Journal of Digital Imaging

Ontology-Assisted Analysis of Web Queries to Determine the Knowledge Radiologists Seek

Daniel L. Rubin,[1] Adam Flanders,[2] Woojin Kim,[3] Khan M. Siddiqui,[4,5] and Charles E. Kahn Jr[6]

Radiologists frequently search the Web to find information they need to improve their practice, and knowing the types of information they seek could be useful for evaluating Web resources. Our goal was to develop an automated method to categorize unstructured user queries using a controlled terminology and to infer the type of information users seek. We obtained the query logs from two commonly used Web resources for radiology. We created a computer algorithm to associate RadLex-controlled vocabulary terms with the user queries. Using the RadLex hierarchy, we determined the high-level category associated with each RadLex term to infer the type of information users were seeking. To test the hypothesis that the term category assignments to user queries are non-random, we compared the distributions of the term categories in RadLex with those in user queries using the chi square test. Of the 29,669 unique search terms found in user queries, 15,445 (52%) could be mapped to one or more RadLex terms by our algorithm. Each query contained an average of one to two RadLex terms, and the dominant categories of RadLex terms in user queries were diseases and anatomy. While the same types of RadLex terms were predominant in both RadLex itself and user queries, the distribution of types of terms in user queries and RadLex were significantly different ($p<0.0001$). We conclude that RadLex can enable processing and categorization of user queries of Web resources and enable understanding the types of information users seek from radiology knowledge resources on the Web.

KEY WORDS: Ontologies, terminologies, vocabularies, RadLex, software tools, controlled vocabulary, natural language processing, web technology

INTRODUCTION

A wealth of radiology information is disseminated on the Web, and radiologists are increasingly searching a variety of Websites in their daily work. It would be helpful to evaluate the quality of Web resources of radiological information. To do this, it would be important to know what information radiologists seek, as coverage of radiologist information needs is an important metric of resource quality.

It is difficult to assess how radiology Web resources are being used and what information radiologists seek from them because these sites receive thousands of queries; reviewing query logs by hand would be a formidable task to undertake; thus, it would not be feasible for answering this question. Computer analysis of query logs could be an approach to this problem; however, computer analysis of query logs is challenging because users type their queries as unstructured text, and there is currently no automated way to deduce the types of information being sought from such raw queries.

Controlled terminologies and ontologies could provide a technological approach to providing

[1]*From the Department of Radiology, Stanford University, Richard M. Lucas Center, 1201 Welch Road, Office P285, Stanford, CA, 94305-5488, USA.*

[2]*From the Thomas Jefferson University, Philadelphia, PA, USA.*

[3]*From the Radiology, University of Pennsylvania, Philadelphia, PA, USA.*

[4]*From the VA Maryland Healthcare System, Baltimore, MD, USA*

[5]*Present Address: From the Health Solutions Group, Microsoft Corporation, Redmond, WA, USA.*

[6]*From the Radiology, Medical College of Wisconsin, Milwaukee, WI, USA.*

Correspondence to: Daniel L. Rubin, Department of Radiology, Stanford University, Richard M. Lucas Center, 1201 Welch Road, Office P285, Stanford, CA, 94305-5488, USA; tel: +1-650-7255693; fax: +1-650-7235795; e-mail: dlrubin@stanford.edu

Online publication 31 March 2010

doi: 10.1007/s10278-010-9289-2

computers the means of automating the process of determining the types of information radiologists seek in their queries of Web resources. Ontologies in particular are potentially useful since they contain knowledge about how terms relate to other terms, such as synonymy and subsumption.[1,2]

RadLex is a controlled terminology for radiology.[3,4] In addition to a list of standard terms, RadLex contains a taxonomy that organizes the terms into category hierarchies. These hierarchies provide subsumption classifications wherein more abstract terms are the parents of more granular terms. The high-level ancestor classes in RadLex are broad categories, such as "anatomy" and "radiology findings." Such categories classify more granular terms—terms actually used in user search. Thus, the term hierarchy of RadLex could be used to classify user search terms to determine the type of searches users are performing. In addition, RadLex contains synonyms—terms related to a preferred term by an "is-synonym" relationship. An application could thus recognize synonyms, map them to the preferred term, and then follow the subsumption hierarchy to deduce the type of search applicable to that synonym. Accordingly, RadLex could be a valuable resource for classifying user queries for this task.

