Thinking PubMed: an Innovative System for Mental Health Domain

Maja Hadzic[1], Russel D'Souza[2], Fedja Hadzic[1], Tharam Dillon[1]
[1]*Digital Ecosystems and Business Intelligence Institute (DEBII), Curtin University of Technology*
[2]*Department of Psychiatry, University of Melbourne*
E-mail: m.hadzic@curtin.edu.au, rdsouza1@bigpond.net.au, f.hadzic@curtin.edu.au, t.dillon@curtin.edu.au

Abstract

Information regarding mental illness is dispersed over various resources but even within a specific resource, such as PubMed, it is difficult to link this information, to share it and find specific information when needed. Specific and targeted searches are very difficult with current search engines as they look for the specific string of letters within the text rather than its meaning.

In this paper we present Thinking PubMed as a system that results from synergy of ontology and data mining technologies and performs intelligent information searches using the domain ontology. Furthermore, the Thinking PubMed analyzes and links the retrieved information, and extracts hidden patterns and knowledge using data mining algorithms. This is a new generation of information-seeking tool where the ontology and data-mining work in concert to increase the value of the available information.

1. Introduction

The recognition that mental health is costly and many cases will not become chronic if treated early has lead to an increase in research in the last 20 years. This has lead to an accumulation of mental health information. This information covers different mental illnesses with a huge range of results regarding different disease types, symptoms, treatments, disease causing factors (genetic and environmental) as well as candidate genes that could be responsible for the onset of these diseases. Information regarding mental illness is dispersed over various resources but even within a specific resource, such as PubMed, it is difficult to link this information, to share it and find specific information when needed. A huge body of the available information is accessible over PubMed database. We need to take a systematic approach to making use of this information that cannot reached its full value unless being systematically analysed and linked with other available information from the same domain. There is a need to increase the transparency of the PubMed database and enable intelligent management, usage and analysis of the mental health information.

Specific and targeted searches are very difficult with current search engines as they look for the specific string of letters within the text rather than its meaning. For example, searching for "chair" returns conference program chair, arm chair, wheelchair, electric chair, professorial chair etc. In search for "genetic causes of bipolar disorder", Google provides 96,500 hits which are a large assortment of well meaning general information sites with few interspersed evidence-based resources. Medline Plus (http://medlineplus.gov/) retrieves 53 articles including all information about bipolar disorder plus information on other mental illnesses. A large number of articles is outside the domain of interest and is on the topic of heart defects, eye and vision research, multiple sclerosis, Huntington disease, psoriasis etc. PubMed (http://www.ncbi.nlm.nih.gov/pubmed) gives the list of 1945 articles. The user needs to select the relevant articles as some of the retrieved articles are on other illnesses such as schizophrenia, autism and obesity.

Wilczynski *et al.* [1]: "*General practitioners, mental health practitioners, and researchers wishing to retrieve the best current research evidence in the content area of mental health may have a difficult time when searching large electronic databases such as MEDLINE. When MEDLINE is searched unaided, key articles are often missed while retrieving many articles that are irrelevant to the search.*" (MEDLINE is part of PubMed)

Most frequently users need to select the relevant articles from the list, read each article individually and establish the links between the selected articles manually. But all this does can be avoided as we already have the technologies that will reduce this extensive workload and enable us to simply see targeted answers to our questions.

Ontology is enriched conceptual model for representing domain knowledge. Ontology captures and represents specific domain knowledge through specification of meaning of concepts including definition of the concepts and domain-specific relationships between those concepts. The precise specifications constrain the potential interpretations of the ontology concepts enabling the ontology to be used by automatic application such as data mining.

Ontologies have been suggested as a mechanism to provide applications with domain knowledge. Ontologies provide a shared common understanding of a domain and can be used to support knowledge integration, use and sharing by different applications, software systems and human resources [2, 3]. The main advantages of

1063-7125/08 $25.00 © 2008 IEEE
DOI 10.1109/CBMS.2008.20

ontologies are that they increase data semantic (i.e, provide context for information); enable knowledge sharing, representation, creation and management; and support intelligent information retrieval and information integration.

