merits of using subsumption information to retrieve medical documents, we compare approaches that score query concepts and their subsumed concepts against approaches that do not consider subsumed concepts. The baseline using no subsumption is indicated no sub., i.e. $\delta(w(c_j, d)) = 0$. Where applicable, parameters were set to the common Okapi values[3].

3.2 Test Collection

We use the TREC 2011 Medical Records Track, a collection of 100,866 clinical record documents taken from U.S. hospitals. Documents belonging to a single patient's admission were concatenated together into a single document called a patient visit document; this is consistent with the unit of retrieval used TREC 2011 MedTrack and collapsing reports to patient visits was a common practise among many TREC MedTrack systems[4].When documents are grouped into visits, the corpus then contains 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 MedTrack'11 corpus of clinical records. Statistics are provided for the original term corpus and subsequent corpus after conversion to SNOMED CT concepts.

The original free-text documents were translated into concept identifiers from the SNOMED CT medical terminology using the information extraction system MetaMap, as suggested by Koopman et al. [5]. Statistics for both the term and concept corpora are provided in Table 1.

The 34 topics from the TREC MedTrack'11 collection were used in the experiments. Retrieval results were evaluated using Bpref and Precision @ 10 in accordance with TREC MedTrack'11. Because the absolute number of judged documents per topic is small, the computation of metrics such as MAP, nDCG, etc. is not meaningful.

3.3 Results

Table 2 outlines the results of the investigated approaches. Results show the effect of different combinations of methods for weighting concepts and adjust the weights of subsumed concepts. For each concept weighting method, the best performances are highlighted in bold. No statistical significant differences are measured between variations of $\delta(w(c_j, d))$ and the corresponding baselines (i.e. no sub.)

Further discussion of the results obtained when considering the concept-based representation and subsumption follows in the next section.

4. CONTRIBUTION OF SUBSUMPTION

The empirical results demonstrate that subsumption relationships supply strong relevant information that can lead to effective retrieval performance.

The use of only subsumed concepts to score documents (sub. only), thereby ignoring matching the query concepts, obtains mixed results based on the employed weighting schemas. However, none of these sensibly improve the corresponding concept baseline.

It is instead when the contribution of subsumed concepts is combined with that of matching query concepts that promising improvements of retrieval performance are witnessed. Specifically, ecfidf used in combination with sqrt($\mathbf{w}(\mathbf{c_j}, \mathbf{d})$) yields the best Bpref values in our experiments. Whereas, ecfidf used in combination with the linear approach yields the highest P@10.

However, no one approach is found that performs the best across the different weighting method $w(c_i, d)$. For example, while using exp($w(c_j, d)$) to weight subsumed concepts obtained the best retrieval performance with cfidf, results obtained with other instantiations of $w(c_i, d)$ do not follow this trend. In particular, when ecfidf is considered, increasing the subsumed concepts' weights using the exponential function actually yields lower B-pref and P@10 than all the other subsumed concept weighting methods.

When ncfidf and ecfidf are considered, both the linear and sqrt($w(c_j, d)$) approaches for adjusting the weights of subsumed concepts yield improvements over the respective concept baselines[5]. But the best function to apply to adjust the weights of subsumed concepts is unclear.

Other approaches for $\delta((w(c_j, d))$ perform lower than the concept baseline (no sub.), with the exception of log($\mathbf{w}(\mathbf{c_j}, \mathbf{d})$) when ecfidf is used: performance increments here are however minimal.

Parallels can be drawn between our approaches, that combine query concepts and subsumed concepts, and the query expansion process [1]. These are similar because they both

[3] $b = 0.75$, $k_1 = 1.2$

[4] `http://trec.nist.gov/pubs/trec20/t20.proceedings.html`

[5] Except the combination of ecfidf and simple, which yields a Bpref lower than that of the concept baseline.

