11-19-2013

Query Based Sampling and Multi-Layered Semantic Analysis to find Robust Network of Drug-Disease Associations

Karththikka Ramani Muthukuri

QUERY-BASED SAMPLING AND MULTI-LAYERED SEMANTIC ANALYSIS TO FIND ROBUST NETWORK OF DRUG-DISEASE ASSOCIATIONS

by

Karththikka Ramani Muthukuri

A Thesis

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Master of Science

Major: Bioinformatics

The University of Memphis

December, 2013

Acknowledgments

I would like to express my deep gratitude to my supervisor Dr. Mohammed Yeasin for his support, encouragement, valuable and constructive suggestions during the planning and the development of this work. His willingness to provide his valuable time so generously is much appreciated. I would also like to express my thanks to Dr. Ramin Homayouni for his guidance. Friday seminars arranged by him shaped me well with the subject knowledge to go further in my research.

Also many thanks are extended to my teachers here at the University of Memphis and special thanks to my committee member Dr. Xiangen Hu for his guidance and suggestions. Special thanks are also extended to Mrs. Tallulah Campbell and Mrs. Becky Ward for their continuous help in the department throughout my graduate studies.

I would also like to thank all the present and past members of the CVPIA lab for their friendship. Special thanks go to Fazle Elahi Faisal, Vida Abedi, Hossein Taghizad, Pratiksha Subedi, and Pouya Bashivan for their valuable comments and suggestions.

Last but not the least I would like to extend my thanks to my dear husband Chidambaram Ramanathan for his continuous encouragement and support all through these years and also my little prince Aashish Chidamabaram for his great patience.

Abstract

Muthukuri, Karththikka Ramani, M.S. The University of Memphis. December 2013. Query-based Sampling and Multi-layered Semantic Analysis to find Robust Network of Association between Drugs and Diseases. Major Professor: Dr. Mohammed Yeasin.

This thesis presents the design and implementation of a system to discover semantically related networks of diseases-drugs associations, called DDNet, from medical literature. A fully functional DDNet can be transformative in identification of "drug targets" and may open new avenues for "drug repositioning" in clinical and translational research. In particular, a local latent semantic analysis (LLSA) was introduced to implement a system that is efficient, scalable and relatively free from systemic bias. In addition, a query-based sampling was introduced to find representative samples from the "ocean of data" to build model that is relatively free from "garbage-in garbage-out" syndrome. Also the concept of mapping ontologies was adopted to determine relevant results and reverse ontology mapping were used to create a network of associations. In addition, a Web service application was developed to query the system and visualize the computed network of associations in a form that is easy to interact. A pilot study was conducted to evaluate the performance of system using both subjective and objective measures. The PharmGKB was used as a gold standard and the PR curve was obtained from a large number of queries at different recall points. Empirical analyses suggest that DDNet is robust, relatively stable and scalable over traditional Global LSA model.

Table of Contents

Chapter		Page
1	**Introduction**	1
	Goals and Objectives	9
2	**Background**	11
3	**Methods**	14
	Drugs/Chemicals Se lection	
	Clustering the biological concept for every model based on concept size	16
	Clustering the biological concept for every model based on concept topic	17
	Data Extraction	19
	Data extraction for every local space	22
	Query Based Sampling	24
	Fuzzy c means clustering	27
	Systemic Bias	29
	Balancing clusters to alleviate bias	30
	Local spaces	31
	Encoding matrix from Local Model Generation And Dictionary reduction	33
	Complete LSA model from local models created	33
	Combining Encoding matrices of local models	33
	Relevance model	35
	DDNet: A Drug-Disease Interaction framework and PubMed Link Tool	36
	Implementation details using software languages	39
	Efficiency of the framework as a result of Pre-computed Results	41
	Network of Drug-Disease association from DDNet using MeSH hierarchal code	42
4	**Results and Discussion**	45
	Drug-Disease Association	45
	Efficiency	47
	Systeimic Bias	47
	Scalability	49
	Robustness	49
	Query Selection for validation and evaluation	50
	PR Curve and Analysis	53
	Mean Average Precision	56

Comparison of the performance of Global LSA and Local LSA model 57
Performance Evaluation based on the degree of Relevancy of retrieved results 62
Highly Relevant Set 62
Reasonably Relevant Set 63
Poor Relevant Set 64

5 Conclusions 67

References 69

Appendices 72

A. Associated drugs/Chemicals for 20 selected queries for Local LSA model with sampled information (Proposed Methodology) 72
B. Associated drugs/Chemicals for 20 selected queries for Global LSA model 81
C. Associated drugs/Chemicals for 20 selected queries for Local LSA model without sampled information 83
D. Cosine values for the associated drugs/chemicals for 20 Selected queries for the proposed LLSA model in the retrieved order 92

List of Figures

Figure		Page
1	Screeen shot showing the numbder of publications in PubMed.	2
2	Screen shot showing the retrieved publications from PubMed for the query “Alzheimer disease”.	3
3	Screen shot showing the retrieved publications from iPubMed and from a network.	4
4	Block diagram representing the workflow to create the literature mining drug disease network using the proposed local LSA models.	14
5	Selected drugs from MeSH categories.	18
6	MySQL database design.	22
7	Pictorial representation describing the balance of local spaces with positive and negative samples of abstracts.	24
8	Diagrammatic representation of tf matrix generation for D01 local space.	27
9	Diagrammatic representation of tf matrix generation for D01 local space after clustering.	28
10	Block diagram showing the steps taken to load eacl local space with unbiased clusters.	29
11	Graph showing varied number of abstracts for every cluster formed.	30
12	Diagrammatic representation of tf matrix generation for D01 local space after redistribution of abstracts in clustering.	31
13	Block diagram showing the creation of final D01 local space (D01 MeSH category); same procedure is done for other 3 local spaces too.	32
14	Encoding matrices U1, U2, U3, U4 placed parallel to form the global encoding matrix; Term 1, … Term 40466 are dictionary terms.	34
15	User interface of DDNet to enter queries.	37

16	The display of ranked Associated Drugs/Chemicals for the user query Alzheimer disease.	38
17	Network of Associations of Drugs/Chemicals for diseases Alzheimer disease and myocardial infarction derived from MeSH hierarchy.	43
18	Graphical representation of the distribution of cosine values for the semantically associated Drugs/Chemicals for the Query Alzheimer disease; shown for both Global and Local LSA models.	48
19	Diagrammatic representation for selection of queries to validate the model.	51
20	Queries categorized under heart related diseases.	52
21	Queries categorized under brain related diseases.	52
22	Queries categorized under cancer related diseases.	52
23	Queries categorized under lung related diseases.	53
24	Averaged 11 point Precision/Recall graph plotted across selected 20 queries. The results were obtained from DDNet.	54
25	Averaged 11-point Precision/Recall graph plotted across selected 20 queries for Local LSA model and Global LSA model.	57
26	Averaged 11-point Precision/Recall graph plotted across selected 20 queries for Local LSA model which has representative information and local LSA model which does not have representative information.	60
27	Averaged 11-point Precision/Recall graph plotted across selected 20 queries for LLSA model with representative information, LLSA model with all possible representative information and Global LSA model.	61
28	Screen shot showing the highly relevant set with Vitamin E at ranks 5 for the query Alzheimer disease.	63
29	Screen shot showing the reasonably relevant set with magnesium at rank 5 for the query Alzheimer disease.	64
30	Screen shot showing the poor relevant set with vitamin K at rank 11 for the query Alzheimer disease.	65

CHAPTER 1

Introduction

The modern advancement in high throughput technology and growth in research capacity resulted in producing large scale biological data. Unlike other research output, which may be preserved as equations or values, biological data are usually preserved in publications discussing them. That led to exponential growth of biomedical literature. This wealth of scholarly knowledge is of significant importance for researchers in making scientific discoveries and healthcare professionals in managing health-related matters. However, the acquisition of such information is becoming increasingly difficult due to its large volume and rapid growth.But, due to massive volume, there is a huge gap between the generated knowledge in the published literature and consumption of that knowledge. PubMed [1] is a free online resource which contains all Medline citations in the field of science. It serves as the primary tool for electronically searching and retrieving biomedical literature and, has currently approximately over 22 million abstracts [2]. This wealth of literature knowledge is of significant importance for researchers in making scientific discoveries and, healthcare professionals in managing health-related matters. Also, PubMed is up to date and it is queried by millions of users around the globe every day. As a negative side, PubMed frequently results in hundreds, thousands or even millions of publications for a single query, as an unranked list as shown in the Figure 1. These many retrievals for a single query is huge enough even for the expertise in the field to read through and seek the necessary information for what he/she is looking. Hence, knowledge gap still exists between the published literature and useful acquisition of it.

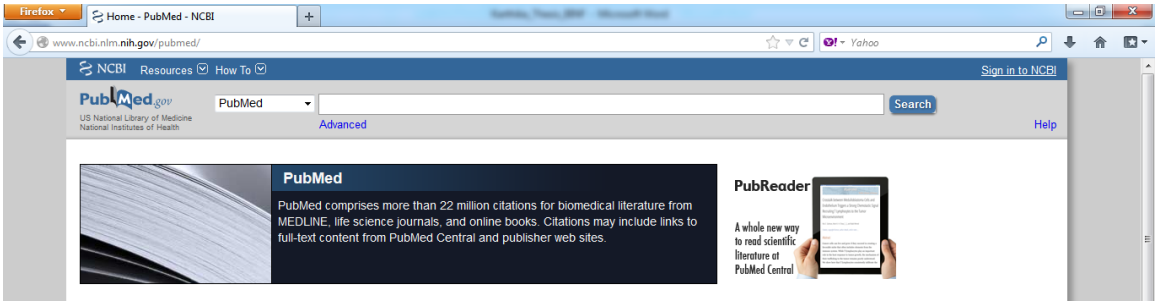

Figure 1 Screen shot showing the number of publications in PubMed.

Lot of web tools has been developed as PubMed search tools, complementary to PubMed. These web tools filtered down the PubMed retrieved publications based on their modeling. iPubMed [3] is one such PubMed search tool which narrows down PubMed results on the basis of user's relevance feedback. For instance, PubMed retrieved 67314 publications for query "Alzheimer Disease", shown in Figure 2 and iPubMed retrieved few hundred publications for the same query, shown in Figure 3.

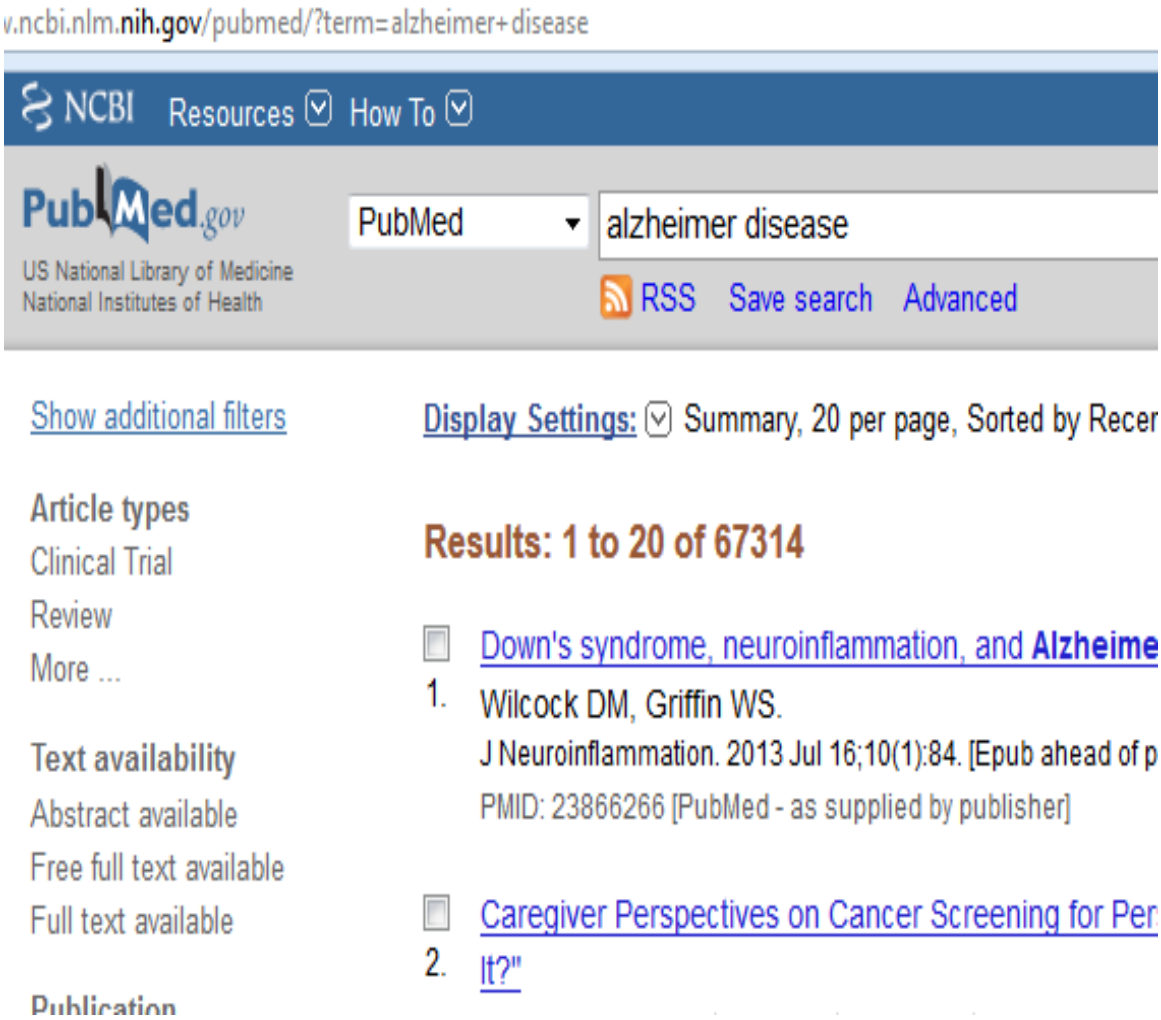

Figure 2 Screen shot showing the retrieved publications from PubMed for the query "Alzheimer disease".

