2005

An ontology based approach for health information discovery

Khin T. Win
University of Wollongong

Minjie Zhang
University of Wollongong, minjie@uow.edu.au

Publication Details

Win, K. T. & Zhang, M. (2005). An ontology based approach for health information discovery. In P. Santiprabhob & J. Daegdej (Eds.), International Conference on Intelligent Technologies (pp. 295-300). Thailand: Faculty of Science and Technology, Assumption University.

An ontology based approach for health information discovery

Abstract
The need of searching and retrieving relevant information has become important in the health care delivery. In this paper, we discuss how domain specific ontologies can support the information selection and introduce an ontology based approach for health information discovery in distributed environments. This research is the preliminary result of a health information retrieval project.

Publication Details
Win, K. T. & Zhang, M. (2005). An ontology based approach for health information discovery. In P. Santiprabhob & J. Daegdej (Eds.), International Conference on Intelligent Technologies (pp. 295-300). Thailand: Faculty of Science and Technology, Assumption University.

An Ontology Based Approach for Health Information Discovery

Khin Than Win and Minjie Zhang
School of IT and Computer Science, University of Wollongong
Wollongong NSW 2522, Australia
{win, minjie}@uow.edu.au

Abstract - **The need of searching and retrieving relevant information has become important in the healthcare delivery. In this paper, we discuss how domain specific ontologies can support the information selection and introduce an ontology based approach for health information discovery in distributed environments. This research is the preliminary result of a health information retrieval project.**

Keywords- health information, ontology, information retrieval,

1. INTRODUCTION

The need of accurate and relevant health information retrieval is important for healthcare providers and healthcare consumers. Accuracy of health information is essential whether the health information is available on the World Wide Web or in different health information systems.

1.1. Health information on the web

There is a need for efficient and reliable information retrieval for health information available for consumers on the web. It was estimated in 2000 that there were over 70,000 healthcare websites [7,3] and it is growing. Studies indicate that consumers surf the web for health information to find out more information about treatments and to assist in healthcare decision making [5,12]. In health domain, inaccurate or error in information can have tremendous impact on a person's health and it is important that the information available is of high quality information. Anyone can publish any information on the web so it is important to ensure the quality of information [22]. It was reported that there were over 1400 suspicious websites in 1996 with an annual increase of 21 percent [19]. According to the survey that was conducted by Harris Poll in USA, about 110 million individuals sought on-line health information in 2002 [23].

1.2. Health information in health record systems

Healthcare institutions around the world are encouraged to use the electronic health record systems (EHRs). Electronic health record is defined as "the electronic longitudinal collection of personal health information, usually based on the individual or family, entered or accepted by health care professionals, which can be distributed over a number of sites or aggregated at a particular source, including a hand-held device" [18].

Therefore, it is important that data contained in EHRs represent the patient's information from cradle to grave. Therefore, integration of data from different EHRs is needed. Nevertheless, correct matching or integration of data is important, as wrong or incomplete data can have tremendous impact on the person's health, or the data of research and public health according to the data involved. There are a lot of issues regarding interoperability in integration of health information systems. The semantic interoperability is one of the important issues [25]. For example, one clinician may note that a patient has "shortness of breath" and another physician may note that as a "dyspnoea". These synonyms are not noted in the searching application so their synonymous condition will not be realised. Therefore, these need to be considered in health information integration, search or discovery. Currently, there are different

standards for the information exchange of standards such as SNOMED CT, LONIC and RxNorm and UMLS contains terminology and information regarding different standards. Health Level 7, HL7 Standards development organisation has been looking into different terminologies and presentations. However, it is important if heterogenous data resources can be queried successfully.

1.3. Importance of health information retrieval

The availability of health information to healthcare provider is important, whether it is from the EHRs or from the web. The healthcare providers should be able to access the relevant information for healthcare decision making and effective treatment. Researchers need patient data in order to undertake the quality research to provide useful knowledge for the future well being of the health care industry.

Evidence-based medicine and evidence-based informatics have been the current trend in health informatics emphasizing safety and quality of healthcare. It has been recommended that medical care should be based as much as possible on the best available evidence from the scientific research rather than on expert opinion or physician's own experience [14,21]. Therefore, literature search from published materials and available information from computerized health record systems is essential for evidence-based practices.

