Research for Research: Tools for Knowledge Discovery and Visualization

Erik M. van Mulligen, PhD, Christiaan van der Eijk, MSc, Jan A. Kors, PhD, Bob J.A. Schijvenaars, PhD, Barend Mons, PhD

Dept. of Medical Informatics, Erasmus University Rotterdam, The Netherlands

ABSTRACT

This paper describes a method to construct from a set of documents a spatial representation that can be used for information retrieval and knowledge discovery. The proposed method has been implemented in a prototype system and allows the researcher to browse interactively and in real-time a network of relationships obtained from a set of full text articles. These relationships are combined with the potential relationships between concepts as defined in the UMLS semantic network. The browser allows the user to select a seed term and find all related concepts, to find a path between concepts (hypothesis testing), and to retrieve the references to documents or database entries that support the relationship between concepts.

INTRODUCTION

Digital dissemination of scientific information in publicly available databases has recently increased the possibility to acquire new insights for a large scientific public. In the medical field, for instance, the MedLine database currently contains the abstracts of over 11 million citations and is growing with an unprecedented speed of 500,000 abstracted articles per year. In addition, molecular databases containing the latest sequence data and other related information grow at an even higher rate. The consequence of this plethora of information is that a researcher is spending more and more time to read articles in order to keep up with his specialty. The researcher may find it easier to generate a stack of potentially relevant articles; the stack is however as much a threat as it is an asset. The accumulated information is hard to digest, let alone to synthesize into a comprehensive picture of its implicit knowledge.

Many research projects have focused on providing support tools for researchers to enable them to analyze the literature more effectively and to support them with hypothesis generation based on the literature. A first line of research focuses on improving information retrieval (IR). One possibility is to create special search engines (in addition to general search engines) that use a domain-specific thesaurus [1]. The thesaurus supports the expansion of a query to also include synonyms and proposes additional terms that could be added as query refinements. Another possibility is to use intelligent agents that search the information sources on behalf of the user; some of these agents can be trained by the user by indicating relevant and irrelevant documents. An alternative approach is predigested information leading to specialized portals that only provide the filtered information[2,3]. The assumption behind this type of research is that researchers spent much time reading articles to discover that they are not relevant. However, filtering and categorizing of information, although reducing the information overflow, also restricts the potential for associative discoveries based on serendipity. Steve Lawrence has exploited the citations in a scientific article to improve the relevance of the set of documents provided to the user. His assumption is that these citations define a context of the document that can help to assess its relevance to a query[4,5].

Another way to support the researcher is to try to build an abstraction of the knowledge represented in multiple articles and to visualize this knowledge in a way that is more comprehensible to the user. The systems developed with this purpose try to detect patterns of words or terms and, based on co occurrence, exploit the relation between these terms [6].

Quite a few current research projects are focusing on the area of (medical) knowledge discovery[7-11]. The assumption here is that relations that have not yet been explicitly described in the literature, can automatically be identified. The researcher can review the potentially interesting relationships and decide on further analysis. The ArrowSmith project is a good example of such a statistical knowledge discovery tool[7]. The system starts a with a seed term (e.g., Raynaud's disease) that is used to query MedLine. From the resulting set of articles, it finds all words that are co-occurring with that term (for performance reasons, only words in the title are used). Next, the set of words is reduced according to a number of heuristics and the remaining set is used to retrieve a second set of articles from MedLine. This set is then analyzed to find frequent words. This set is subsequently pruned again according to some heuristics and the final set of words is presented to the user as a possible set of interesting related terms. Several improvements to the ArrowSmith approach

have been proposed. In his DAD system, Weeber extended the ArrowSmith system to use UMLS concepts for the whole abstract text[8]. Lindsay developed new measures that can be used to find the terms linking two sets of literature[12]. Stapley and Benoit implemented a system that finds relationships between pairs of genes on the basis of their co-occurrence in a subset of MedLine articles[9].

