1-15-2013

ARIANA: Adaptive Robust and Integrative Analysis for finding Novel Associations

Vida Abedi

ARIANA: Adaptive Robust and Integrative Analysis for finding Novel Associations

by

Vida Abedi

A Dissertation

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

Major: Electrical and Computer Engineering

The University of Memphis

May, 2013

Dedication: I would like to dedicate this dissertation to Ariana, my daughter, for bringing me joy every single day; and to my parents, Farah and Majeed, for loving me and believing in me all along.

Acknowledgements

I would firstly thank my supervisor, Dr. Mohammed Yeasin, his valuable knowledge and guidance has provided me with an immense foundation on which to continue my PhD and my career in research. Also many thanks are extended to my teachers here at the University of Memphis and also special thanks to my committee members, Dr. Deaton, Dr. Homayouni, Dr. Eckstein, Dr. Jacobs, Dr. Nolan, and Dr. George, for their time and support.

Secondly, I would like to thank the members of the CVPIA lab for their friendship and continuous support; but I would like to express a special thanks to Fazle E. Faisal for programming support during his two years stay at the University of Memphis; to Geoffrey West and Trinity Brown for help in editing my writing while they were here; to Pratiksha Subedi for her unconditional friendship and help throughout my four years in the college; to Karththikka Ramani Muthukuri mainly for her contribution towards implementation of our web site, Venkata Ananth Choalla for his web development support and Pouya Bashivan for his help during the past few months.

Special thanks are extended to my brother Keyvan Abedi for his constructive feedbacks, critical mind and editing help. Special thanks are also extended to my dear sister, Dona Abedi, without whom I would not be the same person today.

I also would like to extend my thanks to my dear husband and colleague Dr. Ramin Zand for encouraging and supporting me throughout the years in graduate school, and also for his continuous love and friendship.

Abstract

Abedi, Vida. Ph.D. The University of Memphis. 12/2012. ARIANA: Adaptive Robust and Integrative Analysis for finding Novel Associations. Major Professor: Dr. Mohammed Yeasin.

The effective mining of biological literature can provide a range of services such as hypothesis-generation, semantic-sensitive information retrieval, and knowledge discovery, which can be important to understand the confluence of different diseases, genes, and risk factors. Furthermore, integration of different tools at specific levels could be valuable. The main focus of the dissertation is developing and integrating tools in finding network of semantically related entities.

The key contribution is the design and implementation of an Adaptive Robust and Integrative Analysis for finding Novel Associations. ARIANA is a software architecture and a web-based system for efficient and scalable knowledge discovery. It integrates semantic-sensitive analysis of text-data through ontology-mapping with database search technology to ensure the required specificity. ARIANA was prototyped using the Medical Subject Headings ontology and PubMed database and has demonstrated great success as a dynamic-data-driven system. ARIANA has five main components: (i) Data Stratification, (ii) Ontology-Mapping, (iii) Parameter Optimized Latent Semantic Analysis, (iv) Relevance Model and (v) Interface and Visualization. The other contribution is integration of ARIANA with Online Mendelian Inheritance in Man database, and Medical Subject Headings ontology to provide gene-disease associations.

Empirical studies produced some exciting knowledge discovery instances. Among them was the connection between the hexamethonium and pulmonary inflammation and fibrosis. In 2001, a research study at John Hopkins used the drug hexamethonium on a

healthy volunteer that ended in a tragic death due to pulmonary inflammation and fibrosis. This accident might have been prevented if the researcher knew of published case report. Since the original case report in 1955, there has not been any publications regarding that association. ARIANA extracted this knowledge even though its database contains publications from 1960 to 2012. Out of 2,545 concepts, ARIANA ranked "*Scleroderma, Systemic"*, "*Neoplasms, Fibrous Tissue*", "*Pneumonia*", "*Fibroma*", and "*Pulmonary Fibrosis*" as the 13th, 16th, 38th, 174th and 257th ranked concept respectively. The researcher had access to such knowledge this drug would likely not have been used on healthy subjects.

In today's world where data and knowledge are moving away from each other, semantic-sensitive tools such as ARIANA can bridge that gap and advance dissemination of knowledge.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction to Computing Semantically Related Networks of Entities and Novel Associations

Introduction

The effective mining of biological literature can provide a range of services such as hypothesis generation or semantic sensitive retrieval of information. This service helps to understand the potential confluence of various diseases, genes, risk factors as well as biological processes. In the sense of usability and scalability, the utility of semantic-sensitive knowledge discovery tools is the tremendous increases in scientific publications and the diversity of the concepts that can be brought to the attention of the practitioner or researcher.

Exploratory studies and hypothesis generation often begin with searches and study of existing literature to identify a set of factors and their association with diseases, phenotypes, or biological processes. Many scientists are overwhelmed by the sheer volume of literature for a disease as they plan to generate a new hypothesis or study a biological phenomenon. The situation is even worse for junior investigators who often find it difficult to formulate new hypotheses or, more importantly, corroborate whether their hypothesis is consistent with existing literature. It is a daunting task to be abreast with so much being published and also remember or formulate all combinations of direct and indirect associations. Fortunately there is a growing trend of using literature mining and knowledge discovery tools in biomedical research. However, there is still a large gap between the huge amount of effort and resources invested in disease research and the little effort and return in harvesting the published knowledge. To bridge this gap it is imperative to design and implement efficient, robust, scalable, usable and domain specific knowledge discovery tools as well as integration of tools in finding network of semantically related entities. The concept of ontology mapping, semantic analysis and relevance model can be used to design and implement an adaptive robust and integrative analysis in providing a range of services in biomedicine.

In biomedicine, generating disease-models based on literature data is a very natural and efficient way to better understand and summarize the current knowledge about different high-level systems. Identifying connecting elements between diseases can provide a systematic approach to identify missing links and potential associations. Connecting seemingly unrelated entities could also present new opportunities for collaborations and interdisciplinary research. A connection between two diseases can be formalized as a risk factor, symptom, treatment option, side-effect of a drug, any other diseases, or genes. These concepts are intuitive and provide a context to the researcher to perform specialized searches within the networks of associations. However, identification of these concepts is key in providing a precise domain and accurate information extraction tool. Web tools and online databases can be used to extract these concepts through an ontology mapping process. The key idea is not to implement everything in house from scratch, but to use and map the available resources in a meaningful and efficient way in order to provide a knowledge discovery and information retrieval system that is as best as it can be with the current technology.

In order to construct disease interaction networks, it is essential to identify factors (or concepts) associated with each disease independently. If only genetic concepts are to be taken into account, then genomic data could be used; however, if concepts are considered to cover a wider range then other types of data, such as text data, could be used. In essence, to build a high level view of the disease interaction network, it is essential to utilize factors (or concepts) at a higher level of granularity. For instance, instead of carefully analyzing the chemical structure of interacting compounds, it would be more appropriate to use groups of compounds such as “inorganic compounds”, or “heterocyclic compounds”. However, the system should be flexible enough to incorporate new concepts when a significant amount of information becomes available; additionally, it should allow information from Online Mendelian Inheritance in Man (OMIM) database and similar databases to be integrated for further refinement of the system.

A plethora of the state-of-the-art Web applications on improving information retrieval and users' experience was reported in contemporary literatures and was succinctly reviewed in a recent survey by [1]. A total of 28 tools, targeted to specific needs of a scientific community, were assessed to compare functionality and performance. The common underlying goal of them all was to improve the relevance of search results, to provide a better quality of service as well as to enhance the user experience with the PubMed database. Though these applications were developed to minimize “information overload”, the question of scalability and improving knowledge discovery requires further research. Among the reported applications, the EBIMed [2], is the closest to our work in the sense of finding relationships between concepts.

In particular, among the 28 tools, five used clustering to group the search results into topics; another five systems used different techniques to summarize the results and present a semantic overview of the retrieved documents. The following items are a subset illustrating the scope and potential of these Web tools. One of the systems, Anne O’Tate, [3] uses post processing to group the results into predefined categories such as MeSH topics, author names, year of publication. Even though this tool can be very helpful in presenting the results to the user, it does not provide the additional steps needed to extract semantic relationships and a network of associations. The McSyBi [4], clusters the results to provide an overview of the search and to show relationship among the retrieved documents. It is reported that LSA is also used in the backbone of that system; in addition to that feature, the top 10,000 publications are only analyzed. However, a fully integrated system is not available to run any queries from the Web tool. The program XploreMed [5] allows the users to further explore the subjects and keywords of interest. MedEvi [6] provides ten concept variables as semantic queries. XploreMed puts a significant limit (>500) on the number of abstracts to analyze. CiteXplore (http://www.ebi.ac.uk/citexplore/ date last accessed: 19 October 2012) combines literature search and data mining, it also provides information from other sources such as patent records. MEDIE [7] provides utilities for semantic search based on deep-parsing and, returns text fragments to the user. EBIMED [2] extracts proteins, Gene Ontology, drugs and species, and identifies relationships between these concepts based on co-occurrence analysis.

The STRING - a Search Tool for the Retrieval of Interacting Genes/Proteins [8] and iHOP - [9] were not among the 28 tools reviewed by [1]. Both applications translate unstructured textual information into more computable forms and cross-link them with relevant databases. However, the underlying techniques cannot capture the semantic relationship among entities. According to Altman et al., [10] existing techniques still lack the ability to effectively present biological data in easy to use form and thereby further knowledge discovery by integrating heterogeneous sources of data.

To reduce information overload and complement traditional means of knowledge dissemination, it is imperative to develop robust, scalable and highly precise Web-service applications that are versatile enough to meet the “specific” needs of a diverse community. The utility of such a system would be greatly enhanced with the added capability of finding semantically similar concepts related to various risk factors, side-effects, symptoms and diseases. There are a number of challenges in developing such a robust, yet versatile Web-based tool. One of the main challenges is to create a fully integrated and a functional system that is specific to a targeted audience, yet flexible enough to be creatively used by a diverse range of users. To be effective, it is necessary to have a data stratification process that is global and complete. It is also important to map the range of concepts using a set of criteria to a “dictionary” that is specific to the community through an ontology mapping process. Second, it is essential to ensure that the knowledge-discovery process that is: i) scalable with the growing size of input data, ii) effective in capturing the semantic relationships and networks of concepts, and iii) capable of displaying an embedded multi-layer network of those concepts. Third, it is essential to present to the user a data-driven threshold that can be used to classify the extracted concepts at distinct levels of associations. Finally, an easy-to-use interface with proper visualization is critical to the success of such a tool in meeting the needs of consumers with diverse needs and desires.

Key Contributions

The main focus of the dissertation is developing of efficient, robust, scalable, usable and domain-specific knowledge-discovery tools as well as integration of tools in finding networks of semantically related entities. The concepts of ontology mapping, semantic analysis and relevance model were introduced to design and implement an adaptive robust and integrative analysis. The key contributions are**:**

I. A pilot study to implement Hypotheses Generation Framework (HGF)

II. Design and implement the Adaptive Robust and Integrative Analysis for finding Novel Associations (ARIANA). ARIANA was built on top of the HGF using different layers of Ontology Mapping more suitable for integrative analysis.

III. Integration of ARIANA with the OMIM database was performed to include genetic data.

IV. Case studies to showcase the utility of the HGF and ARIANA

V. A fully integrated Web service and visualization to enhance utility and user experience for ARIANA.

The HGF implemented as pilot study shares similar end goals to SWAN [11] but is more holistic in nature and was designed and implemented using scalable and efficient computational models of disease-disease interaction. The integration of mapping ontologies with latent semantic analysis (LSA) is critical in capturing domain specific direct and indirect "crisp" associations, and making assertions about entities (such as: disease X is associated with a set of factors Z). Pilot studies were performed using two diseases. A comparative analysis of the computed "associations" and "assertions" with curated expert knowledge was performed to validate the results. The encouraging results from the HGF framework and its ability to capture “crisp” direct and indirect associations, and provide knowledge discovery on demand planted the seed for developing fully integrated system for integrative analysis.

One of the key contributions is the design and implementation of an Adaptive Robust and Integrative Analysis for finding Novel Associations. ARIANA is a software architecture and a web-based system for efficient and scalable knowledge discovery. It integrates semantic-sensitive analysis of text data through ontology mapping with database search technology to ensure the specificity required to create a robust model in finding relevant results to a query on an ocean of data.

The ARIANA was prototyped using the MeSH (Medical Subject Headings) ontology and PubMed database for biomedical applications and has demonstrated great success as a dynamic data-driven system that has the capability to improve the quality of information retrieval, knowledge discovery and networks of semantically related concepts or entities. ARIANA has five main components: (i) Data Stratification, (ii) Ontology Mapping, (iii) Parameter Optimized Latent Semantic Analysis (POLSA), (iv) Relevance Model and (v) Interface and Visualization.

Based on the domain knowledge and the expert choice of the concepts and entities a very large and broad database is created using a fully automated process. Ontology mapping was performed on the large database to generate a context-specific dictionary for the domain of an application. A Parameter Optimized Latent Semantic Analysis (POLSA) was used to create a model based on the statistical co-occurrences and to rank list the entities that capture the association between the query and the entities/concepts in the database. A relevance model, based on a user query was introduced to translate the ranked list into three categories of connections, namely, strongly related, related and not related. The relevance model is a trimodal distribution, whose parameters are estimated and the cut points are determined dynamically for every query given to the system. The interface and visualization module receives one or more keywords from the user and the output is a multi-layered network that is collapsible /expandable to a level of detail that is user selectable. These features make the Web tool easy to interact with and provide flexibility needed to serve a diverse range of users.

Another key contribution is the integration of ARIANA with Online Mendelian Inheritance in Man (OMIM) database, a flat list of human curated Gene-disease, and with MeSH (hierarchical database) to provide gene-disease associations.

Empirical studies using of ARIANA resulted some exciting knowledge discovery and network of semantically related entities. Among the observations, the connection between drug Hexamethonium and Pulmonary inflammation and Fibrosis deserves special mention. In a research study (2001) at John Hopkins used this drug that ended in tragic death of Ellen Roche, a healthy volunteer, who died only after few days of inhaling this drug. Following her death, she was diagnosed with pulmonary inflammation and fibrosis based on chest imaging and autopsy report. However, this accident might have been prevented if the researcher knew of a case report published in 1955. Furthermore, since the original case report there has not been any new publications regarding the association of Hexamethonium and pulmonary fibrosis. ARIANA was able to extract this information from an ocean of publication even though the 1955 case report was not in the database. Out of 2,545 concepts in the system, ARIANA ranked " *Scleroderma,* *Systemic"* as the 13th ranked concept, *Neoplasms, Fibrous Tissue*" as the 16th ranked concept, "*Pneumonia*" as the 38th ranked concept, "*Neoplasms, Connective and Soft Tissue>Neoplasms, Connective Tissue>Neoplasms, Fibrous>Fibroma* " as the 174th ranked concept, and finally the "*Pulmonary Fibrosis*" as the 257th ranked concept. If the researcher had access to such knowledge, it was clear that this medication would not have been used on healthy subjects without further investigating its safety.

In today's world where large amounts of information are generated each day and these quantities must be reviewed to obtain useful knowledge, semantic-sensitive tools such as ARIANA and integration of complementary computational tools can bridge this gap and advance dissemination of the resulting knowledge.

Key features that distinguish this work from other state-of-the-art solutions are: i) domain specificity and context dependent model, ii) scalability, iii) broad coverage of literature, iv) extraction of direct as well as indirect associations based on higher order co-occurrence analysis among biological entities, v) different layers of integrative analysis at tool level and at data level, and vi) flexible and easy to use interface design and prototype.

Domain specificity is achieved through usage of customized data-driven dictionary using the ontology mapping. The goal of modular, efficient and scalable design was achieved through the integration of a parameter optimized latent semantic analysis (POLSA) based technique with data stratification based on expert knowledge. The broad coverage is achieved by judiciously extracting and stratifying 50 years of literatures to create the core data set. The limitation of the LSA model on single query was addressed using multi-gram-context-specific dictionary. This also enabled the system to capture direct as well as indirect association that is based on higher order statistical co-occurrence. The integration of ARIANA with OMIM has captured the concept of integration of tools to extract indirect gene-disease association. MeSH and PubMed are integrated through the ARIANA's framework in order to extract association among different biological and medical entities or concepts at a data level. Finally, the design and prototype of the interface guarantee a level of flexibility and ease of use to a wide range of users. The main features of the interface are the graphical representation and the collapsibility features in addition to the option of exporting the complete set of results for further analysis.

The remaining chapters are organized as follows: Chapter 2 provides research context and broad ideas of concepts and tools that are involved developing the tools for integrative analysis. Following this in Chapter 3, a detailed description of HGF framework is presented with empirical studies to showcase the utility of the pilot study. Encouraged by the success of HGF a fully integrated system called ARIANA was developed; this effort is discussed in details in Chapter 4. Empirical studies using domain expert were also presented to corroborate the findings obtained from the system. To further advance the cause of integrative analysis, the ARIANA was integrated with OMIM database to capture gene-disease association; that work is described in the Chapter 5 along with case study to showcase the novel associations. Chapter 6 concludes the dissertation with few remarks on key findings, lessons learned, and suggestions for future directions.

Chapter 2: Research Context

In the post-genomic era and with the advances in high-throughput technologies, new doors have been opened to study and map genetic networks such as human diseases [12, 13, 14]. The majority of new research directions have focused on the genetic causes of diseases by looking at one or few diseases at once. It was only in 2007 that Goh et al. [14] took a conceptually different approach; they proposed an interaction between two diseases when both diseases were associated with a common gene. This idea led to construction of *diseasome*, disease-disease interaction network [14]. This higher level of abstraction, moving from one disease and many genes, or gene byproducts, to many diseases and their respective genes or gene byproducts, provided a new outlook to view genetic networks within bioinformatics community. However, the disease-disease interaction network proposed by Goh et al., [14] relied only on gene-disease interaction data. It was only recently that a combination of disease-gene information and protein-protein information [15] was used to enhance the quality of such network. These types of high level analysis provide insights into topological features and functional properties of the disease interaction network. However, diseases can also be connected through non-genetic features such as risk factors, side-effects of drugs and treatments, or signs and symptoms. Therefore constructing a disease network based on genetic and non-genetic factors can be a valuable reference for clinicians and medical researchers.

A network in biology is comprised of a set of nodes representing biochemical or chemical entities and a set of edges representing interaction between those entities. For instance, in a protein-protein interaction network, nodes represent proteins, and edges could be evidence for physical interaction between proteins. Similarly, in a disease-disease interaction network nodes represent diseases and edges could be common genes between diseases. Network analysis of disease-disease interaction network (where edges represented common genes) showed that the vast majority of genes associated with diseases are non-essential and do not tend to encode hub proteins; in addition to that, genes contributing to a common disorder i) have tendency for their by-products to interact with each other through protein-protein interactions, ii) have tendency to be co-expressed, and iii) tend to share Gene Ontology terms [14] . Hence, the results from studying these complex networks furthered our knowledge to a different level of understanding. As a result, scientists no longer attempt to study one gene or one gene byproduct at a time; rather they plan to study a family of genes or even group of genes (using microarrays) that respond to a given perturbation at one time.

In the Human Disease Network (HDN) nodes represent diseases and edges are common genes between diseases [14]; hence if two diseases share at least on common gene, then there will be an edge between the two diseases. The HDN [14] is constructed based on genetic information, is a major step in providing a higher level of abstraction; yet information content is based only on the genetic information from the Online Mendelian Inheritance in Man (OMIM) database (http://www.ncbi.nlm.nih.gov/omim). OMIM is a collection of human genes and genetic phenotypes; the database contains information for over 12,000 genes and is updated on a daily basis. Even though OMIM provides the level of association between the genes and phenotypes, including how the association was found, the HDN does not incorporate this additional information into the network.

A disease network provides a higher level of abstraction when compared to gene regulatory networks, or protein-protein interactomes. This higher level of abstraction facilitates translational research and is instrumental in clinical studies. This type of analysis can provide a valuable reference for clinicians and medical researchers. However, a disease network could also be constructed based on literature data to incorporate a wider range of factors such as side effects and risk factors.

In fact, generating disease-models based on literature data is a very natural and efficient way to better understand and summarize the current knowledge about different high-level systems. Identifying connecting elements between diseases can provide a systematic approach to identify missing links and potential associations, while presenting new opportunities for collaborations and interdisciplinary research. As compared to only common disease-genes a connection between two diseases can be formalized as a risk factor, symptom, treatment option, side-effect of a drug, or any other disease. These concepts are intuitive and provide a context to the researcher to perform specialized searches within the networks of associations. However, identification of these concepts is key in providing a precise domain and accurate information extraction tool. Web tools and online databases can be used to extract these concepts through an ontology mapping process. A key idea is not to implement everything in house, but to use and map the available out of house resources in a meaningful and efficient way in order to provide a knowledge discovery and information retrieval system that is as best as it can be with the current technology.

In order to construct disease interaction networks, it is essential to identify factors (or concepts) associated with each disease independently. If only genetic factors are taken into account, then genomic data could be used; however, if factors are considered to cover a wider range then other types of data, such as text data, could be used. In essence, to build a high level view of the disease interaction network, it is essential to utilize factors (or concepts) at a higher level of granularity. For instance, instead of carefully analyzing the chemical structure of interacting compounds, it would be more appropriate to use groups of compounds such as “inorganic compounds”, or “heterocyclic compounds”. However, the system should be flexible enough to incorporate new concepts when a significant amount of information becomes available; additionally, it should allow information from OMIM database or similar type of databases to be integrated to this network for further refinement of the system.

The availability of huge textual resources provides the scientists with the chance to search for correlations or associations such as protein-protein interactions [16, 17], and gene-disease associations [18, 19]. However, biology and medicine are rich in terminology; for instance, in pathology reports and medical records, 12,000 medical abbreviations have been identified [20]. In addition, this large vocabulary is also dynamic and new terms emerge rapidly. For instance, the same object may have several names, or distinct objects can be identified with the same name; when in the former case the names are synonyms while in the latter case the objects are homonyms [20]. Consequently, literature-mining of biological and medical text becomes a very challenging task and the terms that suffer the most are gene and protein names [21, 22]. Alternatively, to design and implement a more accurate system, it is important to understand and tackle these challenges at their root level. However, even more challenging is the implementation of the information extraction, also known as deep parsing.

Deep parsing is built on formal mathematical models, attempting to describe how text is generated in the human mind (i.e. formal grammar) [20]. Deterministic or probabilistic context-free grammars are probably the most popular formal grammars [22]. Grammar-based information extraction techniques are computationally expensive as they require the evaluation of alternative ways to generate the same sentence. Grammar-based information could therefore be more precise but at the cost of reduced processing speed [20]. An alternative to the grammar-based methods are vector-based methods such as Latent Semantic Analysis (LSA) method. These alternative methods rely on bag-of-words concept, and have therefore reduced computational complexity. In addition to that, LSA technique has the added advantage of extracting direct and indirect associations between entities.

In essence, since the traditional information retrieval framework, which relies on keyword-based approaches, cannot cope with the huge amount of information that is being produced on a daily basis, scientists have focused on more sophisticated techniques such as text-mining [22] coupled with data-mining approaches. This shift has proven to be valuable in many instances. For example, titles from MEDLINE were used to make connections between disconnected arguments: 1) the connection between migraine and magnesium deficiency [23] which has been verified experimentally; 2) between indomethacin and Alzheimer's disease [23]; and finally 3) between *Curcuma longa* and retinal diseases [24]. Hypothesis generation in literature-mining relies on the fact that 'chance' connections can emerge to be meaningful [22]. Use of vector based models such as LSA in hypothesis generation could be very effective due to the reduction of computational complexity. However such a system should be designed with special care and consideration to be as context specific and precise as possible.

Latent Semantic Analysis

Latent Semantic Analysis (LSA) is a well known information retrieval technique which has been applied to many areas in bioinformatics. In LSA framework [25], a word-document matrix (also known as tf-idf matrix) is commonly used to represent a collection of text (corpus). LSA extracts relations based on second order co-occurrence from a corpus, and maps them onto K-dimensional vector space. The discrete indexed words are projected into an *eigen* space by applying singular value decomposition (SVD).

Arguably, LSA captures some semantic relations between various concepts based on their distance in the *eigen* space [26]. The most common similarity measure used to rank the vectors is the linear cosine similarity measure [26] . The three main steps of the LSA are outlined here and can be found in [25]:

I. **Creation of Term-Document matrix**: The text documents are represented using a bag-of-words model. This representation creates a term-document matrix in which the rows are the words (dictionary), the columns are the documents, and the individual cell contains the frequency of the term appearance in the particular document. Term Frequency (TF) and Inverse Document Frequency (IDF) are used to create the TF-IDF matrix.

II. **Singular Value Decomposition (SVD)**: SVD or SparseSVD (approximation of SVD) is performed on the TF-IDF matrix and the *k* largest *eigen*vectors are retained. This k-dimensional matrix (encoding matrix) captures the relationship among words based on first and second order statistical co-occurrences.

III. **Information Retrieval**: Information related to a query can be retrieved by first folding-in the query into the LSA space and then performing a similarity measure between the documents and the query. A Cosine similarity measure is usually used to rank and retrieve the documents.

Parameter Optimized Latent Semantic Analysis

Even though LSA has been applied to many areas in bioinformatics, the LSA models have been based on *adhoc* principles. In a recent work, the parameters affecting the performance of LSA were studied to develop a Parameter Optimized Latent Semantic Analysis (POLSA) [27]. The various parameters examined were corpus content, text preprocessing, sparseness of data vectors, feature selection, influence of the1st *Eigen* vector, and ranking of the encoding matrix. The optimized parameters should be chosen whenever possible.

Improving the Semantic Meaning of POLSA Framework

Methods such as LSA have been successful in finding direct and indirect associations between various entities. However, these methods still use bag-of-words concept; therefore, they do not take into account the order of words and hence the meaning of such words are often lost. Using multi-keyword words would alleviate some of the problems of the bag-of-words model. In a multi-keyword dictionary, the word “vascular accident” (which is a synonym of “stroke”) would be differentiated from “accident” which could also mean car accident in a different context.

However, it is challenging to generate such a dictionary. If all combinatorial words in the English dictionary are chosen, then the size of such dictionary would be considerably larger even if one considers up-to three-gram words. An implication of the larger dictionary is an increased sparsity in the TF-IDF matrix. A possible solution is to construct the dictionary based on combinations of words that are biologically relevant for the case of biological text-mining. Identification of biologically relevant word combinations can be derived from biological ontology such as Gene Ontology (www.geneontology.org) or Medical Subject Headings (http://www.ncbi.nlm.nih.gov/mesh). Using a multi-keyword dictionary could in principle improve the accuracy of the vector-based frameworks, such as the LSA, that rely only on bag-of-words models. Use of multi-gram dictionary provides also a mean of extracting associations based on higher order co-occurrences.

Knowledge-Based Systems

To build a system or tool, it is important to utilize the most appropriate source of data. In biology and medicine, PubMed (http://www.ncbi.nlm.nih.gov/pubmed/) is the main source of text data. PubMed is a public database developed and maintained by the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/), and updated on a daily basis. Currently this database contains more than 22 million citations for biomedical literature. Whenever an article has an abstract, that abstract is published through PubMed. One of the features of retrieving articles through PubMed is the fact that all entries are tagged using the Medical Subject Headings controlled vocabulary. MeSH vocabularies are used to describe the subject of each journal article. MeSH contains approximately 26,000 terms and is updated annually to reflect changes in the medical field. MeSH terms are arranged hierarchically by subject categories and PubMed allows one to view this hierarchy and search the literature using the controlled vocabularies.

Constructing the core database that is based both on PubMed and MeSH through an ontology mapping process is critical for a system that is robust. Because the MeSH vocabulary ensures that articles are uniformly indexed by subject, whatever the author's words, its integration of the MeSH database is extremely important.

Web Services

Building an integrated disease network would be ineffective unless researchers can interact with the system and obtain valuable information directly. Hence, for a system to be used by clinicians and researchers it is imperative to have a robust and practical application tool. Providing a simple web page to display the results or developing a java applet are conventional frameworks with limited potentials. Instead of providing a web page, it is also possible to provide a web service (WS) to the end users. A WS is a software module, available via network (typically internet) which completes a task and returns the results to the user. WS technology is built on the concept of software-as-a-service (SaaS) [28]. In SaaS, software and data are hosted on a central machine, which is usually connected to the internet, and clients query the system through a web browser over the internet. In a Web service environment, all computation and implementations details are hidden from the user (client); hence, client and server interact only through a well defined interface [28].

There are four key advantages in using a Web service framework as compared to a web-based application [28]: 1) Web service can act as client or server and can respond to a request from an automated application without any human intervention. This feature provides a great level of flexibility and adaptability; 2) Web services are modular and self-descriptive: the required inputs and the expected output are well defined in advance; 3) Web services are manageable in a more standard approach. Even when a Web service is hosted in a remote location, accessible only through the network, and is written in an unfamiliar language, it is still possible to monitor and manage it by using external application management and workflow systems; 4) a Web service can be used by other applications when similar tasks need to be executed. This is particularly important as more tools are being developed and soon integrated in order to provide improved services.

Furthermore, one of the main characteristics of a Web service is that it provides a framework that operates reliably and delivers a consistent service at a variety of levels. In addition, each service may offer various choices of quality of service (QoS) based on technical requirements, focusing on both functional and non-functional properties of services. Examples of such QoS are availability, accessibility, integrity, conformance to standards, reliability, scalability, performance and security [28, 29] . Hence, in a WS context, QoS provides assurance on a set of quantitative characteristics. Finally, web services are the next generation of web-based technology and applications. They provide a new and improved way for applications to communicate and integrate with one another [28]. The implications of this transition are profound, especially with the growing body of data and the available bioinformatics tools.

Chapter 3: Hypotheses Generation Framework

Introduction

Summary of the Study

In bio-medicine, exploratory studies and hypothesis generation often begin by searching existing literature to identify a set of factors and their association with diseases, phenotypes, or biological processes. Many scientists are overwhelmed by the sheer volume of literature on a disease when they plan to generate a new hypothesis or study a biological phenomenon. Fortunately there is a growing trend of using literature mining and knowledge discovery tools in biomedical research. However, there is still a large gap between the huge amount of effort and resources invested in disease research and the little effort in harvesting the published knowledge. The proposed hypothesis generation framework (HGF) finds "crisp semantic associations" among entities of interest - that is a step towards bridging such gaps. The proposed HGF shares similar end goals like the SWAN but the goals are more holistic in nature. HGF was designed and implemented using scalable and efficient computational models of disease-disease interaction. The integration of mapping ontologies with latent semantic analysis is critical in capturing domain-specific direct and indirect "crisp" associations, and making assertions about entities (such as disease X is associated with a set of factors Z). Pilot studies were performed using two diseases. A comparative analysis of the computed "associations" and "assertions" with curated expert knowledge was performed to validate the results. It was observed that the HGF is able to capture "crisp" direct and indirect associations, and provide knowledge discovery on demand. The proposed framework is fast, efficient, and robust in generating new hypotheses to identify factors associated with a disease. A full integrated Web service application is being developed for wide dissemination of the HGF. A large-scale study by the domain experts and associated researchers is underway to validate the associations and assertions computed by the HGF.

