MEDINFO 2004
M. Fieschi et al. (Eds)
Amsterdam: IOS Press

Vaidurya – A Concept-Based, Context-Sensitive Search Engine For Clinical Guidelines

Robert Moskovitch, Alon Hessing, Yval Shahar

Department of Information Systems Engineering, Ben Gurion University, Beer Sheva, Israel

Abstract

A major problem in the effective use of clinical guidelines is fast and accurate access at the point of care. Thus, we are developing a digital electronic guideline library (DeGeL) and a set of tools for incremental conversion of free-text guidelines into increasingly machine-comprehensible representations, which support automated application. Even if guidelines are represented in electronic fashion, care providers need to be able to quickly retrieve the guidelines that best fit the clinical situation at hand. We describe Vaidurya, a search and retrieval engine that exploits the hybrid nature of guideline representation in the DeGeL architecture. Vaidurya can use not only free-text keywords, but also multiple semantic indices along which the guidelines are classified, and the mark up of guidelines in DeGeL, using the semantic roles of one or more guideline-representation languages. Preliminary evaluation of Vaidurya in a standard information task and a large guideline repository is encouraging; formal evaluation is under way.

Keywords:

Information Retrieval, clinical practice guidelines, context-sensitive search.

Introduction

Clinical practice guidelines (CPGs) are a powerful method for standardizing the quality of medical care [1]. CPGs are a set of schematic plans for management of patients who have a particular clinical condition (e.g., insulin-dependent diabetes).Unfortunately, most clinical guidelines are represented in free text, whether in paper or in an electronic format. Paper-based guidelines are relatively inaccessible to care providers at the point of care, while the free-text electronic format provides little support for automated retrieval of the guidelines potentially most applicable to the patient at hand, especially when the user is not an expert in the relevant clinical domain, and no support for automated application.

Several representation formats, or *ontologies*, have been proposed in order to represent clinical guidelines in a structured or even machine-comprehensible fashion. Examples include ONCOCIN [2], EON [3], Asgaard [4], PROforma [5], and GLIF [6].

The DeGeL Architecture

The main hurdle preventing fast conversion of free-text guidelines into machine-comprehensible formats is that expert physicians cannot program using the syntax of the current guideline-specification languages, while programmers and knowledge engineers do not understand the clinical semantics of the guidelines. However, text-based representations also have benefits: They are useful for search and retrieval of relevant guidelines, although a formal representation is necessary for a machine- executable format. Thus, in our view, expert physicians should be transforming free-text guidelines into structured, semantically meaningful representations, while knowledge engineers should be converting marked-up segments to a formal, expressive, executable language

To gradually convert clinical guidelines into a machine executable target representation, we have developed a distributed architecture, the Digital electronic Guideline Library, (**DeGeL**) [Shahar et al., 2003a, 2003b], and a set of web-based software tools for incremental conversion of guidelines into multiple guideline-specification representations (Figure 1). (Currently, full specification of procedural aspects is provided only to the Asbru ontology, although specification and retrieval of declarative roles are supported for any ontology).

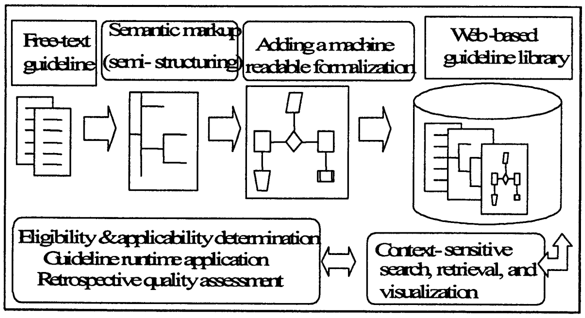

Figure 1 - Gradual conversion of free-text guidelines to an executable format in the DeGeL architecture

The DeGeL architecture uses a clinical-guideline semantic-markup tool, **Uruz**, which combines the expertise of both the expert physician and knowledge engineer to incrementally convert free-text guidelines into a machine-executable format (see Figure 1).

The output of Uruz is a *hybrid representation* of a guideline which, for each guideline, contains one or more of the following three formats.

- *Structured Text* –text fragments assigned to top-level Asbru (or another target ontology) semantic *knowledge-*

roles, such as *intentions*, *effects*, and the procedural *plan body*.

