Domain Specific Multi-stage Query Language for Medical Document Repositories

Aastha Madaan
Database Systems Laboratory
University of Aizu, Aizu Wakamatsu
Fukushima, Japan 965-8580
d8131102@u-aizu.ac.jp

Supervised by: Subhash Bhalla
Database Systems Laboratory
University of Aizu, Aizu Wakamatsu
Fukushima, Japan 965-8580
bhalla@u-aizu.ac.jp

ABSTRACT

Vast amount of medical information is increasingly available on the Web. As a result, seeking medical information through queries is gaining importance in the medical domain. The existing keyword-based search engines such as Google, Yahoo fail to suffice the needs of the health-care workers (who are well-versed with the domain knowledge required for querying) using these they often face results which are irrelevant and not useful for their tasks.

In this paper, we present the need and the challenges for a user-level, domain-specific query language for the specialized document repositories of the medical domain. This topic has not been sufficiently addressed by the existing approaches including SQL-like query languages or general-purpose keyword-based search engines and document-level indexing based search. We aim to bridge the gap between information needs of the skilled/semi-skilled domain users and the query capability provided by the query language. Overcoming such a challenge can facilitate effective use of large volume of information on the Web (and in the electronic health records (EHRs)repositories).

1. INTRODUCTION

The medical domain is complex. Therefore, limited access to target documents by indexing through a standard search engine is not sufficient. The medical information includes both patient-specific information (EHRs) and knowledge-based information (scientific papers and other literature) [7]. The medical knowledge (terminologies and concepts) has evolved over 10s of years. This is available on the Web through Web document repositories (such as, MedlinePlus [15]), popular medical literature related publications (PubMed [17], Medline [14]), other primary and secondary resources and EHRs ([5]). Querying these resources is required by the secondary applications such as evidence-based medicine and secondary use of EHRs to improve the quality of care [7]. Medical information is utilized by a variety of end-users with complex requirements. Practitioners, specialists and researchers are well-versed with the medical knowledge and terminologies. These users have precise queries and expect complete results within time limits (almost real-time). According to Kreshmoi survey [6], the end-user requirements vary on the basis of "level of specialty". Moreover, the issue of trustworthiness of information and authentication of a resource are of major concern for the users.

The widespread use of WWW has given rise to a range of simple query processors, the search engines. These query a database of semi-structured data (the HTML pages). For example, one can use a search engine to find pages containing a word "Villain". However, it is difficult to obtain only pages in which "villain" appears in the context of a character in a "wild west movie" [4]. In the healthcare domain, the end-users vary in their background, experience and have variable needs. These users interact in various contexts, a clinician interacts with the patients during clinical-care. He (or she) may need to query the medical literature and other document repositories to assess the plan for the treatments or patient-diagnosis. For such queries, the users need a query language and a schema. The goal of this study, is to assist the domain experts equipped with domain knowledge but not well-versed to use a query language such as SQL, XQuery to query the Web document repositories. This thesis illustrates the need of in-depth (and granular) querying of medical information on the Web. For example, a physician has an exploratory evidence-based query [1] "cases where helicobacter pylon bacteria causes peptic ulcer" or he or she may have a hypothesis-directed query [2], "Treatment in case of high-fever and dizziness". The conventional search (Yahoo, Google or localized search menu), return all the Web documents which match any of the keywords ranked by a set of criteria defined by the search engine. This may return a large number of documents. The physician has to go over each of the documents to find the one that (exactly) matches his expectations. This task is time-consuming. It may be abandoned by the physician. Both types of queries may involve multiple filtering-conditions (attributes) which make the query complex. To answer such end-user queries there is

Proceedings of the VLDB Endowment, Vol. 6, No. 12

[1]The evidence-based queries are raised during patient-care where the clinicians wish to query the knowledge archives to determine the relevance of signs and symptoms to the potential existence of one or more medical disorders [3].
[2]The hypothesis-directed queries represent the non-diagnostic intent of information about conditions seeking details on potential hypothesis including treatments, cures and outcomes [3].

a need to facilitate DB-style queries, where a user can query "helicobacter pylon bacteria" in context of "symptoms" and "peptic ulcer" in context of "causes" and can append these filtering conditions (attributes) in multiple interactive stages to receive results (in this case, as name of disease).