Key Features of the Work

Key features that distinguish this work from state-of-the-art solutions are: i) domain specificity, ii) scalability, iii) board coverage of literature and, iv) extraction of direct as well as indirect associations among biological entities. Domain specificity is achieved through usage of customized medical dictionary; scalability is achieved through implementation of a LSA based technique further discussed in the methodology section; and broad coverage is based on the fact that 20 years of literature is used to create the data set. Finally, because LSA based method is used, the system is capable of capturing direct as well as indirect association among different entities.

Data Sources and Materials

PubMed Database is a public database developed and maintained by the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/), and updated on a daily basis. Currently this database contains more than 22 million citations for biomedical literature. The core database constructed here is based on the data from PubMed. In the next section the exact procedure to construct the data modules are described.

Hypotheses Generation Framework

The Hypothesis Generation Framework (HGF) has three major modules: Ontology Mapping to generate data-driven domain specific dictionaries, a parameter optimized latent semantic analysis (POLSA), and Disease Model. The schematic diagram of the overall HGF framework is shown in the Figure 1(A). The model is constructed using the POLSA framework, and it is based on the selected documents and the dictionary (Figure 1C). Users can query the model and the output is a ranked list of headings. These ranked headings are grouped into three sets (unknown factors, potential factors, or established factors) using the Disease Model module (Figure 1C and 1D). Analyzing the headings in the three sets can facilitate hypothesis generation and information retrieval based on user query.

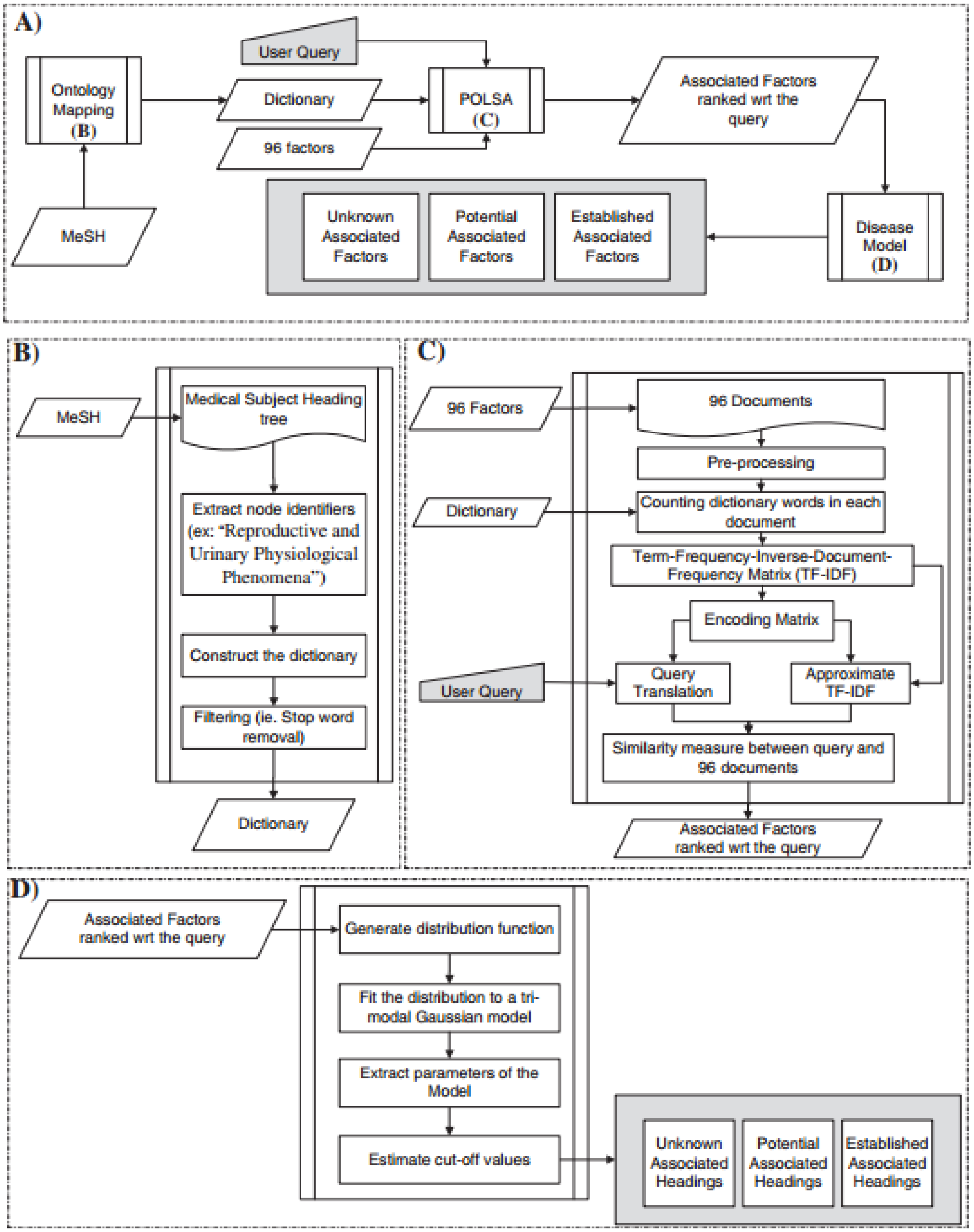

A)
User Query
Ontology Mapping (B)
Dictionary
POLSA (C)
96 factors
Associated Factors ranked wrt the query
MeSH
Unknown Associated Factors
Potential Associated Factors
Established Associated Factors
Disease Model (D)
B)
MeSH
Medical Subject Heading tree
Extract node identifiers (ex: "Reproductive and Urinary Physiological Phenomena")
Construct the dictionary
Filtering (ie. Stop word removal)
Dictionary
C)
96 Factors
96 Documents
Pre-processing
Dictionary
Counting dictionary words in each document
Term-Frequency-Inverse-Document-Frequency Matrix (TF-IDF)
Encoding Matrix
User Query
Query Translation
Approximate TF-IDF
Similarity measure between query and 96 documents
Associated Factors ranked wrt the query
D)
Associated Factors ranked wrt the query
Generate distribution function
Fit the distribution to a tri-modal Gaussian model
Extract parameters of the Model
Estimate cut-off values
Unknown Associated Headings
Potential Associated Headings
Established Associated Headings

Figure 1. Flow diagram of the Hypothesis Generation Framework
A) In a medical and biological setting, Ontology Mapping could use the Medical Subject Heading (MeSH) and generate a context-specific dictionary, which is one of the parameters of the POLSA model. Associated factors are ranked based on a User Query which can be any word(s) in the dictionary. These factors are subsequently grouped into three different bins (unknown factors, potential factors or established factors) based on our Disease Model. B) Ontology Mapping to create domain-specific dictionary. C) Parameter Optimized Latent Semantic Analysis Module. D) Disease Model Module.

Introduction of Ontology Mapping Concept

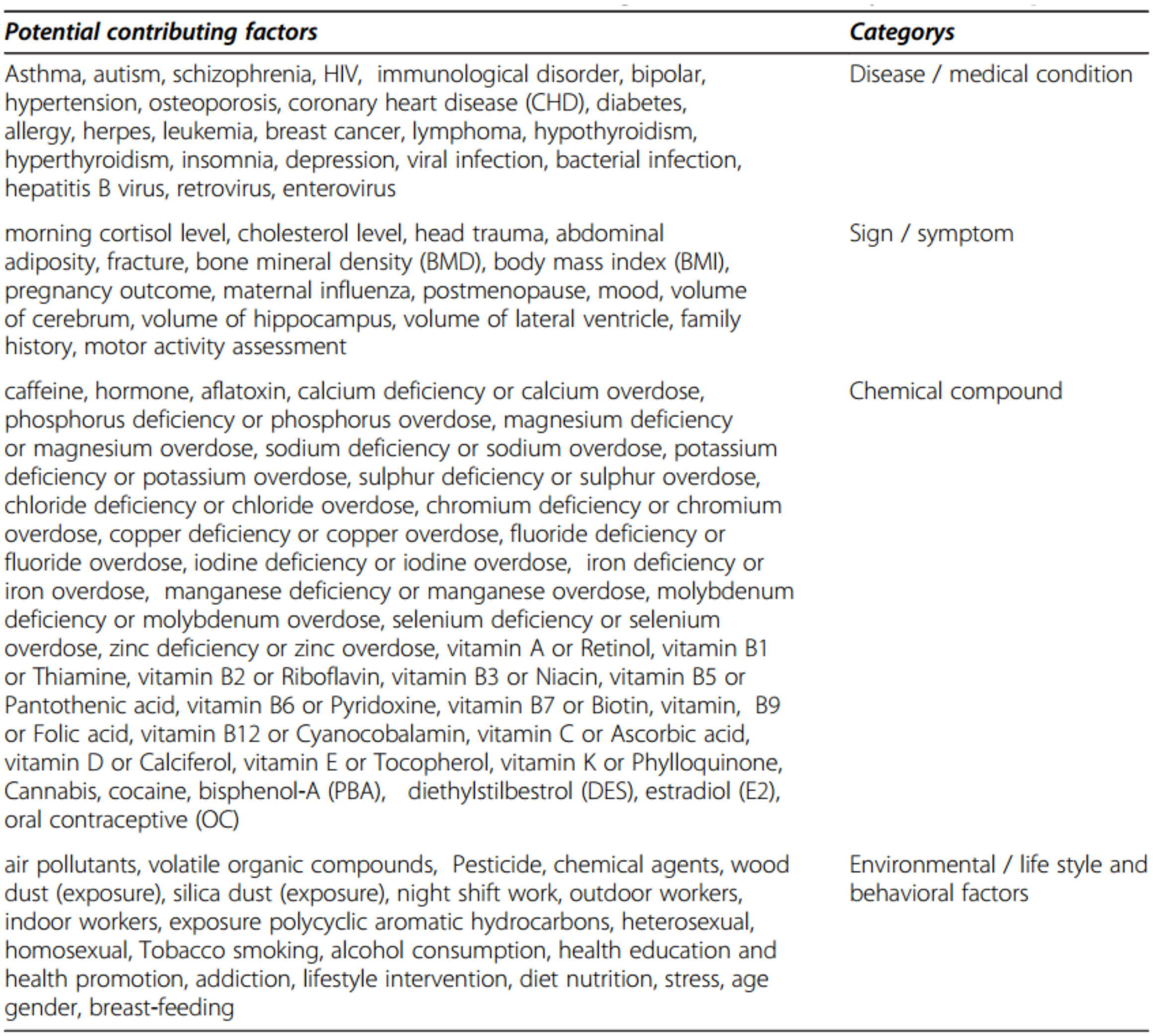
MeSH is used to generate the dictionary in the POLSA model. The mapping of MeSH ontology to create the dictionary for the POLSA significantly enhances the quality of results and provides a crisp association of semantically related entities in biological and medical science. All MeSH headings are reduced to single words to create the context specific and data driven dictionary (see Figure 1B). For instance, “Reproductive and Urinary Physiological Phenomena” is a MeSH term and is reduced to five words in the dictionary (1. Reproductive, 2. and, 3. Urinary, 4. Physiological, and 5. Phenomena). In the filtering step, duplicates as well as stop words such as “and” or words containing two or fewer characters are removed. The final size of this dictionary is 19,165 words. Any dictionary word could be used as a query to the HGF. For instance, the disease “stroke” is a query in this study. The highly ranked factors with respect to a query-disease are considered factors associated with that disease. Cosine similarity measure is used as a metric in the HGF.

POLSA Module

In order to develop an effective literature-mining framework to model disease-disease interaction networks, generate plausible new hypotheses, and support knowledge-discovery by finding semantically related entities, a Parameter Optimized LSA (POLSA)[27] was re-designed and adopted in the proposed HGF framework.

In addition, a set of associated factors was selected to represent interaction between diseases. Ninety-six common associated factors (see Table 1) were selected through a literature review from numerous medical articles by two domain experts. As the first step, a set of articles was selected by querying the PubMed database using a series of diseases and factors. In the second step, the retrieved articles were manually reviewed by domain experts and entities that were associated with diseases or factors were selected. All articles considered for this analysis were peer reviewed articles. In addition, some common diseases such as diabetes and depression were also included in the set of 96 factors, as these are believed to be, in many instances, risk factors to other diseases. Therefore, the set of 96 associated factors represents a wide range of factors including generic factors such as depression and infection as well as specific factors such as vitamin E. As the final step, the set was further revised by an expert in the medical field. Using the improved POLSA technique Yeasin, et al. 2009), meaningful associations from the textual data in the PubMed database are extracted and mined. Furthermore, the factors are ranked based on their level of association to a given query.

Table 1. Potential risk factors and/or contributing factors selected by medical expert



Potential contributing factors	*Categorys*
Asthma, autism, schizophrenia, HIV, immunological disorder, bipolar, hypertension, osteoporosis, coronary heart disease (CHD), diabetes, allergy, herpes, leukemia, breast cancer, lymphoma, hypothyroidism, hyperthyroidism, insomnia, depression, viral infection, bacterial infection, hepatitis B virus, retrovirus, enterovirus	Disease / medical condition
morning cortisol level, cholesterol level, head trauma, abdominal adiposity, fracture, bone mineral density (BMD), body mass index (BMI), pregnancy outcome, maternal influenza, postmenopause, mood, volume of cerebrum, volume of hippocampus, volume of lateral ventricle, family history, motor activity assessment	Sign / symptom
caffeine, hormone, aflatoxin, calcium deficiency or calcium overdose, phosphorus deficiency or phosphorus overdose, magnesium deficiency or magnesium overdose, sodium deficiency or sodium overdose, potassium deficiency or potassium overdose, sulphur deficiency or sulphur overdose, chloride deficiency or chloride overdose, chromium deficiency or chromium overdose, copper deficiency or copper overdose, fluoride deficiency or fluoride overdose, iodine deficiency or iodine overdose, iron deficiency or iron overdose, manganese deficiency or manganese overdose, molybdenum deficiency or molybdenum overdose, selenium deficiency or selenium overdose, zinc deficiency or zinc overdose, vitamin A or Retinol, vitamin B1 or Thiamine, vitamin B2 or Riboflavin, vitamin B3 or Niacin, vitamin B5 or Pantothenic acid, vitamin B6 or Pyridoxine, vitamin B7 or Biotin, vitamin, B9 or Folic acid, vitamin B12 or Cyanocobalamin, vitamin C or Ascorbic acid, vitamin D or Calciferol, vitamin E or Tocopherol, vitamin K or Phylloquinone, Cannabis, cocaine, bisphenol-A (PBA), diethylstilbestrol (DES), estradiol (E2), oral contraceptive (OC)	Chemical compound
air pollutants, volatile organic compounds, Pesticide, chemical agents, wood dust (exposure), silica dust (exposure), night shift work, outdoor workers, indoor workers, exposure polycyclic aromatic hydrocarbons, heterosexual, homosexual, Tobacco smoking, alcohol consumption, health education and health promotion, addiction, lifestyle intervention, diet nutrition, stress, age gender, breast-feeding	Environmental / life style and behavioral factors

Titles and abstracts from PubMed (for the past twenty years) for each of the 96 factors were downloaded on a local system. On average there were 47,570 abstracts per factor; the specific factors such as “maternal influenza” had fewer abstracts associated with them (minimum of 160 abstracts/factor) and the more generic factors such as “hormone” were associated with a greater number of abstracts (a maximum of 557,554 abstracts/factor). The complete collection was then used to construct the knowledge space for the POLSA model. Using a query such as “Parkinson” or “stroke” the 96 factors were then ranked based on their relative level of associations to the query. The distribution of a set of associated factors with respect to a disease was modeled as a tri-modal distribution: a distribution which has three modes. This is due to the fact that some factors are known to be associated with the disease and have high scores. Similarly, some factors are known to be unassociated to the disease and these have negative scores; in addition, some factors may or may not be associated to the disease and these have low similarity scores. Matlab was used to generate two tri-modal distributions based on general Gaussian models for the two distributions obtained from queries “stroke” and “Parkinson”. The model uses the following formulation to describe the tri-modal Gaussian distribution:

$$f = \alpha_1 * \exp\left\{-\left(\frac{(\mathsf{x}-\mu_1)}{\sigma_1}\right)^2\right\} + \alpha_2 * \exp\left\{-\left(\frac{(\mathsf{x}-\mu_2)}{\sigma_2}\right)^2\right\} + \alpha_3 * \exp\left\{-\left(\frac{(\mathsf{x}-\mu_3)}{\sigma_3}\right)^2\right\}, \quad (1)$$

Where α1, α2 and α3 are the scaling factors; μ1, μ2 and μ3 are the position of the center of the peaks, and σ1, σ2, σ3 control the width of the distributions. The goodness of fit was measured using an R-square score.

Disease Model

Using a disease model (see Figure 2), it was possible to map the mixture of three Gaussian distributions into easy to understand categories. The implicit assumption is that if associated factors of a disease are well known, a large body of literature will be available to corroborate the existence of such associations. On the other hand, if associated factors of a disease are not well documented, the factors are weakly associated to the disease with few factors displaying a high level of association (Disease X versus Disease Y as shown in the Figure 2). Since the distribution of association level of factors (including risk factors) will be different in the two scenarios. In the first case (Disease Y) the two dominating distributions are the factors that are associated and those that are not associated with the disease; in the second case (Disease X) the dominating distribution is that of the potential factors. In essence, if one accepts this assumption then the distribution of associated factors follows a tri-modal distribution and it will be intuitive to measure the level of association for different factors with respect to a given disease. Utilization of a disease model (by a tri-modal distribution) allows better identification of the three sets of factors: unknown associations, potential associations and established associations.

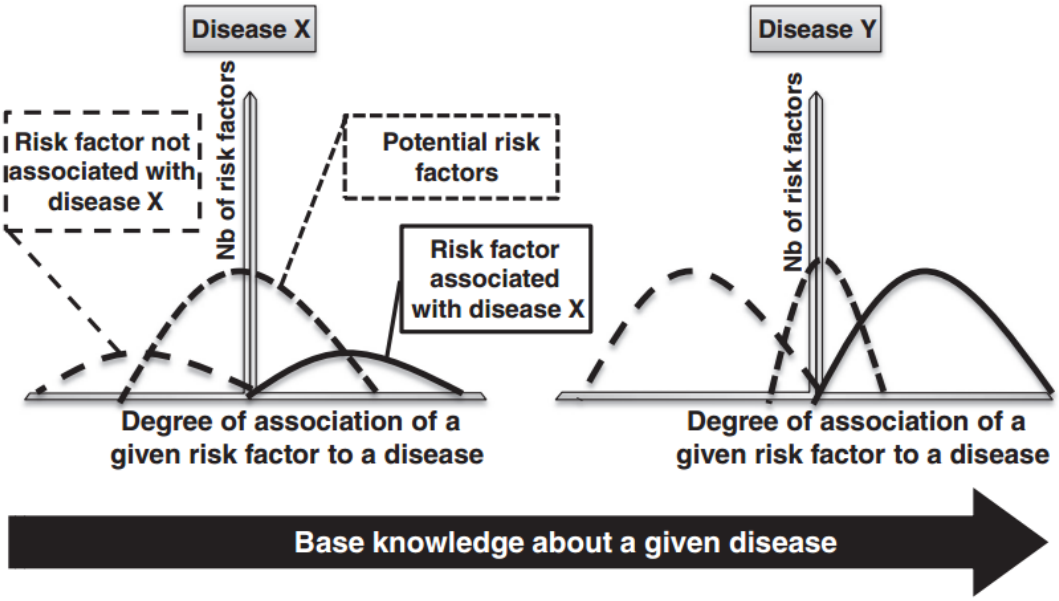

Disease X
Nb of risk factors
Risk factor not associated with disease X
Potential risk factors
Risk factor associated with disease X
Degree of association of a given risk factor to a disease
Disease Y
Nb of risk factors
Degree of association of a given risk factor to a disease
Base knowledge about a given disease

Figure 2. Model for the distribution of associated factors of a given disease
If associated factors – such as risk factors are well known as in the case for Disease Y, then the two dominating distributions are the factors that are associated and those that are not associated with the disease; if on the other hand the associated factors of a disease are not well documented (Disease X) then the dominating distribution is that of the potential factors.

Separating the three distributions allows implementation of a dynamic and data-driven threshold calculation. Hence, the parameters of the distributions can be used to model a cut-off threshold for the factors that are established, potential, or unknown. This method is empirical and provides an intuitive approach to evaluate the results. The score can be further optimized in a heuristic manner with utilization of a large-scale and comprehensive ground truth set. Furthermore, the highly associated factors to the disease are the well known factors; the hidden knowledge on the other hand resides in the region where the associations are positive yet weak.

Results

Two diseases, namely, Ischemic Stroke (IS) and Parkinson's Disease (PD), were used as queries to the hypothesis generation system. The distribution of associated factors is presented in the Figure 3. The results were compared with MedLink neurology (http://www.medlink.com/medlinkcontent.asp), a web resource used by clinicians. Comparative results were summarized in the Figure 4. In the case of IS, most of the associated factors are identified by both systems; however there is a set of factors that have only been identified by the proposed approach. In the case of the PD, a large number of factors have been identified by both systems. However, there are a number of factors that have only been identified by the proposed HGF and only a handful that are mentioned in the MedLink neurology which have positive but low similarity score in the hypothesis generation framework.

The tri-modal distribution model is used to group the associated factors into different levels. The cut-off values to differentiate between different association levels vary slightly depending on the distribution of the similarity scores. The ideal decision boundary can be found if a large number of ground truth cases are available; in this situation the decision boundary is selected intuitively based on the shape of the distributions. For example, in the case of IS, factors are considered highly associated if their cosine score is greater than 0.3, factors are possibly associated if their score is between 0.1 and 0.3 and are possibly not associated if their score is lower than 0.1. In the case of PD, factors are considered highly associated if their cosine score is greater than 0.2, factors are possibly associated if their cosine score is between 0.1 and 0.2 and finally the factors with scores between 0.05 and 0.1 are considered associated at low level, factors with scores lower than 0.05 are considered possibly not associated with the Parkinson’s Disease.

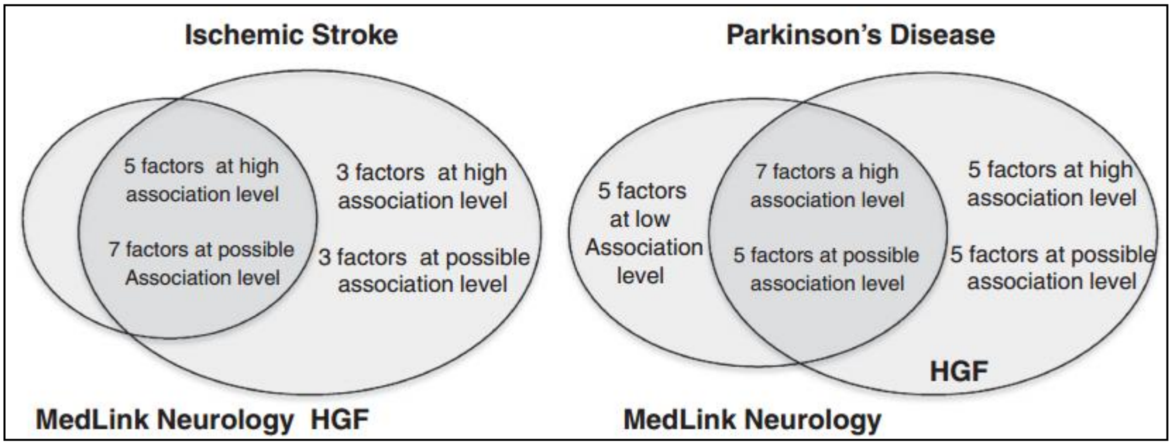

Ischemic Stroke
5 factors at high association level
7 factors at possible Association level
3 factors at high association level
3 factors at possible association level
MedLink Neurology HGF
Parkinson’s Disease
5 factors at low Association level
7 factors a high association level
5 factors at possible association level
5 factors at high association level
5 factors at possible association level
HGF
MedLink Neurology

Figure 3. Number of factors identified by MedLink Neurology and by HGF for Ischemic Stroke (IS) and Parkinson’s disease (PD)
Association levels for IS measured by HGF are high (0.3<cosine score) and possible (0.1 < cosine score < 0.3); associated levels for PD measured by HGF are high (0.2 < cosine score), possible (0.1 < cosine score < 0.2) or low (0.05 < cosine score < 0.1).

In the case of IS, the distribution of known associated factors are more shifted to the right as compared to the factors in PD, hence the separation between the known and unknown factors is more pronounced. In addition to that, associations at both extreme levels (close to +1.0 and −1.0) are likely to be common knowledge; however, the hidden knowledge tends to be captured at similarity scores that are low yet positive. Nonetheless, it is not realistic to compare the precise similarity score values in order to give more importance to one factor versus another factor mainly because there is a systemic bias that is inherent to the biological text data and causes the generic factors to be an underestimate of the true value (data not shown); hence a direct comparison would fail in this case if no additional normalization steps are taken.

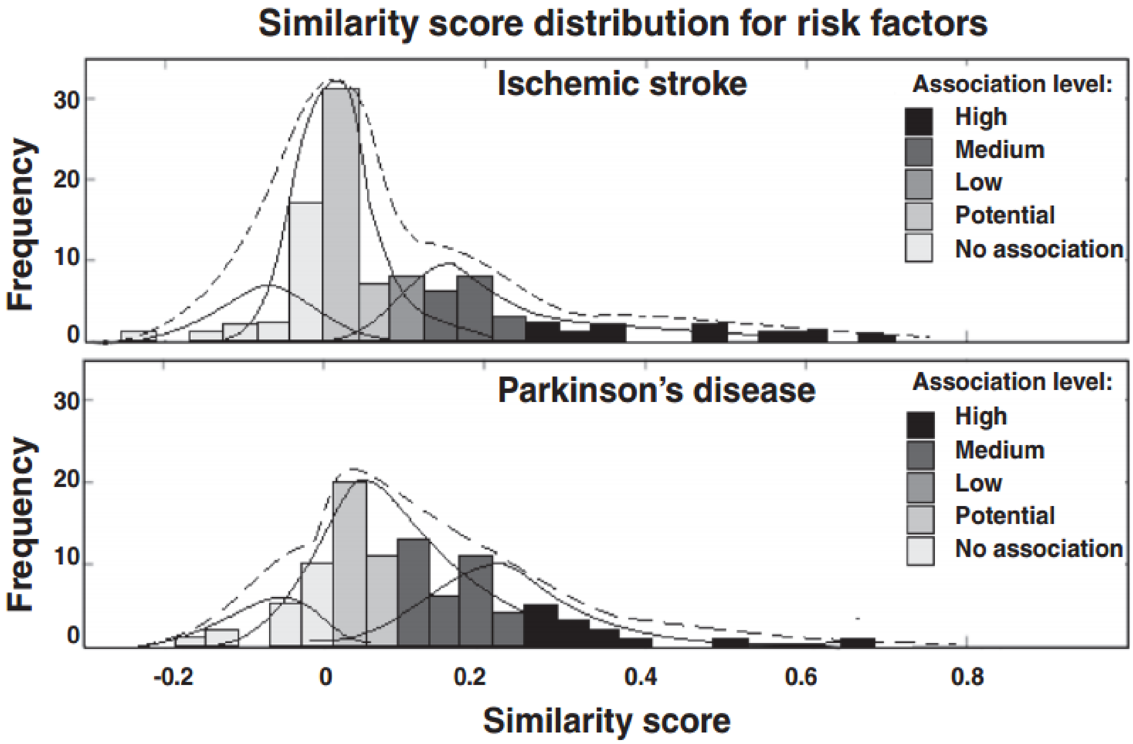


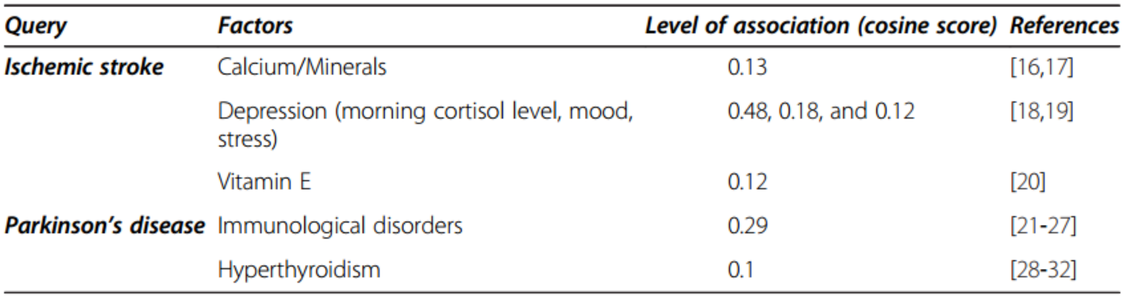
Figure 4. Distribution of cosine similarity score (dashed line) for risk factors associated with Ischemic Stroke (IS) and Parkinson's disease (PD)
The frequency represents the number of factors at each cosine similarity level (-1 to +1). Tri-modal distribution models are represented by solid lines.

Figure 3 summarizes a comparative analysis of MedLink Neurology and HGF for IS and PD. Overall in the case of IS, twelve factors were identified by both systems and six factors were identified by the HGF. In the case of PD, twelve factors were identified by both systems, ten factors were identified by the HGF and five factors were identified by MedLink Neurology. But, these factors had a low association level in HGF. The five factors were either very generic or were not exactly mapped in the set of the 96 factors, hence a direct comparison could not be made. Finally, this small scale comparative analysis corroborates the hypothesis that HGF based on literature can better predict the associated factors for diseases such as IS when the risk and associated factors are well studied and documented. In both cases, MedLink, Neurology, and HGF predicted twelve common associated factors; however, in the case of PD ten new factors were predicted in comparison to six in the case of IS.

Discussion

De novo hypothesis generation can provide an approach on how we design experiments and select the parameters for the study. Interestingly, associations detected by the proposed framework can facilitate extraction of interesting observations and new trends in the field. For instance, it was found that PD could possibly be associated with immunological disorders; this is an intriguing observation. This analysis also facilitates interdisciplinary research and enhances interaction among scientists from sub-specialized fields. A manual review of the literature is performed to find evidences for some of the associations found only by the HGF; Table 2 summarizes these results.

