Feedback-Driven Radiology Exam Report Retrieval with Semantics

Sarasi Lalithsena
Kno.e.sis Center
Wright State University
Dayton, OH, USA
Email: sarasi@knoesis.org

Luis Tari
Knowledge Discovery Lab
GE Global Research
Niskayuna, NY, USA
Email: tari@ge.com

Anna von Reden
UPMC Enterprises
Pittsburgh, PA, USA
Email: vonredenal@upmc.edu

Benjamin Wilson
UPMC Enterprises
Pittsburgh, PA, USA
Email: wilsonbj2@upmc.edu

Brian J Kolowitz
UPMC Enterprises
Pittsburgh, PA, USA
Email: kolowitzbj@upmc.edu

John Kalafut
GE Healthcare
Pittsburgh, PA, USA
Email: john.kalafut@ge.com

Steven Gustafson
Knowledge Discovery Lab
GE Global Research
Niskayuna, NY, USA
Email: steven.gustafson@research.ge.com

Amit Sheth
Kno.e.sis Center
Wright State University
Dayton, OH, USA
Email: amit@knoesis.org

***Abstract*—Clinical documents are vital resources for radiologists to have a better understanding of patient history. The use of clinical documents can complement the often brief reasons for exams that are provided by physicians in order to perform more informed diagnoses. With the large number of study exams that radiologists have to perform on a daily basis, it becomes too time-consuming for radiologists to sift through each patient's clinical documents. It is therefore important to provide a capability that can present contextually relevant clinical documents, and at the same time satisfy the diverse information needs among radiologists from different specialties. In this work, we propose a knowledge-based semantic similarity approach that uses domain-specific relationships such as *part-of* along with taxonomic relationships such as *is-a* to identify relevant radiology exam records. Our approach also incorporates explicit relevance feedback to personalize radiologists information needs. We evaluated our approach on a corpus of 6,265 radiology exam reports through study sessions with radiologists and demonstrated that the retrieval performance of our approach yields an improvement of 5% over the baseline. We further performed intra-class and inter-class similarities using a subset of 2,384 reports spanning across 10 exam codes. Our result shows that intra-class similarities are always higher than the inter-class similarities and our approach was able to obtain 6% percent improvement in intra-class similarities against the baseline. Our results suggest that the use of domain-specific relationships together with relevance feedback provides a significant value to improve the accuracy of the retrieval of radiology exam reports.**

I. Introduction

With the introduction of the Affordable Care Act, the practice of medicine in the United States is changing dramatically. With the focus moving away from volume-based care and towards value-based care, hospitals, health systems and providers are having to redefine the way they deliver care. This is particularly prevalent in the field of radiology. Although traditional imaging workflows often had radiologists working as consultants with the rest of the care team to diagnose a patient, with the rise of digital imaging and electronic medical records (EMRs), radiology as a discipline has become isolated from direct interaction with the patient and their broader care team. This has led to a major impetus for radiology to redefine the value they contribute to the course of a patient's care. To accomplish this goal, imaging specialists must first be able to understand, synthesize and incorporate the patient's complete history into the way they practice medicine.

There is however, at least one major barrier that stands in the way of this transformation: radiologists are frequently presented with images for interpretation without the supporting clinical data needed [1] as a result of fragmented data sources. Without easy access to clinical history, non-radiology reports, or other brief unstructured documents [1] forming a comprehensive clinical review of an imaging study is difficult. Additionally, in those scenarios where the data mentioned above is accessible, navigating it to find information that is relevant to the study at hand is time-consuming and burdensome. This can be tied to the fact that the organization of data within the EMR tends to be oriented towards the producing physician and specialty reports (e.g. cardiology reports, surgical notes, pathology reports, etc.) and not the radiology workflow needs. As a result, radiologists spend a considerable amount of time searching for relevant information.

With the transition away from volume-based metrics, many attempts have been made to quantify the value provided by radiologist's dictated reports. Radiologists typically only spend a few minutes (1-5 minutes for simple X-Ray exams, perhaps 20-30 minutes for more complex Magnetic Resonance exams) reading each exam. The larger the amount of time radiologists spend in EMRs or prior imaging archives to orient themselves to a patient's history, the less time they are able to spend interpreting the new imaging. Radiologists consult prior radiology reports in specific situations such as when: there are incidental findings, the prior diagnosis wasn't clear from images and the imaging does not make sense or is poor quality. A prior report can be relevant to the radiologists based on many factors such as: body region, category of the disease (Trauma, Vascular, Infection, Tumor, Metabolic), and symptoms particularly if acute (nausea, bleeding). Providing a targeted, actionable report of their findings to radiologists is critical, ideally a report that actually answers another physician's specific clinical questions.

With an algorithm to retrieve and prioritize reports of a patient that are more likely to contain valuable context for the current exam, radiologist's time and focus can be more efficiently applied to image analysis and providing a narrower differential diagnosis in their final interpretation. Examples of context include whether or not the exam is a follow-up for an existing condition or an isolated exam for an acute symptom, whether the patient has secondary diagnoses related by anatomical region or disease type, and whether the ordering provider has supplied a specific hypothesis or inquiry elsewhere in the record. Equipped with these answers, radiologists may perceive the images themselves differently and may even notice imaging anomalies or disease progression that otherwise would go unreported.

Lexical and statistical-based methods have been widely adopted in the information retrieval research area in extracting relevant documents based on an information need. However, lexical-based methods fall short in certain cases when there is no lexical similarity between the terms in the query and the documents. For example, suppose foot pain is the patient's reason for exam and the patient might have been diagnosed with diabetes before, radiologists may be particularly interested in realizing such a history of diabetes when performing their diagnoses. But, documents with mentions of *diabetes* would not be retrieved by lexical-based methods with *foot pain* as the retrieval query. An ideal system would need to realize the semantic relations between foot pain and diabetes in order to perform a successful retrieval.

Knowledge-based semantic similarity methods [2], [3] serve as an important component for information retrieval tasks, especially when lexical and statistical methods fall behind. These methods use the semantics of the concepts and their relationships to fill the gap left by the lexical and statistical methods. Researchers in biomedical and healthcare domain have contributed to the development of large number of knowledge bases, such as MeSH (Medical Subject Headings), ICD taxonomy (International Classification of Diseases) and SNOMED CT which makes this domain a suitable candidate to use these knowledge-based semantic similarity methods. [4] proposes an approach based on knowledge-based semantic similarity to compute the document similarity in the context of radiology reports using SNOMED. This work uses only taxonomic relationships and ignores the domain specific relationships. However, domain-specific relationships such as *part-of*, *causes* (between diseases and symptoms) and *treats* (between medications and diseases/symptoms) play a key role in identifying the contextually relevant information.