Data mining is a set of processes that is based on automated searching for actionable knowledge buried within a huge body of data. Data mining help to extract information, to find hidden patterns and to make predictive models for decision making and new discoveries. Data mining helps find patterns and knowledge that are embedded in the data and requires exploration and analysis of large quantities of target data for the purpose of better understanding and deriving knowledge regarding the problem at hand.

Data mining draws work from areas including database technology, machine learning, statistics, pattern recognition, information retrieval, neural networks, knowledge based systems, artificial intelligence, high-performance computing and data visualization [4]. Some advantages of data mining over the traditional approaches include:

- efficient processing of large and complex data (scalability)
- automatically analysing, detecting errors and inconsistencies, classifying, and summarizing the data with no human intervention (automation)
- extracting novel and useful patterns which leads to new knowledge and discoveries (knowledge extractions)
- combining the advantages of various disciplines (multi-disciplinary nature)
- reducing the costs and time associated with the data analysis as a result of its automation (cost and time efficiency).

Integration of ontology and data mining technologies creates a highly efficient and effective system that performs intelligent searchers, and extracts and links the relevant information for the users. In this paper, we present such a system specifically designed to increase the value of the mental health information available through PubMed database.

2. Relevance of the Related Work to Our Vision

The importance of ontologies has been recognised within the medical community and work has begun on developing and sharing biomedical ontologies [5]. A number of biomedical ontologies exists, but we mention as a relevant example Gene Ontology and UMLS. The aim of Gene Ontology (GO) (http://www.geneontology.org/, The Gene Ontology Consortium 2007) project is to enable consistent descriptions of gene products in different databases by using the GO to annotate major repositories for plant, animal and microbial genomes. The Unified Medical Language System (UMLS) [6] project is a collection of many biomedical vocabularies. While there are 1 million biomedical concepts in UMLS, only 135 semantic types and 54 relationships are used to classify these concepts. This is an indication that the UMLS can be seen rather as comprehensive thesaurus then as an ontology. Moreover, as the UMLS is designed as a collection of all biomedical concepts, using the UMLS for a specific domain can be rather inconvenient. UMLS is not specific enough for some domains such as mental health domain, and greater definition of concepts is needed. However, some UMLS concepts can be used to create a Mental Health Ontology and further elaborate the selected knowledge both in terms of set of concepts and in their relationships.

Use of the ontologies in the information retrieval, extraction and data mining has been reported. For the purpose of this paper, we have classified the existing tools into two groups: (1) the tools designed for other than mental health domain such as Textpresso, BioIE, GenIE, GENIA and (2) tools based on UMLS such as MedScan, Medical World Search Engine, BioAnnotator, In the sequel, we introduce some of the existing tools.

Textpresso [7] perform semantic searches through *Caenorhabditis elegans* literature using ontology of 14,500 terms which is based on Gene Ontology. BioIE [8] is an extraction system of the information covering biological interactions. It annotates the results using the terms of Gene Ontology. GenIE [9] extracts information about biochemical pathways, sequences, structures and functions of genomes and proteins. It uses both simple processing techniques as well as syntactic and semantic analysis based on a domain ontology. GENIA [10] corpus covers information about biological reactions concerning transcription factors in human blood cells. All the MEDLINE abstracts on this topic and their titles have been marked-up for biologically meaningful terms, and these terms have been semantically annotated using the GENIA ontology. GENIA ontology is a taxonomy of 47 biologically relevant nominal categories.

The tools of the first group are designed for a different knowledge domain and utilize an ontology or metadata annotation system for that specific knowledge domain so that they cannot be used for the purpose of this project which is the mental health domain. Textpresso, BioIE, GenIE and GENIA work is useful for points of reference but it cannot be used by our systems as it is focused on literature about *Caenorhabditis elegans*, biological interactions, genomics and proteomics, transcription factors in human blood cells respectively, and not on mental health literature.

UMLS have been used to support information retrieval and analysis process. MedScan [11] utilizes a specially developed context-free grammar and lexicon designed from Gene Ontology and UMLS. It processes sentences from MEDLINE abstracts and produces a set of logical

structures corresponding with the meaning of each sentence. Medical World Search Engine [12] uses the UMLS as its built-in knowledge of medical terminology and a selected set of medical resources to perform its searches. BioAnnotator [13] uses domain-based dictionaries (UMLS, LocusLink and GeneAlias) for recognizing known terms within the given text. Due to the incomplete nature of the used dictionaries, a rule engine has been designed for discovering new terms.

