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

Mapping Various Information Sources to A Semantic Network

Wen Ruan[a], Thomas Buerkle[b], Joachim W. Dudeck[c]

[a]*TextWise Labs, Syracuse, NY, USA*
[b]*Institute of Medical Informatics and Biomathematics, University of Muenster, Germany*
[c]*Institute of Medical Informatics, University of Giessen, Germany*

Abstract

Giessen Data Dictionary Server (GDDS) provides context sensitive information services to disparate clinical applications by automatically navigating a semantic network that stores medical knowledge. Mapping multiple information sources to single clinical application becomes a challenge due to different organization of semi-structured information sources. Linking huge unstructured information sources such as medical literature is even more challenging because we need to develop a mechanism to organize unstructured information and take into account the scalability issue. In this paper, two drug information sources have been mapped by developing an independent subnet for each source and interlinking them at proper nodes. For medical literature, we have demonstrated that the semantic network of the Unified Medical Language Systems and human assigned topics to each document can be used to organize the large amount of medical literature into the framework of the GDDS service.

Keywords

Semantic network, Medical Data Dictionary, Medical Data Dictionary Server, knowledge representation, visualization.

Introduction

Semantic network or acyclic graph has been used to represent medical knowledge in building many authoritative medical terminologies such as the Unified Medical Language Systems (UMLS) [1] and the Medical Entity Dictionary [2], etc. The Medical Data Dictionary (MDD) has been developed at Giessen University Hospital to assist decision support applications in the clinical information systems [3]. It has evolved from a rigid tree structure to a semantic network that is more suitable to model real world knowledge. Giessen Data Dictionary Server provides navigation services of medical knowledge stored in the semantic network of the Medical Data Dictionary. The GDDS services have been linked to clinical applications to provide context sensitive help [4] to various clinical applications to meet the information needs of health care providers.

In the semantic network of the MDD, concepts are grouped into classes that are represented as class nodes in the semantic network. For example, *Drug* represents an MDD concept class. It stands for a set of terms representing existing drugs such as Lasix® and Folsan®, which are considered concepts. Each concept maintains an Is_A relation with its concept class. *Drug Substance* refers to terms that are names of drug substances such as furosemide, and *Guideline* the titles of drug therapy guidelines.

Appropriate relations are established between concept classes. For example, *Drug*, of course, contains one or more *Drug Substance*s. For some *Drugs* or *Drug Substances*, therapeutic *Guidelines* exist. Drugs, drug substances and guidelines may have on-line information respectively.

On the concept level, individual concepts may but need not maintain all relationships existing at the concept class level. Some drugs may not have therapeutic guidelines, for example. No concept, however, shall maintain a relationship with another concept if that relationship does not exist between the two corresponding concept classes.

The Giessen Data Dictionary Server has been designed and implemented as an intelligent navigator which needs not know the structure of the semantic network. Instead the GDDS can explore the local paths on its own from any node it starts. It drops dead links, avoids cycles, and returns the final related information with paths it has gone through [5].

If a user submits a search term to the dictionary server, the server locates its concept class first. It then retrieves all possible paths ways to online information at the class level. Finally, it resolves those paths at the concept level for the specific search term and returns the retrieved on-line information with paths to the user. This provides the user a clear view about what information has been found, and how it relates to the original search term.

Up to now, we have successfully linked clinical guidelines to a nursing application and a cancer documentation system respectively and two drug information sources to a drug therapy application within the intranet of the Giessen University Hospital [4, 6]. These information sources are semi-structured and within a comparably limited domain. The main challenge of linking two structurally different information sources is to find out an approach to merge related information from both sources. A real challenge comes if we try to link huge unstructured information sources such as medical literature using the GDDS service.

In this paper, we discuss the experience of linking two structurally different information sources by creating two independent subnets and interlinking them at proper notes. We also present a

solution for integrating medical literature to the framework of the GDDS by adapting the UMLS semantic network and mapping the human assigned topics of medical literature to the semantic network of the GDDS.