Other approaches use templates to find conceptual expression patterns in a corpus of text[13,14]. A set of linguistic rules is defined that are used to distinguish between relevant and irrelevant patterns of concepts.

A last related line of research to be mentioned here is on the identification of new terms in large text corpora. It is clear that in evolving domains (such as genetics) new terms are invented almost daily. In order to be able to use these new terms (and their synonyms) in queries, sophisticated tools have been developed that find new (genetic) terms using rules or frames[15,16].

In this paper, we describe a prototype technology that combines elements of several of the approaches described before and which adds a new dimension to the visual representation of knowledge represented in large sets of literature. The system supports researchers in various ways: (1) to provide them with better navigation tools for browsing large online information sources, (2) to support them in generating new hypotheses , and (3) to facilitate the validation of existing and new information by experts.

METHODS AND MATERIAL

The purpose of our project was to develop a tool that supports researchers in both information retrieval and knowledge discovery. We analyzed ongoing research efforts and concluded that our system should be able to deal effectively with synonyms and homonyms and should allow the use of information from different sources. Our initial interest was directed towards sources containing genetic information (both documents and genetic databases). In the prototype described here, we only used information obtained from documents.

An additional requirement was that the system should not be limited to discovering common terms in two disjunctive sets of literature, but should also be able to generate any connective path between two terms. The biggest challenge here was to contain the combinatorial explosion in the number of paths to consider. Also, to allow interactive inspection of the discovered knowledge, we had to find a different approach than in most systems where term or concept indexes are created when needed. Last but not least, the system should provide a graphical presentation of the discovered relationships.

Architecture

For the prototype development, we selected a set of full text articles in genetics provided by Nature Publishing Group. These articles were fed into the Collexis© conceptual fingerprint system[17] that generated the concepts contained in the text on the basis of the Unified Medical Language System[18] (UMLS© 2001 edition) with for each concept a weight factor attached, indicating the importance of the concept for that particular text. The set of concepts for that text is called a conceptual fingerprint. The system is not limited to the use of UMLS©, but can operate with any (hierarchical) thesaurus. For each full text article in the test set a conceptual fingerprint was obtained.

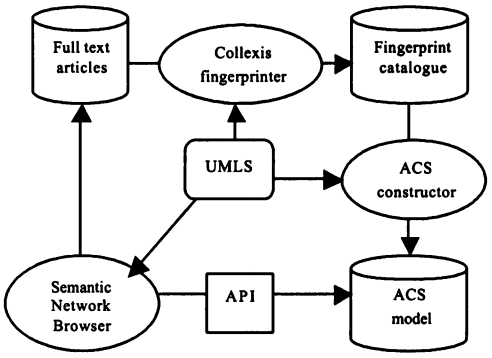

Figure 1. Overview of the architecture. The fingerprints generated by the Collexis software are stored in a fingerprint catalogue. The ACS construction software reads these fingerprints, creates a model and stores the model in a database. The browser will read the model from the database and visualize it.

The fingerprint system scans the documents for terms associated with concepts from the thesaurus using linguistic methods. For each concept it computes a number of statistical information measures: relative frequency, specificity (of the words used for the concept), and textual similarity. These measures are then used to compute a rank for each concept. A clustering algorithm finds the largest number of words in a sliding window that map to a concept in the thesaurus which satisfies the ranking criteria. Furthermore, the Collexis© system exploits the contextual information of homonym concepts to determine what sense is meant in the document. Conceptual fingerprints can be viewed as vectors in a

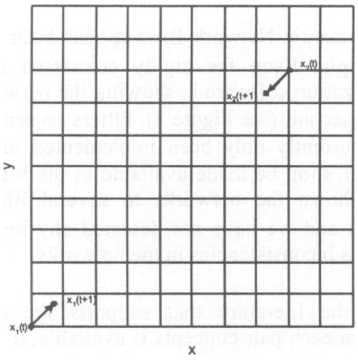

Figure 1. Movement of two concepts following the association rule; the concepts are moving to each other in epoch t+1 due to cooccurrence in a conceptual fingerprint.

high-dimensional space and are used within the Collexis© system to facilitate vector-space based retrieval. All conceptual fingerprints computed are stored in a catalogue for permanent use (see Figure 1 for an architectural overview).