Our hypothesis is that RadLex can enable processing and categorization of user queries of radiology knowledge resources on the Web. Our goal in this work is to (1) demonstrate that it is possible to create an automated computer algorithm to find occurrences of RadLex terms in user queries and (2) show that by evaluating the RadLex terms in user queries, it is possible to classify them, indicating the categorical types of information that users are seeking.

METHODS

Source Data

We obtained the query logs of the ARRS GoldMiner (http://goldminer.arrs.org) and Yottalook (http://yottalook.com) Web search engines. These logs recorded the phrases the users of these resources entered when performing searches, dating back to the inception of their tracking these queries. The Web log data were anonymous; no user information or identifiable IP address was provided. A consecutive sample of search query terms was obtained for each search engine; a total of 30,000 search requests were used for this investigation. In the query logs, each line was a string representing a user query.

We also downloaded RadLex[1] (as of September 1, 2007) in Protégé format. Protégé (http://protege.stanford.edu) is a tool that provides a means to computationally access ontologies and terminologies and use them in applications.[5,6]

Automated RadLex Mapping Algorithm

We created a computer algorithm to find occurrences of RadLex terms in the queries submitted by users of Web resources. The algorithm examines the raw search text and attempts to find the best matching RadLex term(s) using two approaches. First, it looks for an exact match between the user query string and RadLex. Second, the algorithm examines alternate word orderings (permutations) within the user query to find variations in term names that match RadLex. Punctuation is not included in considering search phrases. For example, the algorithm would map both "cavernous hemangioma" and "hemangioma, cavernous" to the same RadLex term. In order to find multiple RadLex terms that may occur in the user queries, the algorithm uses a shifting phrase frame, evaluating each consecutive six-word phrase in the query and permuting subsets of the words in each phrase for a match to RadLex. This permutation-based approach for mapping text to controlled terms is very similar to that described by Shah,[7] in which high precision (86%) was achieved.

All RadLex terms are associated with a particular descriptive category (such as findings, anatomic location, etc.), which can be inferred from the RadLex hierarchy; the categories of terms are the top-level RadLex terms. The particular descriptive category for a given RadLex term was determined in our algorithm by tracing the subsumption relationships in the RadLex hierarchy to these high-level terms. For example, the descriptive category for "Heart" is deduced by the following is-a path in RadLex: Heart >> Mediastinum >> Thorax >> Anatomic Location; thus, the descriptive category for "Heart" is "Anatomic Location." We used the descriptive category associated with RadLex terms to indicate the type of information users were seeking

(e.g., if a user searched for "Heart," then the descriptive category of information the user is seeking is "Anatomic Location").

Analysis of Queries

We implemented the query mapping algorithm and code to calculate term metrics using Python scripts. The scripts accessed RadLex terms using the Protégé suite of tools which provides an interface for running scripts that access ontologies in Protégé.[2,3]

Our scripts counted the number of RadLex terms in each descriptive category found in the user queries. We also counted the number of terms under each descriptive category of RadLex. We compared the frequency distributions of descriptive categories in RadLex with those used in the user queries using the chi square test. The null hypothesis was that there is no difference in these distributions of term categories, and the threshold for significance selected was $p=0.05$.

RESULTS

Composition of RadLex and Its Term Categories

The version of RadLex we used contained 7,413 different terms, and each term belonged to a single category. Most terms were in the Anatomic Location and Diseases categories, though there were also many Radiological Findings/Modifiers terms (Table 1). While there is a category called Image Acquisition, Processing, and Display, there were no RadLex terms of this type.

Table 1. Composition of RadLex and Its Term Categories

RadLex term category	Term count (% of all terms)
Anatomic location	3,874 (52.3%)
Diseases	2,416 (32.6%)
Findings/modifiers	792 (10.7%)
Image acquisition, processing, and display	0
Treatment	290 (3.9%)
Image quality	4
Relationship	15 (0.2%)
Teaching attribute	12 (0.2%)
Uncertainty	10 (0.1%)
All categories	7,413

Tabulation of the categories of RadLex terms and the number (and percentage of the total) of terms in each category

Analysis of User Queries and Occurrences of Types of Terms They Contain

There were 29,669 unique search terms (comprising single words or multiword phrases). Of these, 15,445 (52%) of the user queries could be mapped to one or more RadLex terms by our algorithm. The algorithm mapped a total of 23,036 RadLex terms to the queries, with different proportions among the RadLex descriptive categories (Table 2). More than one RadLex term often was contained in a given query, with each query containing an average of one to two RadLex terms (Table 2).