Methods

Define a Semantic Network for two Drug Information Sources

The Giessen University Hospital Formulary, the so-called House List (HL), is maintained by the in-house pharmacy department. The German Red List (RL) is a formulary issued by the German Pharmaceutical Industry Association that comes in print with a red cover [7]. Both information sources differ in contents and organization of contents. The Red List contains information on a much larger set of drugs and drug substances, but it does not include all drugs available from the in-house pharmacy department. Drugs in the Red List are listed by brand names, while the House List is primarily arranged around drug substances.

The old way to search the House List is to start with a drug substance group, from where find the drug substance of interest, then the user can get to the page about all the drugs that contain this substance. The drug substance group itself is a hierarchy, i.e., a drug substance group may contain multiple subgroups and so on. Guidelines are available to drug groups and may be linked to subgroups at difficult levels according to the scope of individual guideline.

On the contrary, the German Red List references its information based on brand names. These brand name drugs come in various application forms. Drug substances are linked to specific drug application forms. Signatures are available for drug substances. Therefore, browsing the Red List always starts from brand names. Despite these differences, both information sources are heavily used by physicians at the university hospital.

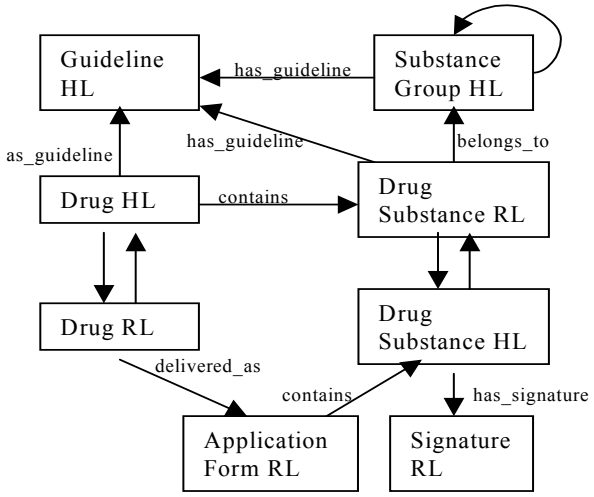

Figure 1 - The semantic network of the two drug information sources

It is not easy to convert the concepts and relations from one information source to another based on the observation. There are overlaps between the drugs and drug substances from both sources though. Considering the maintenance issues that will be encountered later, we have decided to build two independent semantic network for the two information sources. This fits in the scenario of the GDDS because the GDDS is designed to provide various information sources to disparate clinical applications. However these two information sources need to be bridged somehow to provide one-stop information service.

Figure 1 shows the two separate semantic networks for the House List and the Red List respectively. The upper part is the semantic structure of the Giessen House List, the lower part the Red List. We have added labels "HL" and "RL" for each class to identify the source.

In the HL network, *Drug* <contains> *Drug Substance*, which <belongs_to> *Drug Substance Group*. Each of these classes has links to *Guideline*. In the RL network, *Drug* is <delivered_in> *Application Form*, which <contains> *Drug Substance*. *Drug Substance* has links to *Signature*. These relatively independent semantic subnets are bridged with an Is_Synonym relation between drugs respectively drug substances from both formularies.

Mapping Medical Literature to the Semantic Network

Medical literature is an indispensable resource in clinical health care, biomedical education and research. Many clinical information systems provide the functionality of searching the MEDLINE database. However, the visualization of the search results is still as simple as those from normal search engines. It does not provide enough information about how the retrieved documents are related to the query. This is exactly what GDDS service can provide.

However, mapping the medical literature to a semantic network that can be used by the GDDS has several challenges:

1. One document may discuss several topics instead of one. Documents in MEDLINE are normally assigned several main topics using Medical Subject Heading (MeSH) [8]. Therefore, it is impossible to link a document under one concept class.
2. The vast volume of existing medical literature makes it impractical to map each document to the semantic network of the GDDS. In addition, GDDS does not rank documents, which is normally done by search engines.
3. MEDLINE indexes references and abstracts from about 4500 biomedical journals covering almost all the biomedical sub-domains. It is a demanding task to design a semantic network for the concepts covering the whole biomedical domain.