Associative Conceptual Space

The full set of conceptual fingerprints derived from the Nature Genetics articles were fed to a system that constructs an associative conceptual space (ACS) [19]. The algorithm implemented in the system will place each concept in the most appropriate positions in an *n*-dimensional space. The most appropriate position is that position that reflects best the relationships with all other concepts. A relationship between concepts exists if the concepts are both present in a fingerprint. The strength of the relationship is related to the number of times these two concepts co-occur in a fingerprint. Concepts that have a strong relationship are thus positioned close to each other. The advantage of this approach over graphs is that it provides a notion of direction. When trying to find a connection between two concepts in the ACS space, we know in what direction to search. In a graph, the positions of the nodes do not convey any information.

The concepts from the conceptual fingerprints are positioned at random positions in the ACS. When a set of concepts co-occur in a conceptual fingerprint – where we have set a threshold of 0.4 for the weight of the concepts to be included – the concepts are moved to each other (see Figure 2). Non-related concepts are pushed away from the concepts in the fingerprint to prevent contraction of the space. Analysis showed

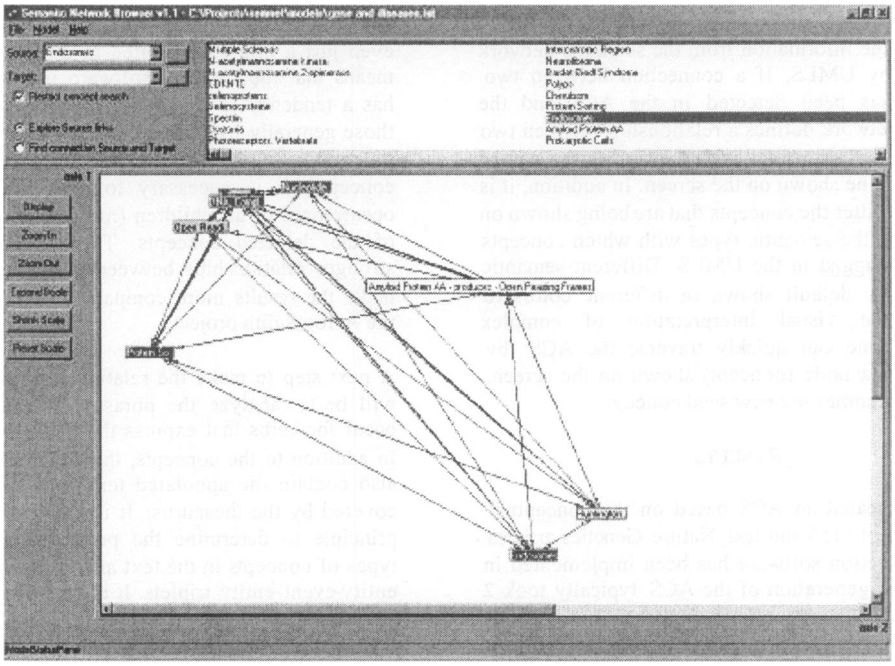

Figure 2. An overview of the semantic network browser with which an ACS can be browsed. Concepts are shown in different colors according to their semantic type. When crossing an edge, the system shows the name of the relationship as obtained from UMLS.

that the ACS model converges to a stable situation after 10 training epochs. The convergence also holds for different number of dimensions.

Connections between concepts were stored in a database together with the coordinates for the concept as computed in the ACS. We have implemented a number of functions that allow us to traverse the ACS. The first function is that we can ask the system to provide all concepts related – i.e., that have a co-occurrence of at least 1 – to a particular seed concept. This function will return a list of vertices and the connections between these vertices. Secondly, we can ask the system to return the strongest pathways between two concepts. Note that these concepts may lie far apart and may have many intermediate concepts on their pathway, but the system can compute long paths of connected concepts between two seed concepts. This function returns the vertices for the path. Thirdly, a function is available that returns the unique document identifiers on which a particular connection between two concepts has been based.