Examination of the frequencies of the RadLex term categories indicated the types of information users were seeking: the most frequent categories of terms in searches related to Diseases (10,882 search terms; 47%), Anatomy (6,153 search terms; 27%), and Image Findings/Modifiers (5,831 search terms; 25%). In addition, there were 4,472 (19%) user queries where findings and anatomy occurred together within the same query, indicating that users were searching for diseases relating to particular anatomy. None of the user queries were related to image quality or uncertainty, and very few were categorized as Relationship and Teaching Attribute. The 10 most frequently searched diseases and findings were cyst, fracture, mass, tumor, carcinoma, calcification, tear, hemorrhage, aneurysm, and hemangioma. The 10 most frequently searched anatomic locations were brain, liver, lung, artery, chest, knee, hip, bowel, pancreas, and kidney.

The dominance of particular categories of RadLex in user queries, such as diseases and anatomy, was similar to the distribution of term categories in RadLex itself (compare Tables 1 and 2). However, the distributions were significantly different: the chi square value was highly significant (42.2, $p<0.0001$), implying we should reject the null hypothesis.

We also analyzed the user queries in terms of RadLex usage and to estimate how much of RadLex is not represented in user queries (Table 3). The distributions of descriptive term categories in user queries and among the unique RadLex terms that mapped to the user queries were not significantly different (compare Tables 2 and 3; chi square=9.2; $p>0.05$). More than half of all RadLex Disease category terms were used in user

Table 2. Occurrences of Search Terms in Radiological Queries

RadLex term category	Phrase count (% of all terms)	Mean number of phrases	SD number of phrases
Anatomic location	6,153 (26.7%)	1.2	0.5
Diseases	10,882 (47.2%)	1.2	0.4
Findings/modifiers	5,831 (25.3%)	1.0	0.2
Image acquisition, processing, and display	0	–	–
Treatment	145 (0.6%)	1.3	0.5
Image quality	0	–	–
Relationship	14 (0.1%)	1.1	0.3
Teaching attribute	11	1.0	0
Uncertainty	0	–	–

Number of unique phrases in user queries that map to RadLex. Table shows the total count of phrases (and percentage of the total phrase count) according to the category of the RadLex term to which they map. The mean and SD of the number of phrases contained in user queries mapping to RadLex are also shown

queries; however, for all the other term categories, most RadLex terms was not used (Table 3).

DISCUSSION

Our work demonstrates the utility of using RadLex to classify user queries that can suggest the types of information they seek. The value of RadLex in providing a controlled terminology for radiology reporting and teaching is well known.[3,4] In this work, we have shown the value of RadLex from the "ontological" perspective—RadLex specifies relationships between terms and categories to which they belong. By mapping user queries to RadLex terms, and then by studying the high-level descriptive categories to which those terms belong, we learned about the types of information users of Web resources are seeking (Table 2). Our use of RadLex is similar to the use of ontologies as knowledge sources in other radiology applications.[8]

We found that the distribution of RadLex term categories for user queries differs significantly from the distribution of those categories in RadLex itself (Compare Tables 1 and 2). This suggests that the choice of terms in user queries carries information—if all queries were simply randomly selected RadLex terms, then the distribution of term categories would have been close to that occurring in RadLex (in general, we would not expect the frequency of terms in a comprehensive terminology to mirror the frequency of search terms in the domain; users generally search for a subset of all available information). On the other hand, there were many RadLex terms that were not used at all in user queries (Table 3). This could be due to the fact that our algorithm missed RadLex terms that could have mapped to the user queries; alternatively, this could indicate that RadLex

Table 3. Occurrences of types of RadLex Terms in Radiological Queries

RadLex term category	Term count (% of all terms)	Mean number of terms	SD number of terms	RadLex terms not used (%)
Anatomic location	471 (24.4%)	1.9	0.7	3,403 (75.6%)
Diseases	1,034 (53.7%)	1.6	0.8	1,382 (46.3%)
Findings/modifiers	367 (19.1%)	1.1	0.3	425 (80.9%)
Image acquisition, processing, and display	0	–	–	0
Treatment	49 (2.5%)	1.4	0.6	241 (97.5%)
Image quality	0	–	–	4 (100%)
Relationship	4 (0.2%)	1.3	0.4	11 (99.8%)
Teaching attribute	2 (0.1%)	1.0	0	10 (99.9%)
Uncertainty	0	–	–	10 (100%)