The tools of the second group are covering an extensive knowledge domain so that specific and targeted searches as well as intelligent analysis of mental health information are practically impossible with these tools. Moreover, some limitations of UMLS have been reported. Suarez *et al.* state that "*Some aspects of the UMLS also decrease its usefulness in information retrieval*", while Mukherjea *et al.* that "*We have seen that many biological concepts are missing from UMLS…*"

Within the health domain, data mining techniques have been predominately used for tasks such as gene expression analysis, drug design, genomics and proteomics. The data analysis necessary for microarrays has necessitated data mining [14].

The use of data mining techniques is encouraging and is useful as a point of reference, but no data mining algorithms have been effectively applied within mental health domain. We aim to bring an innovative breakthrough in this area as we believe that data mining techniques could play a significant role in the study of mental health. The data mining algorithms can be applied to derive patterns specific to mental illness, such as exposing a unique combinations of genetic and environmental factors responsible for onset of an illness in question. The implementation of this system on the PubMed data has the potential to reveal the hidden knowledge and positively transform the way we control and manage mental health.

Our research objectives are similar to the ones accomplished by GOPubMed [15]. The two main differences between GOPubMed and Thinking PubMed are that (1) the GOPubMed is based on Gene Ontology and is not specifically tailored to address the needs of mental health domain, and (2) the GOPubMed does not apply data mining techniques on the retrieved results.As the Gene Ontology has been designed to capture and represent mainly the knowledge specific to genes and as the GOPubMed uses the Gene Ontology to perform its searches, searching for gene-related topics are highly efficient with the GOPubMed. Gene Ontology does not effectively capture and represent mental health information in great detail. For this reason, searching for mental health-related topics using the GOPubMed is not as efficient as they could be with a search engine that uses Mental Health Ontology to perform searches on the PubMed database. While the GOPubMed applies Gene Ontology to perform meaningful searches through PubMed database, we aim to use Mental Health Ontology to perform meaningful searches through PubMed database specific to mental health domain, and go a step further and apply data mining algorithms on the retrieved results.

Wilczynski *et al.* [1] developed search strategies that can help discriminate the PubMed literature with mental health content from articles that do not have mental health content. Our research ideas go beyond this. Our aim is not to improve the current way of retrieving information; we are focused on the transforming the way information is retrieved, transforming the keyword-based searches into ontology-based searches and lists of the retrieved articles into maps of related articles. The new Thinking PubMed is getting a completely different nature from the current search engines.

3. The Thinking PubMed in Tree Steps

The three important parts of the Thinking PubMed are:

- Generic Mental Health Ontology (GMHO), to capture and represent shared knowledge of the mental health domain
- Semi-automated annotation system, to use the GMHO and annotate the PubMed articles, enabling them to be used by automatic processes
- Data mining algorithms, to mine the target information and reveal the knowledge hidden within the corpus of data

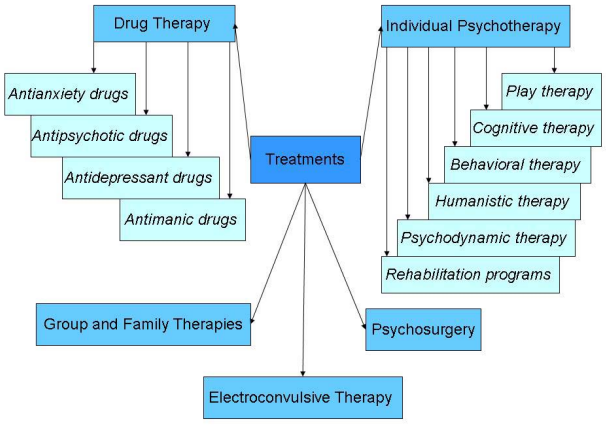

Figure 1. Top-level hierarchy of the 'Treatments' sub-ontology of the GMHO

3. 1 Design of the Generic Mental Health Ontology

The Generic Mental Health Ontology (GMHO) captures and represents generic mental health information. The GMHO has hierarchical structure and is organized according a number of different sub-ontologies, capturing the different aspects of mental health. For example, top-level hierarchy of the 'Treatments' sub-

ontology is represented in Figure 1. Each of the treatment methods, is further classified to represent different kinds of this particular treatment method. For example, Individual Psychotherapy includes Play therapy, Cognitive therapy, Behavioral therapy, Humanistic therapy, Psychodynamic therapy and Rehabilitation programs.