To address this, our approach presented in this paper adopts existing knowledge-based semantic similarity methods to leverage the semantics of multiple relationships defined in the knowledge bases. Furthermore, the proposed approach incorporates the explicit relevant feedback to personalize the results for radiologists. Concrete contributions of this work are as follows:

- Propose an approach to leverage knowledge-based semantic similarity methods with multiple relationships to identify the contextually relevant patient's records.
- Incorporate the explicit relevant feedback to personalize based on user's need.
- Demonstrate our approach for relevancy retrieval and feedback incorporation for a radiology exam corpus.

A popular approach adopted by relevance feedback algorithms is to modify the original query by considering feedback for original results. Feedback can be explicit or implicit. Here, we adopt explicit relevance feedback, particularly Rocchio relevance feedback [5], which is a well-known algorithm for explicit relevance feedback by incorporating feedback into the vector space model through query rewriting. In the biomedical and healthcare domain [6] uses Rocchio to expand the query on top of the MeSH hierarchy to retrieve relevant documents from MEDLINE. Rocchio algorithm is being used by [7] to improve retrieval performance. To our knowledge, our approach is the first to address the different information needs among various radiology specialties by applying semantic similarity to radiology reports together with Rocchio relevance feedback to take into account the explicit feedback by radiologists.

The rest of the paper is organized as follows. Section II describes the related work and Section III describes our approach. Section IV describes the evaluation set up and then discusses the result. Finally, Section IV concludes with suggestions for future work.

II. Related Work

Similarity measures have been applied to various natural language processing tasks that include text segment similarity [2], paraphrase detection [8] and sentence similarity [9]. These methods can be broadly categorized into: a) knowledge-based similarity measures where similarity is based on information gained from semantic network, b) corpus-based similarity measures where similarity is based on information gained from large corpora.

Our work can be categorized as a knowledge-based similarity approach; the remainder of this section will only be contrast to such approaches while readers can refer to more details on semantic similarity methods in these survey papers [10] and [11]. Knowledge-based similarity measures are mainly of two types: a) Path-based e.g., Leacock and Chodorow [12], Wu and Palmer [13]; b) Information-content-based e.g., Resnik [14], Lin [15]. Existing work on measuring the similarity of two medical terms mainly use the hierarchical arrangements of the terms in an knowledge base without [3], [16] or with corpus information [17]. Techniques on term similarity have also been used in various approaches in biomedical and healthcare domains such as document similarity [4] and document clustering [18].

The case-based reasoning system proposed by [19] uses cosine similarity with a domain-specific ontology to measure the similarity of the medical encounters by enhancing term weights using relationships. The inter-patient record similarity technique proposed by [20] uses knowledge-based similarity using shortest path, sub classes and information content. Their evaluation shows that incorporating additional information from a knowledge base improves the similarity calculation. Vaidurya [21] explores the concept hierarchy to identify relevant document for clinical-guideline search engine. In addition to the above work, XOntoRank [22] ranks and returns sub trees of XML documents that either contain or are associated with the query terms through the ontological references. Patient similarities are also being used in predicting the health status of a patient by [23] and their approach represents the patient records as a vector with bag of features based on the features identified.

Among the existing work, the closest to ours is the approach by [4] which uses a vector space model with taxonomic relationships (i.e. *is-a* relationships) based on the SNOMED CT clinical ontology to calculate similarity among patient records. This approach represents each patient record as a vector of concepts from SNOMED CT and weights are assigned to the concepts based on the concepts appearing in the document and parent concepts obtained from the hierarchy of SNOMED CT. Weights are assigned based on the shortest path between the annotated concept and the parent concept. However, domain-specific relationships such as *part-of*, *causes* (between diseases and symptoms) and *treats* (between medications and diseases/symptoms) play a key role in identifying contextually relevant information. Our work differs from [4] in several aspects: a) the use of RadLex ontology[1], a controlled terminology for radiology developed by Radiology Society of North America (RSNA); b) the use of domain-specific relationships, in particular *part-of* relationships; c) the use of feedback mechanism in an attempt to satisfy the different information needs for various radiology specialties.

III. Approach

Our proposed approach identifies prior radiology reports of a patient which are contextually relevant based on the reason for exam. A radiology exam report typically contains multiple sections: (i) *Study Description*; (ii) *History*; (iii) *Comparison*; (iv) *Finding* and (v) *Impression*. Figure 3 shows an example of a radiology exam report. Data contained in the reason for exam field is often brief, consisting of a few keywords and abbreviations, such as *ankle pain*. In some instances, the reason for exam may contain many terms but describe a variety of possible conditions which leads to the need to associate information in a more robust way, such as *tachycardia and short of breath s/p ankle surgery 2 weeks ago; tachycardia; Other dyspnea and respiratory abnormality*. Accurate and comprehensive retrieval of relevant prior radiology reports requires the semantic understanding of the terms appearing in the textual description of the reports and the reason for exam. For example, reports that state "lower extremity" and "leg" can be considered as relevant when realizing that leg is part of lower extremity. This illustrates the need for an approach that goes beyond the lexical understanding of terms in identifying radiology reports that are relevant to the reason for exam.

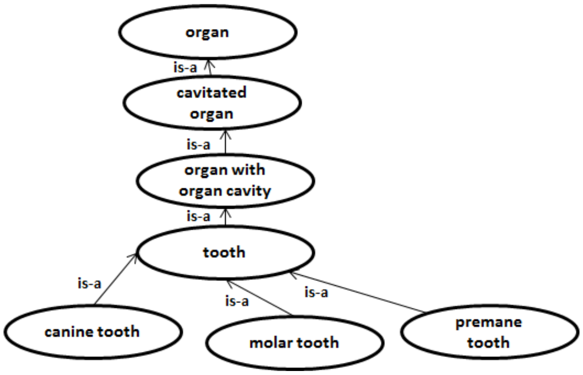


Fig. 2. *The concept* ***tooth*** *and its related concepts through the is-a relationship.*

Semantic understanding in our approach is achieved by processing the textual description of the reason of exam as well as the study description of radiology exam reports using a vector-based semantic similarity with ontologies. The core idea behind our approach is to define a similarity measure that takes advantages of the ontological relations among the terms appearing in the textual description. Examples of ontological relations include subclass or *is-a* relationships, subcomponent relations also known as *part-of* relationships and causal relationships. Fig. 2 shows the concept *tooth* and its related concepts through the *is-a* relationships. According to the figure, concepts *canine tooth*, *molar tooth* and *premane tooth*, *organ with organ cavity*, *cavitated organ* and *organ* are identified as related concepts. By representing text description in the form of vectors, this enables us to encode the concepts that appear in the text, as well as their ontological relations with other related concepts. With the vector representation, it becomes feasible to adopt similarity measures such as cosine similarity and utilize concepts and their ontological relationships to compute similarity efficiently. Such kind of similarity computation can capture the implicit relations between terms that can otherwise be hard to capture by lexical-based approaches.