Table 2. A subset of factors identified only by the hypothesis generation framework



Query	***Factors***	***Level of association (cosine score)***	***References***
Ischemic stroke	Calcium/Minerals	0.13	[16,17]
	Depression (morning cortisol level, mood, stress)	0.48, 0.18, and 0.12	[18,19]
	Vitamin E	0.12	[20]
Parkinson's disease	Immunological disorders	0.29	[21-27]
	Hyperthyroidism	0.1	[28-32]

There are three main limitations in the presented framework. We are currently in the process of finding solutions for these limitations. 1) Manual selection of the factors creates bias in the dataset and also limits its scalability property. To alleviate this problem, MeSH hierarchy will be used to generate the set of factors. MeSH comprises more than 25,000 subjects headings organized in an eleven-level hierarchy. 2) In the set of 96 factors, some factors were very generic and some very specific, therefore, there was a systemic bias in the dataset which caused the score for generic factors to be an underestimate of the true values and factors with limited information to be overestimated (data not shown). To partially solve this technical difficulty, an improved method based on local LSA is being developed in our lab. And finally, 3) looking only at the literature from the past twenty years was not sufficient for the HGF. The expansion of the literature is necessary based on the observation that the association between head trauma and PD was significantly lower than expected.

Generating new hypotheses by mining a vast amount of raw unstructured knowledge from the archived reported literature may help in identifying new research trends as well as promoting interdisciplinary studies. In addition, the presented framework is not limited to uncovering disease-disease interactions; any word from MeSH can be used to query the system, and its associated factors can be identified accordingly. Disease-disease interaction networks, interaction networks among chemical compounds, drug-drug interaction networks, or any specific type of interaction network can be constructed using the HGF. The common basis for all these networks is the knowledge embedded in the literature. Application of this framework is broad as its usage is not limited to any specific domain. For instance, uncovering drug-drug interactions is valuable in drug development and drug administration, uncovering disease-disease interaction is important in understanding disease mechanisms and advancing biology through integrated interdisciplinary research. Even though the framework is not limited to diseases, in this study two neurological diseases were used to test the system and demonstrate the power and applicability of the framework.

In addition to addressing the limitations of the framework, work is in progress to expand the HGF framework to allow the user to generate disease networks based on a number of user-defined queries. Such customized networks can be valuable to a wide range of scientists by promoting a faster identification of associated factors and detection of disease-disease interactions. Disease networks based on genetics and proteomics data display many connections between individual disorders and disease categories [14, 15]. Therefore, as expected, each human disorder does not seem to have unique origins or be independent of other disorders. To uncover potential links between two disorders, knowledge extraction from medical literature could be greatly beneficial and reliable.

Chapter 4: Adaptive Robust and Integrative Analysis for finding Novel Associations

Introduction

Summary of the Study

*Adaptive Robust and Integrative Analysis for finding Novel Associations (*ARIANA), is an efficient and scalable Web-based knowledge discovery tool providing a range of services in the general areas of text analytics in biomedicine. ARIANA's core function is semantic-sensitive analysis of text data through ontology mapping. The ontology mapping is critical for maintaining specificity of the application and ensuring the creation of a representative database from an ocean of data for a robust model. In particular, the Medical Subject Headings ontology (http://www.nlm.nih.gov/mesh/) was used to create a dynamic data-driven dictionary specific to the domain of application, as well as a representative database for the system. The semantic relationships among the entities or concepts are captured through a parameter optimized latent semantic analysis (POLSA). The knowledge discovery and the networks of concepts were captured using a relevance model. Finally, an easy to use interface with a flexible visualization module is implemented to interact with the data at various levels of granularity. The input to the ARIANA can be one or multiple keywords selected from the MeSH ontology and the output is a multi-layered network that is collapsible /expandable at levels of granularity chosen by the user. This feature makes it easy to interact with the Web Tool, and provides the user with the flexibility to focus on the relevant parts of the network and hide other details. The output can also be downloaded as a text file for further processing and experimentation. The interface was designed to improve the user experience by catering to specific needs.

The dynamic data-driven (DDD) concepts were introduced starting from the domain specific “dictionary creation” to the “database selection” and to the “threshold selection” for knowledge discovery using relevance model. The key idea is to make the system adaptive to the growing amounts of data and also to the creative needs of a diverse users. Furthermore, the concepts of relevance model and DDD are critical to provide crisp and meaningful information through an intuitive and easy to use Web service application.
In essence, the ARIANA attempts to bridge the gap between creation and dissemination of knowledge. In addition, case studies were performed to evaluate the accuracy of the computed results.

Key Features of the Work

Key features that distinguish this work are both at the system level and at interface level. At the system level, context specificity and hierarchical structure of a medical ontology (in this case MeSH) are integrated with PubMed at a data level to enhance the quality and specificity of information retrieval. In addition, the broad coverage of the literature along with multi gram and context specific dictionary provides additional means to achieve higher standards in literature mining. Finally, because of integration of LSA based technique and multi gram dictionary, direct as well as indirect associations are captured that are based on higher order co-occurrences. At the interface level, features such a collapsibility/expandability are key for a multi level representational advantage where different users from different fields and with different level of expertise can use the tool with ease.

Data Sources and Materials

As in the case for the HGF, ARIANA is also based on the data from PubMed and MeSH ontology. However, the data extraction procedure is optimized. In addition to that, MeSH vocabularies are used to describe the subject of each journal article. That information is also included in the database. The next section describes the exact procedure to construct the data modules.

ARIANA has five main components: (i) Data Stratification, (ii) Ontology Mapping, (iii) Parameter Optimized Latent Semantic Analysis (POLSA), (iv) Relevance Model and (v) Interface and Visualization. **Figure 5** summarizes the overall architecture of the system.

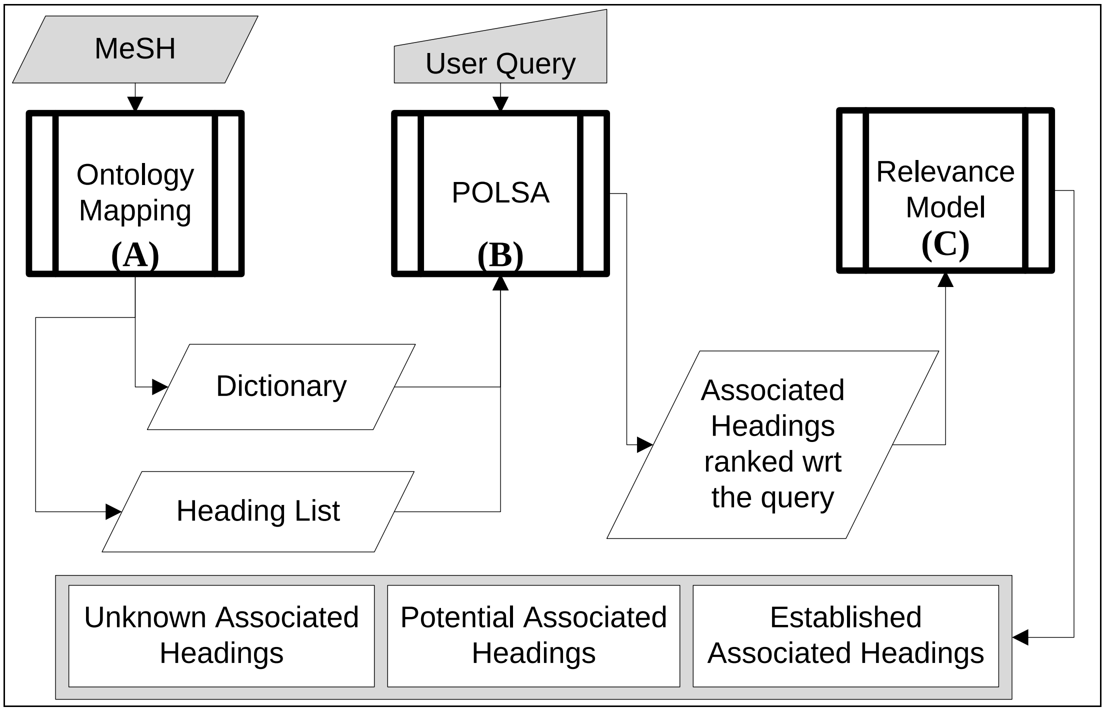

MeSH
User Query
Ontology Mapping
(A)
POLSA
(B)
Relevance Model
(C)
Dictionary
Heading List
Associated Headings ranked wrt the query
Unknown Associated Headings
Potential Associated Headings
Established Associated Headings

Figure 5. Flow diagram of the ARIANA's backbone

Data Stratification

The dataset for the 276 factors is downloaded from PubMed and stored in a MySQL database. The database construction is based on the following design (see **Figure 6**). Using a database to store the data has one key benefit: since the relationship between abstract and Headings is many-to-one, each abstract will only be downloaded once, thus saving significant amount of storage space.

Three tables are used to construct the database for the MeSH-based factors. “Factor” table contains information regarding the 276 MeSH factors, “most recent article (year)” is used to update the entry in the database; “FactorPMID” table contains information need to link the factor to PubMed abstracts using PMIDs (unique identifies of PubMed abstracts); “PMIDContent” table contains information about each abstract. In PMIDContent Mesh Headings are separated by “;”.

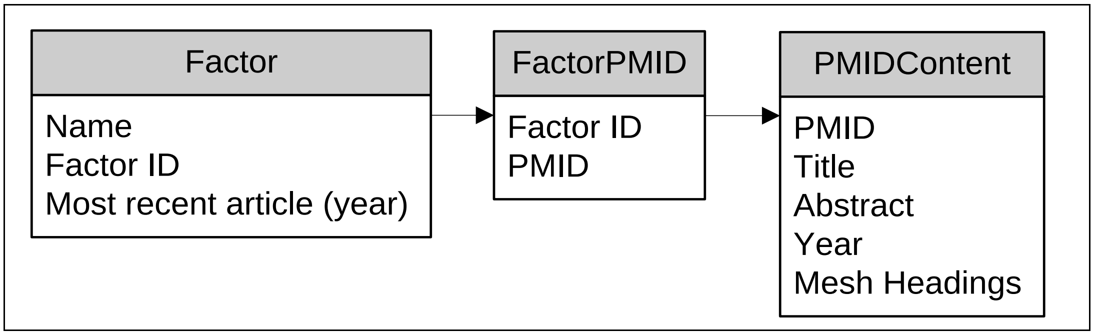
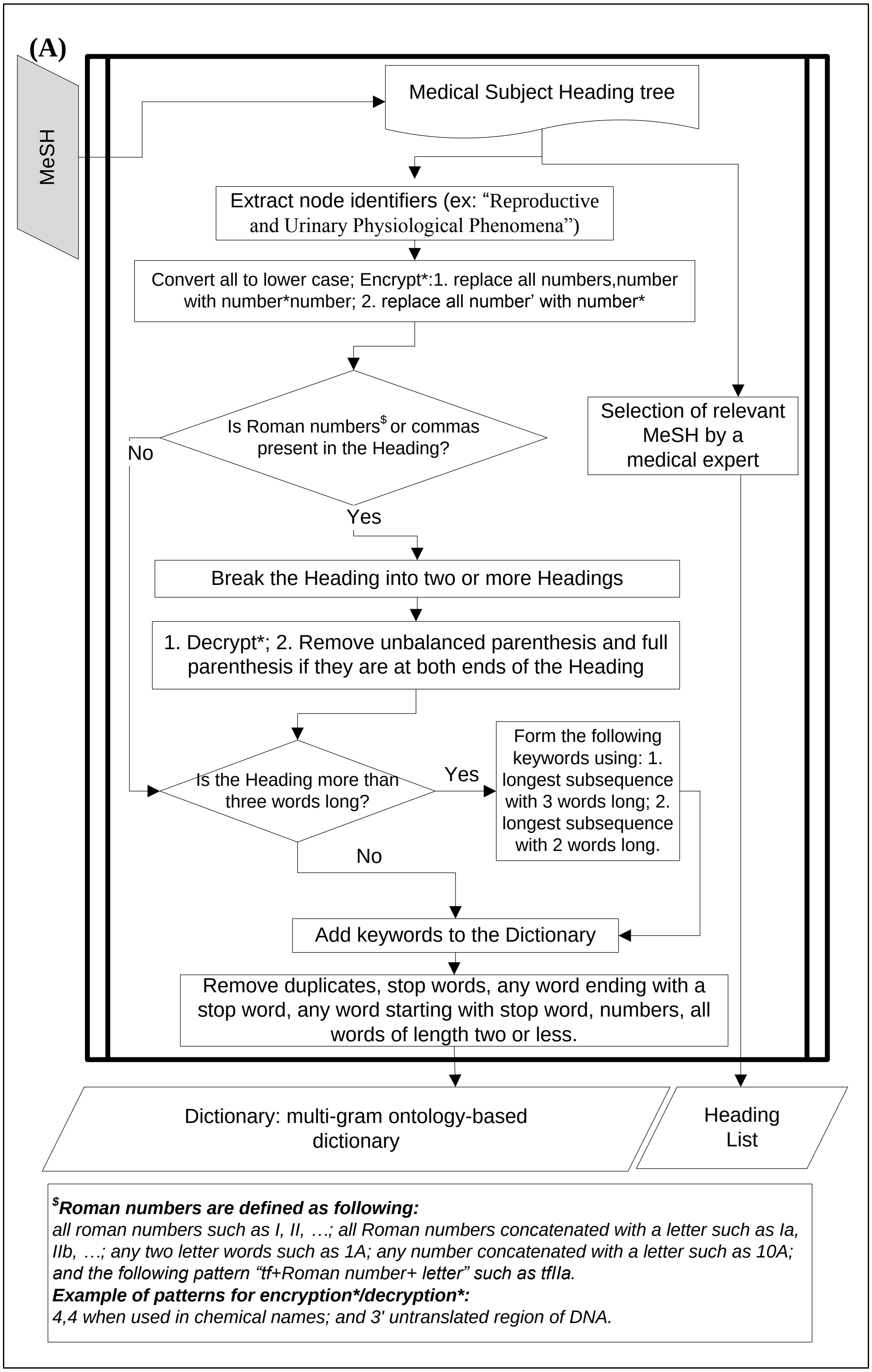


Figure 6. Database system

Ontology Mapping Module

The input to the Ontology mapping process is the MeSH ontology (see **Figure 7**). Using this ontology, two parallel paths are followed to create a multi-gram and context-specific dictionary in addition to a heading list. To create the multi-gram dictionary, first MeSH node identifiers are extracted, and then using a Perl script, the text file containing node identifiers is parsed to construct the mono, bi and tri-gram dictionary. The size of the multi-gram dictionary is 39,107 words. The last filtering step removes duplicates, stop words, words starting with a stop word or number, and all words of length two or less characters. For instance, using the MeSH identifier "Reproductive and Urinary Physiological Phenomena", the followings eight dictionary words are constructed: 1. Reproductive and Urinary, 2. Urinary Physiological Phenomena, 3. Urinary Physiological, 4. Physiological Phenomena, 5. Reproductive, 6. Urinary, 7. Physiological, 8. Phenomena. Two of the eight words are tri-grams, two are bi-grams, and the remaining four words are mono-grams.


(A)
MeSH
Medical Subject Heading tree
Extract node identifiers (ex: "Reproductive and Urinary Physiological Phenomena")
Convert all to lower case; Encrypt*:1. replace all numbers,number with number*number; 2. replace all number' with number*
Is Roman numbers$ or commas present in the Heading?
No
Yes
Selection of relevant MeSH by a medical expert
Break the Heading into two or more Headings
1. Decrypt*; 2. Remove unbalanced parenthesis and full parenthesis if they are at both ends of the Heading
Is the Heading more than three words long?
Yes
Form the following keywords using: 1. longest subsequence with 3 words long; 2. longest subsequence with 2 words long.
No
Add keywords to the Dictionary
Remove duplicates, stop words, any word ending with a stop word, any word starting with stop word, numbers, all words of length two or less.
Dictionary: multi-gram ontology-based dictionary
Heading List

$Roman numbers are defined as following:
all roman numbers such as I, II, …; all Roman numbers concatenated with a letter such as Ia, IIb, …; any two letter words such as 1A; any number concatenated with a letter such as 10A; and the following pattern "tf+Roman number+ letter" such as tfIIa.
Example of patterns for encryption*/decryption*:
4,4 when used in chemical names; and 3' untranslated region of DNA.

Figure 7. Module A - Ontology Mapping

To create the heading list, a careful analysis by a medical expert is performed and a subset of MeSH entries was selected to create the data model. The subset of selected headings, referred to as Heading List, is comprehensive and contains headings from an array of subjects including anatomy, diseases, chemicals, health care, population characteristics and more. A total of 276 headings were selected, and the complete list can be found in the appendices. The main focus when selecting the headings was to include headings that are of general interest and that are relatively specific.

POLSA Module

The Heading List and the Dictionary are the inputs to the POLSA module (see **Figure 8**). Using the Heading List, titles and abstracts of publications are downloaded from the PubMed and stored in a MySQL database on a server (over 8,700,000 abstracts were selected through this process). The number of documents in the corpus will be the same as the number of elements in the heading list, that is 276 headings.

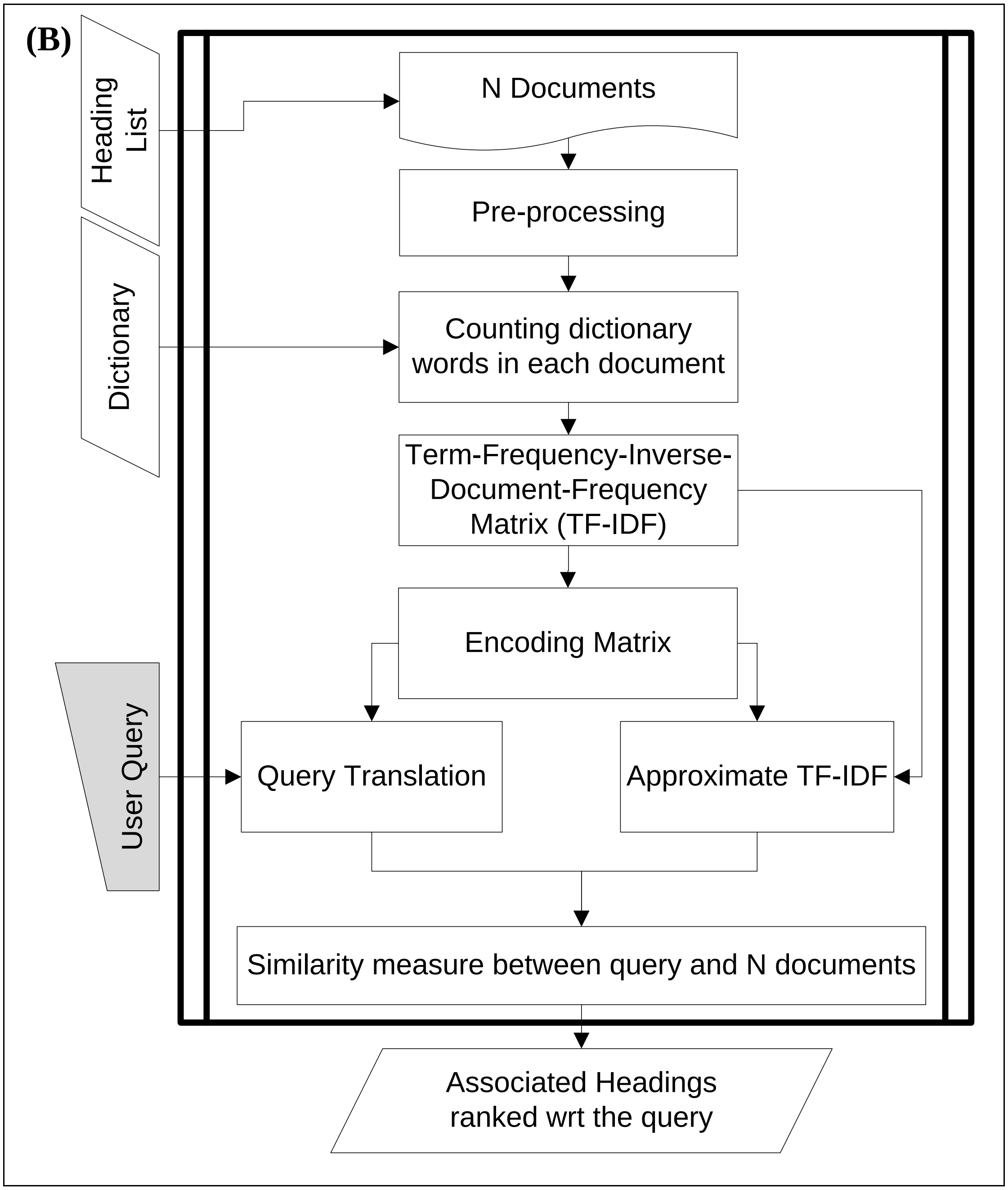

(B)
Heading List
Dictionary
User Query
N Documents
Pre-processing
Counting dictionary words in each document
Term-Frequency-Inverse-Document-Frequency Matrix (TF-IDF)
Encoding Matrix
Query Translation
Approximate TF-IDF
Similarity measure between query and N documents
Associated Headings ranked wrt the query

Figure 8. Module B - Parameter Optimized Latent Semantic Analysis (POLSA)

Each of the 276 documents is parsed to create a term-document-inverse-document frequency (TF-IDF) matrix using the words in the dictionary. The pre-processing step is minimized and does not include stop word removal and stemming as is usually done in text-mining; that is due to the structure of the dictionary as it contains multi-gram words which may have stop words within them. The TF-IDF matrix is then used to create the encoding matrix using singular value decomposition. A user query, which can be any word in the dictionary, will also be an input to this module. Using the encoding matrix, the query is translated into the reduced *eigen* space where direct comparison can be made with the approximate TF-IDF. The approximate TF-IDF is obtained following dimensionality reduction of the encoding matrix. Dimensionality is reduced to cover 90% of the total energy, in this case dimensionality is reduced from 276 to 228 to create the approximate TF-IDF. A comparison between the query and the documents is made by comparing the query in the reduced eigen space and approximate TF-IDF. The cosine similarity measure, which represents the relevancy score, is used to capture the angle between two vectors representing the query and any of the headings. This measure is between +1 and -1. Based on the similarity scores, the headings are then ranked.

Relevance Model

One of the main challenges in knowledge discovery is to present the results in an intuitive and easy to understand form and also allow exploration of results at user defined levels of granularity. The relevance model proposed in this section is a logical extension of disease model originally reported in our previous work [31] also discussed in Chapter 3 of this dissertation. It is an intuitive, simple and easy to use statistical analysis of rank values to compute the strongly related, related, and not related concepts with respect to a user query. Figure 9 illustrates the core concepts of the implemented relevance model. The concepts in this system are a subset of Medical Subject Headings and the user query is constrained by the MeSH ontology. The underlying assumption is if concepts are highly associated then there is a large body of literature available to corroborate existence of their association. On the other hand, if two concepts or biological entities are not well documented then they are only weakly associated. Furthermore, since the distribution of relevance scores is a function of user queries, then the cut-off value to separate highly, possibly and weakly associated entities must be determined dynamically and on the fly. This requires a simplified yet effective model to ensure scalability. To simplify the computation, it was assumed that the distribution of the ranked list can be viewed as a mixture of Gaussian and the partition can be computed using the DDD threshold.

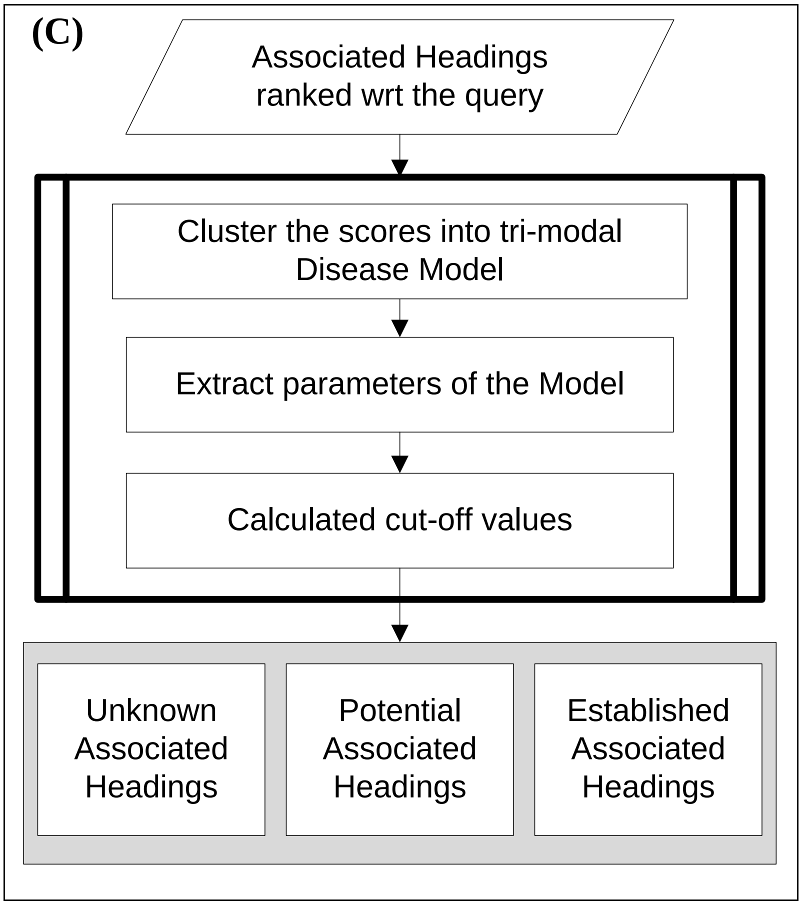

(C)
Associated Headings ranked wrt the query
Cluster the scores into tri-modal Disease Model
Extract parameters of the Model
Calculated cut-off values
Unknown Associated Headings
Potential Associated Headings
Established Associated Headings

Figure 9. Module C - Relevance Model

In particular, the distribution of relevance scores of the Headings for a given query was approximated as a tri-modal Gaussian distribution. The separation of the three distributions allows implementation of the DDD cut-off system. In our previous work [31], also reported in Chapter 3, a curve fitting approach was used to estimate the parameters of the tri-modal distribution and the cut-off values. In this work, fuzzy c-mean clustering approach was implemented to achieve the same goal but in a more robust and scalable manner. This method is much faster and can provide a finely tuned mean to evaluate the results on demand. Furthermore, this dynamic data-driven cut-off value determination can also be integrated in other information retrieval systems.

Even though the relevance model can be highly beneficial for a quality information retrieval system, the ranked list already provides key information about the association between the query and the headings. The top ranked headings are strongly associated with the query and the headings ranked at the bottom do not have significant evidence to support their association to the query. The headings that are between the two extremes are the ones that might or might not be associated with the query as there is some supportive evidence for their association. These weak associations are important in the knowledge discovery process and call for further investigation by domain experts. These weak associations may not always be reported for many cases, and depending on the level of the study, less stringent cut off scores can be considered.

Finally, the computed associations can then be displayed in a network making it easier to analyze than a flat list. If multiple queries are presented to ARIANA, the network for individual queries are computed and displayed. In addition, common associations between queries are also highlighted. Furthermore, the user can hide the queries and their associated Headings that are not of interest and focus on the ones that seem promising and perform additional searches based on the results. The power of this visualization is manifold: it increases the speed for the visual inspection; it facilitates multi query searches; and enhances the quality of the overall search experience.

The fact that the Heading List is from a hierarchical structure, makes it possible to collapse the parts of the network that are of limited interest to the user. Nodes in the network that have common parent nodes (parent-child relationship obtained from the MeSH ontology) can be merged to simplify the network for a better visualization experience. This feature is extremely useful when the number of Heading List increases to a few thousands.

The list of associated Headings that are ranked with respect to a user query is used as input to the Relevance Model (see Figure 9). The top ranked Headings are strongly associated with the query and the Headings ranked at the bottom do not have significant evidence to support their association to the query. The Headings that are between the two extremes are the ones that might or might not be associated with the query as there is some supportive evidence for their association. These weak associations are important in the knowledge discovery process and call for further investigation by domain experts. The similarity measure between the query and all the associated Headings is used to cluster the Headings into three categories. Fuzzy c-means clustering technique is applied to group the associated headings using the MATLAB built-in-function. Using the clustering technique, the scores are first grouped into two clusters. Using the membership values of these two clusters, the following algorithm is used to assign each Heading to one of the three groups in the Relevance Model.

The cosine cut-off values estimated through this process are dynamic and data-driven, hence the cut-offs are subject to change as the dataset expands. The input is the limit that is defined by an expert to separate the known and unknown Headings and place them into the possible Heading group (i.e. the gray zone), a conservative limit threshold of 0.9 was chosen to analyze the results (value of j in Algorithm 1).

```
SET a and b as cluster membership for the headings such that $sum(a) \geq sum(b)$
SET j as the limit to select headings in gray zone
FOR each heading
  IF $a \leq b$            THEN SET c to 1
  END IF
  IF $abs(a-b) \leq j$     THEN SET d to 1
  END IF
END FOR
FOR each heading
  IF $c=1$                 THEN SET group to high_Assoc
  ELSIF $d=1$              THEN SET group to possible_Assoc
  ELSE                     SET group to no_Assoc
  END IF
END FOR
```

Algorithm 1. Grouping of Headings based on fuzzy c-mean clustering

Interface

The interface is an easy to interact web-tool that accepts one or multiple keywords selected from the MeSH ontology and the output is a multi-layered network. Current work is in progress to provide the user with collapsibility/expandability feature as discussed earlier (see also Figure 10). The visualization module is central part of the interface. The primary purpose is to present the associations between the user inputs and the 276 Headings. The proximity matrix generated from the ranked list, represents an adjacency for the associations between the user inputs and the Headings. The user is given with the option of providing a threshold which can be applied to the proximity matrix, to refine the level of association required. In addition, we are also integrating the DDD threshold in the interface.