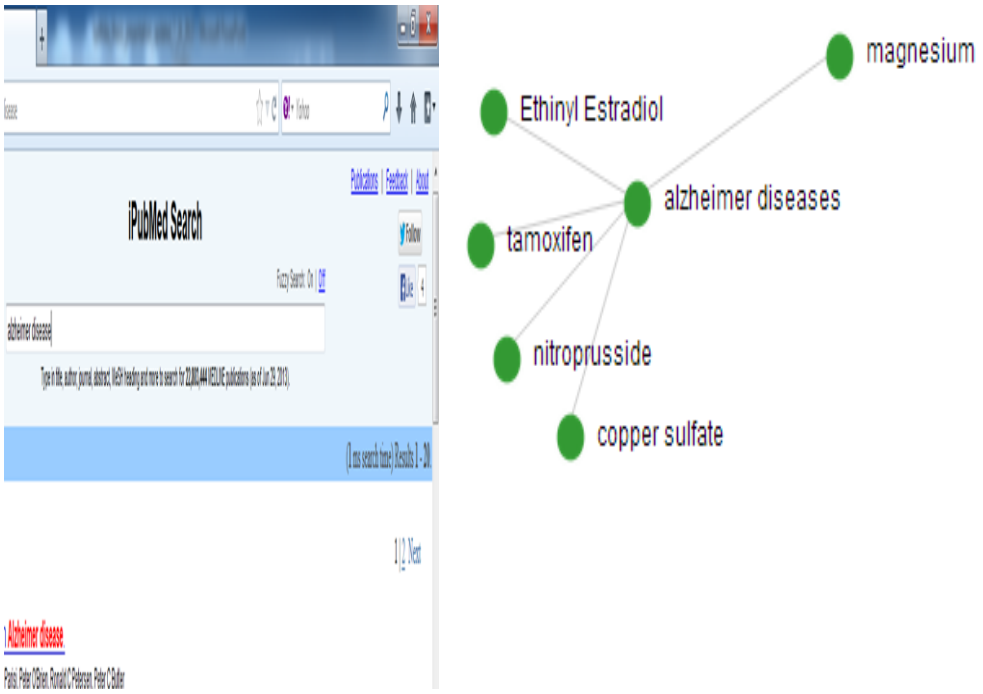

Figure 3 Screen shot showing the retrieved publications from iPubMed and from a network. iPubMed retrieves 20 pages of publications for the query “Alzheimer disease”, which is infeasble for the users to gather information. A network showing the crisp associations of biological concepts for the same query “Alzheimer disease”.

Though PubMed search tools narrowed down the results, still few hundreds will not bridge the knowledge gap. Also, the above mentioned web tools result flat list of unranked publications, relying on keyword based search. There is high probability that human readers may miss the significant information or may miss to gather important biological concept associations. Hence, if a web tool can retrieve network of biological concept associations, users will be benefited by obtaining precise information, as network gathers fragments of information. Hence network would be comparatively an efficient tool to bridge the knowledge gap. The inter-relationships between biological entities drive to create web tools capable of capturing the semantic association between them.

AliBaba [4] and PubMed-EX are geared towards semantic enrichment by identifying biomedical entities from the text. In addition, AliBaba also presents co-occurrence results in a graph. Hence, network of semantically related association of biological concepts can retrieve undiscovered associations, thereby aiding the researchers to generate new hypothesis.

With the nagatives of PubMed and PubMed search tools taken into account and advantages of network of semantic associations of biological concepts, this study is motivated. Developing an Drug-Disease ineraction framework DDNet to aid the process of Drug Repositioning by finding their candidates. Pharmaceutical research and development productivity has significantly declined in recent decade [2] based on the number of drugs approved and the amount of dollars spent for such an approval. To resolve this issue, Drug repositioning has been used as a strategy for decades to get drugs to more patients [5] and exploiting drug–disease relationships would be an efficient way for computational drug repositioning [6]. So, automated discovery of semantically related network of drug- diseases from medical literature can be transformative in clinical and translational research as well as improving health-care delivery by aiding drug repositioning. However the developed networks are mostly based on genomic expression profiles and protein connectivity maps [7, 8]. These networks could also be generated from literature data to cover most of the drug targets as well as to uncover the unknown potential associations.

To construct a drug-disease network from literature data, an efficient and less computational complexity informational retrieval technique is imperative. Latent Semantic Analysis (LSA) [9] is such a widely used semantic information retrieval

technique in the field of Bioinformatics. It identifies and retrieves direct as well as indirect associations by finding higher order co-occurrence of terms in the data. LSA has been shown to be extremely useful in information retrieval but drops its performance when applying to the whole document collection [10]. LSA transforms the original textual data into semantic space by capturing the implicit higher order structure in the association of words through SVD decomposition. When SVD is performed on the term document matrix from the whole data, it pays no attention to the class discrimination and places the documents from different categories near to each other in reduced semantic space. This results in poor performance of the system.

This thesis designs and implements an efficient, scalable, robust and relatively bias free drug-disease interaction framework DDNet. This web tool enhances the literature search by finding semantically related entities through the integration of local LSA and Query Based Sampling. Several parameters have to be taken care of to construct an adaptive, robust, efficient, bias free and scalable web tool of network of semantic Drug-Disease associations. First, data extraction that makes the web tool to be used for targeted audience. Second, domain specific dictionary that defines the global dynamic feature range for the model. Third, semantic analysis model to identify the relavant concepts for queries. Fourth, data driven thresholds to classify the retrieved results at different levels of associations. Finally, an user friendly interface to visualize the results based on the intensity of information need. All five mentioned parameters are critical to the success of web tools to meet the expectation of consumers with different needs and desires. Constructing an interaction network in the specific domain requires a dataset that is wide in its scope and can provide precise biological knowledge. Literature data is the

only source of knowledge with wide extent of information from different sources in the domain of biology. Hence, titles and abstracts of PubMed have been extracted for the selected biological concepts and utilized by the studies to extract the necessary associations. MeSH, controlled vocabulary thesaurus, maintained by NCBI, is used or dictionary creation by possible combination of terms.

Over the past few years, CVPIA Lab has been focused to bridge the knowledge gap by developing network of associations, based on concepts. Domain specific multi gram dictionary terms are created by ontology mapping of MeSH terms in ARIANA [11]. The results showed that quality of results are greatly enhanced by this dictionary. Also, multi gram dictionary, partly alleviates one of the drawbacks of semantic analysis, LSA, losing the biological meaning. Scalable network of drug-disease, requires a domain specific dictionary, out of which more drugs/chemicals can be incorporated into the underlying model. Drugs for this proposed study are derived from MeSH, as PubMed abstracts are annotated with MeSH keywords for easy search and retrieval. Hence, dictionary creation by ARIANA has been adopted for DDNet framework.

A web based tool, PharmNet is developed in CVPIA lab to explore the relationships between pharmaceutical factors such as cellular components, chemical compounds, biological factors, diseases, diagnosis, procedures etc. In this study, constructive research has been done to address one of the drawbacks of LSA. Biological entities for the model are selected from MeSH based on statistical analysis, to alleviate systemic bias problem of LSA. MeSH terms have a heirarichal tree structure with different levels. PharmNet conducted statistical studies to select terms such that they are not too general and not too specific. Too general terms have several publications and will

be retrieved by LSA model, with low association values against the query. On the other side, too specific terms have very few publications and will be retrieved with high association values against the query. One of the vital task of this Drug-Disease network study is to create a relatively bias free semantic model. Hence, selected drugs for PharmNet have been used for defining the feature range of this study. Additonally, care has been taken for the presence of selected drugs in PharmGKB [12], the gold standard, used for validating the drug-disease interaction tool.

DDNet has four major modules: (1) Local LSA models to create scalable and relatively bias free framework. As global or traditional LSA is applied to the entire abstract collection from different classes or categories, it captures irrelevant second order co-occurrence of terms from these varied classes; thereby, capturing misleading higher order semantic patterns of associations in the data. As a result, the retrieved information from the model may not be a precise or even the accurately extracted associations may not be resulted with confidence. Local LSA (LLSA) models are developed and implemented by localizing conceptual relevant entities (drugs in this study) into a separate models, thereby ensuring higher mutual independence between models. The computational complexity will not be high for LLSA models, as the matrices out of which models are generated will be of low dimension. Hence scalability issue can also be resolved by LLSA models. (2) Query Based Sampling (QBS) of abstacts was introduced to define each of the local spaces and thereby the models with representative samples of information. QBS can ensure the models to be free from garbage-in, garbage-out syndrome. (3) Pre-computed results for enhancing the effieiciency (4) Finally, an easy-to-use interface with proper visualization is developed, which is critical to the success of

a web tool. It would be able to retrieve the ranked results for any user query, based on the user's intensity of information need. In this way, the tool would meet the needs of consumers with diverse needs and desires. Relevance model was implemented, to provide range of services to users in biomedicine. It translates the ranked list of results into three categories of connections such as highly related, related and not related. A relevance model was also incorporated into this study, to make the Web tool, created from this study, to provide flexibility needed to serve a diverse range of users. DDNet provides the user the options of choosing highly relevant, reasonably relevant and poorly relevant results based on the intensity of information needs.

Goals and Objectives

Literature data contains redundant information which would degrade the robustness of the information retrieval; also the traditional LSA suffers from its own limitations of scalability and systemic bias. Query Based Sampling is to incorporate relevant information into the model and Local LSA (LLSA) models to address the limitations of traditional LSA model. The goal of this thesis is to develop a scalable, efficient, robust, unbiased, complete and generalized literature mining framework in the Pharmaceutical domain with the underlying proposed LLSA model to model network of semantically related drug-disease associations.

- Optimize the selection of drug entities to define the feature range for developing the model
- Localize the drug entities to each one of the model's region
- Develop the multi gram model to preserve the biological meaning of the terms by creating multi gram dictionary

- Define the regions of local LSA models by sampling and loading it with more relevant representative samples of textual data
- Generate the bias free Global LSA model by combining local models generated
- Develop an interface with the underlying model generated which constructs the drug disease network based on user's query
- Pre-compute the associated factors for all the possible user queries possible with the constraint of within the dictionary range
- Validate the network against the chosen Gold Standard PharmGKB

CHAPTER II

Background

Due to the voluminous biomedical literature there have been lot of effort in developing literature mining techniques by the research community. Shatkey [13] describes some of the literature mining techniques as an overview. Natural Language Processing and machine learning techniques have been applied to unstructured biological text and transform them into structured and computational form to analyze the functional concepts of biological compounds. Also, domain specific search engines have also been built to find the most relevant publications for the user's need.

More interest has been shown to find the associated biological concepts based on semantics rather than keyword based search. FACTA [14] is one such text search engine for MEDLINE abstracts, which retrieves the associated biological concepts based on user's query. It provides the results in a tabular format in the ranked order, where the ranking of the biological concepts are based on co-occurrence of statistics of terms with the user's query. PubMatrix [15] is a simple web tool that mines PubMed using couple of lists of terms and retrieves the co-occurrence terms. Chillibot [16] is content rich software which mines PubMed database to retrieve the relationships between genes, proteins or for any user's information need. The results are displayed graphically, as well as in the form of sentences containing the terms on which the user is interested to look for relationships. GeneIndexer [17] is a robust tool to retrieve and rank the genes based on user's phenotype, cell etc. Parsing is done on full text articles with the hypothesis that biological concepts occurring in the same sentence are somehow associated though biological process. AliBaba [4] is an interactive tool for graphical summarization of search results

extracted on the fly from PubMed query. It parses the abstracts that fit for a PubMed query and presents the extracted information for biological objects such as proteins, diseases and drugs and their relationships as a graphical network. MiSearch [18] is an adaptive literature search tool using implicit relevance feedback, helps users to rapidly find PubMed citations relevant to their specific interests.