The visualization tool (Semantic Network Browser) uses the application programming interface (API) of the ACS system to create a visual presentation of the concepts and their interrelationships. The semantic network browser combines the information from the ACS with the information from the semantic network provided by UMLS. If a connection between two concepts has been detected in the ACS and the semantic network defines a relationship between two semantic types associated with the concepts, a named relation will be shown on the screen. In addition, it is possible to filter the concepts that are being shown on the basis of the semantic types with which concepts have been tagged in the UMLS. Different semantic types are by default shown in different colors to simplify the visual interpretation of complex networks. One can quickly traverse the ACS: by clicking on a node (concept) shown on the screen, this node becomes the new seed concept.

RESULTS

We have created an ACS based on the conceptual fingerprints of 1155 full text Nature Genetics articles. The construction software has been implemented in C^{++} and the generation of the ACS typically took 2 hours on a 1 GHz Pentium (with 10 training epochs for convergence). This model has been stored in a database for browsing. The model contains 1828 different concepts from UMLS2001 with a total of 165961 potential relations (based on co-occurrence in the same article) between these concepts.

The Semantic Network Browser has been developed in Delphi. Even for highly connected nodes, the visualization software is showing the network in less than a second (see Figure 3). Filters on semantic type have currently only been implemented in the ACS, but will soon be made available in the browser. We have shown the networks to several life sciences experts and we have not detected any immediately obvious inconsistencies in the networks.

Since the literature that supports the connection between each pair concepts is available, it is possible to implement an automatic procedure that detects the possible interesting concepts that are referred in two sets of literature, but where the intersection of these two sets of literature is empty or almost empty. In the ACS model these concepts can be found by searching for A-B-C triplets of concepts where concept A is related to B and concept B to C and where the references that support A-B are different from the set supporting B-C.

DISCUSSION

The analysis of full text documents, including the mapping of synonyms to a concept, yields much detailed information on relevant concepts than the word based analysis of only MedLine abstracts or even just a title as exploited in other systems. This means that the Collexis software used in this study has a tendency to find more specific concepts than those generally used in abstracts or titles. In order to give more weight to particular relationships between concepts, it is necessary to also include the co-occurrences of all children (more specific concepts) of the detected concepts. This would then yield stronger relationships between concepts and would make the results more comparable with those from the ArrowSmith project.

A next step to make the relationships more explicit, will be to analyze the phrases in which concepts occur for verbs that express the type of relationship. In addition to the concepts, the Collexis fingerprints also contain the annotated text with the words not covered by the thesaurus. It is therefore possible in principle to determine the proximity of particular types of concepts in the text and to determine typical entity-event-entity triplets. It is thus possible to look at particular verbs in the environment of two concepts, even if these verbs are not yet covered by the thesaurus as part of a known "event" concept. We will use this additional information to move from the form of semantic networks represented here, where the connections between concepts are

essentially representing potential relationships to networks that accumulate knowledge on actual relationships between concepts.

In order to make this approach truly applicable for genetics, we have to augment UMLS with symbols of genes and gene products. The current UMLS coverage of this area is very weak, resulting in many missed concepts .

Finally, we will develop a tool with which bio-medical experts can validate segments of the network and indicate whether connections are valid. In this way, it will be possible to augment the set of validated connections with newly discovered relationships. This will in turn make it easier to improve the information retrieval and the knowledge discovery process.

CONCLUSION

The method presented here has advantages over the traditional knowledge discovery tools that work on titles or abstracts. Secondly, creating a model in advance rather than on demand of the researcher makes it possible to swiftly browse a network and allows the researcher to generate hypotheses on the fly. Obviously multiple models from different selected text corpora can be created on demand.