In the rest of the section, we describe the technical details of our approach. Figure 1 depicts the steps involved in generating the vector representation of radiology exam reports and reason for exam. The subsection III-A describes how our approach represents both prior radiology exam reports and reason for exam as weighted vectors of ontological concepts. This involves the annotation of the terms appearing in the radiology exam reports and the reason for exam using the ontological concepts and then exploring the ontology to identify the related concepts to the actual terms. Such annotation results in vectors of weights representing the involved concepts and their ontological relations. To determine the similarity between

[1] http://rsna.org/RadLex.aspx

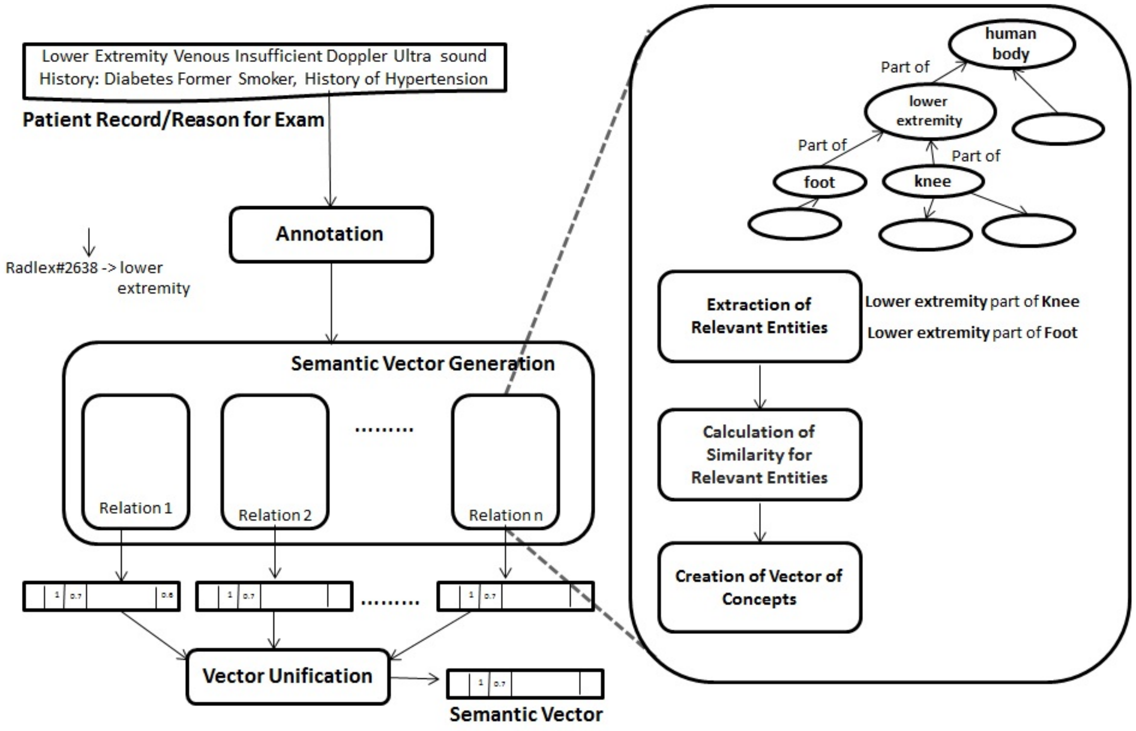

Lower Extremity Venous Insufficient Doppler Ultra sound
History: Diabetes Former Smoker, History of Hypertension
Patient Record/Reason for Exam
Annotation
Radlex#2638 -> lower extremity
Semantic Vector Generation
Relation 1
Relation 2
Relation n
1
0.7
0.6
Vector Unification
Semantic Vector
human body
Part of
lower extremity
Part of
Part of
foot
knee
Extraction of Relevant Entities
Lower extremity part of Knee
Lower extremity part of Foot
Calculation of Similarity for Relevant Entities
Creation of Vector of Concepts

Fig. 1. *Workflow for Vector Generation*

the radiology exam reports and the reason for exam, we adopt the cosine similarity measure and further extend the computation using Rocchio [24] to elicit user feedback for results refinement. Details can be found in subsections III-B and III-C.

A. Semantic Vector Generation

The core component of our approach lies in the vector representation of radiology exam reports and reason for exam that are represented in the form of vectors of concepts. The vector size corresponds to the number of concepts in the ontology used for processing. In this case the RadLex ontology [25] was used in the processing of radiology reports and reason for exam. The RadLex ontology is a controlled terminology for radiology with more than 68,000 concepts C where $C = \{c_1, c_2, ..., c_n\}$ and n is the total number of concepts.

The first text processing step is to apply lemmatization to both the ontology as well as the text fragments that are combined from the various sections for each report. Here, we use ClearNLP[2] for lemmatization and Apache UIMA ConceptMapper[3] for concept annotation. With the use of the RadLex ontology, synonyms are also considered during the concept annotation process. For example "thorax" and "chest" are often used interchangeably and the annotation maps both to the same concept *radlex:RID1243* (thorax) in the RadLex ontology. At the end of this step, we represent each text fragment tf as a set of annotated concepts AC_{tf} where,

$$AC_{tf} = \{c_j | c_j \text{ appears in } tf\}$$

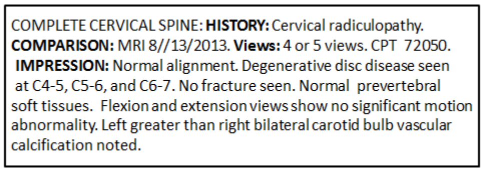

COMPLETE CERVICAL SPINE: **HISTORY:** Cervical radiculopathy. **COMPARISON:** MRI 8//13/2013. **Views:** 4 or 5 views. CPT 72050. **IMPRESSION:** Normal alignment. Degenerative disc disease seen at C4-5, C5-6, and C6-7. No fracture seen. Normal prevertebral soft tissues. Flexion and extension views show no significant motion abnormality. Left greater than right bilateral carotid bulb vascular calcification noted.

Fig. 3. *Sample Radiology Exam Report*

Each text fragment tf is represented as a vector of concepts $\vec{t}$ in $\mathbb{R}^n$ i.e. $\vec{t} = (t_1, \ldots, t_n)$. Traditional vector representation in information retrieval uses the total number of terms in the text as the dimension of the vector and uses tf-idf (term frequency-inverse document frequency) to calculate the relevance of each term to the text. In our case, we use ontology concepts as the dimension of these vectors. In the vector $\vec{t}$, the value of i-th position t_i corresponds to the strength of association of concept c_i to the text fragment tf. Relationships defined in the ontology drives the identification of semantically relevant concepts. A *text fragment relationship vector* $\vec{t_r}$ is generated for each relationship r for a text fragment tf. In our case, we consider the ontology relationships *is-a* and *part-of* for the vector representation.