Effort has been laid in Pharmagenomic literature mining as well. PharmGKB [19-21] is one such comprehensive resource for pharmacogenomics including impact of genetic variations on drug response, biological pathways, relationships between drugs, genes and diseases etc. It is thoroughly a knowledge base on pharmacogenes, their snps, pharmacokinetics and pharmacodynamic pathways to achieve personalized medicine. It contains data on genes (> 20000), diseases (> 3000) and drugs (> 2500), SNPs (450). Sentence level co-occurrence is used to mine and characterize the gene-drug relationships from PubMed abstracts with a recall of 51% and precision of 60% [22]. Semantic networks have been created with pharmacogenomics knowledge [23].

A reasonable number of databases has been to interpret the drug mechanism and their targets as well. DrugBank [24] is bioinformatics-cheminformatics databases that focus on molecular information about drugs and drug targets with hyperlinks to many other reliable databases. Comparative Toxicogenomics Database (TCD) [25] advances the understanding of the effects of environmental chemicals on human health where the researchers manually curated the relations between chemicals, genes and diseases.

Effort has also been put to create Local LSA models with their local regions consisting representative samples. T. Liu et al. proposed a Local Relevancy Weighted LSI method, which distributes the training documents into different classes according the

relevancy to that class and performs SVD separately. It assigned empirical weights to each local semantic space according to its contribution to the global space. There exists tradeoff between different sized local spaces. Large local spaces are capable of discriminating the documents sufficiently. But the model also may contain several non-relevant documents creating noise and systemic bias. On the other hand, small local spaces are less noisy, but observe lack of information in the documents. Hence, there is an issue of the class size parameter tuning.

CHAPTER III

Methods

To achieve the main goal of the study, which is to develop a scalable, unbiased and efficient drug disease interaction network, several important parameters need to be customized. The parameters include drugs selection for this study as input to the model, filtering of representative samples of information to define the local region of every model, generation of local LSA models, complete LSA model by grouping the local models, and relevance model for clustering the results based on their association values against the query, validating the network created against the appropriate Gold Standard, PharmGKB, a manually curated database. Figure 4 describes the procedures to ensure the quality of the network, with the underlying generated model.

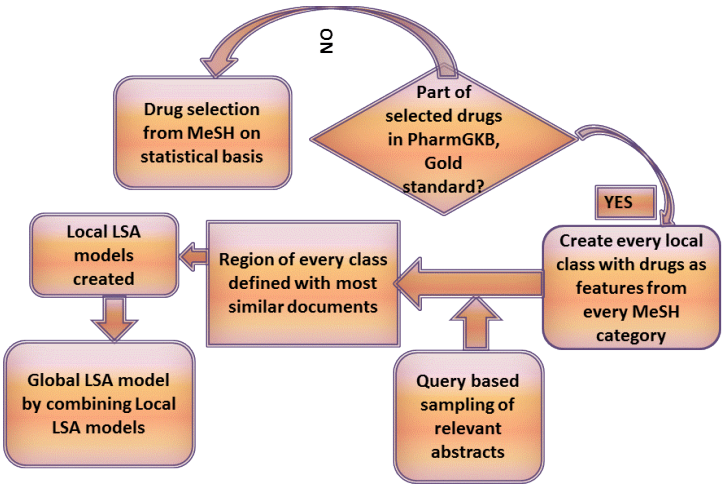

Figure 4 Block diagram representing the workflow to create the literature mining drug disease network using the proposed local LSA models.

Drugs/Chemicals selected based on statistical analsysis are checked for their presence in PhermGKB, the Gold Standard. Local spaces are created for those selected drugs by Query Based Sampling; LLSa models a re-created and integrated to form the complete semantic model.

Drugs/Chemicals selection

Selection of drugs/chemicals is an important criteria for semanic analysis, as a random selection would introduce redundancy into the model. PubMed database in NCBI is used to load the abstracts into the textual corpus from which the local models are generated. So, drugs for the models have been selected from MeSH, a controlled vocabulary database of U.S. National Library of Medicine in NCBI. Medline references in PubMed are cited with MeSH keywords for faster and informative searches in the PubMed database. The distinctive feature of MeSH database is that the terms are categorized with 16 categories and are organized in hierarchal tree structure. The hierarchical levels aid to identify too general and too specific MeSH terms. Terms near the root of the tree are considered too general and the terms near the leaves are considered too specific. For instance, the MeSH term “Alcohol” at the root level is too general with lot of branches and term “sugar Alcohols” at the leaf level is too specific.

For the model underlying the drug disease framework, MeSH category “Chemicals and Drugs” at the root level is chosen which has 16 sub categories at level 1. Specificity function is modeled earlier in PharmNet from our CVPIA lab to select MeSH entities that are neither too general nor too specific. It is because of the fact that too general drug terms will incorporate redundant information into the model and too specific terms will not be sufficient informative to create a complete model. Statistical studies

have been done on different levels, depth, documents ratio of MeSH etc. to select the pharmacological entities, in PharmNet. Drugs derived for PharmNet are incorporated for this study with the constraint that they are present in the gold standard, PharmGKB. Presence of selected drugs in PharmGKB is needed to evaluate the performance of the system. All drug terms form PharmGKB [21] complying within the statistical analysis are not chosen for this pilot study as they fall under different MeSH categories. Because it will increase the computational complexity of the process, when it is scaled with more local models. As the model created, as well as the network is scalable, it can be extended with added local models which may cover up more PharmGKB drug terms. Fifty three drugs have been selected for this pilot study from the above mentioned statistical analysis and PharmGKB's presence.

The drug entities selected on the above stated criteria are derived under 4 different subcategories at level 1 under the root level "Chemicals and Drugs category". The resulting derived subcategories are D01 "Inorganic Chemicals", D02 "Organic Chemicals", D03 "Heterocyclic Compounds and D04 "Polycyclic Compounds". D01 has 7 drugs, D02 has 20 drugs, D03 has 10 drugs and D04 has 16 drugs. The selected entities derived under different categories facilitate the creation of local models. The reason is that the drug entity from each category can define the dynamic feature range for every local model.

Clustering the biological concept for every model based on concept size

Different biological concepts have varied number of literature data from PubMed. Having those biological concepts (drugs/chemicals) in the same model will introduce systemic bias in the model. Concepts having voluminous textual data will be retrieved

with lower association values as weights will be distributed for too many terms and vice versa for the concepts with very few textual data. Though it is computationally feasible, the main drawback of this method lies in its inability to keep relevant concepts in a single cluster and separate irrelevant concepts into different ones.

Clustering the biological concepts for every local LSA model based on topic

In this way of grouping biological concepts to every model, relevant conceptual entities are grouped in a single model thereby ensuring higher mutual independence. So, wrong semantic capturing of terms will be greatly alleviated. But it may undergo the problem of systemic bias introduced by different sized concepts in the same model. If the systemic bias can be taken care, it would be a better way to define the concepts for every local model.

As the systemic bias will be solved (be explained in chapter 2 Methodology section), local spaces are created by clustering the selected drugs based on topic. Below Figure 5 shows the Drugs/Chemicals selected from MeSH based on statistical measures which are to be clustered based on topic.

D01 MeSH category Inorganic Chemicals	D02 MeSH category Organic Chemicals	D03 MeSH category Heterocyclic compounds	D04 MeSH category Polycyclic compunds
• Cisplatin • Lithium • Copper • Magnesium • Nitroprusside • Nitrous oxide • zinc	• Vitamin k • Troleanycophenolic acid • Mycophenolic acid • Memantine • Macrolides • Isoproterenol • Epinephrine • Dicloxacillin • Acetaminophen • Erythromycin • Idarubicin • Lovastatin • Pravastatin • Methdone • Ritonavir • Sirolimus • Tacrolimus • Curcumin • Metoprolol • tamoxifen	• Galantamine • Warfarin • Amoxicillin • Codeine • Morphine • Vincristine • Tacrine • Nicotine • Vitamin E	• Testosterone • Prednisone • Prednisolon • Norethinrone • Mifepristone • Levonogestrol • Hydrocortisone • Ethiny estrdiol • Estrone • Digoxin • Cyclosporine • Calcitriol • Budenosine • Beclamethasone • Aldosterone • triamcinolone

Figure 5 Selected drugs from MeSH categories.

Dictionary creation

Dictionary can be created from the corpus of the selected features but is computationally expensive as well as will result in too voluminous dictionary. That will again lead to computational and storage problems. Also, the dictionary terms have to be domain specific to make the model utilizable to targete audiences. Dictionary, if created from the corpus, will comprise of too general English vocabulary which will introduce noise into the model. With all these constraints, multi gram dictionary is created from MESH terms which will be in the specific biological domain [11]. One gram, two gram, three gram terms are created for dictionary to ensure the biological meaning preserving as LSA will treat each word independently which will likely to lose the meaning. It resulted in 40466 dictionary terms as of year 2013.

Data Extraction

Local spaces, out of which semantic models are to be generated and dictionary have been defined. Now, each of the local spaces have to be loaded with textual information. PubMed is the database from where the abstracts are downloaded through Entrez eutils programming utilities. As data to be extracted is of huge volume, effective and precise tool to serve the purpose is necessary.

An automated tool is developed to extract the necessary dataset from PubMed into a normalized database. The developed script is platform independent and is high computing linux server for maximum efficiency in data extraction. Several input parameters are configured in this scripted tool which are mentioned in a separate text file.

The parameters to be specified for electronic data extraction from PubMed are given below:

- URL: This is the URL of the system from where literature data needs to be extracted. We used the URL http://www.ncbi.nlm.nih/entrez/eutils is the url to be mentioned in the script from where the data is to be extracted
- Database: PubMed is the database in Entrez from where the data has to be extracted
- Starting year: starting year from when the published articles have to be extracted, depending on the amount of information need by the user has to be mentioned. I this study, almost all published articles have been extracted starting from 1950
- Ending year: Ending year also has to be indicated. 2012 is the year in this study
- Maximum number of articles: maximum number of articles to be extracted for each of the entity (Drugs/Chemicals for our case) needs to be indicated; and this parameter is set to be unlimited.
- Block size: The block size of articles fetched from PubMed at a time is also indicated. A block size of 200 is set according to the NCBI rule.

The tool is designed in such a way that it extracts data at slow speed during office hours and high speed during non-office hours, weekends and holidays. This is intentionally done to reduce the data extraction rate during office hours to ensure the safety of PubMed.

The dataset for the 53 factors is downloaded from PubMed and stored in MySQL database. The database construction is based on the following design (see Figure 6).

MySQL Database design:

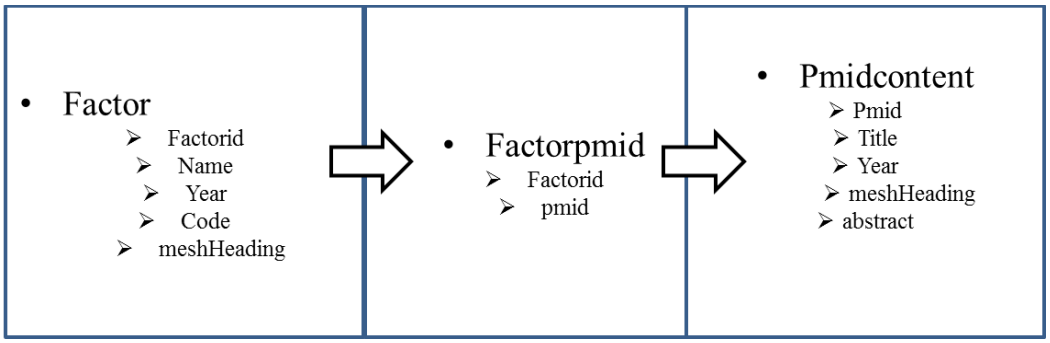

Figure 6 MySQL database design. Three tables are used to construct the database for the MeSH-based factors. Factor table contains 53 MeSH factors, field year in table factor is used to update the recent article for the entity in the database; Factopmid contains information need to link the factor to PubMed abstracts using PMIDs (unique identifier of PubMed abstracts); PMIDContent contains information about each abstract.

The database is designed in such a way that all three tables, factor, factorpmid and pmidcontent are interconnected. Factor and factorpmid are connected through the field "factorid" where every single drug is identified with unique id. Factorpmid table is a many to one table where every factorid has many pmids. Tables factorpmid and pmidcontent are interconnected through field "pmid" where pmid is PubMed id for its publications. In this way of database design, all title and abstracts for single factor (drug in this study) are downloaded from PubMed in the same record. This greatly reduced the storage capacity. Four databases have been created for each local space and data has been extracted for the corresponding drugs.