			$\delta(\mathbf{w}(\mathbf{c_j},\mathbf{d}))$					
		no sub.	sub. only	linear	sqrt($\mathbf{w}(\mathbf{c_j},\mathbf{d})$)	log($\mathbf{w}(\mathbf{c_j},\mathbf{d})$)	pow($\mathbf{w}(\mathbf{c_j},\mathbf{d})$)	exp($\mathbf{w}(\mathbf{c_j},\mathbf{d})$)
$\mathbf{w}(\mathbf{c_i},\mathbf{d})$	cfidf	.3943	.2002	.4080	.4216	.3805	.3791	**.4330**
		.2500	.3147	.3088	.3647	.3500	.2324	**.4206**
	ncfidf	.4430	.4440	**.4544**	.4447	.3831	.4440	.4296
		.3765	.3765	.4265	**.4353**	.3441	.3765	.4176
	ecfidf	.4799	.4691	.4789	**.4814**	.4800	.4691	.4469
		.4941	.4265	**.5147**	.5029	.5000	.4265	.3118

Table 2: Results obtained by the weighing approaches on TREC MedTrack'11, where $\mathrm{w}(\mathrm{c_i},\mathrm{d})$ refers to instantiation of the weighting function for query concepts, and $\delta(\mathrm{w}(\mathrm{c_j},\mathrm{d}))$ refers to instantiations of the weighing function for concepts subsumed by query concepts. For each weighting schema, the first row of results reports the measured Bpref values; the second row reports the corresponding P@10 values. The column labelled no sub. reports the performance of the approaches that do not consider subsumed concepts. The column labelled sub. only refers to results obtained when weighting subsumed concepts only, thus ignoring query concepts.

score documents against the original query and an additional set of terms (concepts) derived from the initial request. However, most query expansion approaches do not weight the expanded terms; weighted query expansion is less common than its unweighted version. In addition, most query expansion techniques rely on corpus statistics to select candidate terms for expansions, and a threshold or limit on number of candidate terms is usually employed. In the approaches proposed in this paper, instead, concepts other than those in the query are selected because their relationship with a query concept present in a document is formally encoded in a domain knowledge source. Corpus statistics are thus used for the weighting process, not for the selection process. In addition, no limit is imposed on the number of additional concepts that are considered when scoring documents, the number of additional concepts is taken from the number of subsumed concepts for a query concept.

5. CONCLUSIONS

This work is an initial investigation on the use of subsumption for concept-based medical information retrieval. Empirical results have shown potential increase in retrieval performance when considering the matching between documents and subsumed concepts alongside with query concepts. The approaches investigated in this paper were based on functions that combine weights of query concepts with those of subsumed concepts; functions that adjust the latter weights were also explored. The best performance was highly dependent either on the specific tf-idf variation considered, or on the specific function used to distinguish the contribution of subsumed concepts, or both. No single approach has provided strong, consistent gains over the concept baselines. How to best combine weights for query concepts and subsumed concepts is an open line of research, but this paper demonstrated promising initial results.

Future work will be directed towards the creation of formal models able to capture the two different matching mechanisms. Specifically, these models may take advantage of additional information regarding the subsumption concepts, for example the distance between the child subsumed concept and the parent concept or the extent the concepts are semantically related [7, 6].

6. REFERENCES

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

[2] M. Douyère, L. Soualmia, A. Névéol, A. Rogozan, B. Dahamna, J. Leroy, B. Thirion, and S. Darmoni. Enhancing the MeSH thesaurus to retrieve French online health resources in a quality-controlled gateway. *Health Information & Libraries Journal*, 21(4):253–261, 2004.

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

[4] B. Koopman, P. Bruza, L. Sitbon, and M. Lawley. AEHRC & QUT at TREC 2011 Medical Track : a concept-based information retrieval approach. In *20th Text REtrieval Conference (TREC 2011)*, pages 1–7, Gaithersburg, MD, USA, Nov. 2011. NIST.

[5] 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.

[6] B. Koopman, G. Zuccon, P. Bruza, L. Sitbon, and M. Lawley. An Evaluation of Corpus-driven Measures of Medical Concept Similarity for Information Retrieval. In *21st ACM International Conference on Information and Knowledge Management (CIKM)*, Maui, USA, 2012.

[7] T. Pedersen, S. V. S. Pakhomov, S. Patwardhan, and C. G. Chute. Measures of semantic similarity and relatedness in the biomedical domain. *Journal of biomedical informatics*, 40(3):288–99, June 2007.

[8] C. Zhai. Notes on the Lemur TFIDF model. Technical report, School of Comp. Sci., CMU, 2001.