Considering the above-mentioned issues, we decided to use a different strategy for delivering GDDS service for medical literature.

First, the GDDS will work on ranked documents retrieved by search engines and provide semantic links for each document showing the topics of each document and how they are related to the query. This solution will leave the burden of indexing, retrieving and ranking documents to commercial search engines.

Second, we will borrow the existing Semantic Network of the Unified Medical Language System as the semantic network for the medical literature in the GDDS. The semantic types in the

UMLS correspond to concept classes in the GDDS. The relationships between each pair of semantic types can be borrowed as well. However, we need to make several adjustments before we can use the UMLS Semantic Network within the framework of the GDDS.

The relations between each pair of semantic types need to be abstracted since there might be multiple relations defined for a pair of semantic types. For example, *Therapeutic or Preventive Procedure* <affects>, or <treats>, or <complicates> *Disease or Syndrome*. Relationship extraction is not the focus of this paper and we are not going to disambiguate relations. Therefore, we will use Associated_with to link any semantic type pairs that have one or more relations defined within the Semantic Network of the UMLS.

Two concepts that belong to a pair of related semantic types respectively may not relate to each other. However, in the context of medical literature, if the two concepts are two topics of a document, for example, two main topics assigned by an indexer, we will infer that these two concepts are related somehow based on the co-occurrence theory in relationship extraction.

The GDDS require that each concept belong to one concept class, which is not true for the UMLS semantic network. The solution to this is that GDDS simply repeats the path searching process with different starting points and finds all possible pathways, then removes the duplicates.

We have used MEDLINE data and the PubMed search engine to test out our solution since MEDLINE documents have human assigned MeSH terms describing the main topics of each document. We only use the terms assigned as main topics. Figure 3 shows parts of several sub-trees of the UMLS semantic network, which are the *Fully Formed Anatomical Structure*, *Organism*, *Pathologic Function*, *Occupational Activity* and *Chemical*. Each box is a semantic type in the UMLS, which is also a note in the GDDS semantic network. The solid lines between semantic types represent the Is_A relation. The dotted lines represent the Associated_with relation. For example, *Disease or Syndrome* <Is_A> *Pathologic Function*. *Disease or Syndrome* is also <Associated_with> *Virus*, *Invertebrate*, *Lab Procedure*, *Immunologic Factor*, etc., which cannot all be shown in the figure.

If a user submits a query about "yellow fever", Figure 4 shows a sample page from PubMed and added feature from the GDDS. It shows results starting from 41 because the top ranked articles are mostly recent and no MeSH terms have been assigned. The last few lines after the PubMed ID are added by the GDDS. For example, the MEDLINE index terms for result 41 include *Dengue/prevention & control, Aedes*, *Insect Vectors* and *Mosquito Control/methods*, which are main topics of this article. The terms after the backslash are MeSH subheadings that supply more restrictive information of the main heading. The GDDS first identifies the concept classes (or the semantic types in the UMLS) for each MeSH term, i.e., *Yellow fever* <Is_A> *Disease or Syndrome, Dengue* <Is_A> *Disease or Syndrome, Aedes* <Is_a> *Invertebrate, Insect Vectors* <Is_A> *Invertebrate, Mosquito Control* <Is_A> *Occupational Activity*. Based on the UMLS semantic network and abstracted relationship defined in GDDS (see figure 3), *Disease or Syndrome* is <Associated_with> *Disease or Syndrome, which is also* <Associated_with> *Invertebrate* and *Occupational Activity.*

Therefore, we get the following reasoning about this document's main theme that *Yellow fever* is somehow related to *Dengue/prevention & control*, which is related to *Aedes*, *Insect Vectors* and *Mosquito Control/methods*. These simple lines give user a clear picture of what this article is about and how it is related to the query.

Results

For the two drug information sources, we mapped about 1,000 drugs and 700 drug substances from the House List to the HL sub-network. Then we mapped approximately 9,500 drugs and 2,600 drug substances from the Red List to the RL sub-network. 700 drugs and 500 drug substances from the two sub-networks are linked to each other.