- *Semi-formal Asbru* – further specification of the structured text towards the formal representation, performed by the knowledge engineer and expert physician
- *Formal Asbru* – the final step of the conversion process, performed by the knowledge engineer, resulting with the guideline converted to a machine executable format

Marking Guidelines in the DeGeL Architecture

Each CPG in the DeGeL library eventually undergoes a *semantic mark up* process, in which the CPG's text is labeled by one of the available ontologies implemented in DeGeL (such as Asbru or GEM). The mark up is made by dragging a text segment from a source CPG to a relevant element in the chosen ontology. The text can then be further modified or extended. The markup is performed mainly by a domain expert. Domain experts are organized by groups, with a corresponding permission model for search and editing [11]. The markup process's main goal is to assist the knowledge engineer in transforming the guideline into a formal machine-comprehensible format, such as full Asbru. these segments of the text representing the relevant context are good for accurate retrieval operations.

The DeGeL Meta-Ontology and the Semantic Axes

The DeGeL library represents guidelines using a meta-ontology format [10]. Ontology independent elements, such as documentary details and semantic classification indices, are common to every guideline, regardless of the lower-level ontology, such as Asbru, used to present the guideline's details.

To classify guidelines, mainly for purposes of efficient retrieval, seven semantic axes are implemented in DeGeL. Each Axis represents a major clinical aspect. Axes include (1) symptoms and signs (e.g., hypertension), (2) diagnostic findings (e.g., blood cells count, electrocardiogram), (3) disorders (e.g., ischemic heart disease, malignant neoplasm), (4) treatment (e.g., antibiotic therapy , abdominal surgery), (5) body systems and regions, or a relevant CPG classification (6) guideline types (e.g., screening, prevention), and (7) guideline specialties (e.g., radiology, internal medicine).

Each Axis is implemented as a hierarchical tree of sub axes. Each CPG is indexed along one or more semantic axes, such as Disorders (e.g., malignant skin melanoma), Guideline specialties (e.g., oncology), etc.

The Vaidurya Search Engine

A major current focus is on the retrieval of CPGs as a valuable tool to improve the adoption and integration of CPGs at the point of care, as part of the evidence based medicine approach. Electronic CPG repositories, such as the National Guideline Clearinghouse (NGC) [8] provide access to electronic guidelines in a free-text or semi-structured format.

We introduce in this paper **Vaidurya**, a powerful search and retrieval tool that uses three types of search: (1) free-text search, using standard key terms; (2) *concept-based search*, which we call also *external search*, which uses a semantic –axes structure that indexes guidelines, as occur in the DeGeL library (and to some extent, in the NGC repository), and (3) *context-sensitive search*, which we call also *internal search*, which exploits the semantic markup performed on guidelines in the DeGeL library. Internal search focuses on searching for key terms only in the context of the text that exists within the scope of a particular semantic knowledge role. (A natural implementation is searching for text within specific predefined XML elements, although other representation formats are potentially possible).

The Vaidurya Search and Retrieval Model

Concept Based Search in Vaidurya

Multiple digital libraries are indexed in a hierarchical structure; examples include the known web portal Yahoo and the NGC library [8]. These sites allow browsing through the categories in the hierarchical structure, starting from the root of the categories tree, and ending at the most specific category, located at the leaves. .The NGC web site has two main Axes based on MeSH[13]: Disorders and Therapy.

Thus, part of the Vaidurya search model includes optional specification of one or more semantic axes or subaxes, and logical operators defining the relations between the axes. The *concept search query* is thus a collection of constraints represented by chosen sub axes and the logical relations between them: conjunction or disjunction..

Context-Sensitive Search in Vaidurya

The context sensitive search exploits the markup process that a CPG goes through in URUZ. Each marked up CPG in DeGeL is represented within the target ontology selected for it by the URUZ user, implemented as an XML structure that depends on the structure of the target ontology..

To implement the Concept-Based Search and the Context-Sensitive Search, we defined two properties for each element in a guideline-representation ontology, *Search Type* and *Search Scope*. These properties, or *aspects*, define the way an element will be indexed, queried and retrieved.

Search Type

The Search Type aspect characterizes the type of an ontology element for the purposes of a search engine. The search-type aspect caters for varying ontologies and for additional ones which DeGeL might need to handle in the future. In order to add a new ontology, one has to define the *Search Type* of each ontology element. Table 1 describes each Search Type and its definitions and properties.