Vector representation consists of annotated concepts and related concepts. Initially, all the vector positions will be assigned to 0 and values for each vector position t_{ri} (i.e. i-th position in vector $\vec{t_r}$) are assigned as follows:

[2] http://www.clearnlp.com

[3] https://uima.apache.org/

$$t_{ri} = \max_{c_m \in AC_{tf}} Sim(c_i, c_m)$$

$Sim(c_i, c_m)$ is defined as the conceptual similarity between the two concepts based on the path length of the two concepts.

$$Sim(c_i, c_m) = \begin{cases} 0 \text{ if } PathLength(c_i, c_m) = 0 \\ \frac{1}{PathLength(c_i, c_m)} \text{ otherwise} \end{cases}$$

$PathLength$ is defined as the number of nodes of the shortest path between c_i and c_m inclusively. For example, the conceptual similarity of *molar tooth* with respect to the *tooth* is 0.5. $PathLength(c_i, c_m)$ is 1 when c_i=c_m and $Sim(c_i, c_m)$ is 0 if a path does not exist between c_i and c_m in the ontology. In calculating the similarity, only those concepts within a path length of 3 from the annotated concepts are being considered.

The last step of the vector representation is to create a unified vector $\vec{t}$ by unifying all relationship-based vectors. The value for vector position t_i is assigned by considering the i-th positions among the text fragment relation vectors $\vec{t_1}, \ldots, \vec{t_k}$.

$$t_i = max(t_{1i}, \ldots, t_{ki})$$

where k is the number of relationship types.

B. Relevant Report Retrieval

The process of semantic vector generation is applicable to both radiology exam reports as well as the reason for exam. In the deployment of our approach, the reason for exam is treated as a query to retrieve the relevant radiology exam reports. This means that the semantic vector generation module is applied to the radiology exam reports as an offline task so that vectors for each patient record are created in advance. During run time, the semantic vector representation module is applied to the reason for exam and the corresponding vector is used to retrieve the relevant radiology exam reports in the form of vectors. The next step is to utilize the semantic vectors to compute the similarity between the reason for exam q and each prior report p based on cosine similarity as defined below in Equation 1.

$$\text{Similarity}(\vec{q}, \vec{p}) = \frac{\sum_{i=1}^{n} q_i * p_i}{\sqrt{\sum_{i=1}^{n} (q_i)^2}\sqrt{\sum_{i=1}^{n} (p_i)^2}} \tag{1}$$

Using the cosine similarity, a ranked list of prior radiology exam reports is then presented to the radiologists.

C. Relevance Feedback

Radiologists working in different specialities may have different opinions as to which exams are most relevant with respect to the reason for exam. Such distinction shows that it is important to adapt the retrieval and ranking process to users' needs. The significance of incorporating implicit or explicit user feedback [26] has been widely discussed with respect to web search ranking. Here, we adopt an explicit relevance feedback in which users mark retrieved documents as relevant or irrelevant. This allows the system to contextualize the search results based on each domain expert's judgment.

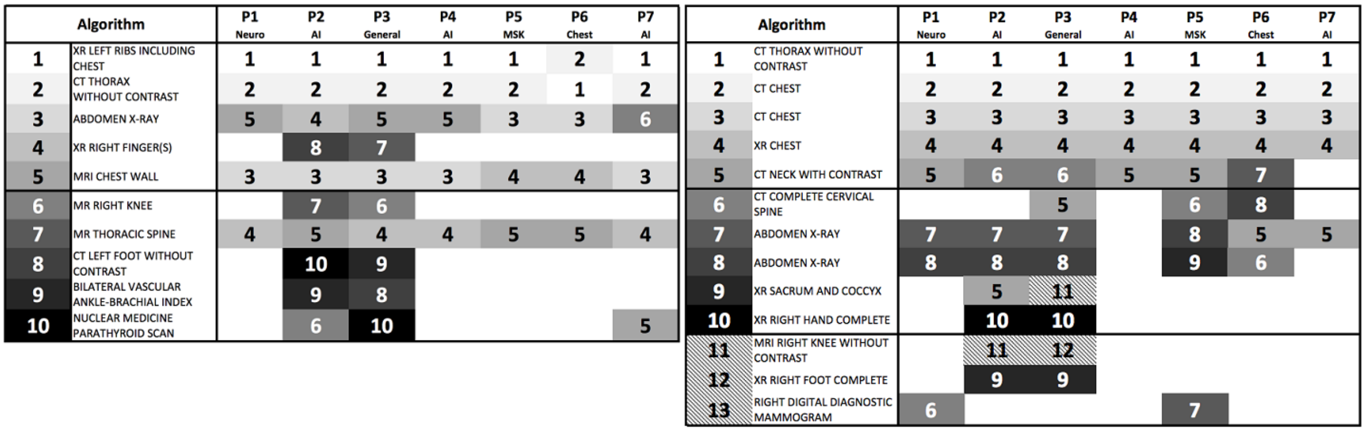
Rocchio algorithm [24] is a commonly used approach for relevance feedback in information retrieval systems. This has been used in a number of previous works, including text categorization [27], [28] and contextual retrieval [29]. Rocchio algorithm works by modifying the query and considering the terms that occur in the documents that are rated as relevant by the users. This simulates the process of a user changing the original query by looking at the results. In our approach, we adopt the Rocchio algorithm to capture the explicit relevance feedback given by the radiologists to the initial result set. The original query is then adjusted based on user feedback so that the modified query is biased towards documents that are determined as similar to the ones that are marked as relevant. The original Rocchio algorithm was designed to work with binary feedback (relevant or irrelevant) from the users. Here, we allow users to rate the records ranging from 1 to 5 (lowest to highest) based on their relevancy. We made a simple extension to the original algorithm to reflect these ratings, as shown in Equation 2.

$$\vec{q_m} = \alpha\vec{q_o} + \beta\frac{1}{|D_r|}\left(\sum_{i=1}^{5} w_i \sum_{\vec{d_j} \in D_i} \vec{d_j}\right) + \gamma\frac{1}{|D_{nr}|}\sum_{\vec{d_k} \in D_{nr}} \vec{d_k} \tag{2}$$

where $\vec{q_m}$ is the modified reason for exam vector, $\vec{q_o}$ is the original reason for exam vector, D_r is the set of relevant patient records according to the feedback, D_{nr} is the set of irrelevant patient records. D_r is defined as $D_1 \cup D_2 \cup D_3 \cup D_4 \cup D_5$, with each D_i representing a set of patient records rated as i. α, β and γ are the weights assigned for the original query vector, relevant set of documents and irrelevant set of documents respectively.

IV. Evaluation

To evaluate our approach, we used a corpus of 6,265 de-identified radiology exam reports and adopted the RadLex ontology for recognizing concepts that include subclasses of Body Parts (RID:13390), Pathophysiologic Findings (RID:4736), Symptoms (RID:39050) and Imaging Modality (RID:10311). Two different evaluation settings were used to assess the quality of our approach: evaluation with domain experts and evaluation based on intra-class and inter-class similarities. The first setting, described in section IV-A, aims to evaluate our approach based on the domain experts' judgments using standard information retrieval measures. The second setting is to evaluate the performance of the similarity approach based on intra-class and inter-class similarities. The intuition behind evaluation based on intra-class and inter-class similarities is that the similarity approach should be able to assign high similarity scores for reports belonging to the same class, and on the other hand low similarity scores should be assigned to reports that are originated from different classes. Such intra-class and inter-class evaluation allows us to evaluate our approach on a large scale when study sessions with radiologists can only be done on a limited basis. We describe intra-class and inter-class evaluation in Section IV-B.