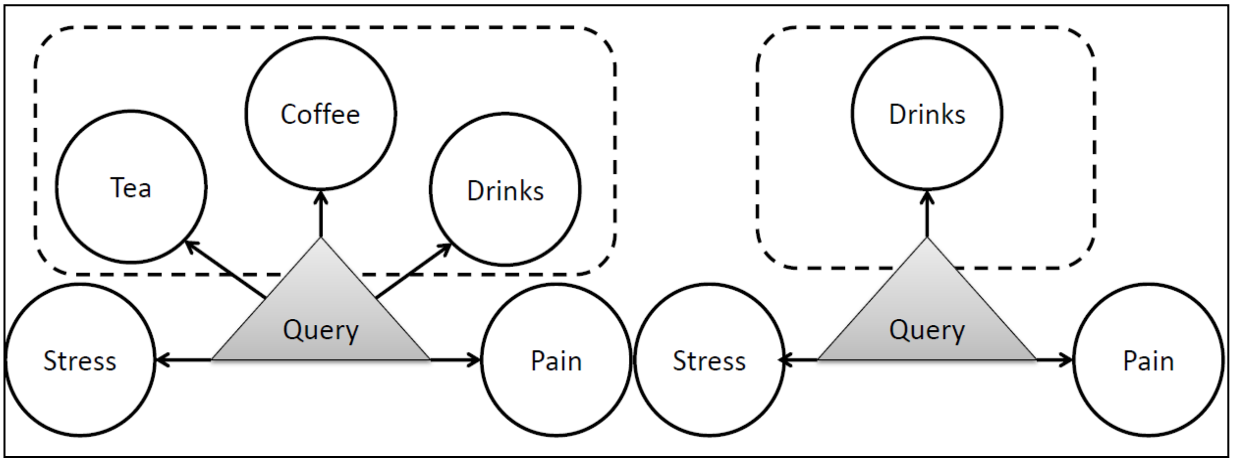


Figure 10. Collapsibility/expandability feature of the network

The user input list and the list of 276 Headings together represent the maximum number of nodes that are present in the graph. Once the threshold is applied, the number of nodes present in the graph is reduced. If no threshold is applied, the maximum number of nodes would be 276 in addition to the 10 user inputs.

To convert the proximity matrix into a graphical network graph, an open source JavaScript library (http://d3js.org/) d3 is used. Force-layout is used to represent the graphs where nodes and links have dynamic properties and can be moved around by the user by stretching or compressing each node using the mouse. When the user relinquishes control, the network goes back to its original shape. The nodes and the links in the force layout have specific properties such as shape, color, text and also physical properties such as friction, charge, gravity and strength. These properties are set appropriately to create a network layout where clutter is minimized. To convert the proximity matrix, a JSON file is written which explicitly identifies the nodes present in the network and the links present among those nodes. A JSON is JavaScript Object Notation which is a light weight data interchange format. A JSON is built on two data structures, one is a collection of key/value (i.e. hash map) pairs and the other is an ordered list of values. For the graph visualization, each of the previously mentioned data structures are used to represent the following components: i) the ordered list representing the list of nodes present in the network; ii) the key/value pairs representing the links in the network. The key corresponds to the source node and the value corresponds to the target node present in the above ordered list. When the user queries the system using more than one query work, the network displays the common Headings (if applicable) separately (see Figure 11 and Algorithm 2 in the appendices).

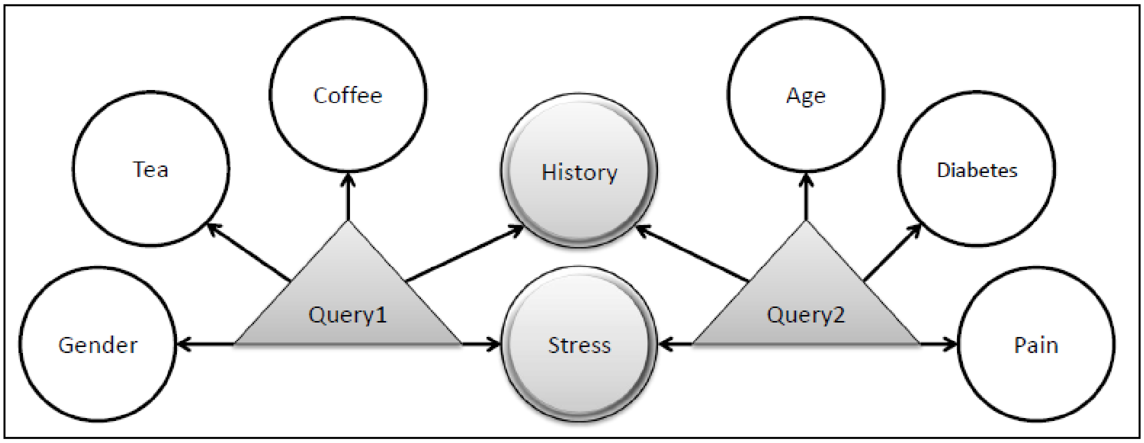


Figure 11. Identification of common nodes for multiple queries.

Evaluation

Evaluation of such analysis is challenging yet very important. The system was evaluated through a comprehensive literature review and then further verified by an expert in the field. The test case was Ischemic Stroke and a board certified physician in Vascular Neurology validated the findings reported here.

Information Retrieval and Knowledge Discovery

A series of queries were used to study and evaluate the potential and scope of the system. The end goal was to detect level of noise and exactitude when running general queries such as common diseases. The results of the analysis are presented in the results section and further discussed in the discussion section of this chapter.

Results

ARIANA’s main objective is to find the semantic sensitive network of associations among concepts and enhance the quality of knowledge discovery and also user experience.

Four different diseases were used as case studies to illustrate the utility and the scope of ARIANA. Diseases used for the study are i) Ischemic stroke (IS), ii) Parkinson's Disease (PD), iii) Lymphoma, and iv) Migraine. Results obtained from the IS simulations are compared with literature and evaluated by a medical expert. Results from PD, Lymphoma and Migraine are displayed and shortly discussed as well. The results presented here are examples and demonstrate how this system can be used to extract information that can be forgotten, and hence bridge the knowledge gap.

Case I: Disease-Heading Network Figure 12 displays the results with "Ischemic Stroke" as query. Table 3 lists the eighteen selected headings with their respective relevance scores. The two cut-off values were obtained by applying the DDD system. The cut-off values place six Headings into the high association and twelve Headings into the possible association group. All the Headings are directly or indirectly associated with stroke. In some cases the indirect association was not obvious and literature search was performed: association between i) stroke and Intermittent Claudication [32], and ii) stroke and Cyanosis [33]. Table 4 lists lower ranking Headings for up to a relevancy score of 0.01. Majority of the Headings listed in table 2 have known association with Ischemic Stroke.

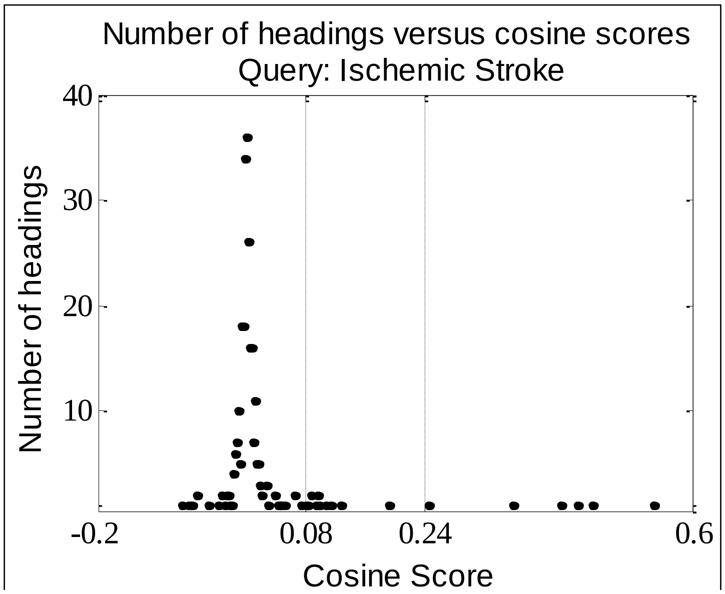


Figure 12. Histogram representation of cosine scores for the 276 Headings (query: Ischemic Stroke)

Table 3. List of Headings ranked and grouped with respect to query "Ischemic Stroke"

Medical Subject Headings	MeSH tree number	Relevancy	
		Score	Level
Cerebrovascular Disorders	C14.907.253	0.550	High
Vascular Diseases	C14.907	0.466	High
Mobility Limitation	C23.888	0.447	High
Myocardial Ischemia	C14.907	0.424	High
Athletes	M01.072	0.359	High
Hemorrhage	C23.550	0.245	High
Mycotoxicosis	C21.613.680	0.191	Possible
Hyperemia	C14.907.474	0.128	Possible
Neuroleptic Malignant Syndrome	C10.720.737	0.116	Possible
Arterial Occlusive Diseases	C14.907.137	0.106	Possible
Pain	C23.888.646	0.099	Possible
Intermittent Claudication	C23.888.531	0.096	Possible
Nervous System Neoplasms	C10.551	0.096	Possible
Personality	F01.752	0.094	Possible
Azotemia	C23.550.145	0.088	Possible
Preconception Injuries	C21.676	0.087	Possible
Cyanosis	C23.888.248	0.082	Possible
Emphysema	C23.550.325	0.081	Possible

Table 4. List of Headings at different relevancy scores with respect to query "Ischemic Stroke"

Medical Subject Headings	Range of relevancy scores
Thyroid Diseases; Metabolic Syndrome X; Hypovolemia; Defense Mechanisms; Neurotoxicity Syndromes; Age Groups; Autoimmune Diseases of the Nervous System	0.08 to 0.041
Neoplasms; Minority Groups; Socioeconomic Factors; Alcohol-Related Disorders; Tumor Virus Infections; Peripheral Vascular Diseases; Hepatitis A; Intestinal Diseases, Parasitic	0.04 to 0.021
Dermatitis, Occupational; Physical Fitness; Neurocutaneous Syndromes; Socialization; Carbon Tetrachloride Poisoning; Mycoses; Muscular Diseases; Immunocompetence; Trauma, Nervous System; Movement Disorders; Bone Diseases, Endocrine; Heart Murmurs; Skin Temperature; Metabolism, Inborn Errors; Quality of Life; Arbovirus Infections; Child, Abandoned; Rheumatic Diseases; Arthritis, Rheumatoid	0.02 to 0.01

Case II: Disease-Disease Network: A number of diseases were used as query to find the network of associations among Headings. The summary of results is presented in the Figure 13 and details are available in the appendices. The majority of the identified Headings are known to be associated with the query. In some cases the association is not strong or it is an indirect association, through other Headings. Only in a few instances the Headings do not have any known association with the query.

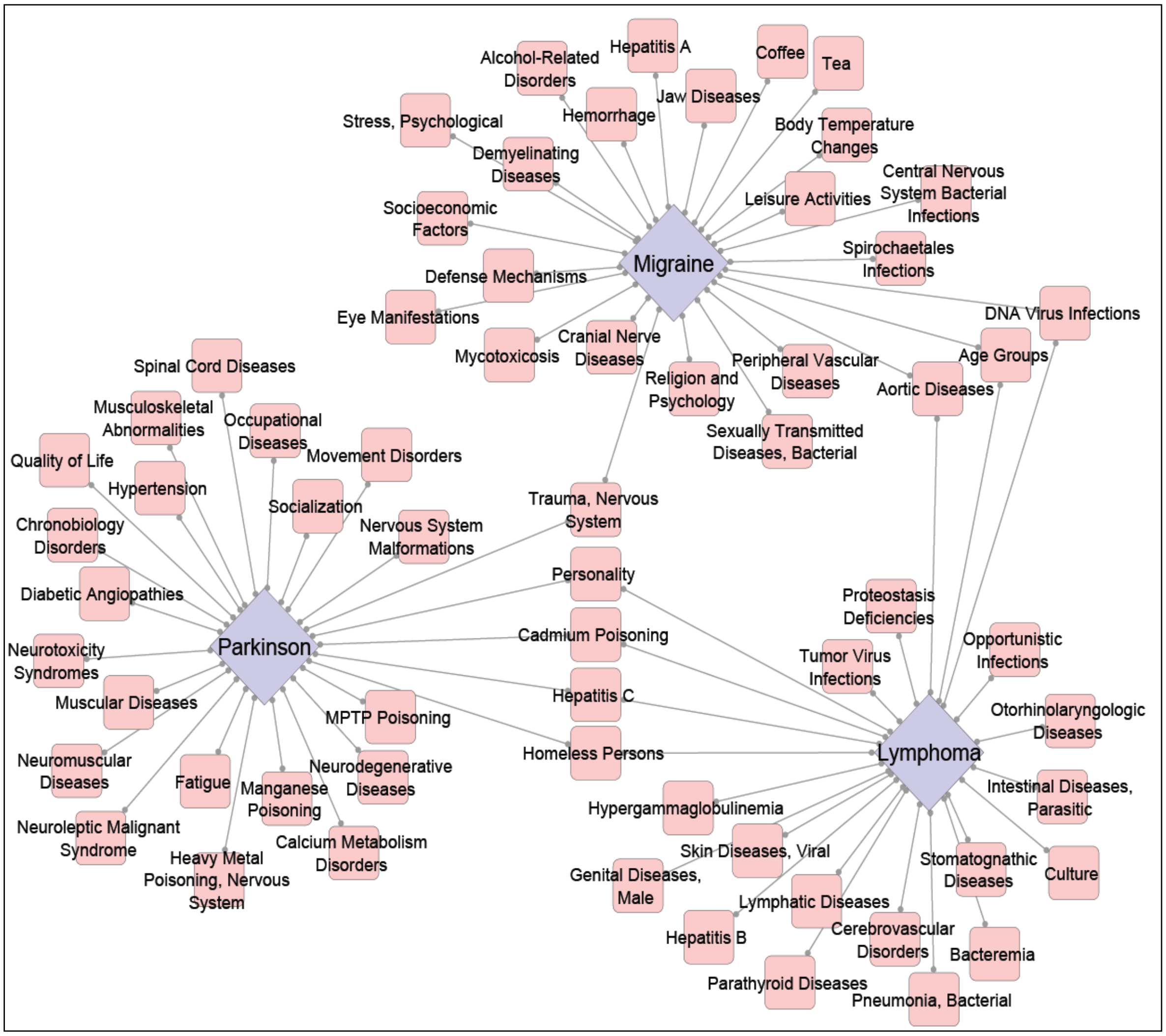


Figure 13. Information retrieval and knowledge discovery using three queries

Parkinson's Disease (PD): The list of ranked Headings for PD highlights the fact that this is a neurological disorder, or more specifically a neurodegenerative disease, affecting movement and muscle functions (see Figure 13). The identified elements also highlight that this disease is likely associated with environmental factors (manganese poisoning, heavy metal poisoning, cadmium poisoning, MPTP poisoning). Case for Migraine: The top three ranked Headings to migraine are: coffee, tea and sexually transmitted diseases (STD) with score of 0.689, 0.592 and 0.286 respectively. The first two associations are expected; however, the association between migraine and STD is less predictable. In a recent investigation [34] 200 HIV/AIDS patients were studied and among them 53.5% reported headache symptoms and 44% were diagnosed with migraine. In addition to that, authors also found a strong correlation between the severity of the HIV disease and the strength and frequency of the migraine attacks. Interestingly, this specific article [34] is not in the current data model, nonetheless the association is detected; hence this is a clear example of knowledge discovery. Case for Lymphoma: Lymphoma, cancer of lymphatic system, begins in the cells of the immune system. Generally, lymphoma seems to be highly associated with different types of infections, that is predictable since patients with a weakened immune system have a higher chance of this cancer. One interesting observation is the association of lymphoma and PD with cadmium poisoning. In fact, the risk of developing childhood acute lymphoblastic leukemia were increased with exposure to cadmium in the drinking water [35]. Some associations, such as cadmium and PD or Lymphoma, are known but can be considered forgotten or buried in the ocean of publications as they are not usually referenced in medical protocols and textbooks. Hence this tool has great potential for data reuse. ARIANA can extract existing associations and improve the quality of information retrieval.

Discussion

ARIANA is a web tool targeting a large scientific community: medical researchers, epidemiologists, biomedical scientific groups as well as junior researchers with focused interests. The tool can be used as a guide to broaden one's horizon by identifying seemingly unrelated entities to the user’s query word. ARIANA computes the networks of semantically meaningful associations from over 8,000,000 documents and provides the relations between query word(s) and the 276 Headings. The guiding principle was to make the design efficient, modular and scalable. The framework can be expanded to incorporate a much larger set of Headings from the MeSH Ontology. In addition, a dynamic data-driven system is implemented to group the ranked Headings into three groups for every query. The DDD system can be applied in other systems to improve the quality of information retrieval. Furthermore, as a consequence of incorporating a context specific multi-gram dictionary, the sparsity of the data model is lower and the size of the dictionary is significantly smaller than if all combination of English words were taken into consideration.

First, ARIANA provides a systematic way of data stratification based on domain knowledge and application constraints. Second, it uses ontology mapping to create a dynamic data driven dictionary, which in turn produces a better model and also helps in finding crisp association of concepts. Third, it computes the network of associations based on higher order co-occurrence analysis and introduction of relevance model to present the results into an easy to use and understandable manner. Finally, a fully integrated tool allows users to interact with the database and computed association at various labels of granularity through an easy to use interface and visualization of expandable/collapsible multi-layer network. In addition, since MeSH provides a hierarchical structure, ARIANA can be expanded to include a very large number of Headings.

Finally, biological entities are diverse in nature and can include different levels of granularity; for instance, these entities can include risk factors, drugs, side effects of treatment and diseases. Furthermore, network of associations between these entities can be useful for management of patients with different conditions. For instance, orthostatic hypotension could be triggered by a number of medications such as agents used in the treatment of hypertension, myocardial ischemia, psychosis and schizophrenia, depression, Alzheimer and Parkinson disease, as well as, a vaccine approved for the prevention of cervical cancer [36]. Therefore, the management of patients with such conditions requires a careful understanding and evaluation of their health record. The scientific knowledge used in the clinical setting will provide new observations (such as publication of new case reports) and can lead to innovative experiments for corroboration of novel instances; thus completing the knowledge discovery cycle.

Chapter 5: Integration of Tools – ARIANA, OMIM and MESH

Introduction

Summary of the Study

In this Chapter, ARIANA is fine tuned and expanded to incorporate 2,545 Headings from the MeSH ontology. The Heading selection process is automated through a fine grain filtering procedure. The modular design and scalability feature of the ARIANA tool allows incorporation of a much larger set of Headings without any technical issue. Following the fine-tuning of the system, a tool-level integration process incorporates a set of genes in the dictionary and an OMIM derived gene-disease association in the network of semantically related entities. OMIM is a flat list of gene disease associations with detailed molecular level information and full references. ARIANA, on the other hand, is a tool that integrates MeSH ontology and PubMed database and can therefore use hierarchical structure of MeSH to organize the disease categories. Mapping of disease in PubMed and disease in MeSH is not a direct mapping, as many disease names are unique to each database. In OMIM, each disease is unique and has no relationship with other diseases. In MeSH, each disease can have a sub class or super class. Equivalence class concept is used to map disease from OMIM and MeSH. The mapping being one to many, a MeSH disease can be mapped to one or more diseases in OMIM (or zero in some cases), where the OMIM database provides additional information regarding gene-disease association. The main functionality of integrated tool will still be finding association among biological entities with the added genetic information from OMIM. In addition, genes that are in the OMIM database are added to the dictionary to target a different layer of semantics as well. This layer of semantics can be complex to analyze and interpret due to the nature of the application.

Finally, we show that fine tuning ARIANA can be valuable for predicting lethal disease-drug association with no citation in PubMed for the years in the database. Integration of genetic information to the ARIANA tool is still experimental and further investigation by field experts is needed.

Key Features of the Work

Key features that distinguish this work from the state of the art works are at the level of tool integration and system enhancement. At the tool integration level, the key contribution is design of an equivalence class for mapping of hierarchical and non-hierarchical databases. At the system enhancement level, the central contribution is fine-tuning of the POLSA system and refining the dictionary, based on analysis of the encoding matrix. Integration of genetic information in the ARIANA tool through addition of gene symbols in the dictionary was also one of the added elements of this work.

Data Source and Materials

As in the case for the ARIANA with the 276 Headings, the expanded version of ARIANA also used PubMed to create the database. In addition to PubMed, MeSH ontology is utilized to make the database as precise as possible. The Heading selection is optimized to provide a better quality of service to the users. In the next section, the exact procedure to construct the data modules is described.

Data Stratification

The fine-tuned and expanded version of ARIANA has seven main components, namely, Database Creation, Ontology Mapping, Parameter Optimized Latent Semantic Analysis (POLSA), Relevance Model, Disease Mapping through equivalence class, Extraction of Gene-Disease association and the Interface. Disease Mapping and Extraction of Gene-Disease information are the two new components that expand the project to a new level. In addition, the Ontology Mapping is fine-tuned through an iterative process. Figure 14 outlines the conceptual model of the new and improved Web-tool.

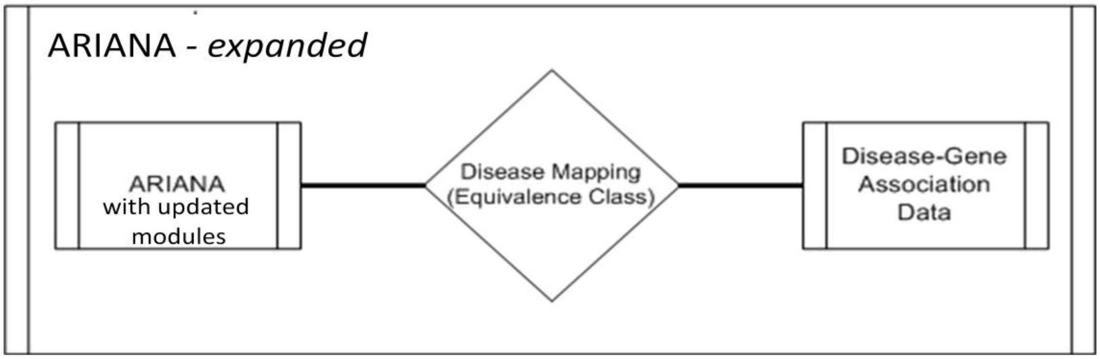

ARIANA - expanded
ARIANA with updated modules
Disease Mapping (Equivalence Class)
Disease-Gene Association Data

Figure 14. Conceptual model of the ARIANA after fine tuning the parameters

Ontology Mapping Components

Based on the domain knowledge and the choice of the concepts and entities, a very large and broad database is created using a fully automated process. In this expanded version, the Heading selection process is also automated, making the system scalable and robust.

One of the key functions of Ontology Mapping is to use the information in the encoding matrix to filter out terms that provide no new information. This refinement process creates an encoding matrix that is not sparse. The multi-gram dictionary is also optimized accordingly. In addition, gene names in the OMIM database are added to the dictionary as well. Figure 15 shows the steps taken in this module.

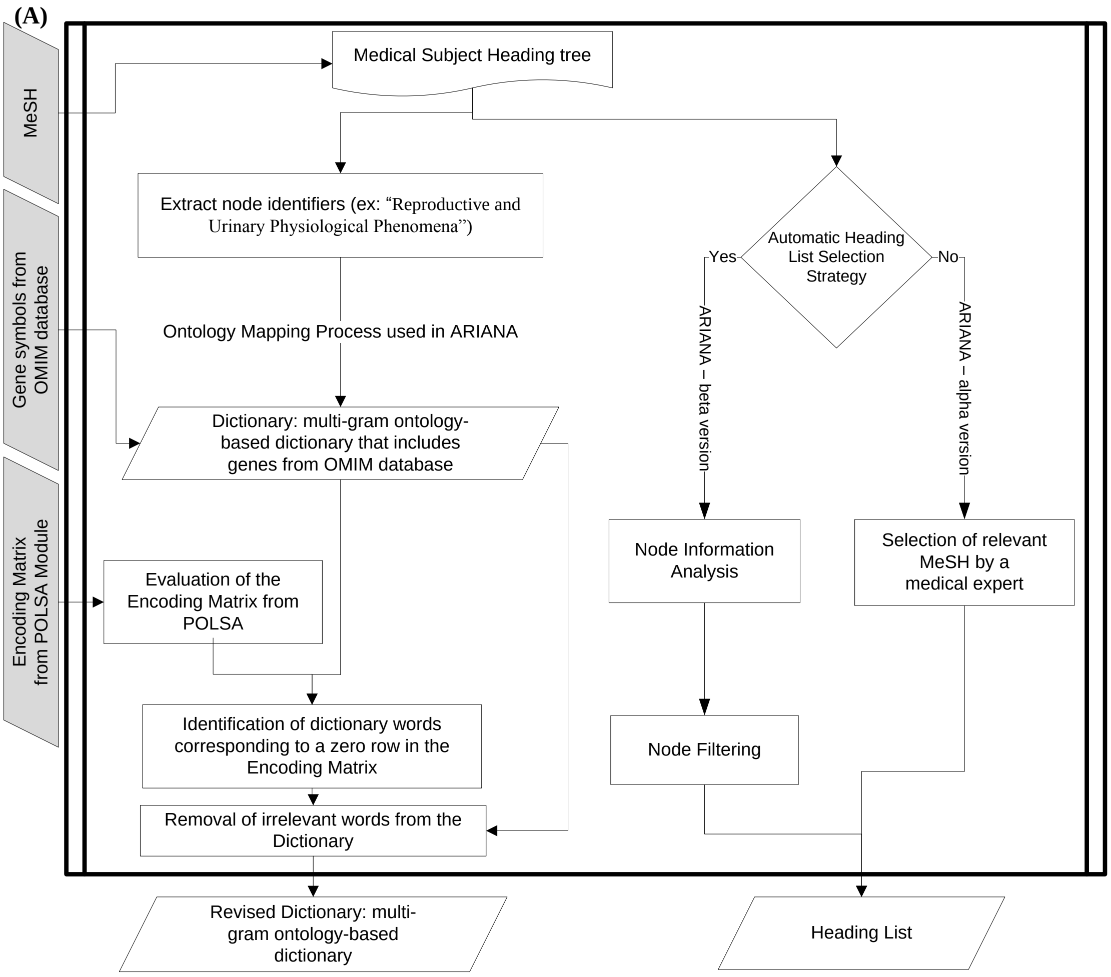
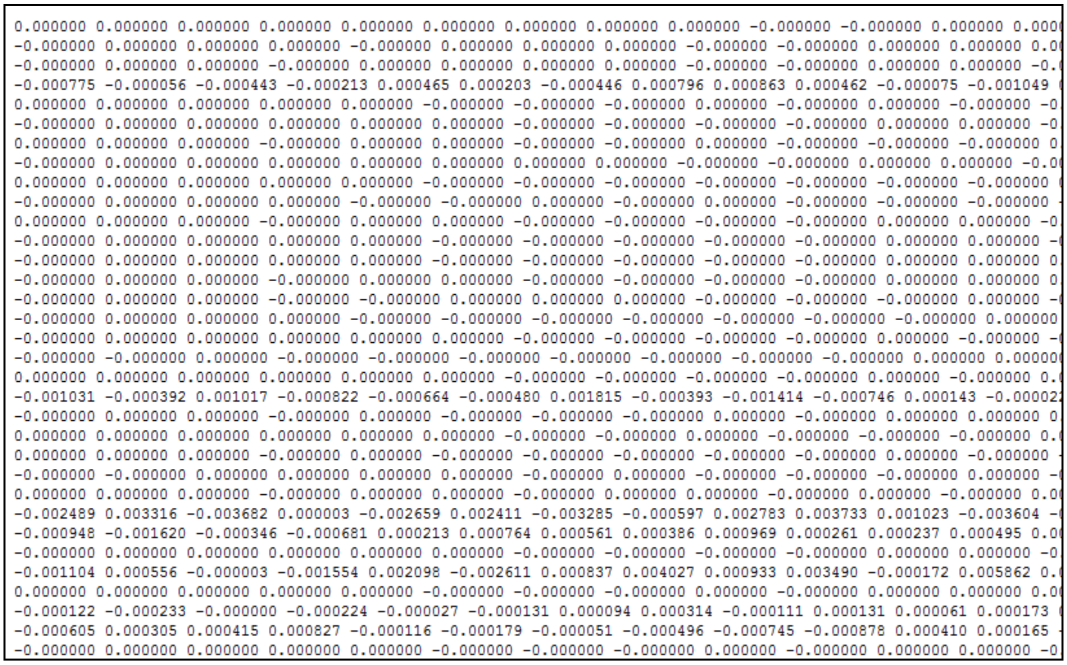

(A)
MeSH
Gene symbols from OMIM database
Encoding Matrix from POLSA Module
Medical Subject Heading tree
Extract node identifiers (ex: "Reproductive and Urinary Physiological Phenomena")
Ontology Mapping Process used in ARIANA
Dictionary: multi-gram ontology-based dictionary that includes genes from OMIM database
Evaluation of the Encoding Matrix from POLSA
Identification of dictionary words corresponding to a zero row in the Encoding Matrix
Removal of irrelevant words from the Dictionary
Revised Dictionary: multi-gram ontology-based dictionary
Automatic Heading List Selection Strategy
Yes
No
ARIANA – beta version
ARIANA – alpha version
Node Information Analysis
Node Filtering
Selection of relevant MeSH by a medical expert
Heading List

Figure 15. Fine tuning of Ontology Mapping in ARIANA

The two key paths in this module are to create a revised multi-gram dictionary that is concise and domain specific, and to create a Heading list to be used in the data extraction process. The Heading list was initially selected by an expert in the field from the MeSH ontology. In the refined version of ARIANA node information is used to automatically extract the best set of Headings. In the next subsection the details of the node filtering procedure is described in depth. To revise the multi-gram dictionary, the encoding matrix obtained from the POLSA module is analyzed. Figure 16 and Figure 17 show a snapshot of a section the encoding matrix before and after the fine-tuning process.