Search Scope

Since CPGs are represented in DeGeL within different hierarchical ontologies, an element in an ontology tree may have descendents. Sometimes there are elements that function as headers and don't have any content; these elements may sometimes use their descendents' contents when being queried Using an element's descendents will be relevant only when its descendents are of the same *Search Type* as the element and when its descendents contents are semantically relevant and have a retrieval value to the element.

Table 1: Search Type definitions

Search Type	Description	Querying Options	Relevance measure
Free Text	An element containing a free text content.	Keywords with disjunction or conjunction logic operator.	Metric
Text Value	An element that may contain only a single fixed string value.	Requested string values with disjunction being the only possible relation.	Boolean
Text Multiple Value	An element containing one or more fixed string values.	Requested string values with conjunction, disjunction relations.	Boolean
Date	An element that its content represents a calendar date.	A date constraint using operators such as '>' or '>=' etc.	Metric
Integer	An element that its content represents an integer value.	An integer constraint using operators such as '>' or '>=' etc..	Metric
Semantic Index	An element represents the conceptualclassification of the guideline	Requested concepts using conconjunction, disjunction operators between indices.	Boolean
Unsearchable	An element that doesn't have content or its content is irrelevant for search.	No query.	Not relevant.

It is possible to query an element's descendents and score-up the result to the queried element. Using the Search Scope approach allows broadening the query when appropriate, and taking advantage of the descendents' contents, a crucial option within a library such as DeGeL, since CPGs are often in the process of being marked up process and the elements' contents aren't necessarily instantiated -- possibly because they weren't marked-up yet.

The Hybrid Query Model

Vaidurya offers a highly variable query structure, in which both very simple and highly sophisticated queries can be built. A query may contain constraints limiting the search in a sub group of CPGs that are indexed by requested semantic axes, such a query defines the logic relations between the sub axes and the axes conjunction or disjunction. The query may include also ontology elements that are of different search *types*; each element can be queried according to its type, as described at Table 1.

Table 2: Search Scope definitions

Search Scope	Description
None	No search at that element Nor at its descendents - elements that don't contain any content, and their descendents contents aren't relevant to them.
SearchMe	Search the element without descendents
OnlyChildren	No search at that element, search only its descendents.
ChildrenInclude	Search both that element and its descendents.

Querying options

There are two main CPG entities implemented currently in DeGeL: Free Text, or guideline source (GLS) – represents a source guideline, Marked Up guideline (GLM) - represents a marked up guideline that was created by the URUZ tool, based on one or more GLSs. A third entity we are currently adding is a guideline represented in a full structured representation, for example Asbru. Currently there are two main search options (1) search for a GLS based on the Source ontology, an ontology containing mainly documentation elements describing the source guideline, and (2) search for a GLM using the meta-ontology, which contains, among other elements, compulsory documentation elements that describe the GLM. One of the meta-ontology elements is the ontology by which the guideline was marked up. Each kind of search retrieves the relevant guideline entity.
A query is a set of constraints on the requested guideline, created by defining constraints on one or more ontology elements. The constraints on each ontology element differ according to its *SearchType*.

The Vaidurya Query Interface

The current Vaidurya query interface is shown in Figure 2. The interface enables users to form query using all the elements of the guideline. However, we are developing a simple-to-use query-interface builder tool that enables customization of the query interface to each user's needs. Our experience indicates that offering a user *all* the elements and semantic axes produces an overloaded, complicated query interface. Furthermore, there is a variety of users for such a search engine : general practitioners, patients, nurses, medical students, physicians at different specialty levels and knowledge, etc. Different users require different levels of complexity in the interface. Thus, we built a query-interface generation tool that enables us to offer different user types different levels of complexity of the query interface, from a simple basic text box to a complex interface. The results returned by Vaidurya are displayed by VisiGuide (figure 3), a sophisticated viewer, which enables users to browse the retrieved guidelines, including their contents and the classification of the returned results. VisiGuide can display any guideline ontology given to it as an XML file.