Data extraction for every local space

Traditional LSA model, with all selected drug entities from different MeSH categories in a single model would be inclined to introduce systemic bias globally because of data imbalance. This is due to the fact that both too general MeSH terms with voluminous number of abstracts and too specific MeSH terms with very little number of abstracts are in a single model thereby weakening the model to be biased towards the specific drug entities. As a result, very specific drug entities will be retrieved with high association when the model is queried with allowable disease terms. Creating Local models with drugs as features from every MeSH category will restrict the bias within the local model. Further pre-processing of abstracts will also alleviate the local bias, also and eventually the combined global model will be far free from systemic bias. The following sections will detail about pre-processing of data. After the feature selection is done for every local space, data extraction has to be done to load the space with textual data.

Data extraction for the local spaces/classes is done by downloading titles and abstracts from PubMed, an online free database developed and maintained by U.S. NLM. Publications in PubMed have been indexed with MeSH terms. It facilitates the complete data extraction for the MeSH terms selected even if synonym terms are missed in the request which is being sent to PubMed electronically.

Loading each local space with only information from the particular MeSH category will limit it with only positive/relevant samples of data with no discrimination information from other local spaces. Vigna proposed a distributed and large scale latent semantic analysis using index interpolation; but the resultant model is Global as it fails to address the discriminative information inside the model. In this study, local models

themselves are created with class discrimination and then combined to form the global model. As a result, each local space will not happen to have closely similar data or information from other classes. It would not be able to capture the higher order semantic structure of terms when the local LSA models developed out of local spaces are combined to form the complete LSA model. Balancing each local class with both positive/relevant samples of data from its own class and also non relevant samples of data from other classes, which are difficult to be distinguished from relevant data, is of greater importance. Moreover, in this work, data/information in the form of abstracts are downloaded from PubMed from years 1950 to 2012. Almost 60 years of research publications will most likely have redundant information which will be introduced as noise into the local regions created. So, extracting even the top ranked positive samples of abstracts from the same class will subjectively eliminate to a greater extent the "noise" into the space.

It is found that equalizing the local spaces with positive and negative samples of information resulted in discarding too much of self-information from the spaces if the statistically selected MeSH terms for that particular class are not too specific. Subjectively, 70% of class is loaded with self-information, and 30% is loaded with discriminative information (shown in Figure 7).

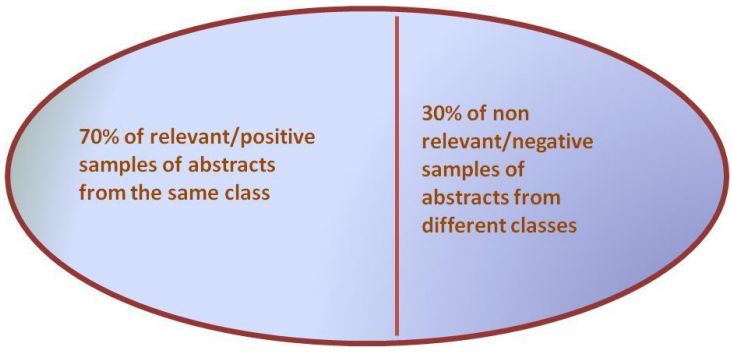

Figure 7 Pictorial representation describing the balance of local spaces with positive and negative samples of abstracts.

Query Based Sampling

Sampling is the process of extracting relevant abstracts from the available ocean by filtering out the irrelevant ones. Query Based Sampling of abstracts (QBS) was introduced for extracting the relevant samples by utliizing LSA. QBS is used to extract and load each of the four local spaces with most relavant textual. The following section will provide a detailed analysis about implementation

LSA is used to retrieve the most relevant samples of abstracts from the same class and non-relevant abstracts from all possible different classes. The foremost reason to choose LSA for the sampling is that it places the semantically similar abstracts close to each other in the reduced Eigen space. As a result, non relevant abstracts from different classes, but similar in concepts with the relevant abstracts, can be captured.

Implementation

Separate databases have been designed for every local space; 4 local spaces are created with abstracts, in this study from D01, D02, D03 and D04 MeSH categories.

When tf matrix is generated from the structured corpus of each database, every column vector is created in such a way that each cell represents the frequency of dictionary terms in every abstract of every drug entity selected (as an instance, tf matrix generation for D01 category shown below in Figure 8). Whereas, in usual implementation of LSA, tf matrix will be generated with every drug as column vector where every cell represents the frequency of terms in each biological factor. The reason for this structuring of matrix is that when SVD is applied on tf-idf matrix, semantically related abstracts will be captured and placed near to one another in Eigen space which will be retrieved on a rank basis when the model is queried with representative terms from other categories. The top ranked abstracts, ranking based on their cosine values with the query vector can be loaded into the local model, where the representative queries came from.

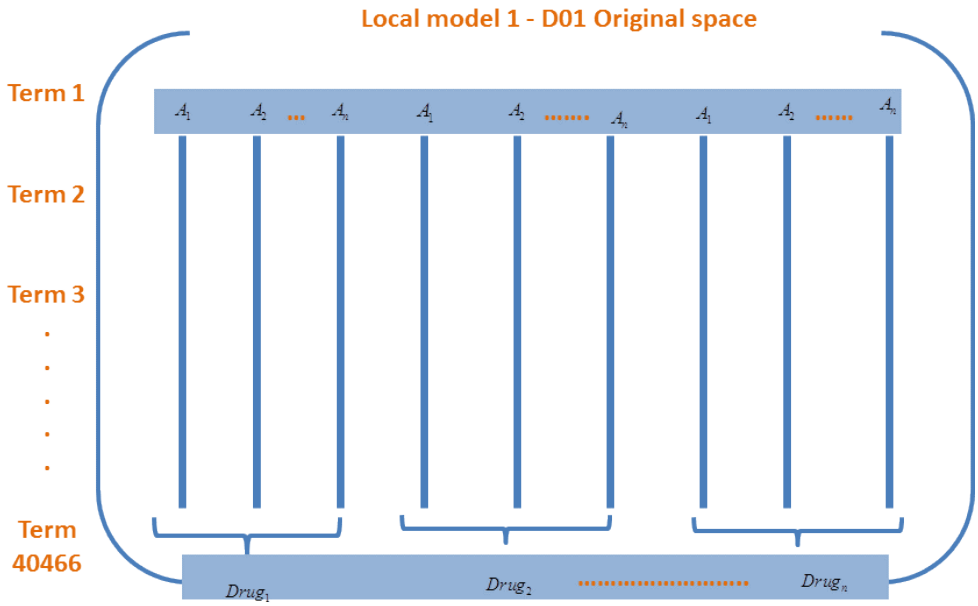

Figure 8 Diagrammatic representation of tf matrix generation for D01 local space. **A_1, A_2…** A_n represents abstracts of every drug like Drug1, Drug2, Drug n. $Term_1$, $Term_2$, Term 40466 are dictionary terms.

Tf matrices are generated for other local spaces too in the same way. Tf-idf matrix is generated and SVD is applied on it. The Encoding matrix U is found to be too sparse with too many zeros in it. The U matrix has to be dense as it captures the information based on the patterns of association of data statistically and contains redistributed weights for every dictionary term given to all the documents. As every column of the original matrix is weight given to abstract, there exists more than 80% zeros in the matrix before SVD is applied. It implies that every column has less information, so, LSA cannot capture higher order co-occurrence of terms leading to the sparseness in the resultant matrix generated and thereby in the U matrix too. To resolve this issue, abstracts in the similar context has to be merged together to increase the information or data amount. Fuzzy c means clustering is used to merge the abstracts by applying clustering on tf matrix vectors.

Fuzzy c means clustering

Fuzzy c means clustering [28] is used to cluster the abstracts based on their high membership values. In fuzzy clustering, each point has a degree of belonging to clusters, rather than belonging to one cluster completely. Thus, points on the edge of a cluster will have lower membership values to the centroid of the cluster and will belong to that cluster in a lesser degree when compared to the points in the center of the cluster. The centroid of a cluster is defined as

$$C_k = \frac{\sum W_k(x) X}{\sum W_k(x)}$$

Where,

$W_k(x)$ is the coefficient of a point describing the degree of it to be in the cluster k, and is inversely proportional to the distance from x to the cluster center.

Initial number of clusters to be given to the fuzzy c means clustering is chosen based on the volume of abstracts in every drug entity. It is applied on the tf matrix, shown in Figure 9, so that vectors close to the centroid of any particular cluster will be clustered together, i.e. abstracts with same concept are merged together (figure shown below).

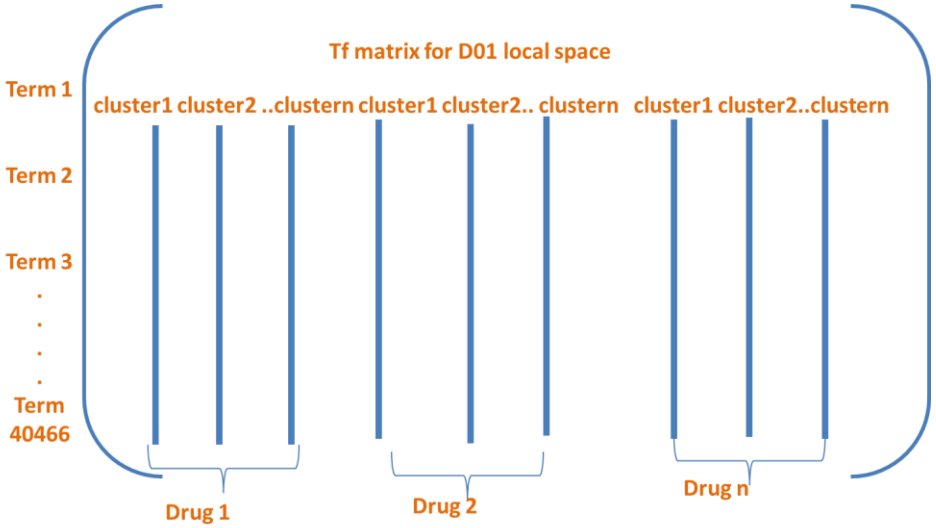

Figure 9 Diagrammatic representation of tf matrix generation for D01 local space after clustering. Cluster1, cluster2 cluster n represents abstracts of every drug like Drug1, Drug2, Drugn. Term1, Term2, .Term40466 are dictionary terms

It is found that some clusters formed are overloaded with abstracts in the same topic and will have overlapping terms; on the contrary, there are some outlier points which will have higher degree of membership to varied clusters resulting in less loaded clusters. Biased clusters will factually introduce into the models, which is greatly alleviated by the process (Figure 10).

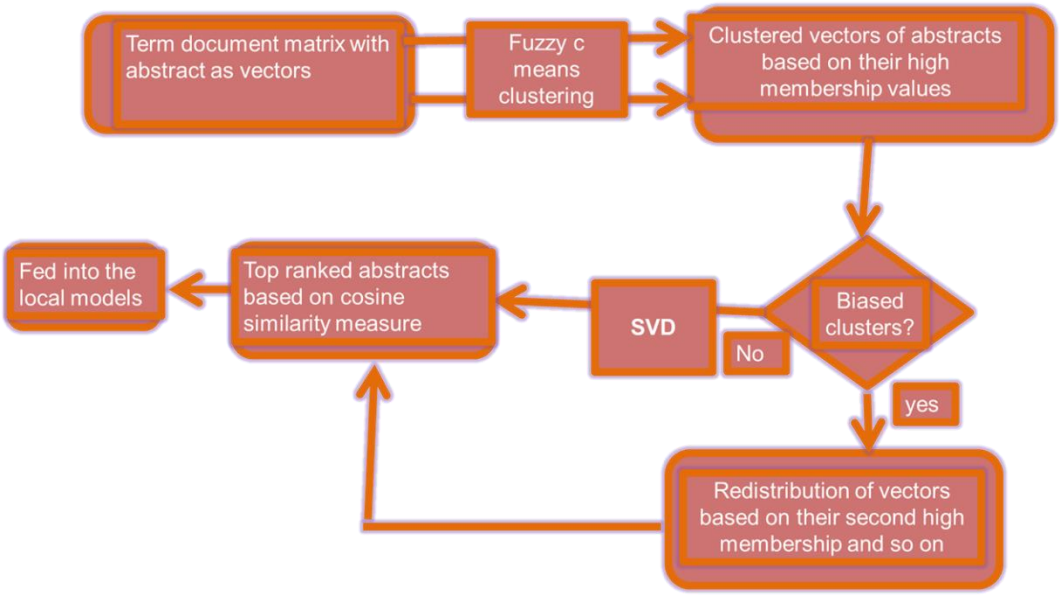

Figure 10 Block diagram showing the steps taken to load each local space with unbiased clusters.