Algorithm		P1 Neuro	P2 AI	P3 General	P4 AI	P5 MSK	P6 Chest	P7 AI
1	XR LEFT RIBS INCLUDING CHEST	1	1	1	1	1	2	1
2	CT THORAX WITHOUT CONTRAST	2	2	2	2	2	1	2
3	ABDOMEN X-RAY	5	4	5	5	3	3	6
4	XR RIGHT FINGER(S)		8	7				
5	MRI CHEST WALL	3	3	3	3	4	4	3
6	MR RIGHT KNEE		7	6				
7	MR THORACIC SPINE	4	5	4	4	5	5	4
8	CT LEFT FOOT WITHOUT CONTRAST		10	9				
9	BILATERAL VASCULAR ANKLE-BRACHIAL INDEX		9	8				
10	NUCLEAR MEDICINE PARATHYROID SCAN		6	10				5

Algorithm		P1 Neuro	P2 AI	P3 General	P4 AI	P5 MSK	P6 Chest	P7 AI
1	CT THORAX WITHOUT CONTRAST	1	1	1	1	1	1	1
2	CT CHEST	2	2	2	2	2	2	2
3	CT CHEST	3	3	3	3	3	3	3
4	XR CHEST	4	4	4	4	4	4	4
5	CT NECK WITH CONTRAST	5	6	6	5	5	7	
6	CT COMPLETE CERVICAL SPINE			5		6	8	
7	ABDOMEN X-RAY	7	7	7		8	5	5
8	ABDOMEN X-RAY	8	8	8		9	6	
9	XR SACRUM AND COCCYX		5	11				
10	XR RIGHT HAND COMPLETE		10	10				
11	MRI RIGHT KNEE WITHOUT CONTRAST		11	12				
12	XR RIGHT FOOT COMPLETE		9	9				
13	RIGHT DIGITAL DIAGNOSTIC MAMMOGRAM	6				7		

Fig. 4. *Results generated by our vector-based semantic similarity algorithm compared by the seven radiologists (P1 to P7) on 2 queries: (left) Q1: XR Chest, Rib Fracture and (right) Q2: CT Chest, Lung Nodule*

A. Evaluation with Domain Experts

In order to measure the accuracy of the algorithm, we compared examples of its output against radiologists' rankings of the same radiology exam reports. Our participants were seven radiologists of varying levels of experience and with a focus in one of several subspecialties (3 Abdominal, 1 Chest, 1 Neuro, 1 Musculoskeletal and 1 Generalist). We selected two exam queries: XR Chest with RFE: Rib Fracture (Q1) and CT Chest with RFE: Lung Nodule (Q2), and used our approach to rank the patients' prior imaging exam records for each query.

We transformed the top 20 records for each query into a series of laminated cards containing a) the exam description (i.e. MR THORACIC SPINE), b) the similarity score in % format (i.e. 57% match), c) the exam "history" text from the record, and d) the exam "impression" text from the record. These items represent the data a radiologist would be most likely to reference when searching for an appropriate prior exam comparison.

We created an additional set of cards that showed the same 20 records, but substituted the algorithm's similarity score with an indication of whether the body part and/or modality matched those in the query (e.g. for Q1, all XR Chests, XRs, and Chest exams would be prioritized). This notion of similarity more closely approximates prior exam relevancy models used by radiologists in existing software. Each radiologist was presented with the top results from both sets of cards side-by-side, and was asked to qualitatively compare the two methods of determining similarity, first for Q1 and then for Q2. The radiologists were then asked to rearrange the cards to show their preferred ranking for the set of exam records presented by our algorithm as shown in Figure 4.

We also captured the qualitative reasoning behind each radiologist's preferred rankings, so as to better understand what features in the records could be targeted to improve the algorithm. It also allowed us to understand the benefits of using a similarity algorithm to rank prior imaging records as an alternative to the more standard anatomy/modality-matching method.

		Q1			Q2	
	RPrec@2	RPrec@5	RPrec@8	RPrec@2	RPrec@5	RPrec@8
BOC	0.5	0.59	0.67	1	0.71	0.85
ISA	0.5	0.73	0.72	1	0.71	0.85
VSS	1	0.73	0.72	1	0.88	0.93

TABLE I
PERFORMANCE COMPARISON AMONG OUR VECTOR-BASED SEMANTIC SIMILARITY APPROACH (VSS), BAG-OF-CONCEPTS (BOC) AND CONCEPTS WITH IS-A RELATIONSHIPS (ISA)

Radiologists' final rankings after discussion were used as a standard to refine and evaluate the algorithm. We adopted *R-Precision* [30], a commonly used IR evaluation metric, as our metric to evaluate the overall performance of our approach. R-Precision is the precision at rank R (denoted as $RPrec@R$), where R is the number of documents relevant to the query. We compute $RPrec@R$ for different values of R,

$$RPrec@R = \frac{r}{R}$$

where r is the number of relevant documents retrieved by the system among top-R documents.

1) Results and Analysis on Vector-based Semantic Similarity: To assess the quality of the initial rankings based on our approach, we compute R-Precision values for the results generated by our vector-based semantic similarity approach (VSS) by considering each radiologist's ratings separately. We compared our approach with two baseline approaches: bag of concepts (BOC) and concepts with *is-a* relationships (ISA) approaches. Figure 4 shows the initial actual rankings of reports that are assigned by the algorithm.

We summarize the above result by taking the average of all radiologists for each $RPrec@R$ for each approach. As shown in Table I, our approach out-performs both BOC and ISA approaches for Q2 and is on par for Q1. Our approach performing on par with the ISA approach for Q1 indicates that concepts added as a result of part-of relationships do not provide extra values in retrieving results for Q1. But note that, even for the Q1 RPrec@2 is higher than baseline approaches,

	Q1						Q2					
	BOC		ISA		**VSS**		BOC		ISA		**VSS**	
	No Feedback	With Feedback	No Feedback	With Feedback	No Feedback	With Feedback	No Feedback	With Feedback	No Feedback	With Feedback	No Feedback	With Feedback
RPrec@2	0.5	1	0.5	0.5	1	1	1	1	1	1	1	1
RPrec@5	0.58	0.56	0.73	0.73	0.73	0.73	0.71	0.74	0.71	0.71	0.88	0.88
RPrec@8	0.67	0.67	0.76	0.76	0.76	0.78	0.85	0.79	0.85	0.85	0.93	0.93