Several recent studies have made an attempt to address the information needs of the end-users within various domains. These include work on granular-level domain-specific search, estimating granularity of information in a document [19], estimating the difficulty of a document, the quality of documents, document summarization and the use of terminology resources for query refinement [7]. Cross-lingual search is of importance for end users at all levels [7]. However, several challenges remain for efficient health-care delivery.

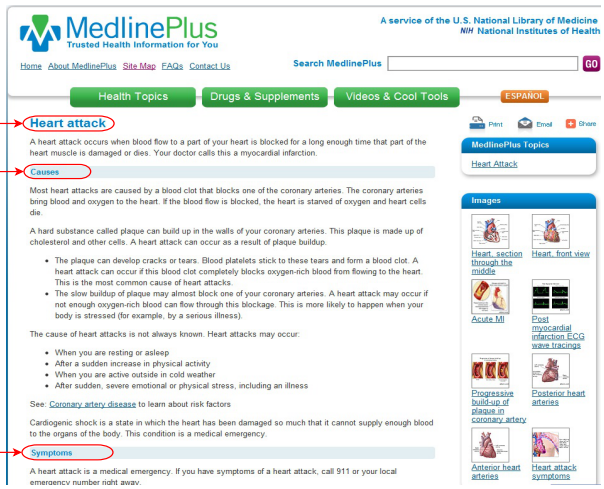

Figure 1: A snippet of the Web document "Heart Attack" from the MedlinePlus medical encyclopedia document repository.

In this study, we consider these challenges and outline the ways in which these can be addressed. Here, we focus on querying the medical information resources which form the expert-users personal and external knowledge base. A multi-stage query language is proposed which provides a user-level query calculator to formulate a query using domain concepts. Overcoming these will simplify the querying tasks for the expert and novice domain users. It will enable them to get the desired results.

2. STATE-OF-ART

In this section we describe the state-of-art of the complexity of medical information. We consider the various end-users and the recent approaches for querying data, considering user's preferences to facilitate the information retrieval and querying from various domain-specific resources. Further, the need for returning granular results to the users for their queries on Web documents is emphasized.

2.1 Query over Data Repositories

Recently form-based interfaces have been proposed for the naive users to allow them to query a database. These work well with the query logic of the end-users. For the complex queries, the number of forms might increase. Considering the approach proposed by Jayapandian [10], where the forms are clustered based on the query-needs of the users, the forms need to be modified to address the queries which cannot be performed otherwise.

2.1.1 *Granular Querying for Web Document Repositories*

The existing (keyword) searches consider the Web documents as bag of words and do not exploit the rich structural information available for the on-line medical information [13]. Varadarajan proposes the need for returning only the relevant segments of the Web documents as results of user search instead of returning complete Web documents for efficient querying and searching [22]. The proposed query language aims to present only the relevant segments of the Web documents as results to the users which satisfy the complete context of user-query. This will make the query results granular and the querying tasks time-efficient for the medical domain experts.

2.2 Complex Medical Information

The major barriers for efficient response to querying of medical information include perceived lack of quality, relevance of the content, inaccessibility, and trustworthiness of resources [6]. Inadequate quality assessment of on-line medical information may lead to misleading information. Therefore, validity and reliability of the Internet resources is often questioned for medical information [6]. Moreover, there is no criteria so as to when a search or query should be terminated. The clinicians often find that the results returned to them are not appropriate as per their requirement [6]. Querying medical information may vary in relation to depending on the role of the clinician and the way information is presented. While computer-based knowledge resources are useful in addressing these needs, ineffective search skills and lack of time are common barriers to information seeking [8]. The major challenges that arise due to the complexity of medical information are:

1. Identification of suitable query methods, and the results searched.
2. Identification of resources to be queried for the results (knowledge of schema).
3. Identification of results presented to the end-users.

Figure 1 represents a snippet of a structured medical encyclopedia document from the MedlinePlus document repository [15]. It contains the "disease name" as the *topic* and the associated concepts such as, "causes" and "symptoms" as *sub-topics*. There may be multiple subtopics within a given topic and each of the documents may contain distinct sub-topics depending upon the topic it represents. Each of the topic and sub-topics labels can be represented as queryable attributes to the end-users. These can significantly ease the task of querying by the health-care experts.