As only a small portion of UMLS is related to mental health, some UMLS concepts specific to the mental health domain can be used in combination with the frequently used concepts from PubMed mental health corpus (corpus of articles on mental health).

The complete PubMed mental health corpus consists of approximately 72500 entries. We will first choose X the most relevant entries. The first 0.9X entries will be used during the GMHO design to identify generic ontology concepts. The other 0.1X entries will represent the corpus of data 'unseen' by the designers and machines, and will be used to evaluate the conceptual coverage of the designed ontology. Where possible, the UMLS identifier will be given for each GMHO concept. Additionally, each GMHO concept will be identified with its synonyms and with grammatical variations of these concepts. For example, 'mental illness' will be identified with 'mental disorder' as its synonym and 'mentally ill' as a grammatical variation of this concept. For the modelling of the ontologies, we will utilise Protégé developed by Stanford University.

GMHO will be a first comprehensive knowledge model of mental health domain. GMHO will be used to annotate the PubMed abstracts. This will enable the automatic applications such as data mining, to 'read' and 'understand' the PubMed content and use the available information constructively. As GMHO will capture and represent knowledge that generally holds to be true within mental health domain, it can be seen as a reusable knowledge component and will be made available to be used in other mental health applications.

3. 2 Design of the Semi-automated Annotation System

The automated annotation will be carried out through the following three steps: tokenization (splitting the sentences into tokens), matching the tokens against the GMHO terms and matching the tokens against the GMHO relationships. The tokenization will happen dynamically and the tokens will be matched against the GMHO terms and relationships against the GMHO relationships until the best fit is found. Use of highly expressive knowledge models such as ontologies enables the machines to view the text not as a set of words but as set of meaningful expressions.

Human intervention is required to check the validity of the annotation performed by machines. It is to be expected that the machine will make some annotation mistakes in the beginning. For this reason, we will implement machine-learning algorithms and strive towards establishing a system that will require minimal human intervention.

3. 3 Implementation of the Data Mining Algorithms

The annotated iinformation can be used in automatic data analyses. We have done some preliminary work on data and ontology mining [16, 17].

A number of articles from the PubMed mental health corpus will be selected. This set will be split into two smaller sets, one for deriving the knowledge model (source set) and one for testing the derived knowledge model (test set). During pattern discovery phase, the source set will be intelligently explored and analysed to extract information, find hidden patterns and knowledge embedded in the data and make predictive models. These kinds of results are hypothesis at this stage. But if supported, could make a significant contribution to the research and study of mental illnesses. During testing and evaluating of the knowledge, the test set will be used to verify the hypothesis so that it can become reliable enough to extend the current knowledge. In this phase, we will also experiment with variation of data mining parameters as their choice can affect the nature and granularity of the obtained results.

Because of the nature of the data we are examining, it is necessary to use powerful data mining techniques that go beyond relational data mining techniques. The nature of the information we are mining requires the use of data mining technologies for complex data. Specifically, we need to use tree-mining technologies for both ordered and unordered trees, induced and embedded trees as well as support definitions which allow for transaction-based, occurrence-match and hybrid support [18, 19].

3. 4 The Three As One

Use of the GMHO within the system will enable effective information retrieval and analysis within PubMed database. GMHO will be used to annotate the content of the PubMed using semi-automated annotation tool and data mining algorithms will be applied on the annotated data. The synergy of the ontology and data mining technologies enables systematic organization of data (through use of ontology) and gives the resulting system reasoning abilities of constructive nature (through implementation of data-mining technologies).

The PubMed database has the potential to become transparent when the GMHO is used in combination with data mining algorithms. The annotation of the PubMed database using the GMHO enables systematic and meaningful organization of the data and gives a clear insight and overview of the available information. Application of data mining algorithms exposes patterns and knowledge buried within this data. The integration of ontology and data mining technologies will result in

integration of data. Our access to the information will not be limited by the access to a specific article but will also allow access to the integrated knowledge. The PubMed corpus on mental health data will become a single large virtual article with a systematically organized content showing the references to specific articles. In this way, the PubMed database become transparent as nothing of the information is unseen or hidden from the user.