Systemic Bias

As every column of tf-idf is a cluster of abstracts, encoding matrix was denser with no redundant information. Representative query terms for each local model is used to query the other models and abstracts from top ‘k’ ranked clusters are sampled based on their dot product between them.

When every local model is defined with sampled abstracts from top ranked cluster retrieved from other local models, it is found that some of the clusters are overloaded with data points and vice versa for remaining others. Systemic bias is more likely to be introduced into the models, as over populated clusters will introduce too much information, and less populated clusters will leave the model with less information resulting in imbalance in the local models. Bias for Erythromycin is shown in Figure 11.

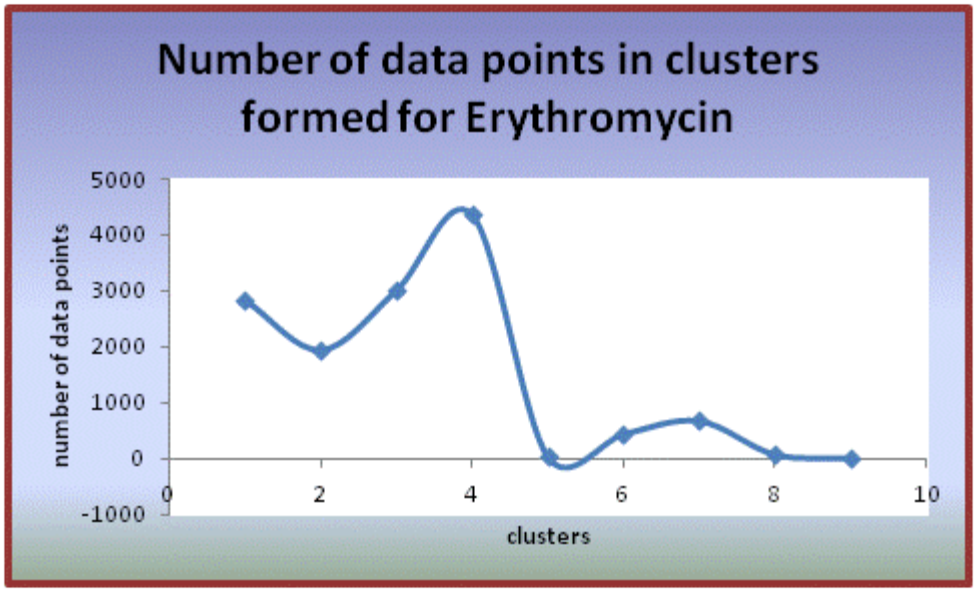

Figure 11 Graph showing varied number of abstracts for every cluster formed. Example shown for one entity erythromycin; cluster is overloaded with nearly 4750 abstracts and cluster 9 is less loaded with 120 absrtacts.

Balancing clusters to alleviate bias

To resolve this issue, clusters are averaged with approximately the same number of abstracts by distributing the points in the overloaded cluster to the less loaded ones, based on their second higher membership values, third higher membership values and so on. As every cluster is averaged with approximately equivalent number of abstracts as shown in the below figure, any cluster retrieved from LSA information retrieval technique will be yielding same amount of information into the querying models.

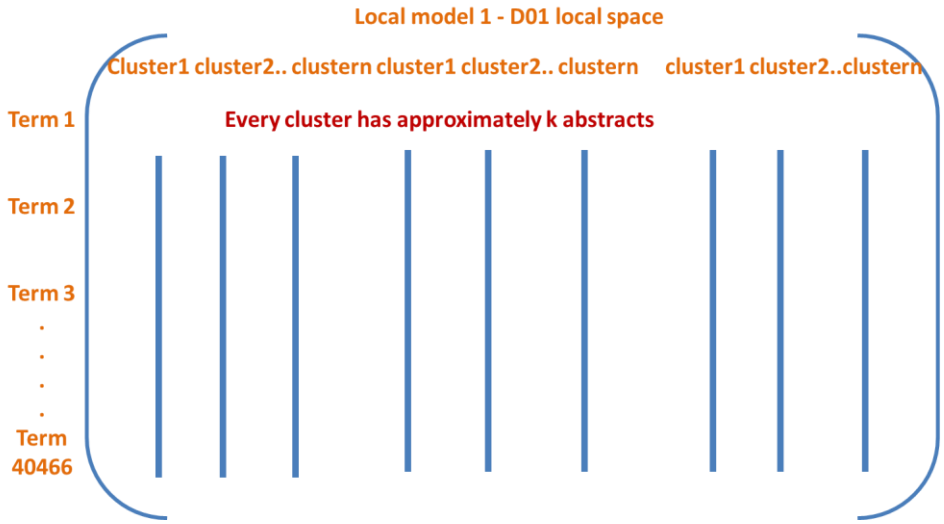

Figure 12 Diagrammatic representation of tf matrix generation for D01 local space after redistribution of abstracts in clustering. cluster1, cluster2 cluster n represents abstracts of every drug like Drug1, Drug2, Drugn, where every cluster is approximately loaded with equivalent abstracts. Term1, Term2, …Term40466 are dictionary terms.

Local spaces

Tf-idf matrix is then generated from the tf matrix with equivalent abstract clusters and SVD is applied on it. Local LSA model is generated for every local space and is queried with representative terms from other models and top ranked clusters are retrieved. The abstracts from which the data points in those top ranked clusters are then loaded into the corresponding queried local space.

Also, care has been taken to create the local spaces with approximately comparable number of abstracts in every space. As a result, local regions are created with balanced self-representative information and discriminative information from other local regions. Local region for D01 category is detailed in the below Figure 13.

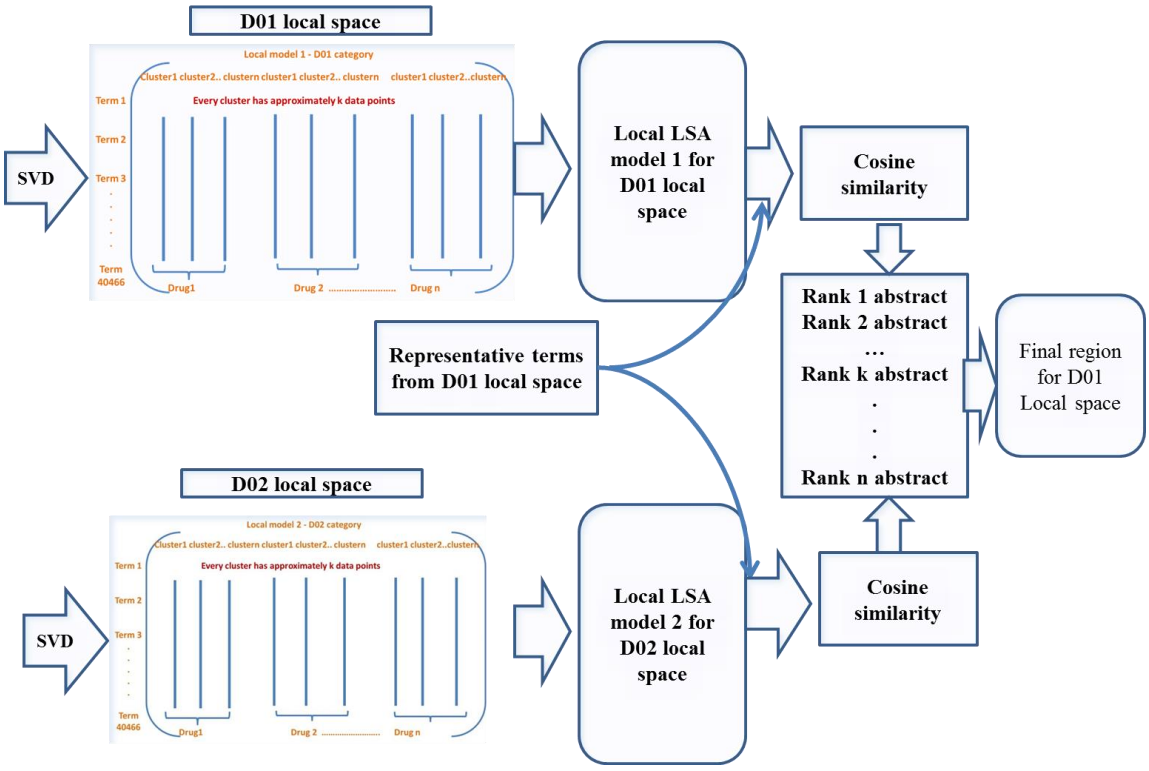

Figure 13 Block diagram showing the creation of final D01 local space (D01 MeSH category); same procedure is done for other 3 local spaces too. Ranked positive samples of abstracts are retieved from local LSA model of D01 original space and inserted into D01 new space. Also, negative samples of abstracts are retrieved from D02 original space by Local LSA model and inserted into D01 new space.

Query based sampling of abstracts is used to load every local spaces with 30% undistinguishable negative abstracts from other local spaces through the integration of LSA. Also, 70% positive abstracts are sampled from own space and loaded into the final local pace of the class. Eventually, all local spaces are created with most relevant self-information and discriminative information.

Encoding matrix from Local Model Generation and Dictionary reduction

Local models are created by applying SVD on the tf-idf matrix from the newly created local space and the resultant Encoding matrix U from each of the local models was sparse. Careful analysis showed that the U matrix has one third of its rows filled with zeros, and the rows are derived from dictionary terms which are inappropriate to the feature range selected for the models. Those terms are discarded from the dictionary which condensed the dictionary size from 40466 to 29915 terms. Further processing of tf matrix from reduced dictionary by applying SVD generated highly dense encoding matrix with no sparseness in it.

Complete LSA model from local models created

Let's say that the encoding matrices generated from the four local models are U_1, U_2, U_3, U_4, and are approximated to U_1^k, U_2^k, U_3^k, U_4^k in the Eigen space where the column dimension of each of U matrices is reduced by k. The dimension k is obtained for every local model by capturing 97% of the information by thresholding their corresponding singular value matrix S_1, S_2, S_3, S_4, so that dimension k is different for each one of them. As the Encoding matrix U_is from SVD captures all information on term concept basis, combining encoding matrices from every local model incorporate the needed information.

Combining Encoding matrices of local models

The global Encoding matrix can be formed by adding the local Encoding matrices element wise by weighting each element of local encoding matrix by relevance score; the relevance score is determined by the contribution of a particular model when combined globally. The problem with this approach is that every local model's

Encoding matrix must be of same dimension so that they can be added element wise. But in our study, as every local space is defined by different number of features which resulted in different column dimension of the encoding matrices even though row dimension is same because of the same dictionary.

In the proposed method to create the global encoding matrix, all encoding matrices are placed parallel to each other as, mentioned above, the dictionary is same for all of them. Let's say the encoding matrix for created four local models are to U_1^k, U_2^k, U_3^k, U_4^k then, the combined local model would be generated, as shown in Figure 14.

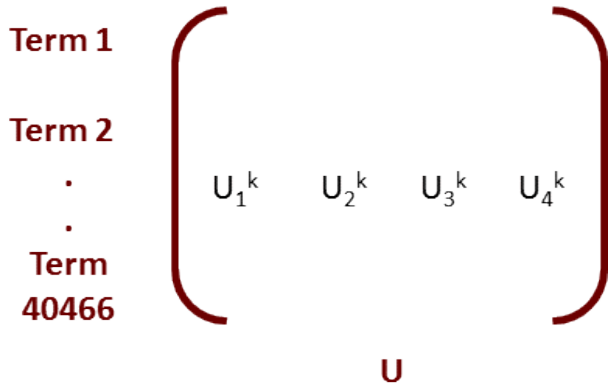

Figure 14 Encoding matrices U1, U2, U3, U4 placed parallel to form the global encoding matrix; Term 1, … Term 40466 are dictionary terms.

The shortcoming of this approach is that the columns of the resultant U will not be orthogonal to each other which are an essential feature to use SVD, as it is supposed to capture unique information in each of the columns. This is resolved by applying SVD again on U to decompose it again to U, S, and V^T. Now the raw data of each local model, i.e. columns of tf-idf matrix is projected onto the reduced Eigen space in such a way that $P_i = U^T A_i$ where A_i is the tf-idf matrix of each local space. The global matrix $P = [P_1\ P_2\ P_3\ P_4]$ is created which can be queried with queries projected onto reduced Eigen space ($Q_p = U^T Q$). The retrieved results for the query are ranked based on cosine similarity measure between the projected query vector Q_p and every column of P. As the systemic bias due to varied sized documents (Drugs/Chemicals here) has been taken care of during Query based sampling, cosine similarity measure is used to measure the similarity between the projected query vector and each column vector of P and the results are retrieved in ranked order.