2.3 The End-Users

The end-users in the medical domain can be classified on the basis of their knowledge and expertise in the field. The first set of users can be termed as the novice-users (or the casual users). These users include the patients and their relatives. They are not well-versed with medical terminologies and concepts and may use keywords which may not be (much) related to the actual terms. They possess low ability

to pose reasonable queries. These users (may or) may not be good with the computer expertise. Hence, they utilize the results that are available from general-purpose search engines.

On the other hand, the doctors and clinicians, are the medical domain-knowledge experts and frequently need to access the knowledge archives (Web document repositories). These users must use reliable sources and require complete context of the results. They may (or may not) be good with the computer expertise but can pose exact and precise queries. Unlike searches on the Web, the domain-experts in the medical domain have in-depth knowledge of the particular information to be located, the details of the resource type and its reliability. They make use of the complex structural information to extract the relevant semantic information.

Therefore, a query language is developed to support the needs of the specialists (domain experts). It can subsequently reduce the learning curve of the novice users. There is a knowledge gap, which is often figured out by the intermediary agent about what is exactly required by the user [11]. It can be eliminated by allowing the user to construct queries interactively on a user-level schema (with query-able attributes).

2.4 Existing Solutions and Issues

The generic search engines may be time-efficient and free but often lack the specificity in providing relevant and high-quality information for the professional use in medical care [6]. Physicians who search using keyword search approach, say, a problem from its conditions may find the most relevant link on the second or third page. Whereas, they wish to find the relevant results for their queries quickly and efficiently. As a result of difficulties, the clinicians tend to believe that the answers to their (complex, specific and specialized) queries does not exist or exists in a fuzzy state [6].

Thus, the quality of results returned by the conventional search engines about medical information suffer in quality of information because of the following reasons:

1. Query with only 1 or 2 terms may not contain enough terms for the search engine to retrieve the desired information to the user.

2. In the document repository of the search engine, there might exist more than thousands of articles matching the query requested. This makes it impossible to locate the desired information by simply browsing through contents of returned results.

3. Conventional search engines focus on generic information search, domain-specific results are usually not taken into consideration during the search. Thus a simple keyword-based search does not produce relevant search results in specific domains such as the medical domain.

3. PROBLEM STATEMENT

The general practitioners need a query tool which gives 5-10 results per page with a simple layout. These users query the health Web document repositories. Although physicians are time-constrained, they are prepared to devote time to complex queries. The immediate need at point-of-care is critical for the physicians. Information on drugs, disease descriptions, treatment and clinical trial information are dominant needs [6]. We describe two key-features of the proposed approach to address this problem statement.

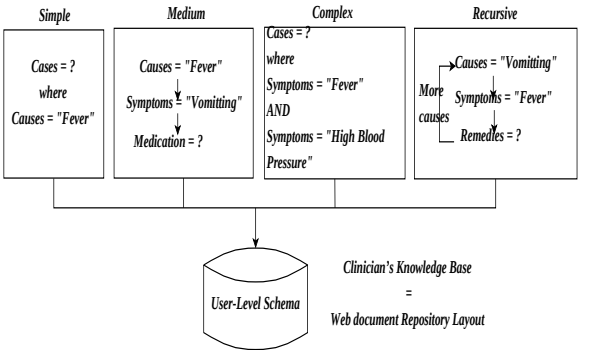

Figure 2: Different levels of queries performed by a clinician during health-care delivery.

The proposed query language attempts to overcome the following shortcomings.

1. Implement the query methodology to seek diagnostic or hypothesis-directed information (followed by a medical domain-expert) and,