The Associative Conceptual Space model allows the researcher to find pathways between concepts that do not have a common neighbor, but are connected through a number of intermediate concepts. The ACS model also improves information retrieval, since researchers can look at the concepts and see what combination of concepts are related. When these concepts are selected, the system automatically retrieves the related set of documents .

REFERENCES

1. Joubert M, Fieschi M, Robert JJ, Volot F, Fieschi D. UMLS-based conceptual queries to biomedical information databases: an overview of the project ARIANE. J Am Med Inform Assoc 1998;5:52-61.
2. Hersh WR, Brown KE, Donohoe LC, et al. CliniWeb: managing clinical information on the World Wide Web. J Am Med Inform Assoc 1996;3:273-80.
3. Detmer WM, Barnett GO, Hersh WR. MedWeaver: integrating decision support, literature searching, and Web exploration using the UMLS MetaThesaurus. In: Proceedings of the AMIA Annual Fall Symposium. 1997:490-4.
4. Lawrence S, Giles CL, Bollacker K. Digital Libraries and Autonomous Citation Indexing. IEEE Computer 1999;32:67-71.
5. Giles CL, Lawrence S. Accessibility of information on the web. Nature 1999;400:107-9.
6. Craven M, Kumlien J. Constructing biological knowledge bases by extracting information from text sources. Proc Int Conf Intell Syst Mol Biol 1999:77-86.
7. Swanson DR, Smalheiser NR. An interactive system for finding complementary literatures: a stimulus to scientific discovery. Artificial Intelligence 1997;91:183-203.
8. Weeber M. Literature-based Discovery in Biomedicine [thesis]. Groningen: Rijksuniversiteit Groningen; 2001.
9. Stapley BJ, Benoit G. Biobibliometrics: information retrieval and visualization from co- occurrences of gene names in Medline abstracts. Pac Symp Biocomput 2000:529-40.
10. Veeling A, Van der Weerd P. Conceptual grouping in word co-occurrence networks. In: Proceedings of IJCAI '99; 1999: Morgan Kaufmann Publishers, San Francisco, USA; 1999. p. 694-9.
11. Ng S-K, Wong M. Toward routine automatic pathway discovery from on-line scientific text abstracts. Genome Informatics 1999;10: 104-12.
12. Lindsay RK, Gordon MD. Literature-based discovery by lexical statistics. J Am Soc Inf Sci 1999;50:574-87.
13. Rindflesch TC, Tanabe L, Weinstein JN, Hunter L. EDGAR: extraction of drugs, genes and relations from the biomedical literature. Pac Symp Biocomput 2000:517-28.
14. Cousins SB, Silverstein JC, Frisse ME. Query networks for medical information retrieval – assigning probabilistic relationships. In: Miller RA, ed. Proceedings of the Fourteenth Annual Symposium on Computer Applications in Medical Care. Los Alamitos CA: IEEE Computer Society Press; 1990:800-4.
15. Sekimizu T, Park HS, Tsujii J. Identifying the interaction between genes and gene products based on frequently seen verbs in Medline abstracts. Genome Inform Ser Workshop Genome Inform 1998;9:62-71.
16. Stephens M, Palakal M, Mukhopadhyay S, Raje R, Mostafa J. Detecting gene relations from MedLine abstracts. Pacific Symposium on Biocomputing 2001;6:483-496.
17. Van Mulligen EM. Diwersy M, Schmidt M, Buurman H, Mons B. Facilitating networks of information. Proc AMIA Symp 2000, p. 868-72
18. Lindberg DAB, Humphreys BL, McCray AT. The unified medical language system. Meth Inf Med 1993;32:281-91.
19. Schuemie M. Associatieve Conceptuele Ruimte, een vorm van kennisrepresentatie ten behoeve van informatie-zoeksystemen [Master thesis]. Rotterdam: Erasmus University; 1998.

Acknowledgments

We kindly acknowledge Nature Publishing Group for granting us to use their articles for this research project. The National Library of Medicine is acknowledged for providing us with the Unified Medical Language System. We thank Collexis for allowing us to use their software.