Relevance Model

As LLSA model is scalable, more features (Drugs/Chemicals in this study) can be added, as a result, evaluation of the model with queries which have voluminous relevant set would be accomplished. So, it is imperative to verify whether the model is capable of pulling high precision in the top ranked result or highly relevant set, when it is scaled. The similarity measure, cosine value in this study ranges differently for different queries, as it is data driven. Hence manual thresholding to determine the highly relevant set is not feasible for every single query. So, a relevance model is developed, to automatically cluster the retrieved results into three groups like highly relevant results, moderately relevant results and low relevant results. Fuzzy c means clustering [28] is used to cluster

the cosine values of retrieved drugs/chemicals for every query. The whole set of retrieved Drugs/Chemicals for every query is categorized into three groups such as highly relevant set, reasonably relevant set, and poor relevant set.

DDNet: A Drug-Disease Interaction Framework and PubMed Link Tool

Tremendous growth in biomedical literature as a consequence of experimental and computational biomedical data drove the scientific community to develop literature mining web tools to find the nuggets of information most relevant and useful for specific analysis tasks. The ultimate goal of this study is to aid the research community to browse through the Drug-Disease associations.

So, a fully integrated, interactive, user friendly, web based framework DDNet is developed and deployed. The underlying LLSA model is used to explore the associated Drugs/Chemicals for any given user query, thereby facilitating the information retrieval. It is broad enough to accept multi gram MeSH terms as queries with the options of visualizing the results based on the intensity of information need. The semantically extracted associated factors are ranked in order based on the similarity measure between the user query and factors in concept level.

Additionally, the ranked factors are clustered as highly relevant set, reasonably relevant set and poorly relevant set as an outcome of Relevance model.

DDNet users have the comfort of seeing narrowed down results depending on the options chosen by them. Naïve users, who want to gain the basic knowledge about drug-disease interactions, might be settled down with highly relevant set, medical researchers or pharmacists might broaden their knowledge with reasonably relevant or poor relevant result set for deeper analysis and thoughts. Uncovering of Knowledge Discovery in the

course may smooth the researcher's progress of generating new hypothesis, which will ease the Drug repositioning, for which this framework is studied and developed.

The user interface of DDNet is show in Figure 15 and the ranked results from the webtool is shown in Figure 16.

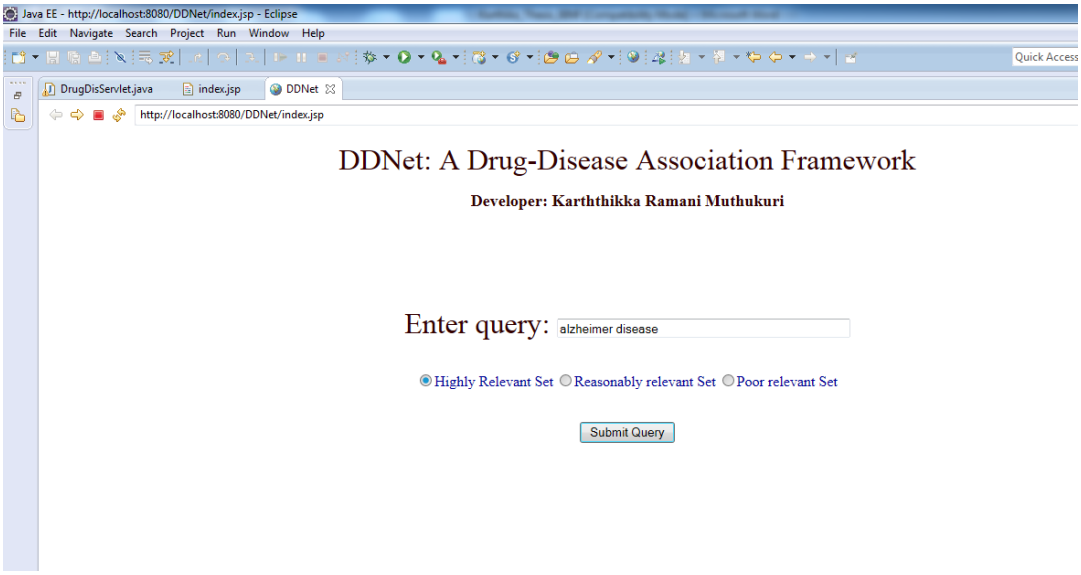

Figure 15 User interface of DDNet to enter queries.

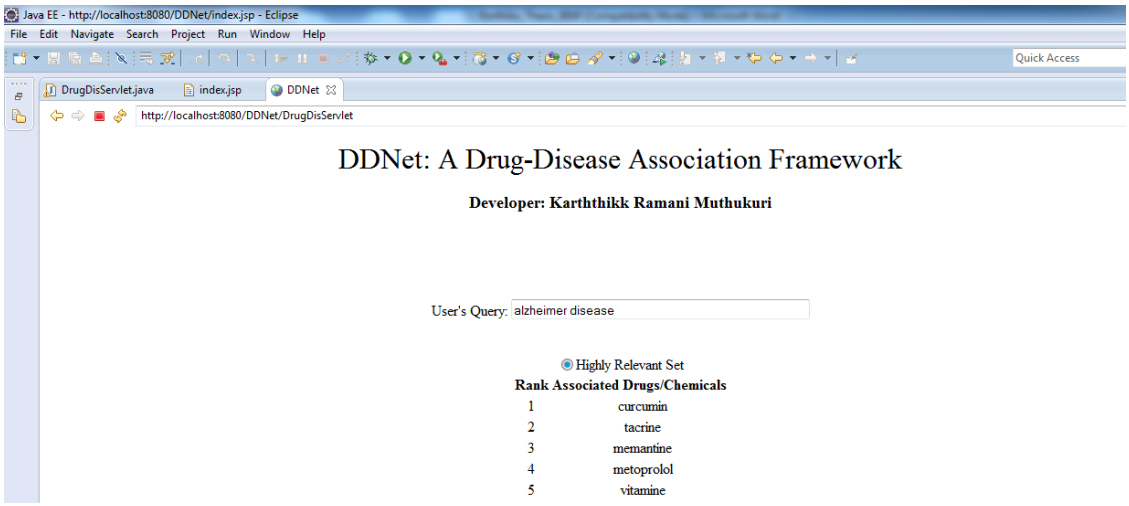

Figure 16 The display of ranked Associated Drugs/Chemicals for the user query Alzheimer disease.

In the DDNet display shown above, the user is displayed with ranked highly relevant set of associated drugs/Chemicals for his query, based on his option of relevance chosen. The time scaled for the user to analyze the results is reduced as it is clearly ranked in the flat file. A plethora of web tools are developed by the research community in biomedical domain with interactive interfaces. As an instance to note, [29] developed a single graph theoretic framework for all known phenotype and disease gene associations which are represented by nodes and edges. In this framework, disorders are represented by nodes which are interconnected with edges; genes are represented by edges with their thickness dente the number of genes in the interconnection.

The interface of DDNet displays the results with ranks which are comparatively easier for the users to analyze rather than analyzing the thickness of edges. Future direction of this work is to expand this framework with increased Biological concepts, allowing multiple queries from the users, displaying the graphical display with interconnected concepts between those queries as edges with queries as nodes.

Implementation details using software languages

The entire web tool is programmed using the object oriented language core java, which is beneficial [30] for the model generation and computing similarity measure between the queries and the biological concepts (Drugs/Chemicals here), Java enterprise technologies for web implementation and MATLAB for handling matrix computations beneath the model.

Reasons for java:

- The main reason for utilizing java is that it is a well suited programming language to develop highly interactive Graphical user interfaces.

- It is platform independent and can be run in any other operating systems in spite of the fact it is programmed in windows.
- Java scripts which are extended nowadays to generated graphical displays where java script objects can be stored as JSON objects and be displayed in nodes and edges
- As future work is intended to graphical network display, java is the well suited programming package to use

Java usage: The front end interface, i.e. query entry and results displays are done with Java server pages, a JEE technology as static pages can be made interactive by bundling them inside java code. Similarity measures are computed in core java by creating Matrix classes and corresponding methods such as transpose multiplication etc. to deal with matrices.

MATLAB usage: The structured data is converted into meaningful numbers though tf-idf matrix generation and LSA model is created by applying MATLAB in built function SVD.

Efficiency of the framework as a result of Pre-computed results

DDNet is very time efficient in retrieving the associated information to the user. For instance, the time taken to retrieve the associated drugs/chemicals to the users query "Alzheimer disease" is just a couple of seconds. The reason behind this time efficiency is that the results are pre-computed for all 29915 dictionary terms (please see section 3.8 for details) which can be given as queries to this network. The associated Drugs/Chemicals for any query is found by cosine similarity between them and ranked in order by sorting. HashMap: Java has been useful in pre-computation part also as the language has tables to store apart from arrays to directly retrieve the required information by indexing. HashMap is one such useful mapping table which allocated separated buckets for each of it collections. HashMap is used over HashTable, which is also a similar package from where HashMap is derived, because HashMap allows null values. The extracted associated drugs/chemicals for some of the queries from DDNet are null as they do not have any associations semantically. So HashMap is the reliable procedure to store the results for our case.

The extracted results for each of the possible queries within the dictionary range are loaded into a hash map with every dictionary term as key and array of its associated concepts as values. When the user hits a query to this web tool, it index the corresponding key and fetch the values for that key and displays it in ranked order based on their similarity measure. As HashMap does not have to iterate through the key collections, and directly index it from the bucket, the time taken to retrieve the results is comparatively less. The framework developed will be time saver for the users as they do not have to wait for seeing the displays as they have to do in front of the frameworks which compute

the results on the fly. This is one of the key advantages of the developed framework DDNet.

Subjective analysis has been conducted on ranking of associations between drugs and disease of DDNet by analyzing the results with published articles for its factualness.

Network of Drug-Disease association from DDNet using MeSH hierarchal code

Network of Drug-Disease associations can be derived from DDNet framework utilizing hierarchal structure of MeSH, from which the domain specific dictionary is created by Ontology mapping [11]. The associated Drugs/chemicals from DDNet, for a disease query can be represented as hierarchal tree by Reverse mapping onto MeSH ontology to the root level. Combining multi queries with their retrieved semantically associated concepts with their tree structure will result in network of interconnected objects through Drugs/chemicals. This hierarchal structure network will be categorically useful for the Pharmacists to derive the hierarchy of drugs/chemicals from which he/she can gain knowledge about the chemical composition of drugs as well as associations between drugs though the diseases. As the results are classified as highly, reasonably and poor relevant set, networks can be derived based on the intensity of information need. As an instance, highly relevant set of Drugs/Chemicals retrieved for queries Alzheimer disease and Cardiovascular disease are shown as interconnected network in Figure 17 with their reverse mapping onto MeSH ontology.

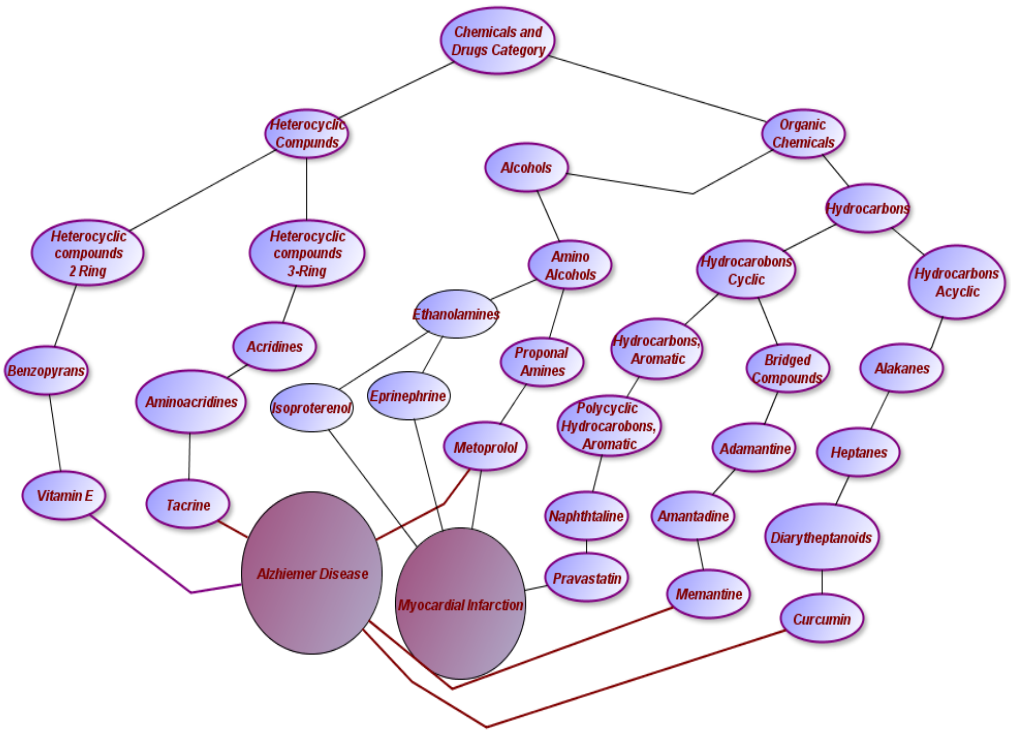

Figure 17 Network of Associations of Drugs/Chemicals for diseases Alzheimer disease and myocardial infarction derived from MeSH hierarchy.