TABLE II
COMPARISION OF R-PRECISION VALUES BEFORE AND AFTER FEEDBACK AMONG THE THREE APPROACHES

	BOC Before Feedback		After Feeback P1		After Feedback P2	
	Rank	Score	Rank	Score	Rank	Score
XR LEFT RIBS INCLUDING CHEST:	1	1	1	1	1	1
ABDOMEN X-RAY	2	0.5	4	0.518	4	0.585
XR RIGHT FINGER(S)	2	0.5	3	0.63	3	0.634
CT THORAX WITHOUT CONTRAST	2	0.5	2	0.74	2	0.707
NUCLEAR MEDICINE PARATHYROID SCAN	5	0	9	0.012	10	0.01
BILATERAL VASCULAR ANKLE-BRACHIAL INDEX	5	0	9	0.012	9	0.06
MR RIGHT KNEE	5	0	7	0.185	7	0.243
MR THORACIC SPINE	5	0	7	0.185	7	0.243
CT LEFT FOOT WITHOUT CONTRAST	5	0	5	0.37	5	0.341
MRI CHEST WALL WITHOUT CONTRAST	5	0	6	0.271	6	0.292

TABLE III
AN EXAMPLE OF THE CHANGED FEEDBACK FOR Q1 QUERY WITH BOC FOR RADIOLOGISTS P1 AND P2

and this is a good indicator that our approach was able to rank relevant result first compared to other approaches.

Domain experts rankings also suggest that the relevance of a prior report goes beyond the modality. Related to our first case *X-Ray Chest, Rib fracture*, almost all radiologists would rank *X-Ray Left Ribs* exams as important/relevant information to X-Ray Chest studies. Most radiologists consider the Abdominal X-Ray as relevant to X-Ray Chest primarily because there's a chance this exam would capture some of the lower ribs better than other modalities' view of the chest for the purposes of identifying a fracture.

2) Results and Analysis on Relevance Feedback: Differences in rating by different radiologists by the previous experiment emphasizes the need to personalize the results for each radiologist. To assess the performance in utilizing relevance feedback, we re-ranked the patient records by extending our semantic similarity algorithm with relevance feedback to adapt to the results rankings given by radiologists. Our algorithm requires the ratings given by the radiologists for the initial results and the weights for the parameters α, β and γ as its input. We performed experiments in determining the optimal values for the parameters α, β and γ. Our experiments showed that the optimal values for α ranges from 0.4 to 1.0, while the value for β is 0.5 and γ is -0.1.

Table II summarizes the results with the R-Precision values before and after feedback incorporation for all three approaches, including our approach. We observed an improvement in the results after they were re-ranked by means of relevance feedback for a couple of cases. While we do not observe a significant performance improvement of relevance feedback approach over the original approach, we can see that the rankings of the reports are re-ordered in a way that is closer to the intent of the radiologists. Table III shows an example of how the results are changed before and after feedback by two radiologists P1 and P2 for query Q1 for the BOC approach. As depicted in Table III, before incorporating the feedback the system ranked *CT THORAX WITHOUT CONTRAST*, *ABDOMEN X-RAY* and *XR RIGHT FINGER(S)* in the same order for Q1. But after feedback, the system was clearly able to make the distinction that CT THORAX WITHOUT CONTRAST is more related to Q1 than the other two records. This shows the promise of relevance feedback which can cater to the information needs of radiologists from different specialties.

Exam Code	**Description**	**No. Reports**
CXR	Chest X-Ray	665
MAMSCRNDIG	Digital Mammogram Screening	599
DEXABOD	DEXA Bone Density Exam	279
USABDCOMP	Complete Abdominal Ultrasound	165
OSRCHESTCR	Outside Chest CR	164
LUMBARSP	Lumbar Spine Exam	117
MAMDIAGDIG	Bilateral Digital Diagnostic Mammogram	108
USSOFTISSU	Soft Tissue Ultrasound	105
CTABDPEL	CT Abdomen and Pelvis with Contrast	97
USBREASTLT	Left Breast Ultrasound	85

TABLE IV
NUMBER OF REPORTS FOR EACH EXAM CODE, ITS DESCRIPTION AND NUMBER OF REPORTS

B. Evaluation on Intra- and Inter-class Similarities

Hosting study sessions with radiologists is an ideal way to obtain feedback and evaluate our approach. However, such study sessions can only be done on a limited basis. We compensate the limited case studies by using *inter-class similarities* and *intra-class similarities* to further examine the effectiveness our approach, similar to the evaluation performed in [4].

1) Experiment Setup: We used exam codes as our classes to calculate the intra-class and inter-class similarities. Exam reports with the same exam code are considered as a single class. We selected 10 exam codes with the highest number of reports covering 2,384 reports. Table IV shows the number of reports for each exam code, resulting in 10 intra-class similarities and 45 inter-class similarities. As an example, for the two classes CXR and MAMSCRNDIG we considered two intra-classes (CXR-CXR and MAMSCRNDIG-MAMSCRNDIG) and one inter-class (CXR-MAMSCRNDIG) to compute similarity. In each of these cases, pairwise cosine similarity is applied across all possible patient record pairs to compute the averaged similarity.

2) Analysis: Table V shows the pairwise similarities for the 10 classes in our experiment setting. As expected intra-class similarities are always higher than the inter-class similarities. In order to illustrate the effect of the relationships we compared our vector-based semantic similarity approach (VSS) with the approaches using bag of concepts (BOC) and is-a relationships (ISA). Also, we expect that with the semantic enhancements intra-class similarities will be increased and inter-class similarities will be decreased. Figure 5 and Figure 6 show the intra-class and inter-class similarities respectively for the three approaches.

On average, intra-class similarities across all classes were able to obtain a 6% increment via adding more relationships as given in Table VI. Out of the ten classes,

1) our approach outperforms the BOC and ISA approaches in three classes A, E, and I.
2) our approach performs in the same way as the ISA approach in five classes B, F, G, H and J .
3) our approach underperforms in two classes C, D.

In certain cases as in (1), it clearly shows the importance of incorporating *part-of* domain-specific relationships. For cases in (2), the *part-of* relationship does not contribute in improving the similarity. This is due to the unavailability of *part-of* relationships for the concepts appearing in the corresponding patient records.

However, for cases presented in (3) the inclusion of *part-of* relationships negatively impacts the results. To analyze this, let's consider the class CXR in which our approach performs well and another class USABDCOMP in which our approach does not. We analyzed the effect of the expanded concepts being included by our approach as compared to the ISA approach by identifying the number of occurrences for these expanded terms across the reports. For this purpose, we take the 100 radiology exam report pairs with highest similarities. We observed that a significant number of the expanded concepts included for CXR reports indeed have a high number of frequencies across the reports, while only a handful of expanded concepts appear across the USABDCOMP reports. For instance, in the case of CXR there were 95 concepts appearing more than 180 times across all reports, and in the case of USABDCOMP there were only 10 concepts appearing more than 60 times. In fact, the maximum frequency of any concept appearing in USABDCOMP was only 79. That means, expanded concepts for USABDCOMP have a relatively low importance for that particular class of records which negatively affects the intra-class similarity. This illustrates that concepts included as a result of taxonomic and domain-specific relationships may have disparate impact on the retrieval process. One potential workaround to improve the performance is to restrict the concept expansion process to certain classes of reports. Another direction could be to only include expanded concepts that have a high number of occurrences in the reports.