For the medical literature, we need to map the 134 UMLS semantic types as concept classes to the GDDS semantic network and mapping MeSH main headings to the concept level with Is_A relation linking to proper semantic types. We don't need to insert relations between concepts because we will infer the relations from the class level relation and the co-occurrence information. Of course, based on this solution, the GDDS navigation part needs a bit adjustments.

Discussion

If building a sophisticated semantic network is a demanding task, maintaining such a semantic network is even more challenging [9]. Besides the structural issues of the two drug information sources, the reason of building separate sub-networks for each source and interlinking them is based on the concern of maintenance issues. When one source is updated, new concepts and relations can be added to the corresponding network, obsolete concepts and relations can be removed logically without breaking the links to other sources. The trade-off is the storage space which is not a concern in today's technology.

For unstructured medical literature, we have borrowed human assigned category information. Otherwise, automatic text categorization technology is needed to organize huge amount of information in order to build a meaningful semantic network. The current solution accepts keyword query and uses straight string matching to get MeSH main headings. The next step could be using natural language processing technology to interpret the semantic links within the natural language query and match to that of retrieved documents so that we can re-rank the documents based on the semantic matching.

The aim of the GDDS service is to provide efficient information access for health care providers. Therefore, user's satisfaction is the goal. Currently we choose to show potential relations between the main topics ignoring the others and isolated islands of topics. User study needs to be done in the future to assess what kinds of topics and relations should be provided or omitted in order to provide efficient information access.

Conclusions

The GDDS framework provided an environment to link various information sources to disparate clinical applications. This paper has shown that multiple information sources can be linked to one clinical application seamlessly provided the information sources are properly mapped to the semantic network inside the MDD. It also demonstrates that huge unstructured information sources can also be integrated with the help of text categorization technology.

References

[1] Lindberg, D.A.B., Humphreys, B.L., McGray, A.T. The Unified Medical Language System. Math. Infor. Med. 32 (1993) 281-291.

[2] Cimino, J.J., Hripcsak, G., Johnson, S.B., Clayton, P.D. Designing an Introspective, Multipurpose, Controlled Medical Vocabulary. In: Kingsland L.C. (Ed) Proceedings of 13th SCAMC, IEEE Computer Society Press, Washington DC (1989) 513-518.

[3] Prokosch, H.U., Bürkle, T., Storch, J., Strunz, A., Müller, M., Dudeck, J., Dirks, B., Keller, F. MDD-GIPHARM: Design and Realization of a Medical Data Dictionary for Decision Support Systems in Drug Therapy. Informatik, Biometrie und Epidemiologie in Medizin und Biologie, 26 3(1995)250-261.

[4] Ruan, W., Bürkle, T., Dudeck, J., A Dictionary Server for Supplying Context Sensitive Medical Knowledge. In Overhage, J.M. (Ed) Proceedings of the AMIA Annual Fall Symposium, Los Angles, November, Hanley & Belfus, Inc. (2000) 719-723.

[5] Ruan, W., Bürkle, T., Dudeck, J. An Object-Oriented Design for Automated Navigation of Semantic Networks inside A Medical Data Dictionary, Journal of Artificial Intelligence in Medicine, 18,1, (2000) 83-103.

[6] Ruan, W., Bürkle, T., Dudeck, J. Context Sensitive Information at the Clinical Workstation – the Role of the Medical Data Dictionary. In Greiser, E. et al (Ed) Methoden der medizinische Informatik, Biometrie und Epidemiologie in der modernen Informationsgesellshaft, MMV Medien & Medizin Verlag, München, (1998) 214-217.

[7] Rote Liste, Rote Liste Service GmbH. http://www.rote-liste.de/

[8] http://www.nlm.nih.gov/mesh/meshhome.html

[9] Cimino, J.J., Clayton, P.D., Hripcsak, G., Johnson, S.B. Knowledge-based Approaches to the Maintenance of a Large Controlled Medical Terminology. JAMIA 1 (1994) 35-50.

Address for correspondence

Wen Ruan,

TextWise Labs, 401 South Salina Street, Syracuse, NY 13202, USA.
Email: wen@textwise.com