The design of Thinking PubMed has been inspired by the way mental health experts read mental health information. Years of research and study give the mental health expert knowledge and understanding of the mental health domain. Thinking PubMed receives this knowledge in the form of Generic Mental Health Ontology. When reading every article, mental health expert uses his mental health knowledge to understand the information they read. In an analogous way, Thinking PubMed uses GMHO to annotate the PubMed articles, and in this way, 'understands' the information written in these articles. Lastly, the mental health expert, establishes links between different articles, analyses and compares the complementary and overlapping information and optionally proposes various hypothesises. Thinking PubMed carries out similar tasks using data mining algorithms. The big difference between the mental health expert and Thinking PubMed is the time required to carry out information retrieval and analysis as well as the amount of information that can be processed together.

4. How Does The Thinking Pubmed Address the Key Problems

The Thinking Pubmed will address a large number of problems slowing advancement in mental health research. These include:

a) keyword-based search engines

The current search engines perform keyword based searches. The retrieved results are often meaningless and/or outside the domain users are interested in. The users still need to manually extract and analyse the retrieved information. With the Thinking PubMed there is no sifting necessary by the researcher; the system will use the ontology to perform intelligent searches by looking for the meaning of the information over its appearance in the text. The ontologies will also enable the system to link complementary information from the different PubMed articles. The added advantage of data mining techniques will expose the knowledge and patterns hidden in the large body of information.

b) the available information is not annotated

PubMed database store over 70000 entries on the topic of mental health. This information is not annotated. The annotation of information makes the information machine 'readable' and 'understandable', and enables the automated applications, such as data mining algorithms, to expose the knowledge and patterns buried in the corpus of data. GMHO will make the annotation possible and open the door for automatic information processing tools to use this information intelligently.

c) increasing body of mental health information

As the research continues, new papers or journals are frequently published and added to the PubMed database. Use of the ontology will enable us to annotate this new information and merge it with the existing body of already annotated knowledge. We gain a better control over the storage, access and retrieval of the data. No information can be lost or escape the notice in this way.

d) information overlaps, redundancies and complementarities

Portions of the PubMed data may be related to each other, portions of the information may overlap, and portions of the information may be semi-complementary between one another. The Thinking PubMed will be able to link the annotated information and spot overlaps, redundancies and complementarities in the data. 'Common knowledge', for example, is going to reduce the possibility of undertaking the same experiments, such as examining the same region of DNA sequence, by different research groups, thus saving time and resources. This is going to create a cooperative environment making big research tasks coherent between different research teams.

e) complexity of mental illnesses

Despite major medical advances, the identification of genetic and environmental factors responsible for mental illnesses still remains unsolved and is therefore a very active research focus today. Most published papers cover only one factor and perhaps one aspect of a mental illness. For example, in the paper "Bipolar disorder susceptibility region on Xq24-q27.1 in Finnish families" (PubMed ID: 12082562), the research team Ekholm *et al.* examined one genetical factor (Xq24-q27.1) for one type of mental illness (bipolar disorder). As mental illnesses do not follow Mendelian patterns but are caused by a number of genes usually interacting with various environmental factors [20], all factors for all aspects of the illness need to be considered. Currently, no tool exists that allows examination and analysis of the different factors simultaneously. The Thinking PubMed will use data mining algorithms to expose the patterns in the data, facilitate the search for the combinations of genetic and environmental factors involved and provide an indication of influence.

5. Conclusion and future work

The Thinning PubMed represents a new generation of information-seeking tool. The totality of the innovation comes by combining ontology with data mining techniques. The system uses ontologies to perform semantic-based meaningful searches, and applies data mining algorithms to find and expose hidden knowledge and patterns within the retrieved information. A number

of similar systems exist within other domains, but the Thinking PubMed is the only one to address the needs of mental health domain.

In our research centre, we have a number of experts working on the Thinking PubMed. The system is in its early implementation phase, and lots of work still remains. The progress will be reported in our future publications.