2. Present the relevant areas (granular results) of the Web document that match the user's query criteria.

3.1 Multi-Stage Visual Query Language

Querying in the medical domain has a natural multi-stage and transitional progression from symptoms, to causes, remedy (and so-on) during patient diagnosis. The medical domain users wish to express these complex semantics of patient care and receive precise and complete answers. Database queries can easily fulfill this requirement as compared to the keyword-based web search. Hence, in this thesis we aim to model the multi-stage process of patient-care as a query language over a database of Web documents. At each stage user can choose an attribute (filter-conditions) that he or she wants to append to the query. The system provides flexibility to the users to form the order of accessing these features and add conditions on these attributes. Figure 2 describes various levels of query-complexity that occur as a result of querying by a domain-expert on his knowledge-base during patient-diagnosis. He may perform a simple query involving a single medical-concept (symptoms or causes), say, "Find diseases, where fever is a *cause*". For a query, "Find *medication* when a patient has vomiting due to fever", a user may query two-concepts,*symptoms* and *causes*. Such queries, can be termed as "medium-level" complex queries. Further, a clinician may have a more complex query. He or she may query the knowledge base after observing fever as a *symptom*, but on further observing the patient may discover "high-blood pressure" also as symptom. In such a case, he may wish to query the*symptoms* (concept) incrementally with two values. In more complex scenarios, a user may recursively query multiple concepts with different values. For example, in Figure 2 for the query "find remedies when a patient is having fever because of vomiting", after reading the remedies, a clinician may further wish to query the causes, with a value, "over-eating".

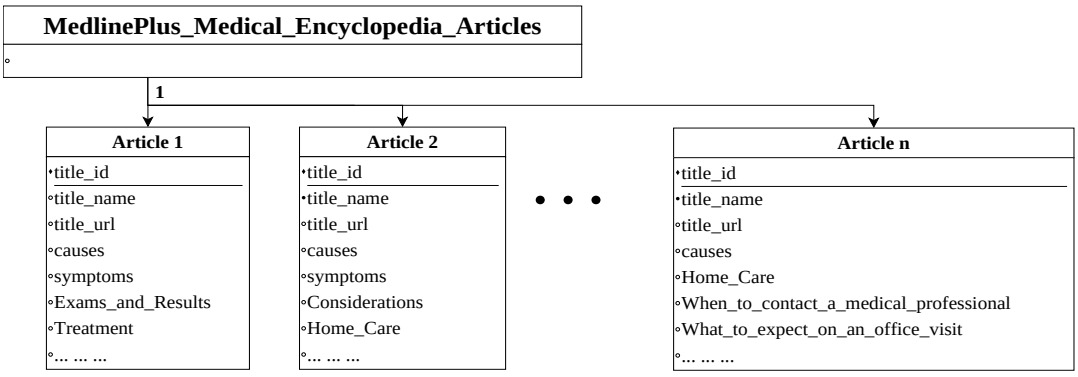

Figure 3: Structure of the user-level schema used for the proposed multi-stage query language.

3.2 User-Level Schema

The concepts and the terminologies in the medical domain have evolved from the domain knowledge and experience of the domain experts (clinicians and researchers) over several years. This information does not change frequently and is enriched over a period of time. With the emergence of the standards for medical terminologies, for example the LOINC [12] for the laboratory tests, ICD (9,10) [9] and SNOMED-CT [21] for disease codes, the medical information is increasingly becoming interoperable across geographically distributed health-care systems. The Web document repository creates a semantic object-level universal schema which can be queried by domain experts and other users. The attributes and data stored in this schema is largely understandable by the users and is easy-to-query. Such a schema can simplify and enrich the querying experience of the end-users. This facilitates a query language for these users to easily formulate their queries interactively in a multi-stage manner. We propose to capture the hierarchical schema (mapped to XML form) for the Web based document repositories. The multi-stage querying and the XML-based schema in terms of the domain-level knowledge of the users eliminates the need to write complex code for the queries (or complex SQL or XQuery expressions). This approach has been earlier proposed in [20], for granular and precise querying of the archetype-based EHRs. Therefore, the integration of the available domain knowledge and programming methodology is required to surmount the difficulty of using the query languages.

Figure 3 describes the structure of the user-level schema which is to be generated for the document repository (in this case, the MedlinePlus medical encyclopedia [15]) after transformations. It represents the concepts from each of the documents of the repository presented to the user as query-able attributes. Figure 4 shows a sample of the XML document corresponding to the "Aarskog Syndrome" document from the MedlinePlus medical encyclopedia [15]. It contains the conceptual-level tags which can be implemented as query-able attributes "symptoms" and "causes".

4. PROPOSED APPROACH: OUTLINE

For the purpose of creating the user-level schema (previous section), the proposed approach maps the domain knowledge of the experts and the concepts represented in a Web document. It extracts the syntactic structure of Web document considering the meta-data (in form of headings and sub-headings). And further, these concepts are used to label the nodes of the hierarchical structure to form a concept-based structure. These node-labels define the attributes for querying. We aim to develop a segmentation algorithm which segments the Web document considering the layout features and the semantic organization of the domain-level concepts. It makes use of the concept of Web page-segmentation, but considers the visual, layout and semantic features as compared to the earlier visual-based approach [2].