Number of unique RadLex terms (and percentage of the total) contained in user queries. The mean and SD of the number of RadLex terms contained in user queries mapping to RadLex are also shown. The last column shows the number (and percentage) of RadLex terms that were not used in user queries

contains many terms not used by users searching radiology Web resources.

Our work is limited in that it is an indirect assessment of user information needs based on the categories of RadLex terms in queries. Ideally, we would have questioned each user about their information needs, though this would have been challenging to do given the number of users of Web resources. A second limitation is that we have not yet compared our results against a gold standard; however, the goal in this work was to show potential utility of RadLex in shedding light on user information needs, a task that would be very difficult by analyzing the raw user queries alone.

A final limitation is that our mapping algorithm is also limited in that it does not account for synonymy and lexical variations. RadLex has links to synonyms, so our method could be improved in the future to exploit this knowledge.

Recently, a newer version of RadLex has been released (http://bioportal.bioontology.org/ontologies/40885) with greatly expanded content, particularly in naming radiology procedures and procedure steps (the version of RadLex used in our study contains terms acquired during the first 2 years of terminology development). RadLex continues to evolve and will likely grow in the future. Evolution of RadLex could certainly affect the specific results of the analysis of user queries reported in our study. For example, we found that no RadLex terms of type "Image Acquisition, Processing, and Display" mapped to the user queries. This is likely due to the fact that in the version of RadLex we used, there were only 20 RadLex terms under this category; in the current release of RadLex, many more terms for radiology procedures have been added, and more user queries would likely be found that map to these new RadLex terms. Our main objective in this paper was to demonstrate that RadLex can enable processing and categorization of user queries of radiology knowledge resources and to indicate the types of information that users are seeking. The analysis of user queries can be updated as RadLex evolves, and this can be semi-automated by using the RadLex mapping algorithm we developed.

We believe the methods we describe in this work are extensible beyond analysis of query logs of Web resources. Mapping free text to RadLex could be useful for indexing and analyzing the text of articles, teaching files, or databases of radiology reports. By categorizing the RadLex terms to higher-level descriptive term categories, users can search for radiology information at varying levels of term granularity.

In conclusion, the RadLex terminology enables analysis of user queries that can suggest the kinds of radiology information they seek. The majority of queries can be indexed with RadLex, and the most common types of queries relate to diseases, and searches for other types of information are less frequent. RadLex may be useful for analyzing and summarizing other types of radiology texts.

ACKNOWLEDGEMENTS

The authors thank the American Roentgen Ray Society and iVirtuoso Inc. for access to the query logs. This work was supported in part by the American Roentgen Ray Society.

REFERENCES

1. Bodenreider O, Stevens R: Bio-ontologies: current trends and future directions. Brief Bioinform 7:256–274, 2006

2. Rubin DL, Shah NH, Noy NF: Biomedical ontologies: a functional perspective. Brief Bioinform 9:75–90, 2008

3. Langlotz CP: RadLex: a new method for indexing online educational materials. Radiographics 26:1595–1597, 2006

4. Kundu S, et al: The IR RadLex project: an interventional radiology lexicon–a collaborative project of the Radiological Society of North America and the Society of Interventional Radiology. J Vasc Interv Radiol 20:433–435, 2009

5. Gennari JH, et al: The evolution of Protege: an environment for knowledge-based systems development. Int J Human-Comput Stud 58:89–123, 2003

6. Rubin DL, Noy NF, Musen MA: Protege: a tool for managing and using terminology in radiology applications. J Digit Imaging 20(Suppl 1):34–46, 2007

7. Shah NH, Rubin DL, Espinosa I, Montgomery K, Musen MA: Annotation and query of tissue microarray data using the NCI Thesaurus. BMC Bioinformatics 8:296, 2007

8. Kahn Jr, CE, Channin DS, Rubin DL: An ontology for PACS integration. J Digit Imaging 19:316–327, 2006