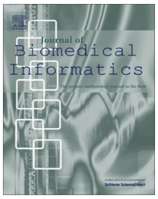
Journal of Biomedical Informatics 46 (2013) 857–868

Contents lists available at SciVerse ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin



An ontology-based similarity measure for biomedical data – Application to radiology reports

Thusitha Mabotuwana [*], Michael C. Lee, Eric V. Cohen-Solal

Philips Research North America, 345 Scarborough Road, Briarcliff Manor, NY 10510, USA

ARTICLE INFO

Article history:
Received 15 January 2013
Accepted 28 June 2013
Available online 11 July 2013

Keywords:
Document similarity comparison
Radiology informatics
Semantic similarity

ABSTRACT

Background: Determining similarity between two individual concepts or two sets of concepts extracted from a free text document is important for various aspects of biomedicine, for instance, to find prior clinical reports for a patient that are relevant to the current clinical context. Using simple concept matching techniques, such as lexicon based comparisons, is typically not sufficient to determine an accurate measure of similarity.

Methods: In this study, we tested an enhancement to the standard document vector cosine similarity model in which ontological parent–child (*is-a*) relationships are exploited. For a given concept, we define a *semantic vector* consisting of all parent concepts and their corresponding weights as determined by the shortest distance between the concept and parent after accounting for all possible paths. Similarity between the two concepts is then determined by taking the cosine angle between the two corresponding vectors. To test the improvement over the non-semantic document vector cosine similarity model, we measured the similarity between groups of reports arising from similar clinical contexts, including anatomy and imaging procedure. We further applied the similarity metrics within a *k*-nearest-neighbor (*k*-NN) algorithm to classify reports based on their anatomical and procedure based groups. 2150 production CT radiology reports (952 abdomen reports and 1128 neuro reports) were used in testing with SNOMED CT, restricted to Body structure, Clinical finding and Procedure branches, as the reference ontology.

Results: The semantic algorithm preferentially increased the intra-class similarity over the inter-class similarity, with a 0.07 and 0.08 mean increase in the neuro–neuro and abdomen–abdomen pairs versus a 0.04 mean increase in the neuro–abdomen pairs. Using leave-one-out cross-validation in which each document was iteratively used as a test sample while excluding it from the training data, the *k*-NN based classification accuracy was shown in all cases to be consistently higher with the semantics based measure compared with the non-semantic case. Moreover, the accuracy remained steady even as *k* value was increased – for the two anatomy related classes accuracy for $k = 41$ was 93.1% with semantics compared to 86.7% without semantics. Similarly, for the eight imaging procedures related classes, accuracy (for $k = 41$) with semantics was 63.8% compared to 60.2% without semantics. At the same *k*, accuracy improved significantly to 82.8% and 77.4% respectively when procedures were logically grouped together into four classes (such as ignoring contrast information in the imaging procedure description). Similar results were seen at other *k*-values.

Conclusions: The addition of semantic context into the document vector space model improves the ability of the cosine similarity to differentiate between radiology reports of different anatomical and image procedure-based classes. This effect can be leveraged for document classification tasks, which suggests its potential applicability for biomedical information retrieval.

1. Introduction

Despite on-going efforts to capture data in a unified manner, much of the data in the medical domain still remains as unstructured free textual content; as seen for instance in office notes, radiology or pathology reports. Healthcare is rapidly moving towards electronic medical records providing access to digital documents and it is becoming increasingly advantageous to be able to compare individual clinical documents with each other in order to facilitate clinical tasks such as retrieval of prior reports, document classification and ranking, improved search, mining of related reports to extract recommendation information, consistency checking during data entry and clinical decision support functionalities.

[*] Corresponding author. Tel.: +1 914 945 6125.
E-mail addresses: thusitha.mabotuwana@philips.com (T. Mabotuwana), michael.c.lee@philips.com (M.C. Lee), eric.cohen-solal@philips.com (E.V. Cohen-Solal).

1532-0464/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jbi.2013.06.013

A number of approaches have previously been proposed to measure similarity between documents, typically using statistical, data mining and machine learning techniques in conjunction with domain corpora. However, many classical techniques do not take semantic relationships into consideration; for example, a *neoplasm* is an abnormal mass of tissue commonly referred to as a tumor, and a *hemangioma* is a benign neoplasm. These concepts are semantically related since they refer to the same conceptual medical idea of tumor, but do not share any commonality from a lexical point of view. To address some of these limitations, there has been a growing trend in recent years, especially for biomedical data, towards approaches using concepts defined in an ontology to define the notion of semantic similarity as the similarity between concepts representing documents to be compared. This semantic comparison between documents goes beyond the lexical level and takes advantage of the relationships between concepts provided by the ontology.