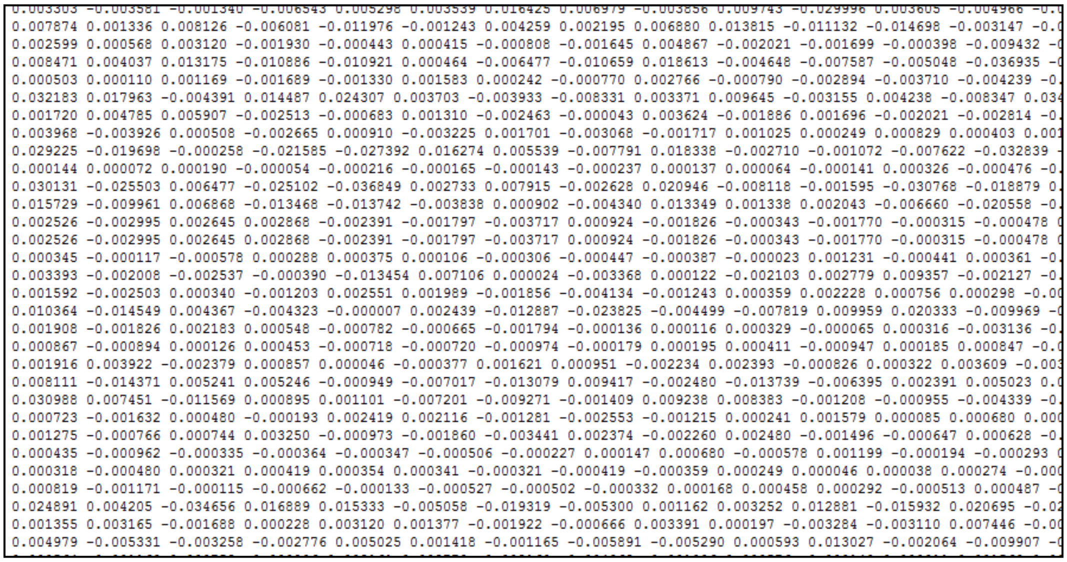
Fine-Tuning of the Multi-Gram Dictionary

Following the analysis of the encoding matrix, it was observed that many of the entries were zero even when full SVD was used. Further investigation showed that removing some of the irrelevant dictionary words reduced the rows of zeros; hence the problem was due to the fact that some dictionary words that were generated through an automated process (see Chapter 4) were irrelevant or did not provide sufficient information to the model. Removing of these dictionary words creates an encoding matrix that is not sparse. Removing the dictionary words that correspond to a row of zeros in the encoding matrix is therefore critical for a robust system. The final size of the multi-gram dictionary is 17,074 words, these include words that are mono, bi, and tri-grams.



```
0.000000 0.000000 0.000000 0.000000 0.000000 0.000000 0.000000 0.000000 0.000000 -0.000000 -0.000000 0.000000 0.000
-0.000000 0.000000 0.000000 0.000000 -0.000000 0.000000 0.000000 0.000000 -0.000000 -0.000000 0.000000 0.000000 0.00
-0.000000 0.000000 0.000000 -0.000000 0.000000 0.000000 0.000000 0.000000 -0.000000 -0.000000 0.000000 0.000000 -0.0
-0.000775 -0.000056 -0.000443 -0.000213 0.000465 0.000203 -0.000446 0.000796 0.000863 0.000462 -0.000075 -0.001049
0.000000 0.000000 0.000000 0.000000 0.000000 -0.000000 -0.000000 -0.000000 0.000000 -0.000000 0.000000 -0.000000 -0.
-0.000000 0.000000 0.000000 0.000000 0.000000 0.000000 -0.000000 -0.000000 -0.000000 -0.000000 0.000000 0.000000 -0.
0.000000 0.000000 0.000000 -0.000000 0.000000 0.000000 -0.000000 -0.000000 0.000000 -0.000000 -0.000000 -0.000000 0.
-0.000000 0.000000 0.000000 0.000000 0.000000 0.000000 0.000000 0.000000 -0.000000 -0.000000 0.000000 0.000000 -0.00
0.000000 0.000000 0.000000 0.000000 0.000000 -0.000000 -0.000000 -0.000000 -0.000000 -0.000000 -0.000000 -0.000000
-0.000000 0.000000 0.000000 0.000000 -0.000000 -0.000000 0.000000 -0.000000 0.000000 -0.000000 -0.000000 -0.000000
0.000000 0.000000 0.000000 -0.000000 0.000000 0.000000 -0.000000 -0.000000 -0.000000 -0.000000 0.000000 0.000000 -0.
-0.000000 0.000000 0.000000 0.000000 0.000000 -0.000000 -0.000000 -0.000000 -0.000000 -0.000000 0.000000 0.000000 -
-0.000000 0.000000 0.000000 0.000000 0.000000 -0.000000 -0.000000 -0.000000 -0.000000 -0.000000 0.000000 0.000000 0.
-0.000000 0.000000 0.000000 -0.000000 0.000000 0.000000 -0.000000 -0.000000 -0.000000 -0.000000 0.000000 0.000000 0.
-0.000000 0.000000 0.000000 -0.000000 -0.000000 0.000000 0.000000 0.000000 -0.000000 -0.000000 -0.000000 0.000000 -
-0.000000 0.000000 0.000000 0.000000 -0.000000 -0.000000 -0.000000 -0.000000 -0.000000 -0.000000 -0.000000 0.000000
-0.000000 0.000000 0.000000 0.000000 0.000000 0.000000 -0.000000 -0.000000 -0.000000 -0.000000 0.000000 -0.000000 -
-0.000000 -0.000000 0.000000 -0.000000 -0.000000 -0.000000 -0.000000 -0.000000 -0.000000 -0.000000 0.000000 0.000000
0.000000 0.000000 0.000000 0.000000 0.000000 0.000000 -0.000000 -0.000000 -0.000000 -0.000000 0.000000 -0.000000 0.0
-0.001031 -0.000392 0.001017 -0.000822 -0.000664 -0.000480 0.001815 -0.000393 -0.001414 -0.000746 0.000143 -0.00002
-0.000000 0.000000 0.000000 -0.000000 0.000000 -0.000000 -0.000000 -0.000000 0.000000 -0.000000 0.000000 0.000000 0.
0.000000 0.000000 0.000000 0.000000 0.000000 0.000000 -0.000000 -0.000000 0.000000 -0.000000 -0.000000 -0.000000 0.0
0.000000 0.000000 0.000000 -0.000000 0.000000 -0.000000 -0.000000 -0.000000 -0.000000 -0.000000 0.000000 -0.000000 -
-0.000000 -0.000000 0.000000 0.000000 0.000000 0.000000 -0.000000 0.000000 -0.000000 -0.000000 -0.000000 0.000000 -
0.000000 0.000000 0.000000 -0.000000 0.000000 0.000000 -0.000000 0.000000 0.000000 -0.000000 0.000000 -0.000000 0.00
-0.002489 0.003316 -0.003682 0.000003 -0.002659 0.002411 -0.003285 -0.000597 0.002783 0.003733 0.001023 -0.003604 -
-0.000948 -0.001620 -0.000346 -0.000681 0.000213 0.000764 0.000561 0.000386 0.000969 0.000261 0.000237 0.000495 0.00
-0.000000 0.000000 0.000000 0.000000 0.000000 0.000000 -0.000000 -0.000000 -0.000000 -0.000000 0.000000 0.000000 -0.
-0.001104 0.000556 -0.000003 -0.001554 0.002098 -0.002611 0.000837 0.004027 0.000933 0.003490 -0.000172 0.005862 0.0
0.000000 0.000000 0.000000 0.000000 0.000000 -0.000000 -0.000000 -0.000000 0.000000 -0.000000 0.000000 0.000000 0.00
-0.000122 -0.000233 -0.000000 -0.000224 -0.000027 -0.000131 0.000094 0.000314 -0.000111 0.000131 0.000061 0.000173
-0.000605 0.000305 0.000415 0.000827 -0.000116 -0.000179 -0.000051 -0.000496 -0.000745 -0.000878 0.000410 0.000165
-0.000000 0.000000 0.000000 0.000000 0.000000 -0.000000 -0.000000 -0.000000 0.000000 -0.000000 0.000000 0.000000 -0.
```

Figure 16. Section of the encoding matrix following singular value decomposition



```
0.003303 -0.003581 -0.001340 -0.006543 0.005298 0.003539 0.016425 0.006979 -0.003856 0.009743 -0.029996 0.003605 -0.004966 -0.
0.007874 0.001336 0.008126 -0.006081 -0.011976 -0.001243 0.004259 0.002195 0.006880 0.013815 -0.011132 -0.014698 -0.003147 -0.0
0.002599 0.000568 0.003120 -0.001930 -0.000443 0.000415 -0.000808 -0.001645 0.004867 -0.002021 -0.001699 -0.000398 -0.009432 -
0.008471 0.004037 0.013175 -0.010886 -0.010921 0.000464 -0.006477 -0.010659 0.018613 -0.004648 -0.007587 -0.005048 -0.036935 -
0.000503 0.000110 0.001169 -0.001689 -0.001330 0.001583 0.000242 -0.000770 0.002766 -0.000790 -0.002894 -0.003710 -0.004239 -0.
0.032183 0.017963 -0.004391 0.014487 0.024307 0.003703 -0.003933 -0.008331 0.003371 0.009645 -0.003155 0.004238 -0.008347 0.034
0.001720 0.004785 0.005907 -0.002513 -0.000683 0.001310 -0.002463 -0.000043 0.003624 -0.001886 0.001696 -0.002021 -0.002814 -0.
0.003968 -0.003926 0.000508 -0.002665 0.000910 -0.003225 0.001701 -0.003068 -0.001717 0.001025 0.000249 0.000829 0.000403 0.001
0.029225 -0.019698 -0.000258 -0.021585 -0.027392 0.016274 0.005539 -0.007791 0.018338 -0.002710 -0.001072 -0.007622 -0.032839 -
0.000144 0.000072 0.000190 -0.000054 -0.000216 -0.000165 -0.000143 -0.000237 0.000137 0.000064 -0.000141 0.000326 -0.000476 -0.
0.030131 -0.025503 0.006477 -0.025102 -0.036849 0.002733 0.007915 -0.002628 0.020946 -0.008118 -0.001595 -0.030768 -0.018879 0.
0.015729 -0.009961 0.006868 -0.013468 -0.013742 -0.003838 0.000902 -0.004340 0.013349 0.001338 0.002043 -0.006660 -0.020558 -0.
0.002526 -0.002995 0.002645 0.002868 -0.002391 -0.001797 -0.003717 0.000924 -0.001826 -0.000343 -0.001770 -0.000315 -0.000478 0
0.002526 -0.002995 0.002645 0.002868 -0.002391 -0.001797 -0.003717 0.000924 -0.001826 -0.000343 -0.001770 -0.000315 -0.000478 0
0.000345 -0.000117 -0.000578 0.000288 0.000375 0.000106 -0.000306 -0.000447 -0.000387 -0.000023 0.001231 -0.000441 0.000361 -0.
0.003393 -0.002008 -0.002537 -0.000390 -0.013454 0.007106 0.000024 -0.003368 0.000122 -0.002103 0.002779 0.009357 -0.002127 -0.
0.001592 -0.002503 0.000340 -0.001203 0.002551 0.001989 -0.001856 -0.004134 -0.001243 0.000359 0.002228 0.000756 0.000298 -0.00
0.010364 -0.014549 0.004367 -0.004323 -0.000007 0.002439 -0.012887 -0.023825 -0.004499 -0.007819 0.009959 0.020333 -0.009969 -
0.001908 -0.001826 0.002183 0.000548 -0.000782 -0.000665 -0.001794 -0.000136 0.000116 0.000329 -0.000065 0.000316 -0.003136 -0.
0.000867 -0.000894 0.000126 0.000453 -0.000718 -0.000720 -0.000974 -0.000179 0.000195 0.000411 -0.000947 0.000185 0.000847 -0.0
0.001916 0.003922 -0.002379 0.000857 0.000046 -0.000377 0.001621 0.000951 -0.002234 0.002393 -0.000826 0.000322 0.003609 -0.003
0.008111 -0.014371 0.005241 0.005246 -0.000949 -0.007017 -0.013079 0.009417 -0.002480 -0.013739 -0.006395 0.002391 0.005023 0.0
0.030988 0.007451 -0.011569 0.000895 0.001101 -0.007201 -0.009271 -0.001409 0.009238 0.008383 -0.001208 -0.000955 -0.004339 -0.
0.000723 -0.001632 0.000480 -0.000193 0.002419 0.002116 -0.001281 -0.002553 -0.001215 0.000241 0.001579 0.000085 0.000680 0.000
0.001275 -0.000766 0.000744 0.003250 -0.000973 -0.001860 -0.003441 0.002374 -0.002260 0.002480 -0.001496 -0.000647 0.000628 -0.
0.000435 -0.000962 -0.000335 -0.000364 -0.000347 -0.000506 -0.000227 0.000147 0.000680 -0.000578 0.001199 -0.000194 -0.000293 0
0.000318 -0.000480 0.000321 0.000419 0.000354 0.000341 -0.000321 -0.000419 -0.000359 0.000249 0.000046 0.000038 0.000274 -0.000
0.000819 -0.001171 -0.000115 -0.000662 -0.000133 -0.000527 -0.000502 -0.000332 0.000168 0.000458 0.000292 -0.000513 0.000487 -
0.024891 0.004205 -0.034656 0.016889 0.015333 -0.005058 -0.019319 -0.005300 0.001162 0.003252 0.012881 -0.015932 0.020695 -0.02
0.001355 0.003165 -0.001688 0.000228 0.003120 0.001377 -0.001922 -0.000666 0.003391 0.000197 -0.003284 -0.003110 0.007446 -0.00
0.004979 -0.005331 -0.003258 -0.002776 0.005025 0.001418 -0.001165 -0.005891 -0.005290 0.000593 0.013027 -0.002064 -0.009907 -
```

Figure 17. Section of the encoding matrix following singular value decomposition after refinement

Heading Selection Process

Automatic Heading selection for ARIANA is achieved through a systematic node filtering process described here. The automatic framework for the selection of a subset of Medical Subject Headings focuses on capturing representative data while creating a balanced dataset from an ocean of unstructured text. Eight sub-categories from the MeSH tree are selected based on the application constraints and domain knowledge: Diseases (C), Chemicals and Drugs (D), Psychiatry and Psychology (F), Phenomena and Processes (G), Anthropology, Education, Sociology and Social Phenomena (I), Technology, Industry, Agriculture (J), Named Groups (M), and Health Care (N). The eight sub-categories are subject to filtering where about 2.5-17% of their descendent nodes are selected to be included in the final Heading List. The Headings are selected in such a way to i) create a balanced representative dataset, ii) remove noise and systemic bias, and iii) remove Headings that are either too generic or too specific. Three features are used in the filtering process: number of abstracts for each Heading, number of descendent node associated with each Heading and ratio (also referred to as fold change) of the number of abstract between child-parent node. These features capture the specificity of the Headings. Finally, 2,846 Headings from a total of 38,618 are selected to populate the database; 61% (1,828 out of 2,846) of the Headings are from the Disease category.

Heading selection rules are progressive rules and are based on the application constraints. The rules for the selection of Headings are adjusted in each sub-category in order to include concepts from a wide range of fields, while keeping a higher number of Headings from the disease class. The disease class includes the MeSH from the C category and the non-disease class contains Headings from all the remaining seven categories.

Furthermore, the inclusion criteria are continuously adjusted to reduce the skewness in the dataset. For instance, some categories like Chemicals and Drugs (D) are very large with over 20,000 sub-headings, while some categories such as Named Groups (M) are very small, with only 190 sub-headings, for this reason the selection criteria is progressively adjusted to reduce the bias in the dataset. A total of 475 out of 20,015 sub-headings were selected from the D category (only 2% coverage), while a total of 13 out of 190 (or 7% coverage) were selected from the M category. The 475 Headings represent less than 50% of the Headings in the non-disease category. To include a higher number of Headings to have a higher coverage from the D category will create a skewed dataset where Chemical and Drugs are over-represented.

The main constraint is to select more than half of the Headings from the disease category. The filtering process is therefore less stringent for the disease category. The three features (number of abstracts for each Heading, number of descendent node that is associated with each Heading and fold change) are used to build three rules to measure the specificity of the Headings and facilitate the selection process. Headings are selected if they satisfy the following three criteria: i) at least one and at most 100 child nodes; ii) at least 1,000 and at most 50,000 abstracts; and iii) at most 10 fold change with respect to their parent node. The empirical thresholds are selected for the selection process with the goal of reducing the systemic bias and noise that is inherent to the biological datasets. In fact, initially the number of child nodes for all the factors in the disease category ranged between 0 - 370 with an average values of 1.7; the average number of child nodes after filtering is increased to 5. Hence, the filtering process removes a large number of leaf nodes and a few generic nodes, where generic nodes are known to be associated with large number of child nodes. Similarly, before filtering, the number of abstracts for each Heading in the disease category ranges between 0 - 563,913 with an average that roses from 5,067 before filtering to 10,309 after filtering. These numbers demonstrate that a large number of specific and a smaller number of generic diseases are removed, thus the systemic bias due to document size is reduced. Lastly, there is also noticeable difference in fold change in this category: before filtering the fold change for each Heading ranges from 1 to 11,674. The average fold change drops from 201 to 3 following the filtering process. A very high fold change can identify Headings that are too specific and could therefore be the cause of systemic bias in the dataset. Finally, a total of 1,828 Headings are selected from the C category, this number represents 17% coverage from the C category and accounts for 64% of the total number of Headings in ARIANA's database.

The Chemical and Drugs category (category D) is one of the largest categories in MeSH with 20,015 headings. The selection criteria for this category are therefore very stringent. One of the main objectives is to select Headings that would represent a maximum of 50% from the non-disease group. A total of 475 headings are selected from the D category, this number represents 47% of the headings from the non-disease group. The number of child nodes for each Heading ranges from 0 - 1,605 with an average value of 2.5. In the filtering process the Headings that have at least one and at most ten child nodes are only considered, this limit removes very specific as well as generic Headings. Furthermore, the number of abstracts for each Heading ranges from 0 to 1,177,960 with an average value of 7,407. These numbers illustrate the range of specificity in the dataset; in fact, a Heading that is associated with over one million abstracts is too generic to be useful. After filtering, the range of abstracts per Heading is significantly reduced (5,000 -

10,000), but the average value is only slightly changed to 7,319. Finally, before filtering the fold change reaches a maximum of 557,279 with an average value of 345; after filtering the fold change is limited to five and the average value is significantly reduced to 2.1.

Category F, also known as Psychiatry and Psychology category has only 1,050 headings. Selection criterion is therefore adjusted to select about 10% of the best representative Headings. The 1,050 Headings have a wide range of specificity, indeed the number of abstracts for each Heading ranges from 1 to 859,564. The filtering process attempts to select the most homogenous Headings to minimize systemic bias and noise. An average value of 11,396 abstracts per Headings is slightly reduced to 11,386; however, the range is significantly reduced (1,000 to 30,000) in this case. Similarly, before filtering the number of child nodes for each heading ranges from 0 to 69 with an average value of 1.04; after filtering the headings that have less than 2 and more than 50 child nodes are removed, bringing the average value to 2.7. Even though the average value is only changed slightly, the Headings that were at both extremes of specificity are removed. Finally, before filtering, the fold change ranges from 1 to 49,761 with an average value of 232, which is significantly reduced to 2.7 following the filtering process during which Headings having more than 10 fold changes were discarded.

The G category (Phenomena and Processes) is a relatively large category with 3,164 Headings. Selection criterion is set to select fewer than 10% of the Headings for the database. A total of 242 Headings are selected from this category to represent 24% of the non-disease class in the database. Before filtering, the number of abstracts per Heading ranges from 0 to 1,266,295; the range is significantly large as some Headings are associated with over one million abstracts and some Headings are associated with fewer than ten abstracts. The selection process guarantees inclusion of Headings that have a minimum of 1,000 and a maximum of 20,000 abstracts; the average number of abstracts per Headings is therefore reduced significantly from 20,374 to 8,755. In the same way, the range of child node per Heading is reduced from [0; 248] to [1; 10] thus providing a mean to select the best representative Headings. Finally, the fold change ranges from 1 to 66,905 before filtering, limiting the range to [1; 10] brings the average value of fold change from 367 to 3.1, hence proving an additional step in the Heading selection process.

The I (Anthropology, Education, Sociology and Social Phenomena) and J (Technology, Industry, Agriculture) categories have similar characteristics with 559 and 558 Headings respectively. The I category has an average of 7,374 and J category an average of 7,290 abstracts per Heading. Similarly, the I category has an average of 1.7 child node while the J category has an average of 1.6 child node. Finally, the I and J category have an average of 114 and 99 fold change per Heading respectively. The selection rules can therefore be adjusted in a similar manner with the ultimate goal to select about 100 nodes to populate roughly 10% of the non-disease category.

To this end, the filtering process is designed in parallel: the filtering process excludes Headings that have i) fewer than 1,000 or greater than 10,000 abstracts per Heading, ii) fewer than 1 or greater than 10 child nodes, and iii) a fold change greater than 10. As a result of this filtering process, a total of 31 headings from the I category and 66 from the J category are selected. The average number of abstracts is reduced to 5,520 in the I category and 4,787 in the J category. Similarly, the average number of child nodes is 2.2

in the I category and 1.6 in the J category; finally, the average fold change per Heading is reduced to 3.5 in the I category and 3.1 in the J category. Together these numbers demonstrate that the filtering process that relies on a the three features can be best as the number of abstract may not in this case be enough and fold change can be a more useful measure. In fact, the average number of abstracts is only reduced by 25%-34% after filtering for I and J category respectively; however, average fold change is reduced by 97% for both I and J category. Hence the fold change in this case has a more discriminative power.

The M category (Named Groupes) is a small category with only 190 Headings. The selection process filters this category in a way to only include a small subset of Headings in the non-disease class. Even though this category has a limited number of Headings, the variation in terms of specificity of the topics is large. The number of abstracts per Heading in this group ranges from 0 - 3,600,540 with an average of 64,924 abstracts. Some of the Headings in this category are very generic with over 3 million abstracts while some Headings are very specific with very limited abstracts; therefore the filtering process can be very useful in filtering out Headings at both extremes. The child node for each heading reaches a maximum of 71 and the fold change ranges from 1 - 15,136. Again, these numbers confirm the extent of variability in specificity of the Headings in this category. The filtering process, limits the number of abstracts from 1,000 - 20,000 and the fold change to a maximum of 5 while selecting Headings that have greater than 1 and fewer than 5 child nodes. After this filtering process there are 13 Headings that are selected to be in the non-disease class with an average number of abstracts of 6,785 per Heading, an average fold change of 2 and average number of child node of 1.5. The inclusion of a small representative sample from this category can be important as these are potentially interesting Headings such as: "Hispanic Americans", "Twins", or "Emergency Responders".

The N category (Health Care) has 2,207 Headings with a large range of specificity. The number of abstracts per Heading ranges from 0 - 3,727,938 and the number of child node per Heading reaches a maximum of 165 with an average value of 1.7; furthermore the fold change per Heading has also a very large range [1 - 121,977]. As with the other categories, the filtering process is critical for the selection of a balanced representative data. The selection process excludes headings that have i) fewer than 5,000 and greater than 10,000 abstracts, ii) fewer than 1 and greater than 10 child nodes, and iii) fold change greater than 10. This filtering process creates a small subset of headings from this category (for a total of 63, or 6% of the non-disease group). The selected subset of Headings has an average number of 7,473 abstracts and an average of 2.3 child nodes in addition to an average of 3 fold change. This selection process ensures the inclusion of Headings that have moderate specificity thus reducing systemic bias in the dataset.

One of the key objectives in the factor selection process is to create a balanced representative dataset across all categories. In fact, after filtering the average number of abstracts, the average number of child nodes and the average number of fold change is within closer range. Before filtering, the average number of abstracts per Heading was 19,451±20,658, this number was reduced to 7,792±2,259; similarly the average number of child node per heading was 1.6±0.5 and this number was increased to 2.6±1.1; finally the average number of fold change per Heading was 221±99 and this number was reduced to 2.8±0.5. If the selected features are good representation of specificity, then the

2,846 factors selected through this process have a comparable specificity and can be use to build a robust model where noise and systemic bias due to dataset characteristics are reduced. The higher the quality of the dataset the higher is the quality of the model and that translates directly to a higher quality of knowledge discovery tool.

Once the Headings are selected, then the duplicates are removed. In MeSH some nodes are duplicates as their parent node are different; however, the documents retrieved for both duplicated nodes are identical, hence duplicates can be removed without causing any inconsistency. A total of 301 Headings are duplicated and are removed in the final stage; these include 218 (or 12%) from the C category, 39 (8%) from the D category, 7 (5%) from the F category, 32 (13%) from the G category, 2 (3%) from the J category and finally 3 (or 4%) from the N category. This reduction is the last step in the Heading selection process, reducing the final Heading List from 2,846 to 2,545.

POLSA

There is no change in the structure of the POLSA module in the expanded version of ARIANA. The main differences are the followings: there are 2.545 Headings in the model; the dictionary is also revised and genes from OMIM are added and irrelevant words are removed based on a first run; the approximate TF-IDF is obtained following dimensionality reduction of the encoding matrix covering 95% of the total energy, or in this case dimensionality is reduced to 1,400 to create the approximate TF-IDF.

Mapping of Equivalence Class and Tool Integration

Any given tool solves a specific problem based on set of constraints and limitations. Also any given tool attempts in general to be as precise as possible while overcoming the many challenges. Integrating such efforts can be beneficial at many levels; however, the main idea is to reduce the inefficiencies due to replication of functionalities. Tool integration can be achieved at different levels yet the main challenge is to create an efficient mapping of classes. For instance, the main difficulty for the integration of Online Mendelian Inheritance in Man (OMIM) database with ARIANA is to map disease classes between the two systems. ARIANA uses the hierarchical disease names from MeSH while OMIM has a flat list of diseases and their associated genes.

Therefore, to use and integrate the gene information to the ARIANA web tool we have to solve the disease-mapping challenge. The OMIM is a database of human genes and genetic disorders (http://www.omim.org/). It is possible to download the full database in a local machine and use that information to display any gene disease association that relates to the user's query. The main challenge is to extract disease names that correspond to the disease names used in MeSH. To be as transparent, we designed a system that displays information that is complete without encapsulating or hiding any intermediate steps. To solve the disease mapping problem we implemented a three step process that would take a disease in MeSH and bring to user's attention a series of genes that may or may not be associated to the disease; the expert can make his or her decision as to follow the lead. These associations are mainly indirect associations.

Step 1: extraction of significant words, representing disease names, from the OMIM database. The significant word is the first word in the multi-gram disease name that is used in the OMIM database. For instance, Alzheimer is the significant word for "Alzheimer disease 1, familial".

Step 2: extraction of disease names from OMIM. This entry represents the full disease name in the database, including identifiers preceeding or following the disease name. In addition, the disease identification number is also kept for furhter reference. OMIM identifiers include the followings (when available):

#: Phenotype mapping method - # appears in parentheses after a disorder. **1**: The disorder is placed on the map based on its association with a gene, but the underlying defect is not known. **2**: The disorder has been placed on the map by linkage; no mutation has been found. **3**: The molecular basis for the disorder is known; a mutation has been found in the gene. **4**: A contiguous gene deletion or duplication syndrome, multiple genes are deleted or duplicated causing the phenotype. For instance, "*Alzheimer disease 1, familial, 104300 (3)"* has a disease identifier that is 104300 and the phenotype mapping method is set to 3.

Note that the full entry in the OMIM database for the example above is: *"Alzheimer disease 1, familial, 104300 (3)|APP, AAA, CVAP, AD1|104760|21q21.3"*
That means that the gene APP (also known as AAA, CVAP or AD1) is associated with the disease 104300. This association is based on a phenotype mapping explained by #3.

Step 3: extraction of genes symbol for the identified diseases. For instance in the case of *"Alzheimer disease 1, familial, 104300"* the corresponding gene symbol is APP.
Figure 18 shows the mapping process described here.

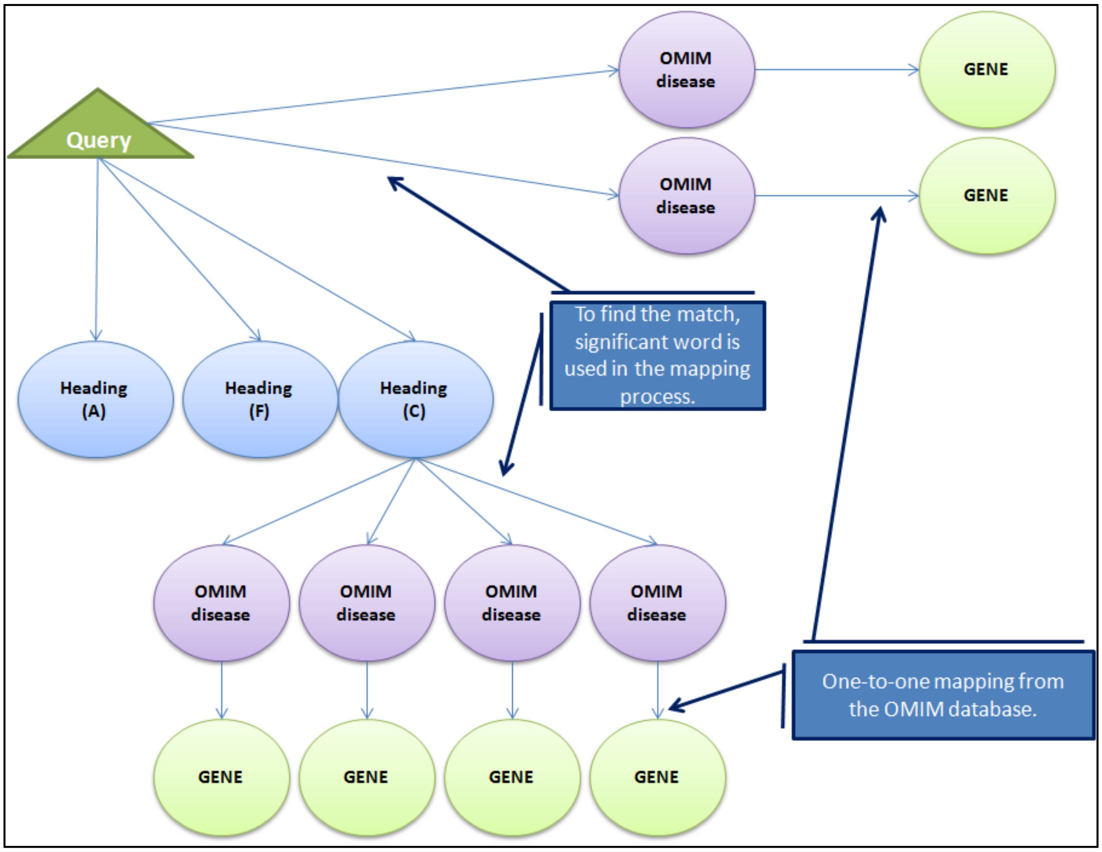


Figure 18. Disease-mapping module for extraction of gene-disease association

Relevance Model

The relevance model is as described in Chapter 4. Here we demonstrate (see Figure 19) the power of the dynamic data driven technique for extraction of information using three simulations: infection, tuberculosis (TB), and multiple sclerosis (MS).