In the above network, semantically related Drugs/Chemicals for disease Alzheimer disease and Myocardial infarction are shown with their reverse mapped MeSH hierarchal structure, back to the root level. It might facilitate the Pharmacists to derive the chemical composition of the drugs thereby analyzing its uses for other diseases or targets apart from what they have been produced for. Also, this network shows that Alzheimer disease and Myocardial infarction are related through Metoprolol, which detailed about the usage of it.

Future Directions in DDNet: Network of Drug-Disease associations are to be developed automatically from the retrieval of DDNet would be accomplished in future to make this tool a complete and informative framework.

Chapter IV

Results and Discussion

Drug-Disease Association

An efficient, robust, scalable and unbiased model for finding the network of semantically related drug-disease associations is built and user friendly interface, DDNet is developed to display the associations in readable manner. LLSA is designed and implemented to achieve this goal. Semantically related associations can be derived from published literature, where huge amount, biological results are preserved as result discussion. A brief overview about the procedure implemented is described below.

Statistical analysis on MeSH terms resulted in 53 drugs/chemicals for this study, and are derived under D01, D02, D03 and D04 MeSH categories. Four local spaces have been defined from the four derived MeSH categories. Textual data has been extracted from PubMed for 60 years for the drugs in each of the local spaces. Around 0.2 million abstracts have been extracted for D01 original local space. Aroung 0.3 million abstracts have loaded into D02 original local space. Approximately 0.15 million abstracts have been extracted for D03 original local space. Approximately 0.25 million abstracts have been extracted for D04 original local space.

Multi gram dictionary is created by the combination of MeSH terms to preserve the biological meaning in the literature when it is modeled.

The local spaces are defined with well relevant representative samples of abstracts from own as well as varied classes to balance the region, through Query Based Sampling method (described in section 3.6 for details). Also, care has been taken to distribute equivalent volume of information in each of the spaces thereby reducing the bias.

Out of 2,04,617 abstracts in D01 original space, 1,35,387 abtracts have been retrieved as relevant samples by QBS and loaded into D01 new local space. Around 69, 000 abstracts have been retrieved from other local spaces such as D02, D03 and D04. Out of 2,70,090 abstracts in D02 original local space, 1,89,063 abtracts have been retrieved as relevant samples by QBS and loaded into D02 new local space. Approximately 81,000 abstracts have been retrieved and loaded into D02 from other local spaces such as D03, D01 and D04. Out of 1,52,599 abstracts in D03 original local space, 1,06,000 abtracts have been retrieved as relevant samples by QBS and loaded into D03 new local space. Half a million abstracts have been retrieved from other local spaces such as D01, D02 and D04. Out of 2,55,369 abstracts in D04 original local space, 1,78,758 abtracts have been retrieved as relevant samples by QBS and loaded into D04 new local space. Ten percent of million abstracts have been retrieved from other local spaces such as D01, D02 and D03.

With the dictionary of size 40466 created and local spaces defined, Local LSA models are created from each of the local spaces. As the encoding matrices are very sparse with zeros in it, irrelevant dictionary terms in those aerp row indeces are discarded resulted in a reduced dictionary od 29915 terms. LLSA models relieve the system to a major extent from bias, scalability issue; the traditional global model suffers.

This chapter describes about the objective evaluation of the LLSA in finding the associations which is underlying the interface. A number of empirical studies have been conducted to study about the system's efficiency, scalability, bias and robustness which concluded the overall performance of the system.

Efficiency

To ensure about the efficiency of the network, time taken for the LLSA system to retrieve the associated Drugs/Chemicals with the supportive evidence of PubMed IDs links for any user query is noted down. The results are pre-computed for all the possible 29915 queries, stored in Hashtable and, can be extracted by indexing when the system is queried (Please see Section 3.11for details). As a constructive consequence of pre-computation, the observed time taken by the system for a query is approximately 25milliseconds which is far less than any other system which computes the results on the fly. Hence, pre-computation greatly enhanced the efficiency of the system which is one of the parts of the goal in this study.

Systemic Bias

Systemic Bias, which is a general occurrence in Global LSA model, is greatly lessened by Query Based Sampling of Abstracts (see section 3.6 for details). Every LLSA model is created with approximately equivalent amount of information so that each of the retrieved concepts is predicted to have association values within narrow range for any given query. Cosine similarity values, which yielded the association, are analyzed for the LLSA model's non-biasedness. Cosine values for associated Drugs/Chemicals for the Query "Alzheimer Disease" are taken for analysis and are plotted for comparison with the values from Global LSA model (see Figure 18).

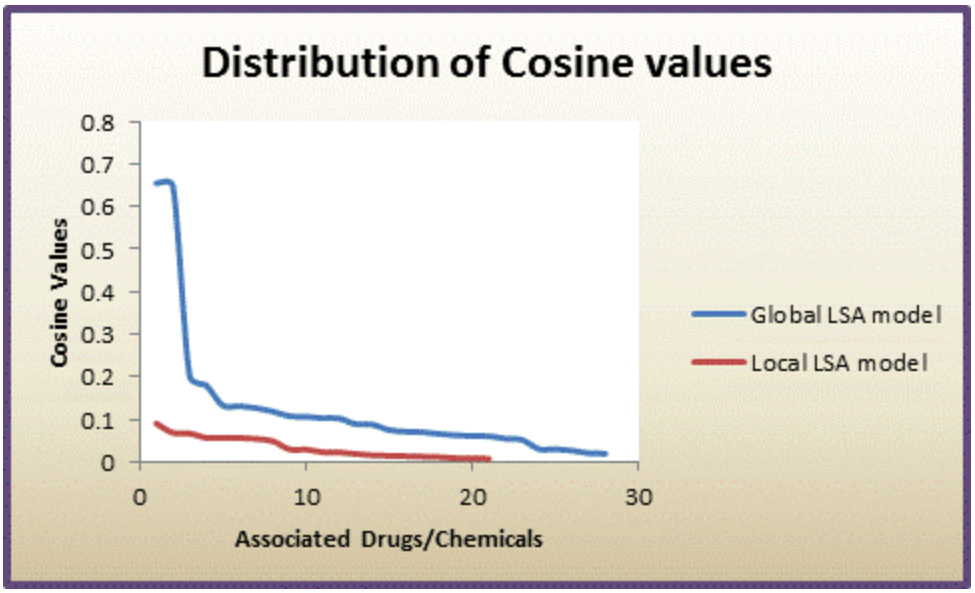

Figure 18 Graphical representation of the distribution of cosine values for the semantically associated Drugs/Chemicals for the Query Alzheimer disease; shown for both Global and Local LSA models.

From the above graph, it is evident that cosine values for Local LSA model fall within a very narrow range, whereas cosine values for Global LSA model fall within a comparatively wider range. For LLSA model, the maximum of the cosine values is 0.0903 and the minimum is 0.0071 with **standard deviation of 0.024,** whereas for Global LSA model, the maximum of the cosine values is 0.6547 and the minimum is 0.0191 with **standard deviation of 0.153**. Lower standard deviation obtained for LLSA implied that the vectors of cosine values of semantically related Drugs/Chemicals for the query "Alzheimer disease" are placed near to each other. This meant that all the retrieved Drugs/Chemicals contain approximately amount of derived useful information in the model from the corpus and eventually resulted in a bias free model. For Global LSA model, the standard deviation of cosine values is higher which implied that

Drugs/Chemicals with voluminous amount of information are retrieved with lower cosine values and those with less information are retrieved with higher cosine values. This clearly indicated that Global LSA model is inclined to bias.

LLSA is relieved from systemic bias by Query Based Sampling of Abstracts which loaded the local spaces with equivalent amount of positive and negative samples of information. Cosine values of associated retrievals for other queries are given in Appendix section for further analysis.

Scalability

In Global LSA model, all the features selected from MeSH terms, for this domain, have to be incorporated into a single model, which would result in bulky amount of textual data. The computational complexity is high as the model has to be computed from a very high dimensional tf-idf matrix. Even, loading of very high dimensional matrix into MATLAB for further computations is infeasible as MATLAB is restricted with 40000X5000 matrixes. Concluding, Global LSA model is not scalable because of computational complexity.

The proposed model is scalable as MeSH terms can be incorporated into it as unique local models. The added features are localized into separate models with limited textual data in each of the local spaces. Hence, tf-idf matrices generated from each of the spaces is of low dimension and thereby compatible to be computed in MATLAB for model creation. LLSA has an added improvement of scalability over traditional Global LSA model.

Robustness

Recall and Precision curve/PR curve [10], the standard measure to determine the effectiveness of an information retrieval system are used to evaluate the robustness of the IR system developed. The results are retrieved from DDNet, an LLSA system, and to look how precision varies for every recall, a Gold Standard is needed to evaluate the retrieved results against the relevant.

PharmgGKB: The Pharmacogenetics and Pharmacogenomics Knowledge Base, PharmGKB [19], a manually curated comprehensive database is used as Gold standard to validate the web tool developed. It is a web based public repository of genotype and phenotype information relevant to pharmacogenetics, developed at Stanford University with the funding from NIH. It has varied data from research, clinical outcome etc. to catalyze the research in the field of Personalized medicine. It had an Excel file with relationships between drug-drug, drug-disease and disease-gene, which is downloaded in 2012 to evaluate the performance of the system.

Query selection for validation and evaluation

Queries, used to plot PR curve is an important parameter as random selection without sufficient relevant set would result in incorrect low recall. So, they are carefully selected based on relevant drugs/chemicals which are common in both the model and PharmGKB, the Gold Standard, used for validation. If subset from relevant set of results for any particular query is present in the model, then that particular query is selected for validation purposes. If none of the relevant results for a query is present in the model, validating the system with that query would be unfair and also will end up in zero recall and precision. In the Figure 19 shown below, query 1 is selected for validation as part of

relevant set of drugs/Chemicals is in the model also and Query 2 is discarded as there is no common items. This resulted in the selection of 20 queries.

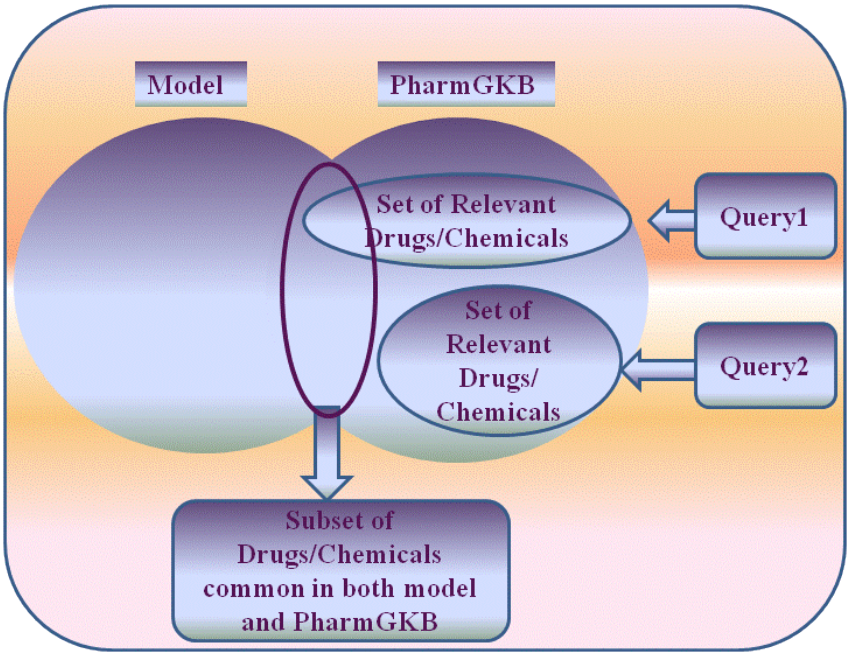

Figure 19 Diagrammatic representation for selection of queries to validate the model. Query1 is selected and Query2 is discarded.

Also, care has been taken that selected queries are not random so that the model could have been analyzed for its performance by queries falling under specialized categories of diseases. This categorized query selection will formulate the model to be credible and complete enough for any specific medical set. Yet again, with inclusion of more biological concepts (Drugs/chemicals in our case), the model can be enhanced to

be complete enough for extended medical groups too, thereby generalizing the network developed. Twenty queries selected fall under 4 broad categories such as Heart related diseases (Figure 20), Brain related diseases (Figure 21), cancer related disease (Figure 22), Lung related diseases (Figure 23).