The Vaidurya Data Structures and Ranking Algorithm

A CPG is represented in our repository by its ID number and the ontology ID (e.g., GEM, Asbru). Thus, each ontology element, of any Search Type, is indexed as a pair, which differs from traditional textual databases represented by an Inverted File of one dimension that represents the document free text [7], which in our system is a private case of that would be implemented as an ontology that has only one element of *Free Text* search type. A pair of document *i* and a context *c* (*i*,*c*) represents a multiple dimensional Inverted File. Each pair represents an CPG's ontology element content, which indexed according to its Search Type.

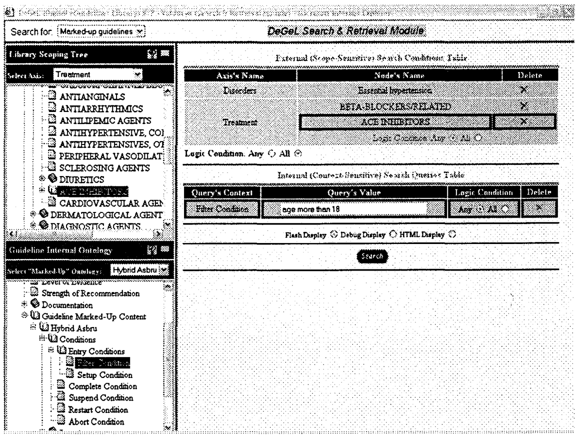

Figure 2 - A full query interface

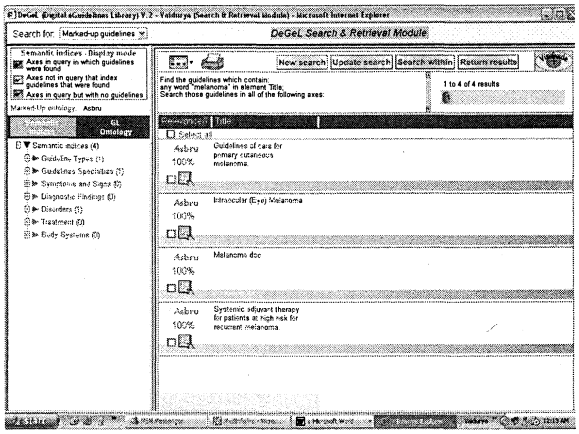

Figure 3 - Display and retrieved guidelines by VisiGuide

When a query is processed and operated against the knowledge in the repository represented by the pairs (*d,c*) and its value, data structures that represent vectors and matrices are built at query time, these allowing the process of evaluating the CPGs relevancy to the query and ranking them.

Each *Search Type* element has a different importance in the ranking process. For example a query constraint created on a date type element will contribute more than a Free Text constraint, that is because it is easier to define a query date constraint properly than a Free Text constraint. Also among the same search Type elements there are elements with clearer content and less clear for retrieval, we implemented these differences by giving each *Search Type* a weight that represents its value in the retrieval task, and the same for each ontology element.

Implementation of such algorithm is based on basic linear operations, let CM be a matrix which one dimension is the CPGs and the second is a group of queried elements, which are contexts, of the same *Search Type I;* for example all the elements that chosen in a query from the *Free Text* search type. The value of each entry within CM represents the degree of matching of a CPG's content in context *c*. The number of the CM matrices is the number of the queried search types. Let qST be a vector that represents the elements of the search type *I* and their weights, where each entry represents the weight of element *c*. Let $stRank_i$ be a vector which is a result of the matrix-vector multiplication of CM_i and qST_i as shown in Equation 1.

$$stRank_i = CM_i \times \overrightarrow{qST_i}$$

After a computation of *stRanks* was made we know the score each CPG had for each Search Type*i*. Let *STM* be a matrix of the unification of all the vectors $stRank_i$, each entry $STM_{d,i}$ represents the score of a CPG *d* at the search type *I;* let q be the vector that represents the weights of each *search type.* Let *R* be the vector representing the rank each CPG scored at the ranking process; each entry at the vector R is the final rank of a CPG. *R* is the result of the matrix-vector multiplication of the matrix *STM* and the vector *q* as shown in Equation 2.

$$R = STM \times \vec{q}$$

Results

Initial results using the TREC 6 collection[12], a benchmark textual database of queries and judgments and the NGC repository are encouraging. Formal evaluation is underway, We are planning to evaluate the concept based and context sensitive search methods on a guidelines repository.