1.4. Health information search

In healthcare domain, the Unified Medical Language System (UMLS) is a knowledge representation framework for biomedical research queries, which includes over 100 medical terminology sources [13]. Metathesaurus and Semantic Network are the two major sources for the UMLS. The Metathesaurus contains a large collection of concepts and the Semantic Network contains the semantic types that form an abstraction of the Metathesaurus [4]. Semantic matching is not generally available [4], although there is mapping of free text to UMLS concepts [1] and UMLS based applications that allow natural language inputs [10].

Earlier efforts for health information retrieval also include the metadata for web documents such as the Dublin Core Metadata Initiatives (DCMI) for effective information retrieval proposed by Malet et al [16]. It involves a unique resource identifier and character string for information retrieval. Authors of web sites need to describe their subjects according to the MCM-MeSH metadata tag. However, the actual usage of DCMI tags in web pages is very low, estimated at only 0.3 % [15]. As there is a very low percentage of actual usage, searching of documents through DCMI initiatives is not very useful.

Health information retrieval can be conducted by using a domain specific thesaurus by using intelligent agents to retrieve the information on behalf of the user and providing information through specialised portals that only provide filtered information [24]. Current health information retrieval systems use manual indexing to filter health information and this is a significant drawback of these systems.

With UMLS, disease database users can query according to their needs. However, the search is based on the single word search and it will not allow fever, pain and cough to reflect. It is not an intelligent search tool to include differential diagnosis [17]. The disease database listed different diseases and symptoms. If different concepts are queried together, the search will be based on one concept at a time and results will be the search outcomes from 2 different queries. Information retrieval through ontology-based approach will solve the unpredictable queries from users. It could also assist in integration of heterogenous health information system for interoperability. Therefore, consumer oriented information discovery is important.

In healthcare domain, inaccuracy or error in information can have tremendous impact on a person's health and it is important that the information available is high quality information. Previously, until November 2003, CliniWeb from the Oregon Health Sciences University provided quality health information search targeted to consumers. However, the user cannot search more than one disease at a time. CliniWeb required ongoing human indexing [11]. Because of discontinuation in funding, CliniWeb has discontinued and is not available on the Web anymore [9].

1.5. Problems in information source selection

In www environment, there are two key problems in current information source selection mechanisms: (i) Imprecision and ambiguity of user queries and (ii) the development of suitable search mechanisms which guarantee to select relevant information sources as many as possible.

Current information retrieval systems mostly employ the keyword-based approaches to perform a full text analysis based on word occurrence based on information resources. Only the information sources where the user specified words that frequently occur in documents will be chosen [27]. Current information retrieval mechanisms provided by the search engines are based on either the keyword-based search (e.g. Lycos) or the directory-based search (e.g. Yahoo).

As described in section 1.2, there is semantic heterogeneity in health information sources and semantically related information sources will be missed because they do not contain the keywords specified by the user. To overcome this problem, we propose a content-based approach in this paper for the matching between the query and information sources by using ontology management. Semantic interoperability is achieved by using semantic relations among terms in the ontology. Therefore, it allows implicit information sources beyond the capacity of the keyword-based approach.

The rest of the paper is managed as follows. Domain specific ontology is introduced in Section 2; How to decrease ambiguity is described in Section 3; the calculation of similarity scores of information sources to query is introduced in Section 4. Finally, the conclusion and further work are outlined in Section 5.

2. ONTOLOGY

An ontology is a shared understanding in a knowledge domain that can be communicated across people and systems [8]. Therefore, ontologies can be used as a language for communicating between different systems in a distributed, heterogeneous environment. It enables a solution to the problem of semantic heterogeneity. In this research, we focus on the use of domain-specific ontologies for resource discovery. The ontologies serve as a means for establishing a conceptually concise basis for communicating knowledge. A domain specific ontology is a shared and common understanding of a particular information domain.

Figure 1 illustrates an example of Disease ontology. Concepts are linked with the lines of different shapes that denote various kinds of relationships. A domain specific ontology specifies a conceptualization of a domain in terms of concepts. Each concept represents a class for a specific set of entities.

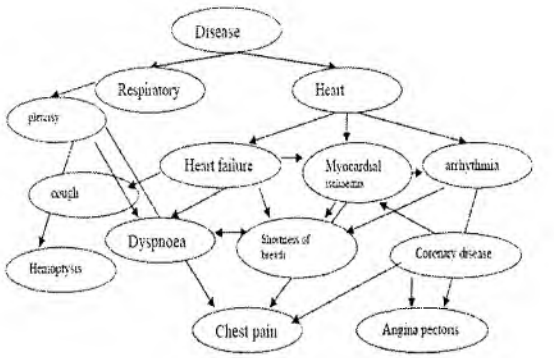

Figure 1: Disease Ontology

It is characterized by a unique label name in the ontology, and is usually expressed as a word with synonyms. For example, concept 'Myocardial ischemia' has a synonymous list, which consists of 'Coronary Disease', and 'Angina Pectoris'.