Using the default setting, we compare the results of the simulations. For the case of infection disease, scores higher than 0.06 are considered highly associated (for a total of 358) and scores between 0.062 and 0.026 are considered possibly associated (for a total of 255), scores below 0.026 are clustered in the unknown group. For the case of TB scores higher than 0.124 are considered highly associated (for a total of 35 Headings), and scores between 0.124 and 0.059 are considered possibly associated (for a total of 40 Headings). Score lower than 0.059 are clustered in the unknown group. For the case of MS, scores higher than 0.11 are considered highly associated (for a total of 55 Headings), and scores between 0.11 and 0.0048 are considered possibly associated (for a total of 40 Headings). Score lower than 0.0048 are clustered in the unknown group.

In the first case, infection disease is well studied with over one million publications in PubMed. The cut-off values obtained in this case are 0.062 and 0.026 respectively. There are over 300 Headings in the first cluster (high association) and over 200 Headings in the second cluster (possible association). Using the same setting, the simulation for TB returns different cut-off values: 0.124 and 0.059 respectively. TB is less studied, when compared to infection, with over 200,000 publications in PubMed. Finally, there are even fewer PubMed entries for MS, and for that reason we expect the cut-off values to be different. In fact, with the default setting, the two cut-off values for MS are 0.11 and 0.0048, putting 55 entries in the first cluster and 40 entries in the second cluster. This analysis demonstrates the importance of a dynamic data driven threshold calculation.

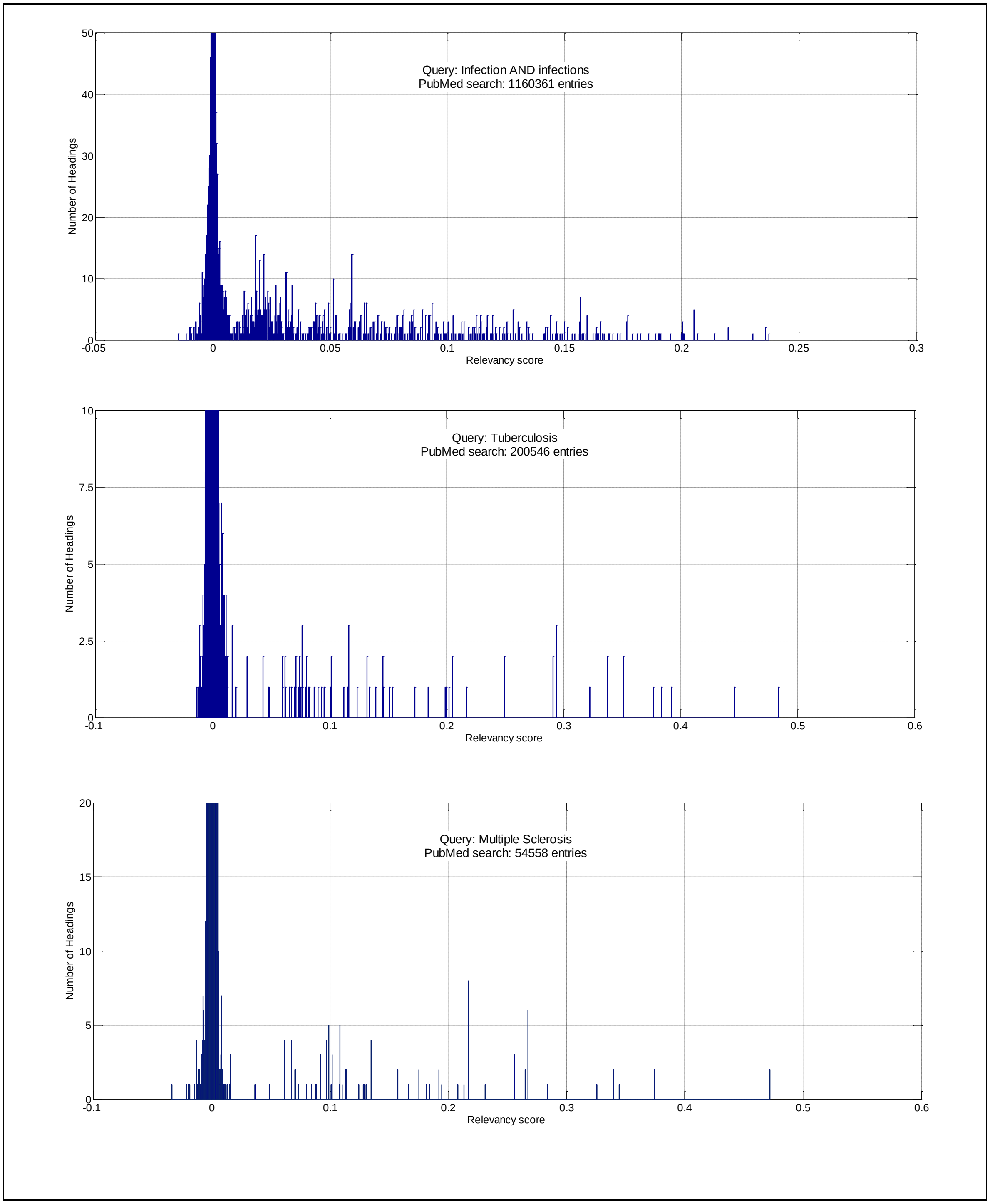


Figure 19. Histogram representation of three different queries

Results

Knowledge Discovery - Case for "Pulmonary Fibrosis" and "Hexamethonium"

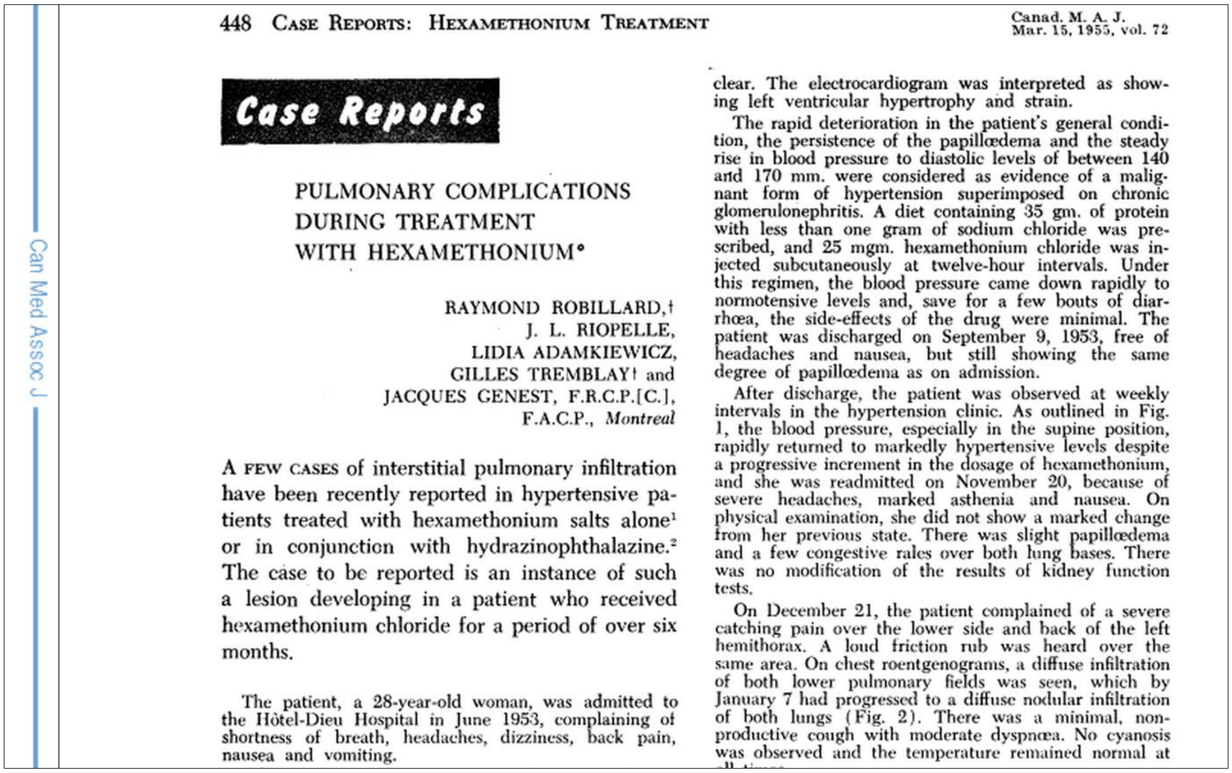
The drug Hexamethonium is a drug that can be used to treat chronic hypertention, of the peripheral nervous system; however, the non-specificity of its action led to discontinuing its use. In 2001, a research study in John Hopkins used this drug to induce asthma in healthy research objects. During the course of the study, a healthy volunteer, Ellen Roche, died only after few days of inhaling this drug. She was diagnosed with pulmonary inflammation and fibrosis based on chest imaging and autopsy report following her death (http://www.hopkinsmedicine.org/press/2001/july/report_of_internal_investigation.htm). In fact in the autopsy report it was stated the following facts: "The microscopic examination of the lungs later revealed extensive, diffuse loss of alveolar space with marked fibroris and fibrin thrombi involving all lobes. There was also evidence of alveolar cell hyperplasia as well as chronic inflammation compatible with an organizing stage of diffuse alveolar damage. There was no evidence of bacteria, fungal organisms, or viral inclusions on routine or special stains." (http://www.hopkinsmedicine.org/press/2001/july/report_of_internal_investigation.htm).

This study was headed by Dr. Alkis Togias, who made a "good-faith effort" to research the drug's (in this case hexamethonium) adverse effects. Dr. Togias search mainly focused on a limited number of resources, including PubMed database, and the ethics panel subsequently approved the safety of the drug. This tragedy highlights the importance of literature search for designing experiments and enrolling healthy individuals in control groups.

The volunteer was a healthy person with no lung or kidney problems; however, because of inhaling a "believed to be safe" chemical she lost her life. One day after enrolling in the study she developed a dry cough and dyspnea. Two days after she developed flu-like symptoms and her FEV1 was reduced. On May 09, 2001 she was febrile and was admitted to the Johns Hopkins Bayview Medical Center (JHBMC). Her chest X-ray revealed streaky densities in the right perihilar region and her arterial oxygen saturation fell to 84% after walking a short distance. She was referred to ICU on May 12, 2001. She was subsequently intubated and ventilated, suffered bilateral pneumothoraces, and presented a clinical picture of adult respiratory distress syndrome (ARDS). She died on June 02, 2001.

The main message is that this tragic accident could have been prevented. In fact there was a case report (Figure 20) of the toxicity of this drug in a 28 year old woman [36]. However, when Dr. Togias did his PubMed search, the PubMed database was searchable only back to 1966; therefore a basic PubMed search missed this important evidence. Professional medical librarians could have found this evidence through other online databases such as TOXNET, the Toxicology Data Network resource provided by NLM (http://toxnet.nlm.nih.gov/index.html). Following this incident, the federal Office of Human Research Protection (OHRP) suspended the University's Federal licence (MPA) to conduct human research involving human subjects. "All Federally funded research was suspended at: the Johns Hopkins University School of Medicine (JHUSM), the Johns Hopkins University School of Nursing, the Johns Hopkins Hospital, the Johns Hopkins

Bayview Medical Center, the Gerontology Research Center of the National Institute of Aging-Bayview Campus, the Kennedy-Krieger Institute, and the Applied Physics Laboratory. This made the international news headlines, as how a preventable research casualty could happen in the this century, where technology and data are at everyone's fingertip.


Can Med Assoc J

448 CASE REPORTS: HEXAMETHONIUM TREATMENT

Canad. M. A. J.
Mar. 15, 1955, vol. 72

Case Reports

PULMONARY COMPLICATIONS
DURING TREATMENT
WITH HEXAMETHONIUM°

RAYMOND ROBILLARD,†
J. L. RIOPELLE,
LIDIA ADAMKIEWICZ,
GILLES TREMBLAY† and
JACQUES GENEST, F.R.C.P.[C.],
F.A.C.P., *Montreal*

A FEW CASES of interstitial pulmonary infiltration have been recently reported in hypertensive patients treated with hexamethonium salts alone[1] or in conjunction with hydrazinophthalazine.[2] The case to be reported is an instance of such a lesion developing in a patient who received hexamethonium chloride for a period of over six months.

The patient, a 28-year-old woman, was admitted to the Hôtel-Dieu Hospital in June 1953, complaining of shortness of breath, headaches, dizziness, back pain, nausea and vomiting.

clear. The electrocardiogram was interpreted as showing left ventricular hypertrophy and strain.

The rapid deterioration in the patient's general condition, the persistence of the papillœdema and the steady rise in blood pressure to diastolic levels of between 140 and 170 mm. were considered as evidence of a malignant form of hypertension superimposed on chronic glomerulonephritis. A diet containing 35 gm. of protein with less than one gram of sodium chloride was prescribed, and 25 mgm. hexamethonium chloride was injected subcutaneously at twelve-hour intervals. Under this regimen, the blood pressure came down rapidly to normotensive levels and, save for a few bouts of diarrhœa, the side-effects of the drug were minimal. The patient was discharged on September 9, 1953, free of headaches and nausea, but still showing the same degree of papillœdema as on admission.

After discharge, the patient was observed at weekly intervals in the hypertension clinic. As outlined in Fig. 1, the blood pressure, especially in the supine position, rapidly returned to markedly hypertensive levels despite a progressive increment in the dosage of hexamethonium, and she was readmitted on November 20, because of severe headaches, marked asthenia and nausea. On physical examination, she did not show a marked change from her previous state. There was slight papillœdema and a few congestive rales over both lung bases. There was no modification of the results of kidney function tests.

On December 21, the patient complained of a severe catching pain over the lower side and back of the left hemithorax. A loud friction rub was heard over the same area. On chest roentgenograms, a diffuse infiltration of both lower pulmonary fields was seen, which by January 7 had progressed to a diffuse nodular infiltration of both lungs (Fig. 2). There was a minimal, nonproductive cough with moderate dyspnœa. No cyanosis was observed and the temperature remained normal at

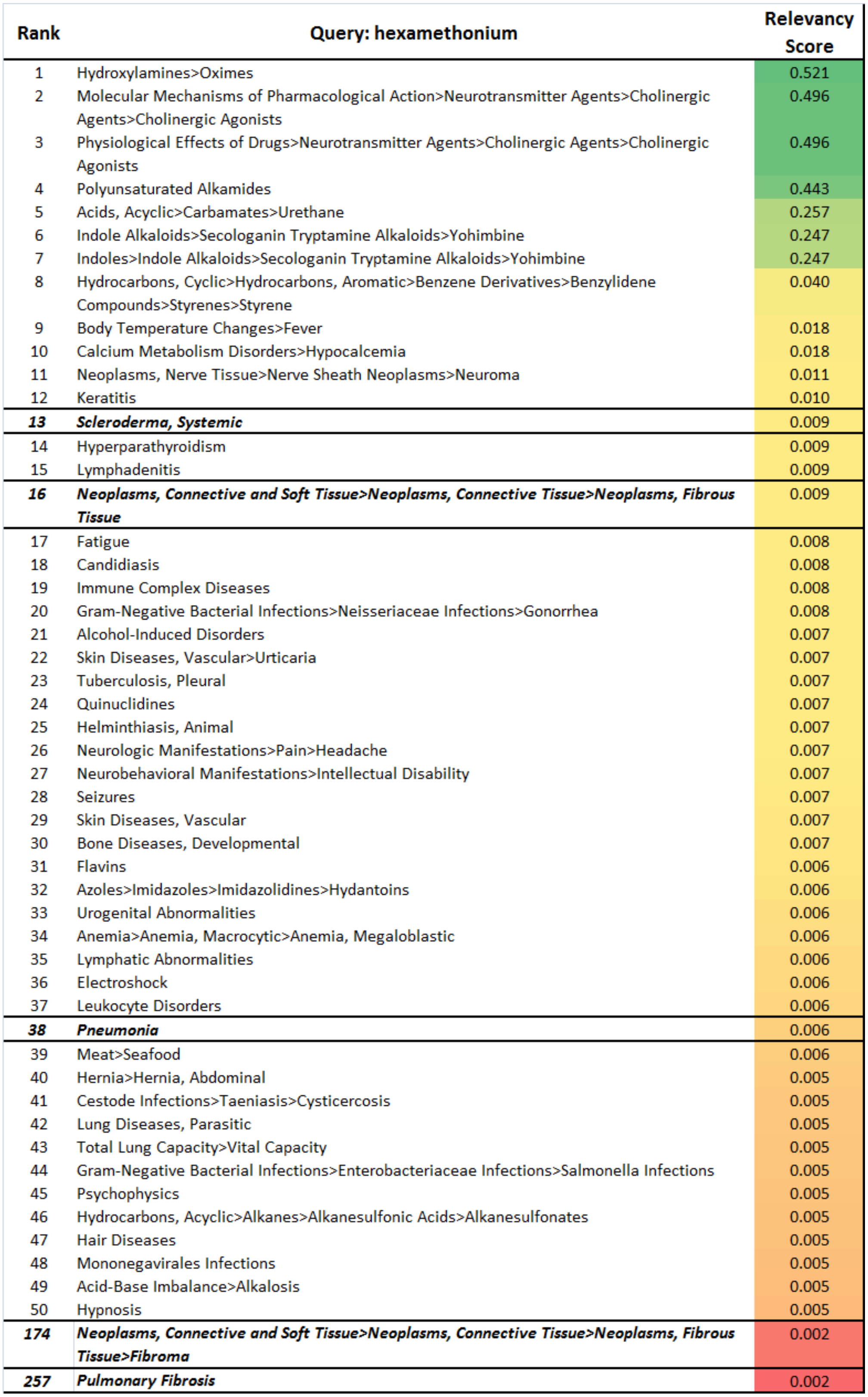
Figure 20. Case report published in Can Med Assoc J. in 1955

Since the original case report there has not been any new publications regarding the association of Hexamethonium and pulmonary fibrosis, in fact searching the PubMed will not identify any relevant material for that pourpose. A PubMed search of :

hexamethonium and "pulmonary fibrosis" returns four hits, none of them with available abstracts online. One of the returned publications is in Russian language published in 1967. The other three are publishe in 1979, 1956 and 1967 [37, 38, 39, 40]. Note that, PubMed search also includes seraching through the MeSH entires, therefore searching the PubMed is not only a keyword-based search. Searching individual entries return 19516 record for "pulmonary fibrosis" and 7026 entries for Hexamethonium; therefore the number of publication on each topic is not the limiting factor. Therefore, still to this date there is very limited evidence of the toxicity of this drug in PubMed. The PDF of the case report published in 1955 can be found in PubMed today; however, many data mining tools, including ARIANA, do not take into account PDFs of articles, especially when they are published in the 50s or even 60s.

We tested ARIANA using the keyword "Hexamethonium" to see whether our system can find any association between Hexamethonium and fibrosis /pulmonary fibrosis. Figure 21 highlights the top ranked Headings and their relevancy scores. The Headings in bold and italic are clear indication of the association between what caused the death of Ellen and the drug Hexamethonium. We show that the ARIANA tool can detect this association and provide a number of clues of the danger of this drug. ARIANA is based on higher order co-occurrence analysis, therefore even though there is no published evidence of this association the system could find such relations.

Searching ARIANA with "Hexamethonium" as query produces a ranked list of 2,545 Headings. The 13th Heading in the list is "*Scleroderma, Systemic*", the 16th Heading is "*Neoplasms, Connective and Soft Tissue>Neoplasms, Connective Tissue>Neoplasms, Fibrous Tissue*", the 38th Heading is "*Pneumonia*", the 174th Heading is "*Neoplasms,* *Connective and Soft Tissue>Neoplasms, Connective Tissue>Neoplasms, Fibrous>Fibroma* " and finally the 257th Heading is "*Pulmonary Fibrosis*". Since the health of human subjects is at stake, the researcher should, when analyzing the data, look at least at the top 500 (or top 20%) ranked Headings. Figure 22 shows the cluster membership when relevancy model is applied and the respective cosine score for the top 25 ranked Headings. Based on the relevancy model analysis the top 7 Headings are classified as being highly associated with the drug, the remaining Headings have unknown association level.



Rank	Query: hexamethonium	Relevancy Score
1	Hydroxylamines>Oximes	0.521
2	Molecular Mechanisms of Pharmacological Action>Neurotransmitter Agents>Cholinergic Agents>Cholinergic Agonists	0.496
3	Physiological Effects of Drugs>Neurotransmitter Agents>Cholinergic Agents>Cholinergic Agonists	0.496
4	Polyunsaturated Alkamides	0.443
5	Acids, Acyclic>Carbamates>Urethane	0.257
6	Indole Alkaloids>Secologanin Tryptamine Alkaloids>Yohimbine	0.247
7	Indoles>Indole Alkaloids>Secologanin Tryptamine Alkaloids>Yohimbine	0.247
8	Hydrocarbons, Cyclic>Hydrocarbons, Aromatic>Benzene Derivatives>Benzylidene Compounds>Styrenes>Styrene	0.040
9	Body Temperature Changes>Fever	0.018
10	Calcium Metabolism Disorders>Hypocalcemia	0.018
11	Neoplasms, Nerve Tissue>Nerve Sheath Neoplasms>Neuroma	0.011
12	Keratitis	0.010
13	***Scleroderma, Systemic***	0.009
14	Hyperparathyroidism	0.009
15	Lymphadenitis	0.009
16	***Neoplasms, Connective and Soft Tissue>Neoplasms, Connective Tissue>Neoplasms, Fibrous Tissue***	0.009
17	Fatigue	0.008
18	Candidiasis	0.008
19	Immune Complex Diseases	0.008
20	Gram-Negative Bacterial Infections>Neisseriaceae Infections>Gonorrhea	0.008
21	Alcohol-Induced Disorders	0.007
22	Skin Diseases, Vascular>Urticaria	0.007
23	Tuberculosis, Pleural	0.007
24	Quinuclidines	0.007
25	Helminthiasis, Animal	0.007
26	Neurologic Manifestations>Pain>Headache	0.007
27	Neurobehavioral Manifestations>Intellectual Disability	0.007
28	Seizures	0.007
29	Skin Diseases, Vascular	0.007
30	Bone Diseases, Developmental	0.007
31	Flavins	0.006
32	Azoles>Imidazoles>Imidazolidines>Hydantoins	0.006
33	Urogenital Abnormalities	0.006
34	Anemia>Anemia, Macrocytic>Anemia, Megaloblastic	0.006
35	Lymphatic Abnormalities	0.006
36	Electroshock	0.006
37	Leukocyte Disorders	0.006
38	***Pneumonia***	0.006
39	Meat>Seafood	0.006
40	Hernia>Hernia, Abdominal	0.005
41	Cestode Infections>Taeniasis>Cysticercosis	0.005
42	Lung Diseases, Parasitic	0.005
43	Total Lung Capacity>Vital Capacity	0.005
44	Gram-Negative Bacterial Infections>Enterobacteriaceae Infections>Salmonella Infections	0.005
45	Psychophysics	0.005
46	Hydrocarbons, Acyclic>Alkanes>Alkanesulfonic Acids>Alkanesulfonates	0.005
47	Hair Diseases	0.005
48	Mononegavirales Infections	0.005
49	Acid-Base Imbalance>Alkalosis	0.005
50	Hypnosis	0.005
174	***Neoplasms, Connective and Soft Tissue>Neoplasms, Connective Tissue>Neoplasms, Fibrous Tissue>Fibroma***	0.002
257	***Pulmonary Fibrosis***	0.002

Figure 21. list of Headings with query "Hexamethonium"

Ellen died of *pulmonary fibrosis* and having access to this tool could have prevented her death. Pulmonary Fibrosis ranked in the top 10 percentile of the Headings, giving clear indication of the potential danger of this drug. In addition, Scleroderma is a chronic systemic autoimmune disease characterized by fibrosis, vascular alterations and autoantibodies, and having "Scleroderma, Systemic" as one of the top ranked Headings should also give a clear indication of potential risks of the drug. This indication can trigger the need for further investigation by the researcher and highlight the importance of seeking professional help to assure the safety of the drug when using in clinical trials.

Information Retrieval and Data Reuse - Case for "Alzheimer Disease"

To show the efficacy of the tool in knowledge extraction we experiment with a common disease name: "alzheirmer". A search in PubMed of the keyword "alzheirmer" returns 65,122 entries (search done on 12/06/2012). Of coarse going through all the 65 thousands entries is going to be impossible by a single (or a team) person. On the other hand, as we have seen for the case of Hexamethonium, there might be one single case report or some key papers that one would be interested to see before designing or implementing any research project. We want to test ARIANA's capability in terms of information extraction for a simple search.

For a any query, ARIANA will rank the 2,545 Headings with respect to the relevancy score. Figure 23 list the 25 top ranked Headings for the query Alzheimer. Figure 24 shows the distribution of the scores. As it is expected most of the Headings have a low score, that is close to zero. There a 15 Headings with relevancy score above the first automatic threshold (between 0.16 and 0.15), these are categorized as highly associated with the query, there are also two Headings (16th and 17th Headings) that are classified as possibly associated with the query (second automatic threshold is between 0.14 and 0.02). All the remaining Headings are classified as having unkown association with the query. Being classified in the unknow group does not mean that there is no evidence for the association, it is just an indication of the strenght of the association, for the reason this analysis will focus on the top 3% (or top 80) of the ranked Headings. Figure 25 shows the results of relevancy score as well as the cluster membership when relevance model is applied for the top 25 ranked Headings. Based on the relevancy model analysis the top 15 Headings are classified as being highly associated, the 16th and 17th Headings are classified as possible Headings and the remaining Headings have unknown association level. However, depending on the purpose of the study, the researcher may want to explore more or fewere of the Headings for potential association.

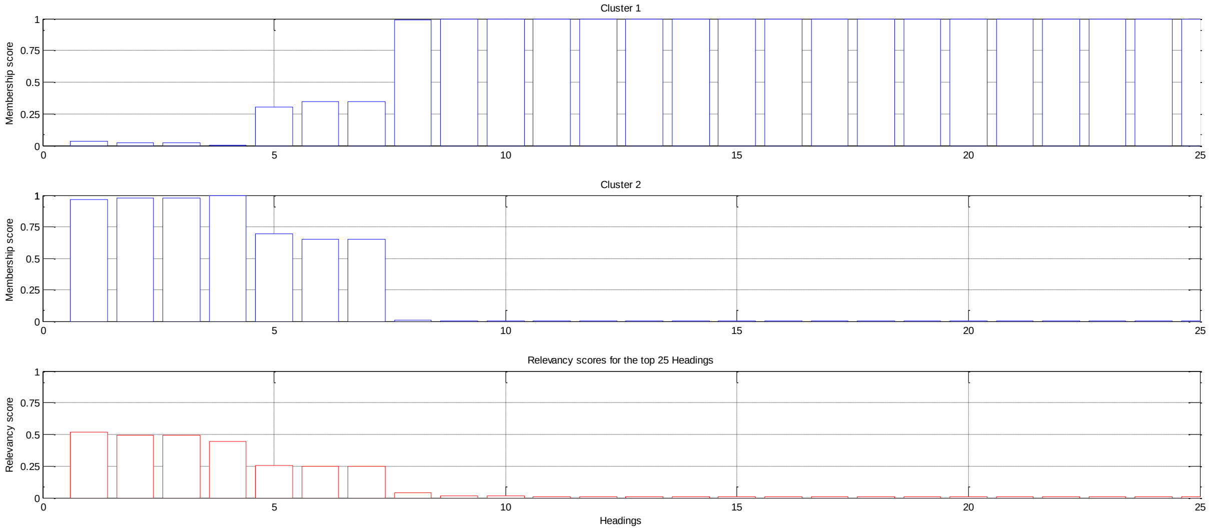
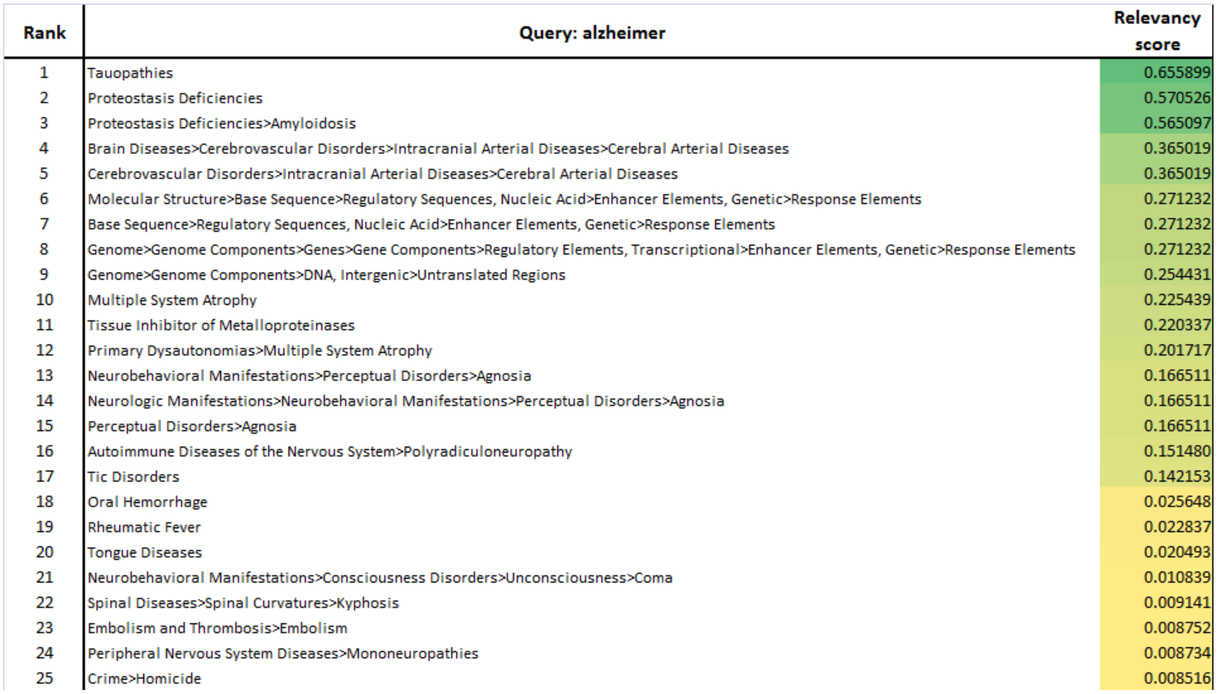


Figure 22. Relevance model applied to the results obtained from simulation of query: "Hexamethonium" for the top 25 Headings.