Conclusions

The mark up process a CPG goes through in DeGeL gives us the outstanding opportunity to query by the context sensitive methods, we had learned already that this querying approach allows more specific search that may limiting the search and avoid irrelevant results which most of the search engines suffer from. The disadvantages are a complicated query that needs an advanced familiarity with the query interface and time that users wouldn't like to spend sometimes. The concept based search allows limiting the search to a specific subset of the repository which ascends the precision of the returned results, the potential disadvantages in this approach is the manual classification of a CPG in DeGeL, this a subjective issue that may harm the precision of the results.

Future Work

We are on the procedure of building a gold standard of queries that will be the base of the formal evaluation. Preliminary experiments had shown that the query interface is very overloaded, we are working on new interfaces that will be simpler and more user friendly. This GUIs will offer derived version customized by the user. We are developing also a semi natural language search interface, since the classification of each CPG is doen by expert it may be subjective, we are developing an automatic classifier that will classify a CPG along few axes based on a given knowledge-base of classified CPGs.

Acknowledgments

This research is supported partly by the NIH/NLM award LM-06806. We thank our colleagues at Standford University Drs. Mary Goldstein, Susana Martins, and Soroka Hospital ,Prof Eitan Lunenfeld, were extremely helpful in assessing and developing Vaidurya..
we would like to thank also to all the students who made their senior projects at the Vaidurya research at the Department of Information Systems and all the physicians and users that gave their important comments.

References

[1] Grimshaw, J.M. and Russel, I.T. (1993). Effect of clinical guidelines on medical practice: A systematic review of rigorous evaluations. *Lancet*, 342: 1317–1322.

[2] Tu, S.W., Kahn, M.G., Musen, M.A., Ferguson, J.C., Shortliffe, E.H., and Fagan, L.M. (1989). Episodic Skeletal-plan refinement on temporal data. *Communications of ACM* 32: 1439–1455.

[3] Musen, M.A., Tu, S.W., Das, A.K., and Shahar, Y. (1996). EON: A component-based approach to automation of protocol-directed therapy. *Journal of the American Medical Infor-mation Association 3*(6): 367–388.

[4] Shahar, Y., Miksch, S., and Johnson, P. (1998). The Asgaard project: A task-specific framework for the application and critiquing of time-oriented clinical guidelines. *Artificial Intelligence in Medicine,* 14: 29-51.

[5] Fox, J., Johns, N., and Rahmanzadeh, A. (1998). Disseminating medical Knowledge: the PROforma approach. *Artificial Intelligence in Medicine,* 14: 157-181.

[6] Peleg M, Boxwala A. A., Omolola O., Zeng Q., Tu, S.W, Lacson R., Bernstam, E., Ash, N., Mork, P., Ohno-Machado, L., Shortliffe, E.H., and Greenes, R.A. (2000). GLIF3: The Evolution of a Guideline Representation Format in *Overhage M.J., ed., Proceedings of the 2000 AMIA Annual Symposium (Los Angeles, CA, 2000),* Hanley & Belfus, Philadelphia.

[7] G. Salton, A. Wong, and C.S. Yang. A vector space model for automatic indexing. *Communications of the ACM, 18:613 620, 1975*

[8] National Guideline Clearinghouse, www.ngc.org

[9] Purcell, G. P., Rennels, G. D., and Shortliffe, E. H. (1997). Development and Evaluation of a Context-Based Document Representation for Searching the Medical Literature. *International Journal of Digital Libraries* 1:288-296.

[10]Shahar Y, Young O., Shalom E., Mayaffit A., Moskovitch R., Hessing A., and Galperin M. (2003a). DEGEL: A hybrid, multiple-ontology framework for specification and retrieval of clinical guidelines. *Proceedings of the 9th Conference on Artificial Intelligence in Medicine—Europe (AIME) '03,* Protaras, Cyprus.

[11]Shahar Y, Shalom E., Mayaffit A., Young O., Galperin M., Martins S.B., and Goldstein, M.K. (2003b). A distributed, collaborative, structuring model for a clinical-guideline digital-library. *Proceedings of the 2003 AMIA Annual Fall Symposium*, Washington, DC.

[12]Text REtrieval Conference (**TREC**), http://trec.nist.gov

Address for correspondence

robertmo@bgumail.bgu.ac.il
Department of Information Engineering,
Ben Gurion University of the Negev, Israel.
P.O.B. 653, Beer Sheva 84105, Israel.