The concepts are typically organized into a taxonomy tree where each node represents a concept. Concepts are linked together by means of their semantic relationships. The set of concepts together with their links form a semantic network. Various kinds of semantic relationships are maintained among the concepts. Among these, the most relevant to our purpose is Part-Of (Subsumption) relationship which allows a set of concepts to be organized according to a generalization hierarchy. For instance, concept 'Disease' is more general than its subclass concept 'Heart Disease'.

We have chosen the concept terminology used to describe and represent domain of knowledge. The following three major factors need to be considered in the construction of domain specific ontologies.

1. To identify the key concepts in a domain. These concepts are reflected in the representational vocabulary of domain background knowledge, which preserve

coverage of the domain while constructing an ontology as compact as possible.

2. To clearly and accurately arrange the concepts in terms of semantic relationships. The relationships at the concept level reflect various strength of correlations among concepts.
3. To assign proper names to the concepts in a domain.
   Consensus on the concept names is important for the use of the ontology when the users or the information sources adopt a set of previously defined concept names to formulate the user request or represent the web contents.

In the context of multiple, distributed domain specific ontologies, it is impractical to create these ontologies from scratch due to the complexity in the domain specific ontologies. The reuse of existing ontology is an alternative. Moreover, there is a UMLS for medical data so we have chosen to examine UMLS to construct our domain specific ontologies for health information discovery.

3. QUERY DISAMBIGUATION IN THE PROCESS OF QUERY FORMULATION

In query formulation, query ambiguity is a generally acknowledged issue in the information retrieval, where users usually present their queries with one or two words. Such short queries are unclear to clarify the users' underlying intention. It is therefore imperative that a query model designed for users to request information should be functionally powerful enough to assist users in avoiding ambiguous queries [26].

In this research, we explored an ontology-based query model that resolves ambiguity in the process of query formulation. After a user poses a short query, which is ambiguous, the query is reformulated in terms of domain specific ontologies. The model extracts a list of terms that will be semantically related to the query and then presents the term list to the user as options to consider. These additional terms provide sufficient context to clear up ambiguity.

According to our observation, there are two important factors that affect the precision of a formulated query: one is the terms (words) that describe what the user is interested in. The term level ambiguity arises from multiple different meanings (word senses) that query terms might have; the other is the semantic relation implicated by the query terms. The relation-level ambiguity occurs when a semantic gap exists between the formulated query and the intention of the user. The clarity of a query can be hurt by either of these two factors. In order to achieve the disambiguation of queries, we investigate a disambiguation approach which consists of two aspects: query modification and query refinement. Query modification is motivated by the term-level ambiguity. In this case, a set of terms that occur together with the ambiguous query term within a particular context are provided to aid the user to determine the appropriate sense of this query term. At the relation-level ambiguity, query refinement is concerned with adding semantic relationships by which the ambiguous term is related to other concept in the ontology so as to remedy the semantic gap. In the remainder of this section, we show some examples that illustrate how our proposed disambiguation approach can reduce ambiguity.

At the relation-level ambiguity, it is possible that a particular keyword is associated with one or more concepts in the ontology. For example, the term 'chest pain' might be an instance of both concept 'coronary disease' and concept 'heart disease' in the ontology. Although a user probably requests the information on chest pain regarding the coronary disease, the query 'chest pain' will result in the retrieval of irrelevant information such as the detailed information regarding diseases such as pleurisy or pericarditis. As a result, retrieval precision will be affected. In order to refine the query, we need to specify some semantic relationships (so-called semantic constraints) of the query term with other concept in the ontology in order to improve the semantic gap. In this example, if we add an 'Instance-Of' relationship associated with the query, the ambiguity will be greatly reduced. In summary, using domain-specific ontologies, ambiguous queries can be clarified either through the process of resolving different word meanings with query terms or through the process of refinement, depending on the nature of the ambiguity.