Approach	**Average Intra-class Similarity**	**Average Inter-class Similarity**
BOC	0.47239	0.154
ISA	0.5046	0.178
VSS	0.53694	0.162

TABLE VI
AVERAGE INTRA-CLASS AND INTER-CLASS SIMILARITIES AMONG THE THREE APPROACHES

In the case for inter-class similarity, lower similarity scores would indicate better performances. Our vector-based semantic similarity approach (VSS) achieves an average similarity score of 0.162 as opposed to 0.154 and 0.178 for BOC and ISA approaches. While VSS achieves a better performance than the ISA approach, we did further analysis to investigate the performances comparing between our VSS approach and BOC approach. Out of the 45 cases of inter-class similarity, VSS achieves lower similarity scores, or in other words better performance, than the BOC approach for 28 cases. For the rest of the 17 cases, we observed that some of the pairs of classes are indeed related to each other. For instance, MAMSCRNDIG-MAMDIAGDIG are both classes of reports on mammograms but done in different protocols. MAMDIAGDIG-USBREASTLT and MAMSCRNDIG-USBREASTLT are classes of reports that refer to the same body part, and these pairs are assigned with higher similarity scores by our VSS approach as compared to BOC.

V. Conclusion and Future Work

In this paper, we present an approach using domain-specific relationships particularly *part-of* relationships to retrieve the relevant patient records via knowledge-based semantic similarity methods. Our evaluation shows the importance of these relationships to improve the accuracy of the retrieval process compared to the approaches based on bag-of-concept and taxonomic relationships. Furthermore, our approach to incorporate explicit relevance feedback has the capability to adapt the result based on radiologists' individual needs.

We believe that the work contained in this study provides a foundation that can be refined and adapted to answer several

	A	B	C	D	E	F	G	H	I	J
(A) CXR	**0.5265**	0.1022	0.0902	0.0930	0.5193	0.1218	0.1183	0.1289	0.1383	0.1081
(B) MAMSCRNDIG	0.1022	**0.7147**	0.0841	0.2655	0.1070	0.0805	0.3519	0.2117	0.0862	0.3450
(C) DEXABOD	0.0902	0.0841	**0.6031**	0.0582	0.0914	0.1923	0.1410	0.1409	0.0766	0.1381
(D) USABDCOMP	0.0930	0.2655	0.0582	**0.4496**	0.0962	0.0959	0.1925	0.2059	0.1697	0.2101
(E) OSRCHESTCR	0.5193	0.1070	0.0914	0.0962	**0.5280**	0.1257	0.1231	0.1266	0.1361	0.1139
(F) LUMBARSP	0.1218	0.0805	0.1923	0.0959	0.1257	**0.3482**	0.1074	0.1021	0.1123	0.1003
(G) MAMDIAGDIG	0.1183	0.3519	0.1410	0.1925	0.1231	0.1074	**0.5954**	0.2742	0.1364	0.5165
(H) USSOFTISSU	0.1289	0.2117	0.1409	0.2059	0.1266	0.1021	0.2742	**0.5782**	0.1364	0.2916
(I) CTABDPEL	0.1383	0.0862	0.0766	0.1697	0.1361	0.1123	0.1364	0.1364	**0.4995**	0.1474
(J) USBREASTLT	0.1081	0.3450	0.1381	0.2101	0.1139	0.1003	0.5165	0.2916	0.1474	**0.5263**

TABLE V
INTER AND INTRA CLASS SIMILARITIES FOR THE TEN CLASSES

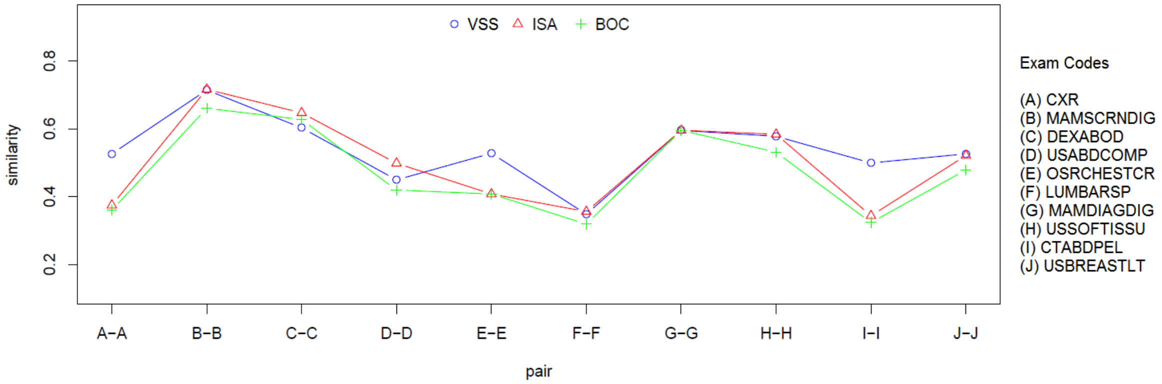


Fig. 5. *Pairwise intra-class similarity comparing our approach (VSS) with the two baseline approaches BOC and ISA*

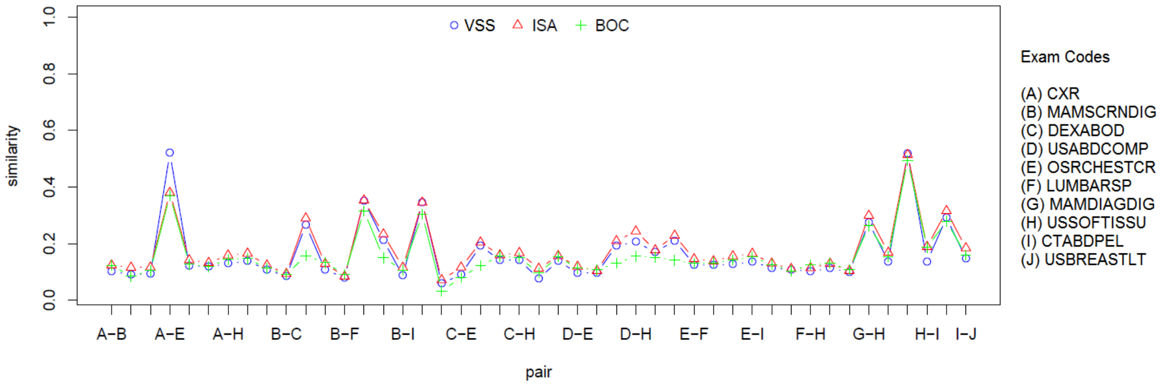


Fig. 6. *Pairwise inter-class similarity comparing our approach (VSS) with the two baseline approaches BOC and ISA*

key clinical questions that radiologists may have about incoming imaging exams. We next intend to extend our data beyond imaging reports, applying a similar reason for exams, body part, and modality-based prioritization of EMR data in order to further reduce radiologists orientation time and increase diagnostic specificity. Eventually, leveraging a combined set of imaging and non-imaging documentation, our goal is to lay the foundation for an imaging-centric, longitudinal view of a patient's historical (and future) treatment for all diagnoses related to an exam's target anatomical region. Furthermore, we plan to improve our evaluation with domain experts by using more queries. We are also interested in applying our approach to other corpora outside of radiology.