Rank	Query: alzheimer	Relevancy score
1	Tauopathies	0.655899
2	Proteostasis Deficiencies	0.570526
3	Proteostasis Deficiencies>Amyloidosis	0.565097
4	Brain Diseases>Cerebrovascular Disorders>Intracranial Arterial Diseases>Cerebral Arterial Diseases	0.365019
5	Cerebrovascular Disorders>Intracranial Arterial Diseases>Cerebral Arterial Diseases	0.365019
6	Molecular Structure>Base Sequence>Regulatory Sequences, Nucleic Acid>Enhancer Elements, Genetic>Response Elements	0.271232
7	Base Sequence>Regulatory Sequences, Nucleic Acid>Enhancer Elements, Genetic>Response Elements	0.271232
8	Genome>Genome Components>Genes>Gene Components>Regulatory Elements, Transcriptional>Enhancer Elements, Genetic>Response Elements	0.271232
9	Genome>Genome Components>DNA, Intergenic>Untranslated Regions	0.254431
10	Multiple System Atrophy	0.225439
11	Tissue Inhibitor of Metalloproteinases	0.220337
12	Primary Dysautonomias>Multiple System Atrophy	0.201717
13	Neurobehavioral Manifestations>Perceptual Disorders>Agnosia	0.166511
14	Neurologic Manifestations>Neurobehavioral Manifestations>Perceptual Disorders>Agnosia	0.166511
15	Perceptual Disorders>Agnosia	0.166511
16	Autoimmune Diseases of the Nervous System>Polyradiculoneuropathy	0.151480
17	Tic Disorders	0.142153
18	Oral Hemorrhage	0.025648
19	Rheumatic Fever	0.022837
20	Tongue Diseases	0.020493
21	Neurobehavioral Manifestations>Consciousness Disorders>Unconsciousness>Coma	0.010839
22	Spinal Diseases>Spinal Curvatures>Kyphosis	0.009141
23	Embolism and Thrombosis>Embolism	0.008752
24	Peripheral Nervous System Diseases>Mononeuropathies	0.008734
25	Crime>Homicide	0.008516

Figure 23. Ranked list of the top 25 Headings with query "Alzheimer"

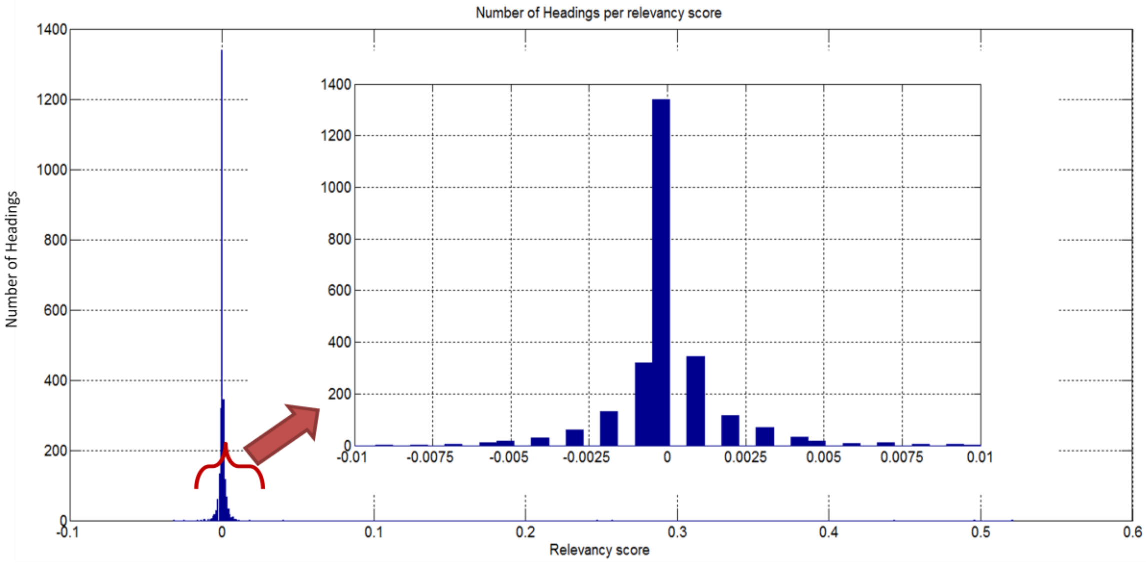


Figure 24. Histogram representation of relevancy score for the query "Alzheimer"

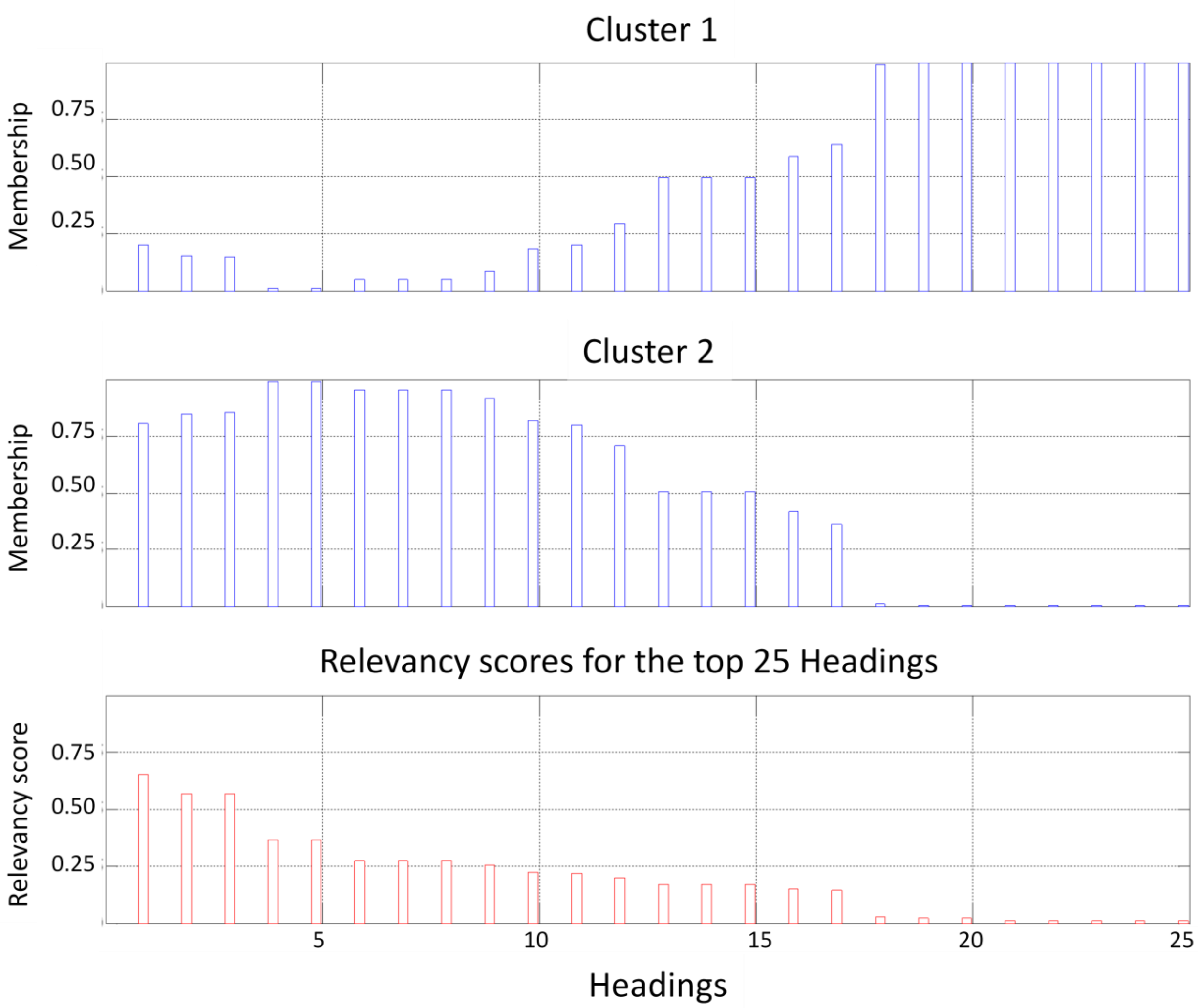
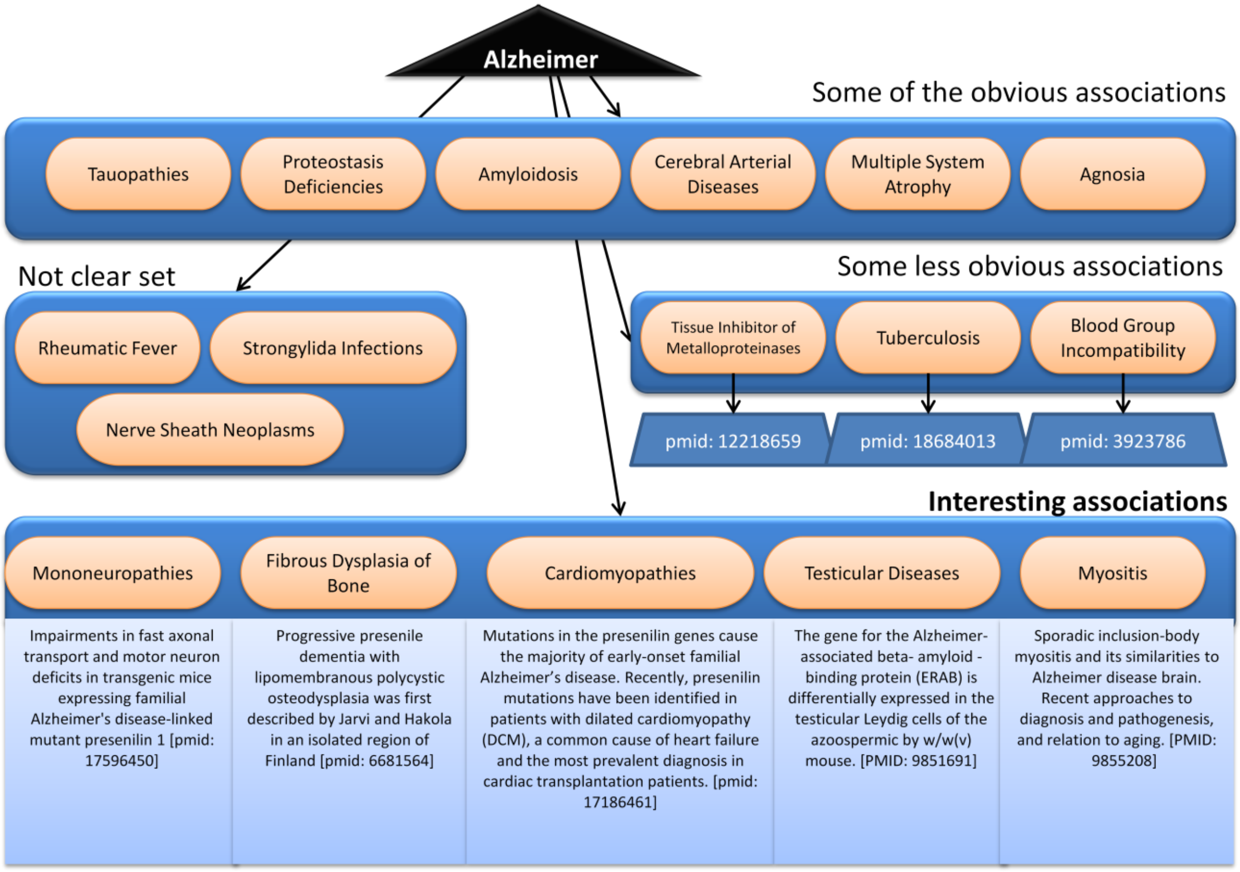


Figure 25. Relevance model applied to the results obtained from simulation of query: "Alzheimer" for the top 25 Headings

The analysis presented here was performed with the help of a board certified neurologist (Dr. Ramin Zand). Figure 26 summarizes the findings. In essence, analysis of the top 80 Headings reveals interesting observations: i) as expected there are a large number of identified Headings that are associated with Alzheimer there include Tauopathies, Proteostasis Deficiencies, Amyloidosis, Cerebral Arterial Diseases, Multiple System Atrophy, and Agnosia; ii) some associations are less obvious, however, a PubMed search clarifies the reason for their high rank, and these include Tissue inhibitor of Metalloproteinases [41], Tuberculosis [42], Blood Group Incomptibility [43]. Finally, there are few Headings that it is not clear why they are highly ranked and these include Rheumatic fever, Strongylida infections, Nerver sheath neoplasms. Yet, the most interesting finding is identification of a set of Headings that are associated with Alzheimer but are not part of a general knowledge of a neurologist. There is evidence for these associations in PubMed; however, bringing these to the attention of a medical expert or researcher could have downstream consequences in their practice. These Headings are Mononeuropathies [44], Fibrous Dysplasia of Bone [45], Cardiomyopathies [46], Testicular Diseases [47] and finally Myositis [48].



Impairments in fast axonal transport and motor neuron deficits in transgenic mice expressing familial Alzheimer's disease-linked mutant presenilin 1 [pmid: 17596450]

Progressive presenile dementia with lipomembranous polycystic osteodysplasia was first described by Jarvi and Hakola in an isolated region of Finland [pmid: 6681564]

Mutations in the presenilin genes cause the majority of early-onset familial Alzheimer's disease. Recently, presenilin mutations have been identified in patients with dilated cardiomyopathy (DCM), a common cause of heart failure and the most prevalent diagnosis in cardiac transplantation patients. [pmid: 17186461]

The gene for the Alzheimer-associated beta- amyloid - binding protein (ERAB) is differentially expressed in the testicular Leydig cells of the azoospermic by w/w(v) mouse. [PMID: 9851691]

Sporadic inclusion-body myositis and its similarities to Alzheimer disease brain. Recent approaches to diagnosis and pathogenesis, and relation to aging. [PMID: 9855208]

Figure 26. Mapping results for query: Alzheimer

Subsequently, the user may search the system with a number of queries. For instance, one may be interested to find common Headings between Alzheimer and Myositis because in an initial simulation that association was interesting to the researcher. Figure 27 demonstrates an example of such search. In that example "Dysostoses" is a common Heading when three queries (Alzheimer, Myositis and Dysplasia of bone) are searched simultaneously. The co-analysis is performed on the top 3% ranked Headings.

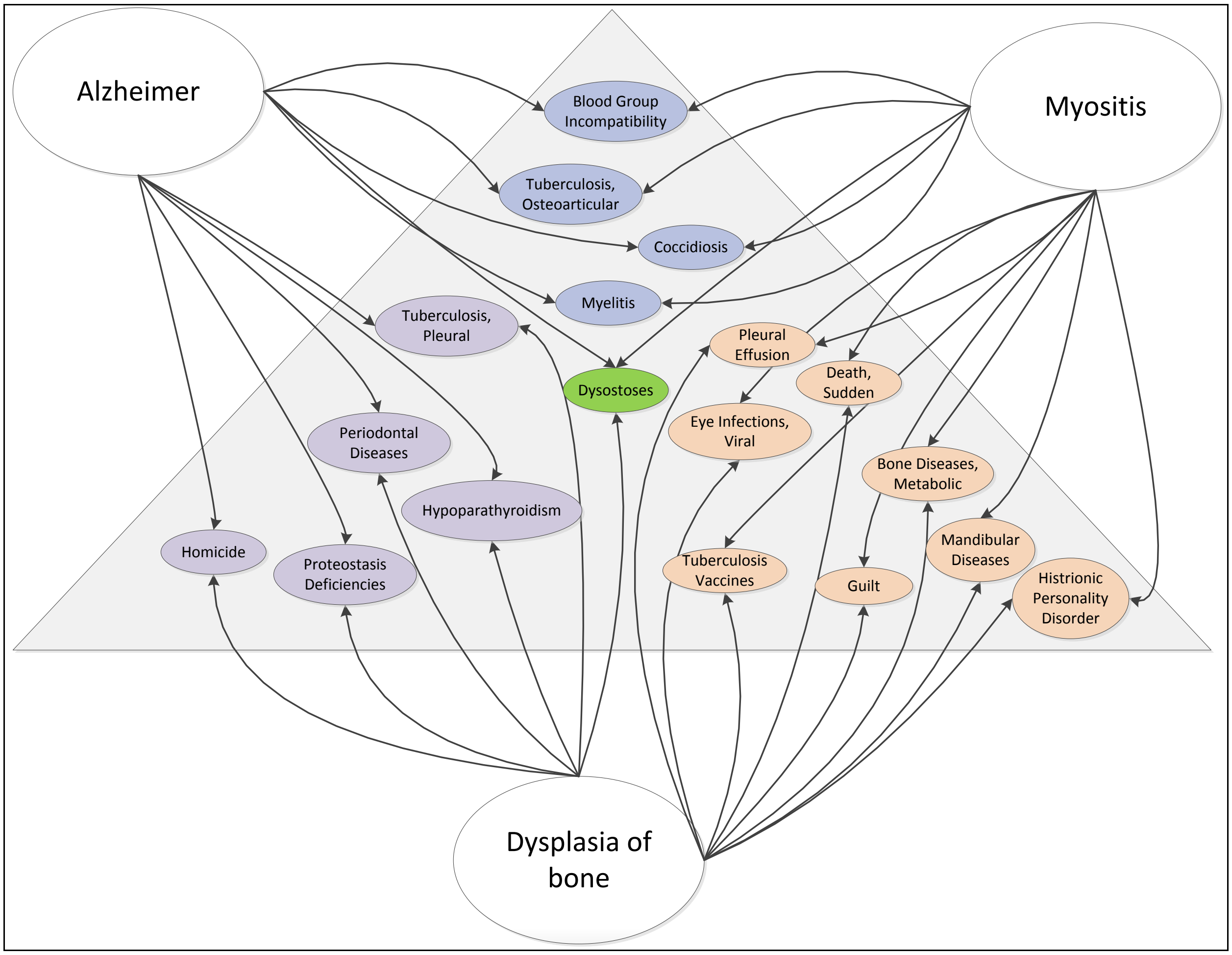


Figure 27. Co-analysis of the top 3% (80/2545) of the three related Headings

The next step is to test the Disease Mapping strategy, to extract gene-disease associations, with Alzheimer disease. The disease mapping gives rise to direct as well as indirect and potential gene-disease associations. Figure 28 and Figure 29 demonstrate the results of this analysis. In the first step, Alzheimer is searched in the list of OMIM diseases; therefore, all the OMIM listed diseases that start with word "Alzheimer" are selected. A total of 22 diseases are selected through this process. These include mainly different types of Alzheimer disease. For each of these diseases, there is one or a number of corresponding genes. Figure 28 shows the result of this first step. Most of the identified genes are the AD genes in addition to other related genes: AD5, AD6, AD7, AD8, AD9, AD10, AD11, AD12, AD13, AD14, AD15, AD16, NOS3, PLAU, APOE, APP, SORL1, APBB2, A2M, ACE, BLMH, HFE, MPO, PACIP1, PSEN1 and PSEN2. The second step in the gene identification process, we are looking at related diseases to bring the potential indirect genes to the user's attention. Many of the diseases have no identified equivalent disease in the OMIM database; however, there are also a number of diseases that have a related disease in the OMIM database. Figure 29 shows these diseases and their corresponding genes. For instance Amyloidosis, Cerebral Arterial Disease, and Multiple System Atrophy have all one or multiple matching diseases in the OMIM database. The corresponding disease for Amyloidosis in OMIM is "Amyloidosis, Secondary, sesceptibility to" and that disease has APCS as related gene. One could investigate the potential association of APCS with Alzheirmer disease by exploring other sources of data. In addition to that, it was found that Alzheimer is associated with "Cerebral Arterial Disease". Through disease mapping, "Cerebral infarction, susceptibility to" is obtained from the OMIM database, which has a mapping to PRKCH gene. Finally, Alzheimer is also associated with "Multiple System Atrophy". For this last case the disease mapping process identifies eight matching diseases (see Figure 29). These lead to a total of eight new genes that may be indirectly associated with Alzheimer.

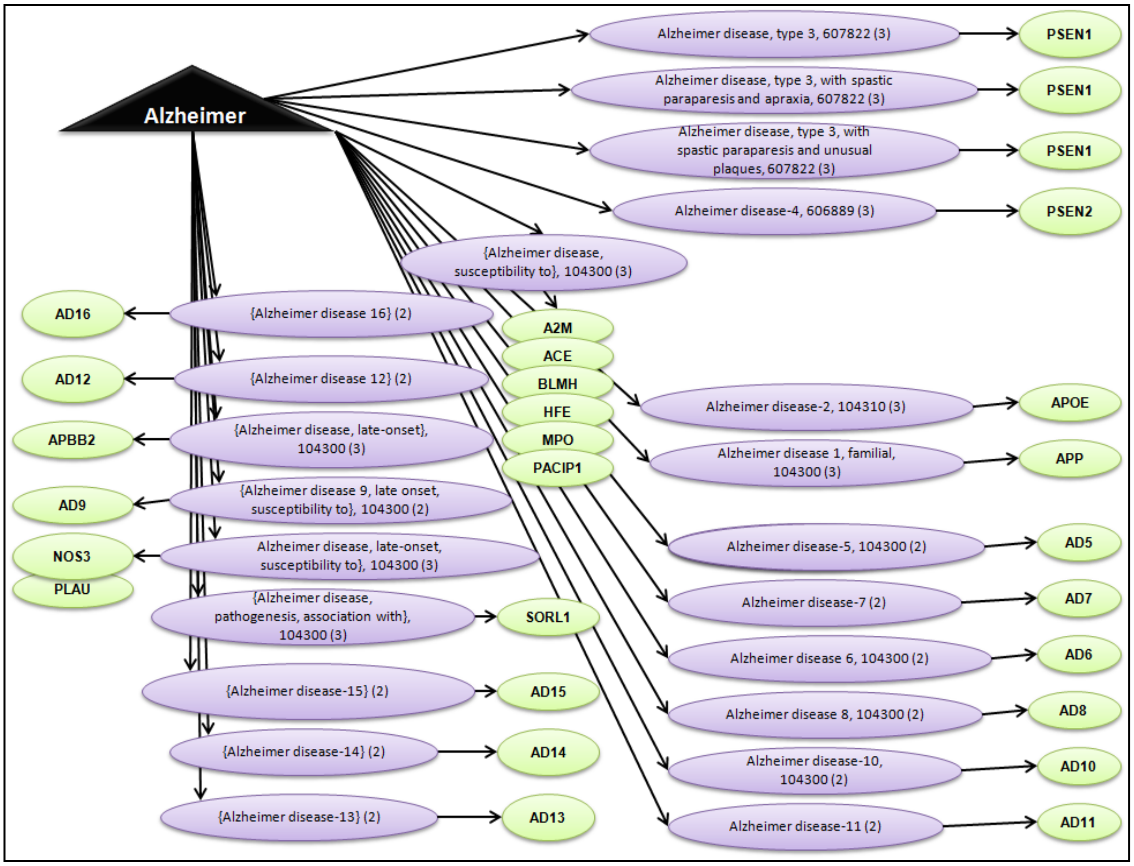

Alzheimer
Alzheimer disease, type 3, 607822 (3)
PSEN1
Alzheimer disease, type 3, with spastic paraparesis and apraxia, 607822 (3)
PSEN1
Alzheimer disease, type 3, with spastic paraparesis and unusual plaques, 607822 (3)
PSEN1
Alzheimer disease-4, 606889 (3)
PSEN2
{Alzheimer disease, susceptibility to}, 104300 (3)
AD16
{Alzheimer disease 16} (2)
A2M
ACE
BLMH
HFE
MPO
PACIP1
AD12
{Alzheimer disease 12} (2)
Alzheimer disease-2, 104310 (3)
APOE
APBB2
{Alzheimer disease, late-onset}, 104300 (3)
Alzheimer disease 1, familial, 104300 (3)
APP
AD9
{Alzheimer disease 9, late onset, susceptibility to}, 104300 (2)
NOS3
PLAU
Alzheimer disease, late-onset, susceptibility to}, 104300 (3)
Alzheimer disease-5, 104300 (2)
AD5
{Alzheimer disease, pathogenesis, association with}, 104300 (3)
SORL1
Alzheimer disease-7 (2)
AD7
Alzheimer disease 6, 104300 (2)
AD6
{Alzheimer disease-15} (2)
AD15
Alzheimer disease 8, 104300 (2)
AD8
{Alzheimer disease-14} (2)
AD14
Alzheimer disease-10, 104300 (2)
AD10
{Alzheimer disease-13} (2)
AD13
Alzheimer disease-11 (2)
AD11

Figure 28. Gene Mapping for Alzheimer

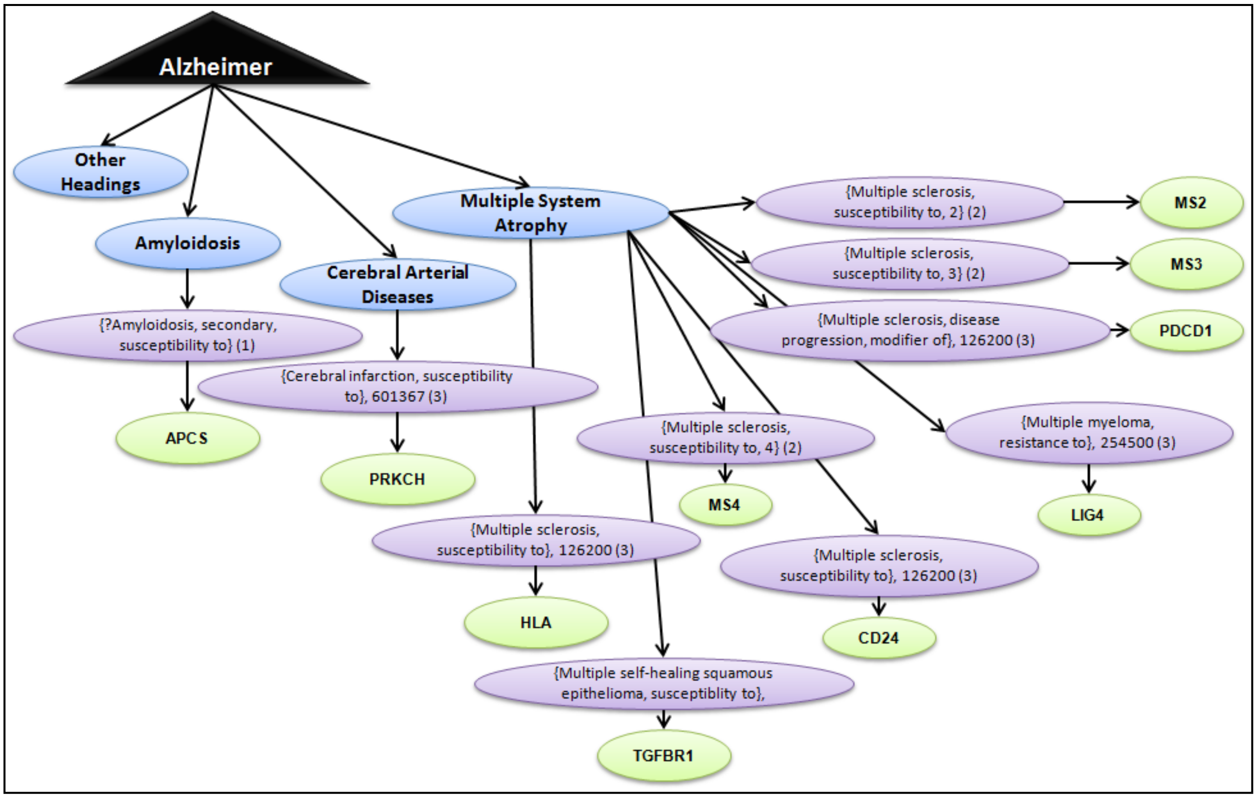

Alzheimer
Other Headings
Amyloidosis
Cerebral Arterial Diseases
Multiple System Atrophy
{?Amyloidosis, secondary, susceptibility to} (1)
APCS
{Cerebral infarction, susceptibility to}, 601367 (3)
PRKCH
{Multiple sclerosis, susceptibility to, 2} (2)
MS2
{Multiple sclerosis, susceptibility to, 3} (2)
MS3
{Multiple sclerosis, disease progression, modifier of}, 126200 (3)
PDCD1
{Multiple myeloma, resistance to}, 254500 (3)
LIG4
{Multiple sclerosis, susceptibility to, 4} (2)
MS4
{Multiple sclerosis, susceptibility to}, 126200 (3)
HLA
{Multiple sclerosis, susceptibility to}, 126200 (3)
CD24
{Multiple self-healing squamous epithelioma, susceptiblity to},
TGFBR1

Figure 29. Gene Mapping for diseases related to Alzheimer through the disease mapping module

The gene disease identification process is accurate as there is no hidden layer between the query and the final genes. It is up to the user to further investigate these findings. Integration of genes in the data-mining module of the ARIANA needs significant analysis and a number of steps. In fact a total of 192 genes symbols are presently in the dictionary and one could query these keywords in order to find potential associations; however, the level of non-specificity is still great due to the nature of the literature. In fact many of the gene symbols have in one way or another some other meaning in the medical field. For instance, AD is also abbreviation for adenovirus, therefore searching the AD genes will return association that are not only related to Alzheimer but also related to a completely different field. Identification of these elements and expansion of the query, when needed, is crucial for a precise information retrieval and knowledge discovery tool. A second example where confusion may arise is for the family of F genes (ie. F2, F5, F7, F8, F9); these may be easily confused with generation of hybrids.

Discussion

ARIANA can be integrated with other Web Tools or databases. In here we have shown that to extract gene-disease association OMIM database can be used. OMIM does not have a hierarchical structure; hence, we have to solve the problem of disease mapping. In addition, we have also demonstrated the integration of OMIM at a data level, where gene symbols from the gene-disease association are added to the ARIANA's dictionary to capture additional semantic information. However, since gene symbols are not unique words, using just that information can lead to complex analytics and confusing associations. Hence, for extraction of crisp associations using LSA based techniques, there is a need to incorporate additional information. Finally the main message here is that integration at data level or tool level can be very beneficial; however, this process needs to be designed and implemented attentively at every level in order to provide a system that is as accurate as possible.

The type of work presented in Chapters 3, 4 and 5 are difficult to validate; however, it is critical to find examples of cases where the system would perform exceptionally well but also cases where the system would generally fail. We have tried to exercise this idea here. We show that ARIANA can be great at finding hidden links among entities derived from MeSH hierarchy, but could easily fail when queries are gene symbols such as F2, or AD1.

Chapter 6: Conclusion

The main goal of this dissertation is to develop an effective mining of biological literature to provide a range of services such as hypothesis generation as well as finding network of semantically related concepts to understand the confluence among various concepts within a specific domain or across the disciplines. To be useful, such knowledge discovery tools must be efficient and scalable with the growing data size. In addition, to effectively reduce information overload and complement traditional means of knowledge dissemination, it is imperative to develop robust, scalable and easy to use Web service applications that are versatile enough to meet the “specific” needs of a diverse community. The utility of such a system would be greatly enhanced with the added capability of finding semantically similar concepts related to various risk factors, side-effects, symptoms and diseases. Such systems are expected to bridge the gap between the effort and resources invested in acquiring knowledge and their effective usage.