Much of the semantic similarity related work outside the biomedical domain has been carried out using general purpose knowledge sources such as SemCor which is a sense-tagged corpora [1] and WordNet which is a large lexical knowledgebase consisting of over 150,000 English words along with semantic relations [2]. However, research has shown that such general purpose knowledge sources perform poorly on biomedical data since coverage of specialized concepts is rather limited [3]. The biomedical community has addressed this knowledge gap by developing specialized controlled terminologies containing biomedical terms such as International Classification of Diseases (ICD) to capture information such as diseases, abnormal findings, signs and symptoms, and Medical Subject Headings (MeSH) for indexing journal articles and books in the life sciences. Ontologies represent another type of knowledge source where medical terms are organized around concepts and relationships between these concepts where multiple terms can be associated with the same biomedical concept. For instance, Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT) captures diseases, findings, procedures, microorganisms and substances (among others) which also contains synonyms and a wide range of relations between concepts. Unified Medical Language System (UMLS) on the other hand aggregates concepts from different terminologies and ontologies to unify the coding of the many systems for its use in biomedical information systems and services.

Various methodologies have been proposed to-date to determine similarity between concepts. These various methods fall broadly into two categories – knowledge-based approaches and corpora-based approaches [4]. The knowledge-based approaches may use techniques such as rule-based heuristics with custom knowledge dictionaries, but typically attempt to exploit the hierarchical structure of an ontology, are machine processable and the domain knowledge is modelled explicitly via concepts and semantic relationships between them allowing various inferences to be made about this knowledge. These approaches typically determine the distance between two concepts of interest using techniques like edge counting, shortest path, ontological depth and Lowest Common Subsumer (or a combination of these). Similarity is then determined to be equal to inverse of distance in its simplest form, or some other mathematical function based on ontological distance. On the other hand, corpora-based approaches, also loosely referred to as information content based approaches, use a (large) corpus of domain-specific text to determine the information value of a concept. This is determined by calculating the frequency of each concept in the corpus – less frequent concepts are seen as more informative than common ones. See reviews by Pedersen et al. [5] and Batet et al. [6] for a more detailed discussion on various semantic similarity measures.

1.1. Related work

There are a number of related efforts to add semantic similarity to document comparisons. Corley and Mihalcea [7] proposed a corpus and knowledge-based approach for measuring semantic similarity of text which uses word-to-word similarity to determine a text-to-text semantic similarity metric by pairing up those words that are found to be most similar to each other, and weighting their similarity with the corresponding specificity score. Specificity of a word is determined using the inverse document frequency (IDF) according to a large corpus. Mihalcea et al. [4] successfully used these algorithms to identify if two text segments are paraphrases of each other. The SemKPSearch tool [8] is a more recent application of these algorithms where the tool attempts to extend search and browsing capability over a document collection to increase the number of related results returned for a keyphrase query. Vaidurya is another search engine that supports multiple-ontology, concept-based, context-sensitive search of clinical practice guidelines to improve performance on free text search retrieval [9]. Another similar system, but with a focus on medical documents, is the XOntoRank system [10] which provides SNOMED aware keyword search of XML documents.

Pivovarov and Elhadad [11] combine ontological and corpus based approaches by proposing a hybrid scheme where similarity is determined using context vectors by combining information from usage patterns (based on IDF) in clinical notes of patients with chronic kidney disease and from SNOMED ontological knowledge; however, the note-based similarity measure is not readily generalizable and is dependent on the annotated corpus and heuristics. Melton et al. [12] explored the use of five similarity measures to determine inter-patient similarity. A database of patient electronic charts consisting of discharge summaries, operative notes, radiology and pathology reports, diagnoses and other information was made available to experts to manually assess which patient documents where similar. The authors concluded that ontology principles and information content provide useful information for similarity metrics but currently fall short of expert performance indicating that still there is no gold-standard to determine document similarity.

The most commonly used measure to determine similarity between documents is perhaps the ‘bag-of-words’ or ‘bag-of-concepts’ approach coupled with the document vector space model which served as the baseline comparison similarity model in this work. Again, one key drawback of this approach is that it does not take ontology-based semantics into consideration. As a result, two documents that are contextually related may still have a zero similarity unless they had specific words in common. To address this limitation Ganesan et al. [13] proposed a Generalized Vector Space Model based on a Lowest Common Subsumer/Ancestor technique which uses hierarchical domain structure in order to produce more intuitive similarity scores. Other recent approaches are also extensions of the document vector space model, with the main difference between different algorithms being the way the document vectors are populated, how the weights are calculated and how the final similarity is computed.

Much of the prior research has focused primarily on determining similarity between individual concepts. In practice, especially in the medical domain, it is useful to compare different clinical documents where a single document is described using multiple clinical concepts. Therefore, in this paper we describe and apply a document similarity measure based on the semantic distance between sets of concepts instead of individual concepts. These algorithms are tested in the context of document-to-document similarity in radiology reports and the use of these similarity measures in classification of radiology reports into anatomy and procedure-based groups. We conclude with a discussion on how this