As part of this thesis, we aim to develop a domain-specific multi-stage query language described in the previous section. Figure 5(a), represents the query formulation process for clinical queries through the proposed multi-stage query language. At each stage of query formulation the user can dynamically select a medical-concept to query. Assign a value for it and then either execute the query or further refine the query by adding another attribute(s) and view results. The query is executed on the user-level schema. It provides the users with the segment-level results. Hence, the proposed query language can allow the user to formulate complex DB-style queries using a simple interface and understandable attributes. Figure 5(b), query formulation for the query, "Cases where a patient has fever due to affliction of pneumonia and tuberculosis" using the proposed multi-stage query language. The query fetches precise results by querying only specific contexts or segments (Causes or Symptoms).

4.1 Tree-structured Repository: Data Model

Let D= d_1, d_2,..., d_n, be a set of Web document from a Web document repository R, where n is the total number of documents in R. For the MedlinePlus medical encyclopedia repository, n is nearly 4000 [15]. For each document d_i, let H be the heading of the document d_i and S = s_1,s_2,. . ., s_n denote the set of sub-headings contained in the document, d_i. Let c_l represents the content enclosed between the two sub-headings or a heading and sub-heading of the document, d_i. The hierarchical structure corresponding to the Web document d_i, representing the coherent segments, can be defined as, let H be the root and s_j first-level of internal nodes. Each of the internal nodes s_j along with the following content c_l forms a segment f_k. The depth of the tree is determined by the number of levels of headings and sub-headings. Each of the node-labels of the hierarchical

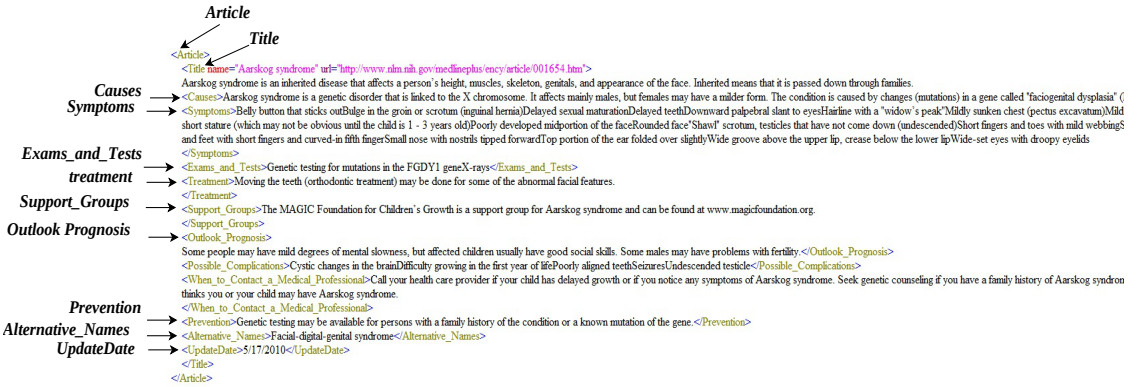

Figure 4: A sample document in user-level (XML) schema.

structure has 1:1 mapping to a medical concept. Each segment f_k, can be represented as $f_k = (s_j + c_j)$. It defines the granularity of the proposed query language.

Query Results. A query Q, can be represented as a combination of item(s) of concern k_i and the concept con_j (disease, causes or symptoms). The item(s) of concern is queried within the concept. The result returned is a fragment f_k, which satisfies the user query criteria. Each of the segment is demarcated by the headings or subheading label(s). The results to user queries may belong to the same segment within the same document (intra-segment query) or it may belong to different segments within the same document (inter-segment query). The result may also belong to different topics (Web documents). Such queries can be considered as inter-topical queries.

5. NEXT STEPS

This thesis will propose query methods (by using a user-level schema) for the medical domain users. The following steps will be needed to complete this thesis.