References

[1] N.L. Wilczynski, R.B. Haynes, T. Hedges, "Optimal search strategies for identifying mental health content in MEDLINE: an analytic survey", *Annals of General Psychiatry*, vol. 5, 2006.

[2] A. Gómez-Pérez, "Towards a framework to verify knowledge sharing technology", *Expert Systems with Applications,* vol. 11, no. 4, 1996, pp. 519-529.

[3] A. Gómez-Pérez, "Knowledge sharing and reuse", *The Handbook on Applied Expert Systems*, CRC Press, pp. 1-36, 1998.

[4] S. Sestito, T.S. Dillon, *Automated knowledge acquisition*, Prentice Hall of Australia, Sydney, 1994.

[5] W. Ceusters, P. Martens, C. Dhaen, B. Terzic, "LinkFactory: an advanced formal ontology management System", *Proceedings of interactive tools for Knowledge Capture* (KCAP 2001), 2001.

[6] O. Bodenreider, "The Unified Medical Language System (UMLS): integrating Biomedical terminology", *Nucleic Acids Research*, vol. 32, no. 1, 2004, pp. 267-270.

[7] H. Muller, E. Kenny, P. Sternberg, "Textpresso: An ontology-based information retrieval and extraction system for biological literature', *PLoS Biology*, vol. 2, no. 11, 2004, p.e309.

[8] J. Kim, J. Park, "BioIE: Retargetable information extraction and ontological annotation of biological interactions from the literature", *Journal of Bioinformatics and Computational Biology*, vol. 2, no. 3, 2004, pp. 551-568.

[9] P. Cimiano, U. Reyle, J. Saric, "Ontology-driven discourse analysis for information extraction", *Data and Knowledge Engineering*, vol. 55, 2005, pp. 59-83.

[10] J. Kim, T. Ohta, Y. Tateisi, J. Tsujii, "Genia corpus – semantically annotated corpus for bio-textmining", *Bioinformatics*, vol. 19, 2003, pp. i180–182.

[11] S. Novichkova, S. Egorov, N. Daraselia, "Medscan, a natural language processing engine for Medline abstracts", *Bioinformatics*, vol. 19, no. 13, 2003, pp. 1699-1706.

[12] H. Suarez, X. Hao, I. Chang, "Searching for information on the internet using the UMLS and medical world search", *Proceedings of the Annual AMIA Fall Symposium*, America, 1997, pp. 824-828.

[13] S. Mukherjea, L. Subramaniam, G. Chanda, S. Sankararaman, R. Kothari, V. Batra, D. Bhardwaj, B. Srivastava, "Enhancing a biomedical information extraction system with dictionary mining and context disambiguation", *IBM Journal of Research and Development*, vol. 48, no. 5/6, 2004, pp. 693-701.

[14] G. Piatetsky-Shapiro, P. Tamayo, "Microarray data mining: Facing the challenges", *SIGKDD Explorations*, vol. 5, no. 2, 2003, pp. 1-6.

[15] A. Doms, M. Schroeder, "GOPubMed: exploring PubMed with the Gene Ontology", *Nucleic Acid Research*, vol. 33, 2005, pp. W783-786.

[16] M. Hadzic, F. Hadzic, T.S. Dillon, "Tree mining in mental health domain", *Proceedings of the Hawaii International Conference on System Sciences (HICSS-41)*, USA, 2008.

[17] M. Hadzic, F. Hadzic, T.S. Dillon, "Mining of Health Information from Ontologies", *Proceedings of the International Conference on Health Informatics (HEALTHINF2008)*, Portugal, 2008.

[18] F. Hadzic, H. Tan, T.S. Dillon, "UNI3 – Efficient Algorithm for Mining Unordered Induced Subtrees Using TMG Candidate Generation", *IEEE Symposium on Computational Intelligence and Data Mining*, USA, 2007.

[19] F. Hadzic, H. Tan, T.S. Dillon, E. Chang, "Implications of frequent subtree mining using hybrid support definition, *Proceedings of the Data Mining & Information Engineering*, UK, 2007.

[20] D.G. Smith, S. Ebrahim, S. Lewis, A.L. Hansell, L.J. Palmer, P.R. Burton, "Genetic epidemiology and public health: hope, hype, and future prospects", *The Lancet*, vol. 366, no. 9495, 2005, pp. 1484-1498.