This dissertation started with a pilot study to build a framework for hypothesis generation by mining existing literature to identify a set of factors and their association with diseases, phenotypes, or biological processes. The key idea was to develop a knowledge discovery system that can find novel associations so that scientists or researchers can use them to generate a new hypothesis or study a biological phenomenon. In addition, The HGF was designed especially to help junior investigators who often find it difficult to formulate new hypotheses or, more importantly, corroborate their hypothesis with existing literature. The key design concern was always to make the system efficient, robust, scalable and practical for a diverse group of users. To make it more effective, the whole development was done to discover domain specific knowledge as well as integration of tools. The concept of ontology mapping, semantic analysis and relevance model can be used to design and implement an adaptive robust and integrative analysis in providing a range of services in biomedicine as well as text analytics.

Based on the preliminary success, a comprehensive system called, ARIANA was developed by expanding the HGF framework. ARIANA is a software architecture and a web-based system for efficient and scalable knowledge discovery. ARIANA integrates semantic-sensitive analytics of context specific text data through ontology mapping and tool integration strategy. ARIANA ensures the specificity required to create a robust model from an ocean of data. The framework is scalable with the growing size of literature data. ARIANA can find novel associations with no direct citation in the database hence, its utility in knowledge discovery and hypotheses generation. At a tool integration level, OMIM and ARIANA are layered in a cohesive and unified way to present to the user both semantic association and genetic information.

ARIANA is prototyped using MeSH ontolgy and PubMed database for biomedical and scientific applications. ARIANA has five main modules, namely: (i) Data Stratification, (ii) Ontology Mapping, (iii) Parameter Optimized Latent Semantic Analysis (POLSA), (iv) Relevance Model and (v) Interface and Visualization. At each level, the focus is to take into account the specificity of the application and cater to the needs of a wide range of users. For instance, the dictionary is customized from the medical language and fine-tuned subsequently to provide only relevant information. In addition, dynamic data driven technique is used at every possible level to customize each unique search. In fact, a dynamic data driven threshold calculation is proposed and implemented that takes into account the distribution and scores of ranked Headings to determine the best cut-off values to separate the highly, possibly and not likely associated Headings on demand. Finally, the interface and visualization module are an integral part of this work, as they provide a mean of exploring the tool. As ARIANA targets a wide range of users, the interface and visualization modules have focused to bring forward a modular design that is scalable with an even large system.

Empirical results demonstrate the usability and potential of ARIANA in knowledge discovery and crisp information retrieval. Empirical analyses show that ARIANA was able to capture the direct/indirect association that was critical in finding connections between Hexamethonium and pulmonary fibrosis, migraine and sexually transmitted disease, or association of both lymphoma and PD with cadmium poisoning.

Finally, ARIANA can be integrated with other Web Tools or databases to enhance expanded network of associations. For example, ARIANA was integrated with OMIM database to capture gene-disease association that can help a wide range of users, from junior research scientists to experienced policy making advocates and to medical professionals. At one level, users can explore the system to extract general information fast and efficiently (for instance, finding what can be associated with migraine). At a second level, users can explore the tool to find novel associations; this task maybe more time consuming, but for an expert in the field it can be a very interesting and exploratory task, as he or she explores PubMed for different highly ranked Headings. It was through this process that the association between migraine and STD came to our attention. The supporting evidence for that association is recently published and is not part of ARIANA's database. At a higher level, the tool can be used to bring forward possible associations and warrant further investigation by professionals before starting an experiment. For instance, association between Hexamethonium and pulmonary fibrosis can be suggested by ARIANA and evidence can be found through other sources such as TOXNET. At an abstract level, ARIANA can be used by team of professionals who make important decision about policies and allocate future funding in health care and social sciences. For instance identification of elements that directly or indirectly affect "adolescent health" or "weight gain" in children can be of tremendous help in designing programs to target those niche areas.

References

[1] Z. Lu, "PubMed and beyond: a survey of web tools for searching biomedical literature," *Database (Oxford),* vol. 2011, p. baq036, 2011.

[2] D. Rebholz-Schuhmann, H. Kirsch, M. Arregui, S. Gaudan, M. Riethoven and P. Stoehr, "EBIMed--text crunching to gather facts for proteins from Medline," *Bioinformatics,* vol. 23, no. 2, pp. e237--244, 2007.

[3] N. R. Smalheiser, W. Zhou and V. I. Torvik, "Anne O'Tate: A tool to support user-driven summarization, drill-down and browsing of PubMed search results," *J Biomed Discov Collab,* vol. 3, p. 2, 2008.

[4] Y. Yamamoto and T. Takagi, "Biomedical knowledge navigation by literature clustering," *J Biomed Inform,* vol. 40, no. 2, pp. 114-130, 2007.

[5] C. Perez-Iratxeta, P. Bork and M. A. Andrade, "XplorMed: a tool for exploring MEDLINE abstracts," *Trends Biochem. Sci.,* vol. 26, no. 9, pp. 573-575, 2001.

[6] J. J. Kim, P. Pezik and D. Rebholz-Schuhmann, "MedEvi: retrieving textual evidence of relations between biomedical concepts from Medline," *Bioinformatics,* vol. 24, no. 11, pp. 1410-1412, 2008.

[7] T. Ohta, Y. Tsuruoka, J. Takeuchi, J.-D. Kim, Y. Miyao, A. Yakushiji, K. Yoshida, Y. Tateisi, T. Ninomiya, K. Masuda, T. Hara and J. Tsujii, "An intelligent search engine and GUI-based efficient MEDLINE search tool based on deep syntactic parsing," 2006.

[8] D. Szklarczyk, A. Franceschini, M. Kuhn, M. Simonovic, A. Roth, P. Minguez, T. Doerks, M. Stark, J. Muller, P. Bork, L. J. Jensen and C. v. Mering, "The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored," *Nucleic Acids Res.,* vol. 39, no. Database issue, pp. D561--568, 2011.

[9] J. M. Fernandez, R. Hoffmann and A. Valencia, "iHOP web services," *Nucleic Acids Res.,* vol. 35, no. Web Server issue, pp. W21--26, 2007.

[10] R. B. Altman, C. M. Bergman, J. Blake, C. Blaschke, A. Cohen, F. Gannon, L. Grivell, U. Hahn, W. Hersh, L. Hirschman, L. J. Jensen, M. Krallinger, B. Mons, S. I. O'Donoghue, M. C. Peitsch, D. Rebholz-Schuhmann, H. Shatkay and A. Valencia, "Text mining for biology--the way forward: opinions from leading scientists," *Genome Biol.*, vol. 9 Suppl 2, p. S7, 2008.

[11] Y. Gao, J. Kinoshita, E. Wu, E. Miller, R. Lee, A. Seaborne, S. Cayzer and T. Clark, "SWAN: A distributed knowledge infrastructure for Alzheimer disease research," *Web Semant.,* vol. 4, no. 3, pp. 222-228, 2006.

[12] J. Pasternak, An Introduction to Human Molecular Genetics: Mechanisms of Inherited Diseases, John Wiley \& Sons, 1999.

[13] U. Broeckel and N. J. Schork, "Identifying genes and genetic variation underlying human diseases and complex phenotypes via recombination mapping," *J. Physiol. (Lond.),* vol. 554, no. Pt 1, pp. 40-45, 2004.

[14] K.-I. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal and A.-L. Barabási, "The human disease network," *Proceedings of the National Academy of Sciences,* vol. 104, no. 21, pp. 8685-8690, 2007.

[15] X. Zhang, R. Zhang, Y. Jiang, P. Sun, G. Tang, X. Wang, H. Lv and X. Li, "The expanded human disease network combining protein-protein interaction information.," *Eur J Hum Genet,* vol. 19, no. 7, pp. 783-8, 2011.

[16] C. Blaschke, L. Hirschman and A. Valencia, "Information extraction in molecular biology," *Brief. Bioinformatics,* vol. 3, no. 2, pp. 154-165, 2002.

[17] H. H. T. A. T. T. Ono T., "Automated extraction of information on protein-protein interactions from the biological literature," *Bioinformatics,* vol. 12, pp. 155-161, 2001.

[18] Y. Hao, X. Zhu, M. Huang and M. Li, "Discovering patterns to extract protein\–protein interactions from the literature: Part II," *Bioinformatics,* vol. 21, no. 15, pp. 3294-3300, 2005.

[19] H. W. Chun, Y. Tsuruoka, J. D. Kim, R. Shiba, N. Nagata, T. Hishiki and J. Tsujii, "Extraction of gene-disease relations from Medline using domain dictionaries and machine learning," *Pac Symp Biocomput,* pp. 4-15, 2006.

[20] J. J. Berman, "Pathology abbreviated: a long review of short terms," *Arch. Pathol. Lab. Med.,* vol. 128, no. 3, pp. 347-352, 2004.

[21] L. Hirschman, A. A. Morgan and A. S. Yeh, "Rutabaga by any other name: extracting biological names," *J. of Biomedical Informatics,* vol. 35, no. 4, pp. 247-259, 2002.

[22] W. J. Wilbur, G. F. Hazard, G. Divita, J. G. Mork, A. R. Aronson and A. C. Browne, "Analysis of biomedical text for chemical names: a comparison of three methods," *Proc AMIA Symp,* pp. 176-180, 1999.

[23] S. N. Swanson D., "Assessing a gap in the biomedical literature: magnesium deficiency and neurologic disease.," *Neurosci Res Commun,* vol. 15, pp. 1-9, 1994.

[24] P. Srinivasan and B. Libbus, "Mining MEDLINE for Implicit Links between Dietary Substances and Diseases," *Bioinformatics,* vol. 20, pp. 290-296, 2004.

[25] T. K. Landauer and S. T. Dutnais, "A solution to Plato’s problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge," *Psychological review,* pp. 211-240, 1997.

[26] M. W. Berry and M. Browne, Understanding search engines: mathematical modeling and text retrieval, Society for Industrial and Applied Mathematics, 1999.

[27] M. Yeasin, H. Malempati, R. Homayouni and M. S. Sorower, "A systematic study on latent semantic analysis model parameters for mining biomedical literature.," *BMC Bioinformatics,* vol. 10, no. S-7, 2009.

[28] M. P. Papazoglou, Web Services: Principles and Technology, Pearson, Prentice Hall, 2008.

[29] A. Mani and A. Nagarajan, *Understanding quality of service for Web services,* 2002.

[30] V. Abedi, R. Zand, M. Yeasin and F. E. Faisal, "An automated framework for hypotheses generation using literature," *BioData Min,* vol. 5, no. 1, p. 13, 2012.

[31] P. L. Antignani, "Treatment of chronic peripheral arterial disease," *Curr Vasc Pharmacol,* vol. 1, no. 2, pp. 205-216, 2003.

[32] E. D. Dominicis, M. Boschello, G. Trevisan and R. D. Nardis, "Threatened paradoxical embolism: its direct visualization by two-dimensional echocardiography," *G Ital Cardiol,* vol. 25, no. 6, pp. 733-736, 1995.

[33] K. E. Kirkland, K. Kirkland, W. J. Many and T. A. Smitherman, "Headache among patients with HIV disease: prevalence, characteristics, and associations," *Headache,* vol. 52, no. 3, pp. 455-466, 2012.

[34] C. Infante-Rivard, E. Olson, L. Jacques and P. Ayotte, "Drinking water contaminants and childhood leukemia," *Epidemiology,* vol. 12, no. 1, pp. 13-19, 2001.

[35] A. D. Mosnaim, R. Abiola, M. E. Wolf and L. C. Perlmuter, "Etiology and risk factors for developing orthostatic hypotension," *Am J Ther,* vol. 17, no. 1, pp. 86-91, 2010.

[36] R. ROBILLARD, J. L. RIOPELLE, L. ADAMKIEWICZ, G. TREMBLAY and J. GENEST, "Pulmonary complications during treatment with hexamethonium," *Can Med Assoc J,* vol. 72, no. 6, pp. 448-451, 1955.

[37] D. E. Stableforth, "Chronic lung disease. Pulmonary fibrosis," *Br J Hosp Med,* vol. 22, no. 2, pp. 132-135, 1979.

[38] A. Brettner, E. R. Heitzman and W. G. Woodin, "Pulmonary complications of drug therapy," *Radiology,* vol. 96, no. 1, pp. 31-38, 1970.

[39] L. T. Malaia, A. A. Shalimov, S. A. Dushanin, M. M. Liashenko and V. V. Zverev, "Catheterization of veins and selective angiopulmonography in comparison with several indices of the functional state of the external respiratory apparatus and blood circulation during chronic lung diseases," *Kardiologiia,* vol. 7, no. 7, pp. 112-119, 1967.

[40] F. J. COCKERSOLE and W. W. PARK, "Hexamethonium lung; report of a case associated with pregnancy," *J Obstet Gynaecol Br Emp,* vol. 63, no. 5, pp. 728-734, 1956.

[41] P. A. S. J. G. L. K. E. S. G. P. G. M. A. d. Q. D. B. C. U. D. L. U. B. S. D. N. H. K. H. T. J. H. P. T. H. C. N. R. Wollmer MA, "Genetic polymorphisms and cerebrospinal fluid levels of tissue inhibitor of metalloproteinases 1 in sporadic Alzheimer's disease," *Psychiatr Genet,* vol. 12, no. 3, pp. 155-160, 2002.

[42] M. S. S. M. E. D. R. R. Wang L, "Bacterial inclusion bodies contain amyloid-like structure," *PLoS Biol.,* vol. 6, no. 8, p. e195, 2008.

[43] E. Renvoize, "ABO and Rhesus blood groups in Alzheimer's disease," *Age Ageing,* vol. 14, no. 1, pp. 43-45, 1985.

[44] M. G. P. G. G. A. C. X. R. J. H. H. B. S. S. S. Lazarov O, "Impairments in fast axonal transport and motor neuron deficits in transgenic mice expressing familial Alzheimer's disease-linked mutant presenilin 1," *J Neurosci.,* vol. 27, no. 26, pp. 7011-7020, 2007.

[45] K. R. L. B. V. B. T. D. Bird TD, "Lipomembranous polycystic osteodysplasia (brain, bone, and fat disease): a genetic cause of presenile dementia," *Neurology,* vol. 33, no. 1, pp. 81-86, 1983.

[46] P. S. K. J. N. D. B. D. L. S. P. J. N. R. A. C. I. R. J. P. L. M. H. R. Li D, "Mutations of presenilin genes in dilated cardiomyopathy and heart failure," *Am J Hum Genet.,* vol. 79, no. 6, pp. 1030-1039, 2006.

[47] J. D. S. A. B. K. I. R. Hansis C, "The gene for the Alzheimer-associated beta-amyloid-binding protein (ERAB) is differentially expressed in the testicular Leydig cells of the azoospermic by w/w(v) mouse.," *Eur J Biochem.,* vol. 258, no. 1, pp. 53-60, 1998.

[48] E. W. Askanas V, "Sporadic inclusion-body myositis and its similarities to Alzheimer disease brain. Recent approaches to diagnosis and pathogenesis, and relation to aging," *Scand J Rheumatol.,* vol. 27, no. 6, pp. 389-405, 1998.

Appendices

Table 5. Medical Subject Headings selected by expert and the corresponding MeSH tree number

Medical Subject Headings	Tree Number
Bacterial Infections	C01.252
Bacteremia	C01.252.100
Central Nervous System Bacterial Infections	C01.252.200
Endocarditis, Bacterial	C01.252.300
Eye Infections, Bacterial	C01.252.354
Fournier Gangrene	C01.252.377
Gram-Negative Bacterial Infections	C01.252.400
Gram-Positive Bacterial Infections	C01.252.410
Pneumonia, Bacterial	C01.252.620
Sexually Transmitted Diseases, Bacterial	C01.252.810
Skin Diseases, Bacterial	C01.252.825
Spirochaetales Infections	C01.252.847
Mycoses	C01.703
Zoonoses	C01.908
Arbovirus Infections	C02.081
Bronchiolitis, Viral	C02.109
Central Nervous System Viral Diseases	C02.182
DNA Virus Infections	C02.256
Eye Infections, Viral	C02.325
Fatigue Syndrome, Chronic	C02.330
Hepatitis A	C02.440.420
Hepatitis B	C02.440.435
Hepatitis C	C02.440.440
Hepatitis D	C02.440.450
Hepatitis E	C02.440.470
Opportunistic Infections	C02.597
Pneumonia, Viral	C02.705
RNA Virus Infections	C02.782
Sexually Transmitted Diseases	C02.800
Skin Diseases, Viral	C02.825
Slow Virus Diseases	C02.839
Tumor Virus Infections	C02.928
Viremia	C02.937
Zoonoses	C02.968
Central Nervous System Parasitic Infections	C03.105

Table 5. Medical Subject Headings selected by expert and the corresponding MeSH tree number

Medical Subject Headings	Tree Number
Eye Infections, Parasitic	C03.300
Helminthiasis	C03.335
Intestinal Diseases, Parasitic	C03.432
Liver Diseases, Parasitic	C03.518
Lung Diseases, Parasitic	C03.582
Mesomycetozoea Infections	C03.600
Parasitemia	C03.695
Protozoan Infections	C03.752
Skin Diseases, Parasitic	C03.858
Zoonoses	C03.908
Neoplasms	C04
Bone Diseases	C05.116
Cartilage Diseases	C05.182
Fasciitis	C05.321
Foot Deformities	C05.330
Hand Deformities	C05.390
Jaw Diseases	C05.500
Joint Diseases	C05.550
Muscular Diseases	C05.651
Musculoskeletal Abnormalities	C05.660
Rheumatic Diseases	C05.799
Digestive System Diseases	C06
Stomatognathic Diseases	C07
Respiratory Tract Diseases	C08
Otorhinolaryngologic Diseases	C09
Autoimmune Diseases of the Nervous System	C10.114
Autonomic Nervous System Diseases	C10.177
Encephalomyelitis	C10.228.440
High Pressure Neurological Syndrome	C10.228.470
Movement Disorders	C10.228.662
Spinal Cord Diseases	C10.228.854
Chronobiology Disorders	C10.281
Cranial Nerve Diseases	C10.292
Demyelinating Diseases	C10.314
Nervous System Malformations	C10.500
Nervous System Neoplasms	C10.551
Neurocutaneous Syndromes	C10.562
Neurodegenerative Diseases	C10.574

Table 5. Medical Subject Headings selected by expert and the corresponding MeSH tree number

Medical Subject Headings	Tree Number
Neuromuscular Diseases	C10.668
Botulism	C10.720.150
Heavy Metal Poisoning, Nervous System	C10.720.475
MPTP Poisoning	C10.720.606
Neuroleptic Malignant Syndrome	C10.720.737
Sleep Disorders	C10.886
Trauma, Nervous System	C10.900
Genital Diseases, Male	C12.294
Urogenital Abnormalities	C12.706
Urogenital Neoplasms	C12.758
Urologic Diseases	C12.777
Kidney Diseases	C12.777.419
Urinary Bladder Diseases	C12.777.829
Urinary Tract Infections	C12.777.892
Urolithiasis	C12.777.967
Female Urogenital Diseases	C13.351
Pregnancy Complications	C13.703
Cardiovascular Abnormalities	C14.240
Cardiovascular Infections	C14.260
Vascular Diseases	C14.907
Aortic Diseases	C14.907.109
Arterial Occlusive Diseases	C14.907.137
Arteriovenous Malformations	C14.907.150
Arteritis	C14.907.184
Cerebrovascular Disorders	C14.907.253
Diabetic Angiopathies	C14.907.320
Hyperemia	C14.907.474
Hypertension	C14.907.489
Hypotension	C14.907.514
Myocardial Ischemia	C14.907.585
Peripheral Vascular Diseases	C14.907.617
Vasculitis	C14.907.940
Venous Insufficiency	C14.907.952
Hematologic Diseases	C15.378
Lymphatic Diseases	C15.604
Connective Tissue Diseases	C17.300
Acid-Base Imbalance	C18.452.076
Calcium Metabolism Disorders	C18.452.174

Table 5. Medical Subject Headings selected by expert and the corresponding MeSH tree number

Medical Subject Headings	Tree Number
DNA Repair-Deficiency Disorders	C18.452.284
Glucose Metabolism Disorders	C18.452.394
Iron Metabolism Disorders	C18.452.565
Lipid Metabolism Disorders	C18.452.584
Malabsorption Syndromes	C18.452.603
Metabolic Syndrome X	C18.452.625
Metabolism, Inborn Errors	C18.452.648
Mitochondrial Diseases	C18.452.660
Phosphorus Metabolism Disorders	C18.452.750
Porphyrias	C18.452.811
Proteostasis Deficiencies	C18.452.845
Wasting Syndrome	C18.452.915
Water-Electrolyte Imbalance	C18.452.950
Hypervitaminosis A	C18.654.301
Infant Nutrition Disorders	C18.654.422
Malnutrition	C18.654.521
Overnutrition	C18.654.726
Wasting Syndrome	C18.654.940
Adrenal Gland Diseases	C19.053
Bone Diseases, Endocrine	C19.149
Diabetes Mellitus	C19.246
Dwarfism	C19.297
Gonadal Disorders	C19.391
Parathyroid Diseases	C19.642
Pituitary Diseases	C19.700
Thyroid Diseases	C19.874
Autoimmune Diseases	C20.111
Addison Disease	C20.111.163
Antiphospholipid Syndrome	C20.111.197
Arthritis, Rheumatoid	C20.111.199
Glomerulonephritis, IGA	C20.111.525
Hepatitis, Autoimmune	C20.111.567
Lupus Erythematosus, Systemic	C20.111.590
Purpura, Thrombocytopenic, Idiopathic	C20.111.759
Thyroiditis, Autoimmune	C20.111.809
Hypersensitivity	C20.543
DNA Damage	C21.111
Occupational Diseases	C21.447

Table 5. Medical Subject Headings selected by expert and the corresponding MeSH tree number

Medical Subject Headings	Tree Number
Agricultural Workers' Diseases	C21.447.080
Dermatitis, Occupational	C21.447.270
Inert Gas Narcosis	C21.447.426
Persian Gulf Syndrome	C21.447.653
Pneumoconiosis	C21.447.800
Poisoning	C21.613
Argyria	C21.613.068
Arsenic Poisoning	C21.613.097
Bites and Stings	C21.613.127
Cadmium Poisoning	C21.613.165
Carbon Tetrachloride Poisoning	C21.613.177
Fluoride Poisoning	C21.613.380
Gas Poisoning	C21.613.455
Lead Poisoning	C21.613.589
Manganese Poisoning	C21.613.618
Mercury Poisoning	C21.613.647
Mycotoxicosis	C21.613.680
Neurotoxicity Syndromes	C21.613.705
Plant Poisoning	C21.613.756
Psychoses, Substance-Induced	C21.613.809
Water Intoxication	C21.613.932
Preconception Injuries	C21.676
Alcohol-Related Disorders	C21.739.100
Amphetamine-Related Disorders Amphetamine-Related Disorders OR Cocaine-Related Disorders OR Marijuana Abuse	C21.739.225 OR [C21.739.300] OR [C21.739.635]
Tobacco Use Disorder	C21.739.912
Wounds and Injuries	C21.866
Arrhythmias, Cardiac	C23.550.073
Ascites	C23.550.081
Azotemia	C23.550.145
Dehydration	C23.550.274
Emphysema	C23.550.325
Hemorrhage	C23.550.414
Hyperammonemia	C23.550.421
Hyperbilirubinemia	C23.550.429
Hyperuricemia	C23.550.449
Hypovolemia	C23.550.455
Leukocytosis	C23.550.526

Table 5. Medical Subject Headings selected by expert and the corresponding MeSH tree number

Medical Subject Headings	Tree Number
Menstruation Disturbances	C23.550.568
Muscle Weakness	C23.550.695
Nerve Degeneration	C23.550.737
Body Temperature Changes	C23.888.119
Body Weight	C23.888.144
Cardiac Output, High	C23.888.176
Cardiac Output, Low	C23.888.192
Chills	C23.888.208
Cyanosis	C23.888.248
Eye Manifestations	C23.888.307
Fatigue	C23.888.369
Flushing	C23.888.388
Heart Murmurs	C23.888.447
Hot Flashes	C23.888.475
Hypergammaglobulinemia	C23.888.512
Intermittent Claudication	C23.888.531
Mobility Limitation	C23.888.550
Pain	C23.888.646
Inorganic Chemicals	D01
Organic Chemicals	D02
Heterocyclic Compounds	D03
Polycyclic Compounds	D04
Macromolecular Substances	D05
Complex Mixtures	D20
Biomedical and Dental Materials	D25
Defense Mechanisms	F01.393
Human Development	F01.525
Personality	F01.752
Appetite	F02.830.071
Sleep	F02.830.855
Stress, Psychological	F02.830.900
Religion and Psychology	F02.880
Resilience, Psychological	F02.940
Body Fat Distribution	G03.180.134
CD4-CD8 Ratio	G12.248
Immunocompetence	G12.460
Immunocompromised Host	G12.470
Sweating	G13.750.829.855

Table 5. Medical Subject Headings selected by expert and the corresponding MeSH tree number

Medical Subject Headings	Tree Number
Skin Temperature	G13.750.844
Refraction, Ocular	G14.760
Vision Disparity	G14.930
Visual Acuity	G14.940
Quality of Life	I01.800
Culture	I01.880.143
Hierarchy, Social	I01.880.298
Minority Groups	I01.880.371
Social Class	I01.880.552
Social Welfare	I01.880.787
Socialization	I01.880.813
Socioeconomic Factors	I01.880.840
Education	I02
Human Activities	I03
Exercise	I03.350
Leisure Activities	I03.450
Physical Fitness	I03.621
Travel	I03.883
Household Products	J01.516
Alcoholic Beverages	J02.200.100
Carbonated Beverages	J02.200.300
Coffee	J02.200.325
Milk	J02.200.700
Milk Substitutes	J02.200.712
Mineral Waters	J02.200.806
Tea	J02.200.900
Food	J02.500
Age Groups	M01.060
Alcoholics	M01.066
Athletes	M01.072
Caregivers	M01.085
Child, Abandoned	M01.097
Child, Exceptional	M01.102
Child of Impaired Parents	M01.106
Child, Orphaned	M01.108
Child, Unwanted	M01.111
Consultants	M01.120
Crime Victims	M01.135

Table 5. Medical Subject Headings selected by expert and the corresponding MeSH tree number

Medical Subject Headings	Tree Number
Criminals	M01.142
Disabled Persons	M01.150
Drug Users	M01.169
Emigrants and Immigrants	M01.189
Homebound Persons	M01.276
Homeless Persons	M01.325
Medically Uninsured	M01.385
Prisoners	M01.729
Refugees	M01.755
Single Person	M01.785
Students	M01.848
Terminally Ill	M01.873
Socioeconomic Factors	N01.824
Environment	N06.230

```
Initialize Heading List
Initialize Look Up Array List

For each Heading
    if relevance score with corresponding input query >= Threshold
    {
        add Heading to Heading List;
        add query number to corresponding Look Up Array List;
    }
End;

Writing the JSON file

Step 1. Write all elements in the Heading List to nodes list structure.
Step 2. Write all elements in the input query list to nodes list structure.
Step 3. For each element in the Heading List
    {
        Generate source target links from corresponding look up list;
    }
    End;
```

Algorithm 2. Generating graph data structure from relevance score given a threshold value.

Table 6. Associated headings for lymphoma, parkinson and migraine and their respective scores

Query node	Associated Headings	Score
Lymphoma	Tumor Virus Infections	0.7283
	DNA Virus Infections	0.6099
	Lymphatic Diseases	0.4247
	Hypergammaglobulinemia	0.4086
	Opportunistic Infections	0.3561
	Skin Diseases, Viral	0.2927
	Bacteremia	0.1639
	Proteostasis Deficiencies	0.1564
	Stomatognathic Diseases	0.1521
	Otorhinolaryngologic Diseases	0.1511
	Intestinal Diseases, Parasitic	0.0215
	Homeless Persons	0.0187
	Hepatitis C	0.0164
	Culture	0.0156
	Age Groups	0.0152
	Aortic Diseases	0.0133
	Personality	0.0129
	Cerebrovascular Disorders	0.0123
	Parathyroid Diseases	0.0122
	Cadmium Poisoning	0.012
	Hepatitis B	0.0112
	Genital Diseases, Male	0.0105
	Pneumonia, Bacterial	0.0104
Parkinson	MPTP Poisoning	0.5658
	Neurodegenerative Diseases	0.5598
	Movement Disorders	0.5417
	Manganese Poisoning	0.5237
	Heavy Metal Poisoning, Nervous System	0.4538
	Neuroleptic Malignant Syndrome	0.3299
	Chronobiology Disorders	0.243
	Neuromuscular Diseases	0.2363
	Muscular Diseases	0.1102
	Neurotoxicity Syndromes	0.0573
	Homeless Persons	0.0454
	Socialization	0.0357
	Cadmium Poisoning	0.0298
	Fatigue	0.0248
	Trauma, Nervous System	0.0235

Table 6. Associated headings for lymphoma, parkinson and migraine and their respective scores

Query node	Associated Headings	Score
Parkinson	Personality	0.0228
	Hypertension	0.0194
	Quality of Life	0.019
	Musculoskeletal Abnormalities	0.0178
	Spinal Cord Diseases	0.0163
	Hepatitis C	0.0158
	Occupational Diseases	0.0134
	Calcium Metabolism Disorders	0.0126
	Nervous System Malformations	0.0123
	Diabetic Angiopathies	0.01
Migraine	Coffee	0.689253
	Tea	0.591746
	Sexually Transmitted Diseases, Bacterial	0.286119
	Cranial Nerve Diseases	0.280735
	Spirochaetales Infections	0.274333
	Mycotoxicosis	0.263275
	Eye Manifestations	0.076593
	Age Groups	0.044755
	Defense Mechanisms	0.039117
	Socioeconomic Factors	0.036901
	Trauma, Nervous System	0.025156
	Stress, Psychological	0.023227
	Hemorrhage	0.022374
	Alcohol-Related Disorders	0.020972
	Demyelinating Diseases	0.020954
	Hepatitis A	0.018597
	Central Nervous System Bacterial Infections	0.017044
	DNA Virus Infections	0.0151
	Peripheral Vascular Diseases	0.01285
	Religion and Psychology	0.012325
	Leisure Activities	0.011235
	Body Temperature Changes	0.010895
	Aortic Diseases	0.01057
	Jaw Diseases	0.010016