1. **XQBE for on-line medical document repositories.** At present, we attempted to use the XQBE graphical query language [1], on the user-level schema generated by the proposed approach. In this approach, the topic and sub-topic labels (query-able attributes) form the intermediate nodes of XQBE structure. The values of the attributes are added as leaf nodes. Since the user-interface is graphical, the users can easily drag and drop the nodes adding the filtering conditions without any prior knowledge of the schema structure. They can specify the attributes (segments) they wish to receive as results . The usability studies are planned to show that this kind of query language is easier to use and return relevant results to the users.

2. **Multi-stage Query-by- Concept Query language for on- line medical documents.** In this thesis, we further consider about relating the user-level schema with a multi-stage query-by-object query language. The term "object" can be defined as a unique identifiable or query-able entity in specialized domains. In this case, object is defined as a medical-concept, topic or subtopic-label in a medical document repository. The user can dynamically select an attribute to query on the user-interface (UI) and subsequently enter the value for that attribute. Further the user can append more attributes to it to formulate the desired query. The interface allows the user to query the knowledge archives and knowledge base without the need to understand the structure of the underlying schema or having any technical expertise [18], [23].

6. EVALUATION

This section presents the planned experimental evaluation for the effectiveness of the proposed query language.

6.1 Datasets and Queries

For the experimental evaluation of the proposed query language, the documents of MedlinePlus web document repository has been downloaded and pre- processed [15]. It contains 900+ documents describing the health-topics, 4000+ entries for the medical encyclopedia and 12000+ documents about drugs. About 1350+ organizations which provide this information and around 18000+ links are provided by it for authoritative medical information. All the original links to the repository are preserved during the preprocessing of the data. A set of 50 test queries (multi-staged) related to various medical concepts (such as diseases and medication) is formulated to test on the above dataset. Further, sample queries have been selected from various discussion forums and blogs on the topic [16]. Along with this a comparative analysis is planned to test the efficiency and effectiveness of the queries with the advanced keyword search provided by the MedlinePlus document repository [15].

6.2 Experimental Evaluation

In order to access and evaluate effectiveness and efficiency of the proposed multi-stage query language, we aim to perform a number of experiments on the real-world data (Section 6.1). The first category of experiments aim to evaluate the effectiveness of the proposed approach to create the conceptual user-level schema from the Web document repository. For this, we plan to measure the accuracy of segmentation of the Web document through quantitative measures of precision and recall. These measures can be calculated in terms of number of accurate, coherent segments discovered and the total number of segments. The second of category of experiments aim to test the effectiveness of the query language for the actual end-users. For this the quantitative measures can be used to calculate the search space used

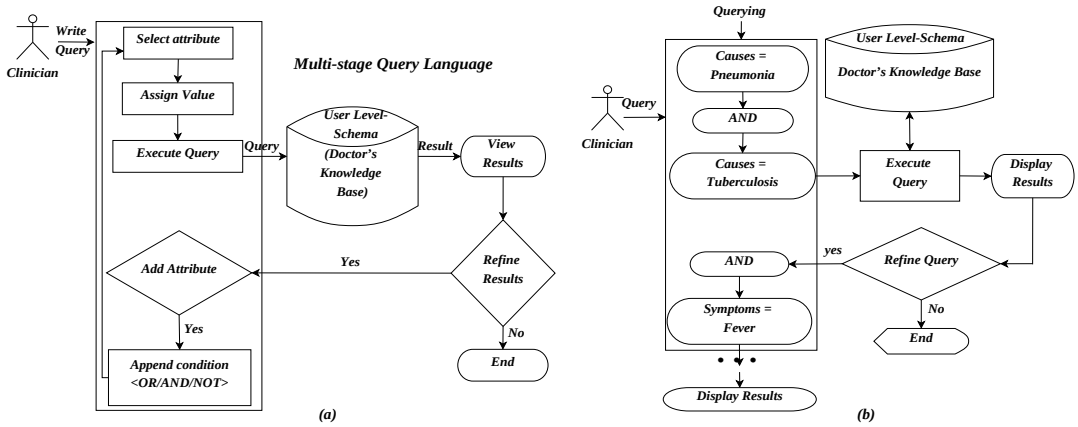

Figure 5: (a) Query formulation through the proposed multi-stage query language on the user-level schema (Figure 3), (b) A sample query formulated using the multi-stage query language.

by the proposed method as compared to the advanced keyword search of Web document repository [15]. We also aim to conduct usability studies with the actual end-users (the clinicians). These can establish effectiveness and efficiency of the query language along with the qualitative dimensions of ease-of-use and understandability for the users.