In order to select appropriate information sources, it is necessary to find relevant concepts in ontologies that match the query terms in the query. Assume that a user query Q consists of a set of query terms, $Q = \{q_1, q_2, \ldots q_i\}$. Note that, here, the query Q has been preprocessed by stemming and removing stopwords. To overcome semantic heterogeneity (e.g., using

different names to express the same intended meaning), we will use concepts in domain-specific ontologies to represent the query instead of query terms. As explained previously, a concept typically has a label name and a list of synonyms. We treat the label name and the synonymous list as the textual content of this concept. In addition, a set of possible concept instances associated with this concept will be additional information for consideration. Since a concept instance is only an example of concept specialization, the terms in the instance set are far less important than ones in the label name or the synonymous list during query matching. Therefore, we assign lower weights to the terms in the instance set. Then, the text content of a concept C can be described as

$$C = \{t_1 w_1, t_2 w_2, \ldots . t_u w_u\} \qquad (1)$$

where term t_j ($1 \leq j \leq u$) is a word which occurs in the text content of the concept C, and w_j is the relevant weight associated with the term t_j. Note that w_j is normalized and $\sum_j w_j = 1$.

So the relevance score of a concept C to a query Q can be calculated as

$$\text{relevance\_score}(Q, C) = \sum_{i=1}^{i=t} q_i w_i$$

where w_i is the weight associated with the query term q_i which occurs in the text content of the concept C. If the relevance score is greater than a relevance threshold τ, this concept C will be selected as a query concept with respect to the query Q.

4. THE CALCULATION OF SIMILARITY SCORES OF INFORMATION SOURCES TO THE QUERY

Once the relevant query concepts have been identified from domain-specific ontologies, the next step is to calculate the similarity scores of information sources to the query concepts. Content-related metadata in information sources play an important role in measuring the similarity scores of information sources to the query. In this paper, our content-related metadata extraction method is text-based, which mainly focuses on the content-related information found in HTML tags such as the title or a heading element, and metatags for keywords and descriptions. They are always the primary source of text features. Sometimes, when the information in the above HTML tags is insufficient, body text of the web page will be considered for analysis. In addition, the hyperlink structure of the web can also be exploited by using the anchor text and the metatag contents from linking documents as another source of text features [2]. All extracted text features are concatenated into a single representative document as the resource description of the information source. Note that all the text features here have been preprocessed by stopword removal and stemming.

To obtain the similarity score of an information source to the query, we use the textual contents of query concepts to match the resource description of the information source. Similarity measurement between the resource description X of an information source and the textual content of query concepts Y is calculated using the Dice Coefficient [20]:

$$\text{Simi}(X,Y) = 2\,\frac{X \cap Y}{X \cup Y}$$

The more the words in the textual contents of query concepts occur in the resource description, the greater the similarity score will become.

Finally, information sources are ranked by the similarity score, and those top-ranking ones will be chosen as relevant to the query.

5. CONCLUSION AND FUTURE WORK

In this paper, we have introduced the novel approach for intelligent selection of health information resources. We also discussed the importance of an ontology-based approach for the health information discovery. Our approach aimed to discover alternative query approach from CliniWeb and allows direct text word searching and will use the Boolean operators to connect multiple text words using UMLS metathesaurus. Future work will be discovering domain specific query by use of test data. Our current work is at the preliminary stage of a health information retrieval project. We will apply agent-based technology for indexing and searching of appropriate information in real applications. In summary, we have identified a system for consumer oriented information discovery in healthcare.

REFERENCE

1. Aronson, A., Meta-map: Mapping Text to the UMLS Metathesaurus, Semantic Knowledge Representation

Research Information Project, available at http://skr.nlm.nih.gov/papers

2. Attardi, G., Gull A. and Sebastiani, Automatic Web Page Categorization by Link and Context Analysis, *Proceedings of THAI-99, European Symposium on Telematics, Hypermedia and Artificila Intelligence*, Vareses, Italy, 105-119, 1999.

3. Benigeri, M. and Pluye, P., Shortcomings of Health Information on the Internet, *Health Promotion International*, Vol. 18, No. 4, 381-386, 2003.

4. Caviedes, J.E. and Cimino, J., Towards the Development of a Conceptual Distance Metric for the UMLS, *Journal of Biomedical Informatics*, Vol. 37, pp.77-85, 2004.

5. Charnock, D. and Shepperd, S., Learning to DISCERN Online: Applying an Appraisal Tool to Health Websites in a Workshop Setting, *Health Education Research*, Vol. 19, No. 4, 440-446, 2004.