REFERENCES

[1] J. W. Nance Jr, C. Meenan, and P. G. Nagy, "The future of the radiology information system," *American Journal of Roentgenology*, vol. 200, no. 5, pp. 1064–1070, 2013.

[2] R. Mihalcea, C. Corley, and C. Strapparava, "Corpus-based and knowledge-based measures of text semantic similarity," in *AAAI*, vol. 6, 2006, pp. 775–780.

[3] M. Batet, D. Snchez, and A. Valls, "An ontology-based measure to compute semantic similarity in biomedicine," *Journal of Biomedical Informatics*, vol. 44, no. 1, pp. 118 – 125, 2011, ontologies for Clinical and Translational Research. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1532046410001346

[4] T. Mabotuwana, M. C. Lee, and E. V. Cohen-Solal, "An ontology-based similarity measure for biomedical data application to radiology reports," *Journal of Biomedical Informatics*, vol. 46, no. 5, pp. 857 – 868, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1532046413000889

[5] G. Salton, "The smart retrieval systemexperiments in automatic document processing," 1971.

[6] K. Shin, S.-Y. Han, A. Gelbukh, and J. Park, "Advanced relevance feedback query expansion strategy for information retrieval in medline," in *Progress in Pattern Recognition, Image Analysis and Applications*. Springer, 2004, pp. 425–431.

[7] M. Daoud, D. Kasperowicz, J. Miao, and J. Huang, "York university at trec 2011: Medical records track." in *TREC*, 2011.

[8] S. Fernando and M. Stevenson, "A semantic similarity approach to paraphrase detection," in *Proceedings of the 11th Annual Research Colloquium of the UK Special Interest Group for Computational Linguistics*. Citeseer, 2008, pp. 45–52.

[9] L. Han, A. Kashyap, T. Finin, J. Mayfield, and J. Weese, "Umbc ebiquity-core: Semantic textual similarity systems," in *Proceedings of the Second Joint Conference on Lexical and Computational Semantics*, vol. 1, 2013, pp. 44–52.

[10] T. Pedersen, S. V. Pakhomov, S. Patwardhan, and C. G. Chute, "Measures of semantic similarity and relatedness in the biomedical domain," *Journal of biomedical informatics*, vol. 40, no. 3, pp. 288–299, 2007.

[11] M. Batet, D. Sánchez, and A. Valls, "An ontology-based measure to compute semantic similarity in biomedicine," *Journal of biomedical informatics*, vol. 44, no. 1, pp. 118–125, 2011.

[12] C. Leacock and M. Chodorow, "Combining local context and wordnet similarity for word sense identification," *WordNet: An electronic lexical database*, vol. 49, no. 2, pp. 265–283, 1998.

[13] Z. Wu and M. Palmer, "Verbs semantics and lexical selection," in *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 1994, pp. 133–138.

[14] P. Resnik, "Using information content to evaluate semantic similarity in a taxonomy," *arXiv preprint cmp-lg/9511007*, 1995.

[15] D. Lin, "Extracting collocations from text corpora," in *First workshop on computational terminology*. Citeseer, 1998, pp. 57–63.

[16] B. T. McInnes, T. Pedersen, and S. V. Pakhomov, "Umls-interface and umls-similarity: open source software for measuring paths and semantic similarity," in *AMIA Annual Symposium Proceedings*, vol. 2009. American Medical Informatics Association, 2009, p. 431.

[17] R. Pivovarov and N. Elhadad, "A hybrid knowledge-based and data-driven approach to identifying semantically similar concepts," *Journal of biomedical informatics*, vol. 45, no. 3, pp. 471–481, 2012.

[18] X. Zhang, L. Jing, X. Hu, M. Ng, and X. Zhou, "A comparative study of ontology based term similarity measures on pubmed document clustering," in *Advances in Databases: Concepts, Systems and Applications*. Springer, 2007, pp. 115–126.

[19] S. Begum, M. U. Ahmed, P. Funk, N. Xiong, and B. Schéele, "Similarity of medical cases in health care using cosine similarity and ontology," in *ICCBR*, 2007, pp. 263–272.

[20] G. B. Melton, S. Parsons, F. P. Morrison, A. S. Rothschild, M. Markatou, and G. Hripcsak, "Inter-patient distance metrics using {SNOMED} {CT} defining relationships," *Journal of Biomedical Informatics*, vol. 39, no. 6, pp. 697 – 705, 2006. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1532046406000207

[21] R. Moskovitch and Y. Shahar, "Vaidurya: a multiple-ontology, concept-based, context-sensitive clinical-guideline search engine," *Journal of Biomedical Informatics*, vol. 42, no. 1, pp. 11–21, 2009.

[22] F. Farfán, V. Hristidis, A. Ranganathan, and R. P. Burke, "Ontology-aware search on xml-based electronic medical records," in *Data Engineering, 2008. ICDE 2008. IEEE 24th International Conference on*. IEEE, 2008, pp. 1525–1527.

[23] F. Wang, J. Hu, and J. Sun, "Medical prognosis based on patient similarity and expert feedback," in *Pattern Recognition (ICPR), 2012 21st International Conference on*. IEEE, 2012, pp. 1799–1802.

[24] R. Baeza-Yates, B. Ribeiro-Neto *et al.*, *Modern information retrieval*. ACM press New York, 1999, vol. 463.

[25] D. L. Rubin, "Creating and curating a terminology for radiology: ontology modeling and analysis," *Journal of digital imaging*, vol. 21, no. 4, pp. 355–362, 2008.

[26] E. Agichtein, E. Brill, and S. Dumais, "Improving web search ranking by incorporating user behavior information," in *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2006, pp. 19–26.

[27] T. Joachims, "A probabilistic analysis of the rocchio algorithm with tfidf for text categorization." DTIC Document, Tech. Rep., 1996.

[28] ——, *Text categorization with support vector machines: Learning with many relevant features*. Springer, 1998.

[29] C. Jordan and C. Watters, "Extending the rocchio relevance feedback algorithm to provide contextual retrieval," in *Advances in Web Intelligence*. Springer, 2004, pp. 135–144.

[30] J. A. Aslam and E. Yilmaz, "A geometric interpretation and analysis of r-precision," in *Proceedings of the 14th ACM international conference on Information and knowledge management*. ACM, 2005, pp. 664–671.