7. REFERENCES

[1] D. Braga, A. Campi, and S. Ceri. Xqbe (xquery by example): A visual interface to the standard xml query language. *ACM Trans. Database Syst.*, 30(2):398–443, June 2005.

[2] D. Cai, S. Yu, J.-R. Wen, and W.-Y. Ma. Extracting content structure for web pages based on visual representation. In *Proceedings of the 5th APWeb*, pages 406–417. Springer-Verlag, 2003.

[3] M.-A. Cartright, R. W. White, and E. Horvitz. Intentions and attention in exploratory health search. In *Proceedings of the 34th Intl. ACM SIGIR conference*, pages 65–74, New York, NY, USA, 2011. ACM.

[4] S. Cohen, Y. Kanza, Y. Kogan, W. Nutt, Y. Sagiv, and A. Serebrenik. Equix-a search and query language for xml. *Journal of the American Society for Information Science and Technology*, 53:2002, 2000.

[5] S. M. Freire, E. Sundvall, D. Karlsson, and P. Lambrix. Performance of XML Databases for Epidemiological Queries in Archetype-Based EHRs. In *Proceedings Scandinavian Conference on Health Informatics 2012*, volume 70 of *Linkping Electronic Conference Proceedings*, pages 51–57. Linkping University Electronic Press, 2012.

[6] M. Gschwandtner, M. Kritz, and C. Boyer. Requirements of the health professional research. In *Technical Report D8.1.2*. Khresmoi Project, 2011.

[7] A. Hanbury. Medical information retrieval, an instance of domain. In *SIGIR'12*. ACM, August 2012.

[8] S. Hunt, J. J. Cimino, and D. E. Koziol. A comparison of clinicians's access to online knowledge resources using two types of information retrieval applications in an academic hospital setting. *J Med Libr Assoc*, 101(1):26–31, 2013.

[9] `http://www.who.int/classifications/icd/en/`, 2011.

[10] M. Jayapandian and H. V. Jagadish. Automating the design and construction of query forms. *ICDE*, page 125, 2006.

[11] F. Li and H. V. Jagadish. Usability, databases, and hci. *IEEE Data Eng. Bull.*, 35(3):37–45, 2012.

[12] `http://loinc.org/`, 2011.

[13] A. Marian and W. Wang. Flexible querying of personal information. *IEEE Data Eng. Bull.*, 32(2):20–27, 2009.

[14] `http://www.nlm.nih.gov/bsd/pmresources.html`, 2011.

[15] `http://www.nlm.nih.gov/medlineplus/`, 2009.

[16] `http://www.linkedin.com/groups/Choice-OpenEHR-persistence-layer-144276.S.208531138?qid=208adbca-fc26-4ada-bf02-7efe5a9e5661&trk=group_most_recent_rich-0-b-ttl&goback=%2Egmr_144276`, 2013.

[17] `http://www.ncbi.nlm.nih.gov/pubmed`, 2011.

[18] S. A. Rahman, S. Bhalla, and T. Hashimoto. Query-by-object interface for information requirement elicitation in m-commerce. *Int. J. Hum. Comput. Interaction*, 20(2):135–160, 2006.

[19] X. Y. Raymond, Y. Lau, D. Song, X. Li, and J. Ma. Toward a semantic granularity model for domain-specific information retrieval. *ACM Trans. on Information Systems.*, 29(3), July 2011.

[20] S. Sachdeva and S. Bhalla. Implementing high-level query language interfaces for archetype-based electronic health records database. In *COMAD*, 2009.

[21] `http://www.ihtsdo.org/snomed-ct/`, 2011.

[22] R. Varadarajan, V. Hristidis, and T. Li. Beyond single-page web search results. *IEEE Transactions on Knowledge and Data Engineering*, 20(3):411–424, 2008.

[23] A. Yasir, M. Kumara Swamy, P. Krishna Reddy, and S. Bhalla. Enhanced query-by-object approach for information requirement elicitation in large databases. In *Big Data Analytics*, volume 7678 of *Lecture Notes in Computer Science*, pages 26–41. Springer, 2012.