6. Eysenbach, G., Sa, E.R. and Diepgen, T.L, Shopping Around the Internet Today and Tomorrow: Towards the Millennium of Cybermedicine, *British Medical Journal*, Vol. 319, pp. 1294–1298, 1999.

7. Grandinetti, D. A., Doctors and the Web, Doctors and the Web: Help Your Patients Surf the Net Safely. *Medical Economics*, April 28-34, 2000.

8. Gruber, T. R., Towards Principles for the Design of Ontologies Used for Knowledge Sharing. In: N. G. a. R. Poli, eds. *Formal Ontology in Conceptual Analysis and Knowledge Representation*. Kluwer Academic Publishers, Deventer, The Netherlands, 1993.

9. Hersh, W., CliniWeb International, available at http://www.ohsu.edu/cliniweb.

10. Hersh, W., Ball, A., Day, B., Masterson, M., Zhang, L. and Sacherek, L., Maintaining a Catalog of Manually-Indexed, Clinically Oriented World Wide Web Content, *Saphire International '98 Web Site*. Available at http://www.ohsu.edu/cliniweb/saphint

11. Hersh, W. R., Brown, K. E., Donohoe, L. C., Campbell, E. M. and Horacek, A. E., CliniWeb: Managing Clinical Information on the World Wide Web, *Journal of American Medical Informatics Association*, Vol. 3, No. 4, pp. 273-280, 1996.

12. Huang, Q R, Creating informed consumers and achieving shared decision making, *Australian Family Physician*, Vol. 32, No. 5, pp. 335-341, 2003.

13. Humpreys, B., Lindberg, D., Schoolman, H. and Barnett, G., The Unified Medical Language: Informatics Research Collaboration. *JAMIA,* Vol. 5, No. 1, pp 1-11, 1998.

14. Institute of Medicine, Crossing the quality chasm: A new health system for the 21st Century: National Academy Press; 2001.

15. Lawrence, S., Giles, C. and Bollacker, K., Digital Libraries and Autonomous Citation Indexing. *IEEE Computer*, Vol. 32, pp 67-71, 1999.

16. Malet, G., Munoz, F., Appleyard, R. and Hersh, W., A Model for Enhancing Internet Medical Document Retrieval with "Medical Core Metadata", *Journal of the American Medical Informatics Association*, Vol. 6, No. 2, pp. 163-172, 1999.

17. Medical Object Oriented Software Enterprises Ltd, http://www.diseasesdatabase.com/

18. National Electronic Health Record Taskforce, A Health Information Network for Australia, Commonwealth of Australia, 2000.

http://www.health.gov.au/healthonline/ehr_rep.pdf accessed October 2000

19. Rigby, M., Forsstrom, J., Roberts, R., and Wyatt, J., Verifying Quality and Safety in Health Informatics, *British Medical Journal*, Vol. 323, pp. 552-556, 2001.

20. Rijsbergen, C.J.V., Information Retrieval, Second Edition ed, 1991.

21. Sim, I., Olasov, B. and Carini, S., An Ontology of Randomized Controlled Trials for Evidence-Based Practice: Content Specification and Evaluation using the Competency Decomposition Method, *Journal of Biomedical Informatics,* Vol. 37, No. 2, pp. 108-119, 2004.

22. Stolberg,, S.G., Trade Agency Finds Web Slippery with Snake Oil. *New York Times*, p. A16. 1999.

23. Taylor, H., Cyberchondriacs Update, The Harris Poll #21, available at http://www.harrisinteractive.com/harris_poll/index.asp?PID=299 accessed September 2005

24. Van_Mulligen, E. M., van-der-Eijk, C, Kors, J. A., Schijvenaars, B.J.A. and Mons, B., Research for Research Tools for Knowledge Discovery and Visualization, *Proceedings of the AMIA 2002 Annual Symposium*, pp 835-839, 2002.

25. Win, K. T., Cooper, J. and Alcock, C., Risk Assessment of Electronic Health Record System, *Proceedings of CollECTeR 2004*, Adelaide, Australia, 2004.

26. Yang, H. and Zhang, M., Intelligent Search for Distributed Information Sources Using Heterogeneous Neural Networks. In *Web Technologies and Applications, LNAI 2003,* Vol 2642, *Lecture Notes in Computer Science,* Springer Verlag Publishers, pp. 513-524, 2003.

27. Yang, H. and Zhang, M., An Ontology Based Approach for Resource Discovery, In *Proceedings of the International conference on Intelligent agents, Web technologies and Internet Commerce,* Gold Coast, Australia, pp 306-317, 2004.