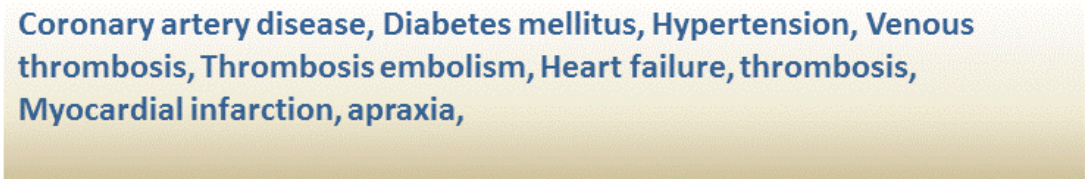

Figure 20 Queries categorized under heart related diseases.

Figure 21 Queries categorized under brain related diseases.

Figure 22 Queries categorized under cancer related diseases.

Figure 23 Queries categorized under lung related diseases.

PR Curve and Analysis

As the results are retrieved in rank order from our system, for any user query, recall and precision can be calculated for each set of top k retrieved items (drugs/chemicals in our study) and precision-recall curve can be plotted for each set. For a single query, if an item added is relevant both recall and precision will increase and if the added item is not relevant, recall will remain the same and the precision will decrease. Hence the PR curve would follow a saw tooth shape with too many jiggles in it. To remove these jiggles and to interpret the PR curve effectively, precision is to be calculated only whenever there is an increase in the recall. The judgment is that user would be ready to look at few more items if it would increase the percentage of the retrieved set which is relevant. The traditional way of doing this is 11 point average precision. For every query, precision has to be calculated at 0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0 recall levels. For multiple queries, average precision is to be calculated for each query for 11 point recall levels and the PR curve can be plotted for analysis. To analyze the PR curve evidently, 11 point averaged precision-recall curve is plotted for multiple queries, shown in Figure 24.

Factually, if the size of the relevant set for any query is huge, the retrieved relevant set from the system will be distributed with slowly decreasing precision for increasing recall rates. So, as an initial step, all the retrieved Drugs/Chemicals for each

of the above selected 20 queries are taken into account and average precision of all queries has been calculated and plotted for every recall rate.

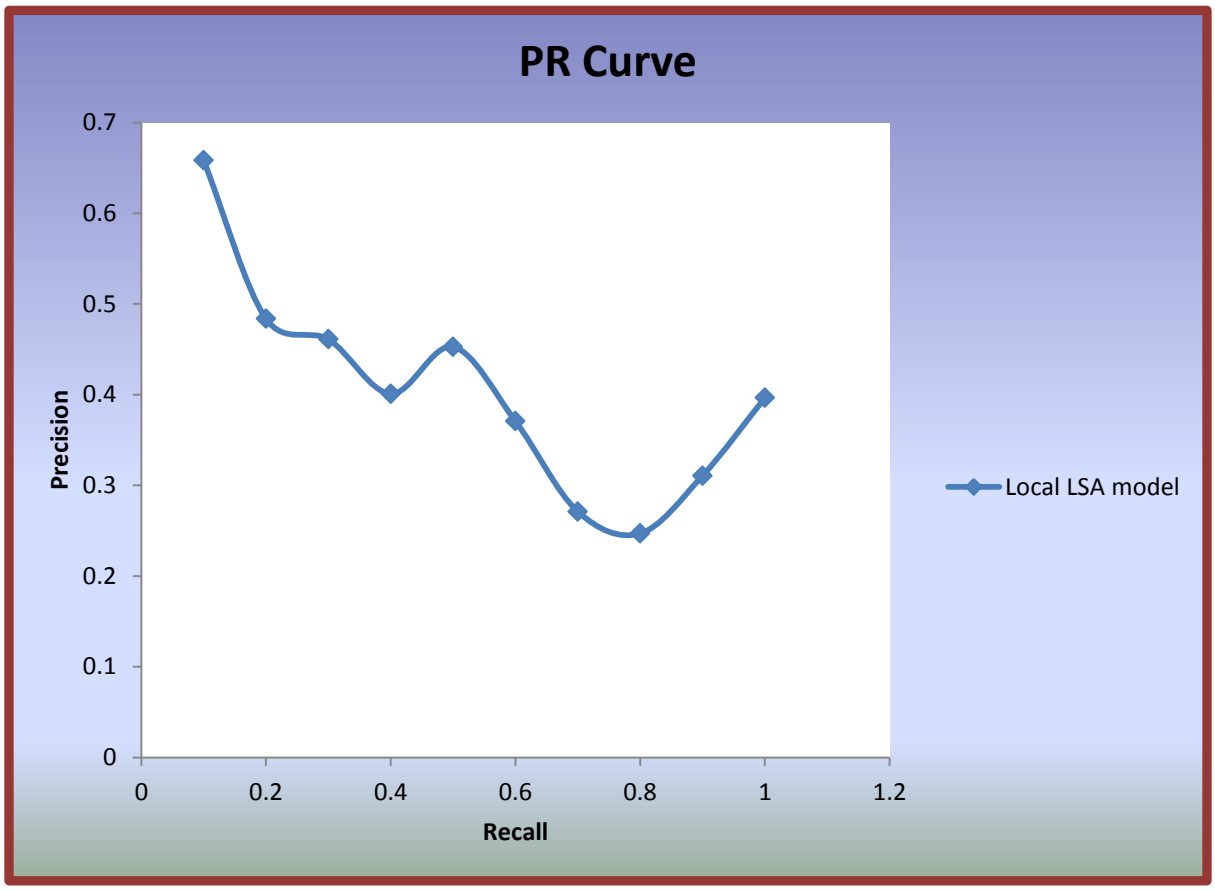

Figure 24 Averaged 11 point Precision/Recall graph plotted across selected 20 queries. The results were obtained from DDNet.

Precision gradually decreases with increase in recall, except for recall level of 0.5, which shows the consistent performance, i.e. robustness of the underlying Local LSA model of the system as relevant results are retrieved at approximate regular intervals. It can be seen that the precision at recall rate of 100% is little higher because there are few queries which has only one relevant drug/chemical present in the model; so, precision goes to 1 when recall increases from 0 to 1. Another important observation from the validation is that except for two queries like Asthma and Alzheimer disease, recall is 100%.

Generally, for any IR system, Recall and Precision are important over one another. For instance, web surfers would like to see only relevant items in their first page and may not be interested to know about every relevant item possible, i.e. high precision even at low recalls, whereas researchers would like to see almost all relevant items possible with the tolerance of having some false positives, i.e. high recall with low precision. Hence, Recall and Precision values need to be high based on the information need of the user. For the utilization of the network developed specifically for some domain specific researchers, high recall is expected even with little less precision.
As this work and the developed web tool and Network visualization intends to facilitate the researchers in this domain and medical specialists, it is imperative for the model to retrieve the relevant results as much as possible, so that the specific users can gain the complete knowledge about their information need. Thus, 100% recall retrieved by our LLSA system validated that the system is accurate for domain specific researchers as all relevant results are retrieved from the model which most researchers/medical experts would be longing to look for.

Mean Average Precision

A single measure of quality to trade off precision versus recall for ranked results is Mean Average Precision. Average precision is the arithmetic mean of precision values obtained for the set of top k retrieved documents for every increase in recall and this averaged precision is again averaged over all possible queries.

Mean Average Precision, MAP (Q) = $\frac{1}{|Q|}\sum_{j=1}^{|Q|}\frac{1}{m_j}\sum_{k=1}^{m_j}\Pr ecision(R_{jk})$

Where,

q_j - every single j^{th} query

d_{mj} - set of relevant documents

R_{jk} - set of ranked retrieval results from the top result until it gets to document d_k

For a single query, average precision approximates the area under the PR curve and so MAP is the average area under the PR curve for the set of queries. Using MAP, recall level is not fixed, so each information need/query is given equal weighting as some queries have many relevant results and some may have very few relevant results. So, recall level will be different for different queries. MAP is calculated just as the average of arithmetic mean of precision values of every query for their individual recall levels. Selected 20 queries have different recall levels from our model also, so, MAP would be the appropriate choice as the single figure measure of quality. MAP for the model and thereby the system developed is 0.2331. To ensure whether the obtained MAP value is rationally a good figure, Mean Average Precision of Global LSA model is computed and checked with the value gotten from Local LSA model.

Comparison of the performance of Global LSA model and Local LSA model

To determine the better performance of the Local LSA model, its performance is compared with the Global LSA model. The same 53 drugs from 4 different MeSH categories are chosen and data/abstracts are extracted from PubMed to define the global space and the global model is created from the space as it is described in section 1.1. The selected 20 queries are used and average precision across all queries is plotted for 11 point recall rates, shown in the below figure.

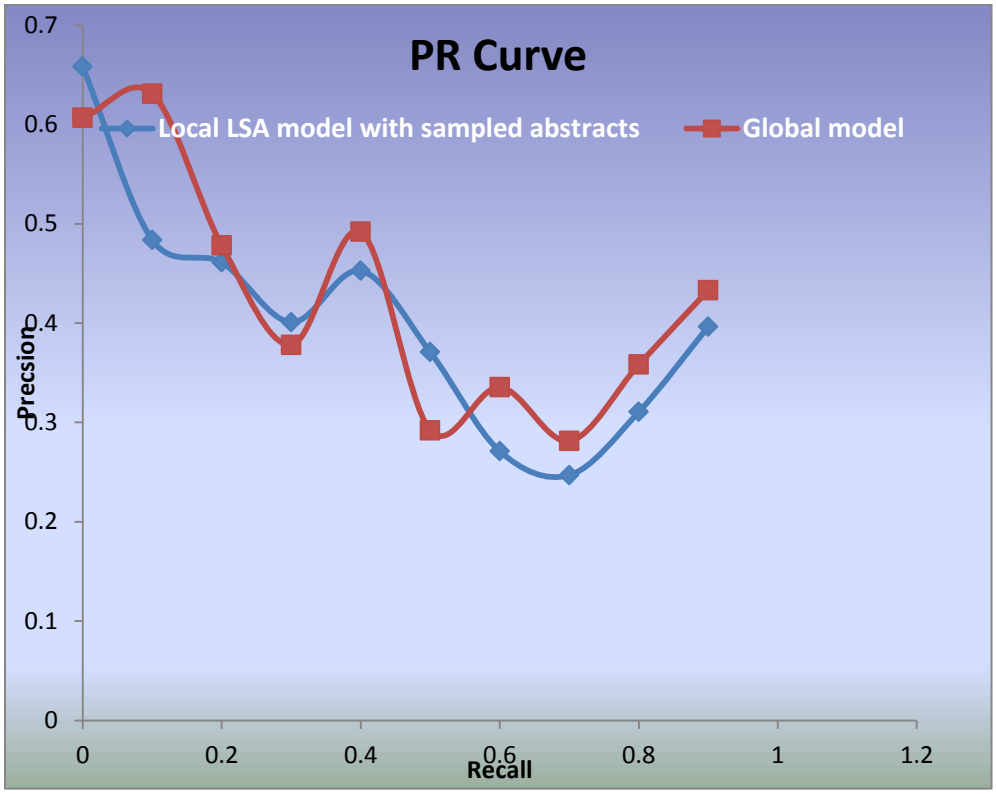

Figure 25 Averaged 11-point Precision/Recall graph plotted across selected 20 queries for Local LSA model and Global LSA model.

From the above PR curve, it can be seen that PR curve varies a lot for Global LSA model showing the inconsistent behavior of the model. The curves revealed clearly that LLSA model is robust than Global LSA model. Also, the precision is comparatively low at recall rate of 1 which again indicates that the model could not retrieve the relevant drugs/chemical for those queries with only one relevant result from the Gold standard, as it could be remembered from section 4.1. It could be clearly stated that Global model could not result in a recall of 100% for many queries as it was the case in Local LSA model. Again the PR curve made evident that LLSA is more an appropriate model to retrieve more relevant results for researchers. Also, MAP for Global LSA model is **0.1874** which is approximately 5% less than MAP of Local LSA model. Though the improvement is very minuscule, it shows that little improvement in the developed model over the traditional Global LSA model.

Future Directions: As the above analysis is based on a pilot study with very less amount of Drugs/chemicals in the Local model, many relevant items from PharmGKB, the Gold Standard are not included. It is very evident that the model showed less precision as relevant items cannot be retrieved because of their non-existence in the model. If the Local model developed could be optimized with much more Drugs/chemicals, it can be indubitably expected that its performance would be higher with higher precision.

Performance evaluation of the Local LSA model developed against Local LSA model without Query Based Sampling of abstracts

While generating the Local LSA models, local regions are shaped up with much care so that it is packed with self-representative information from its own space and discriminative information from other local spaces. Query based sampling of abstracts, described in 3.6 is used to define and classify each of the local space. Each local model developed is an outcome from the local spaces defined, sampled abstracts, through Query Based Sampling, from their own region and other regions play a vital role in the models generated. To check whether the essence of Query Based Sampling is a merit or demerit for the model, the performance of the developed model is being been compared with Local LSA model developed without Query Based Sampling using PR curve (see Figure 26).

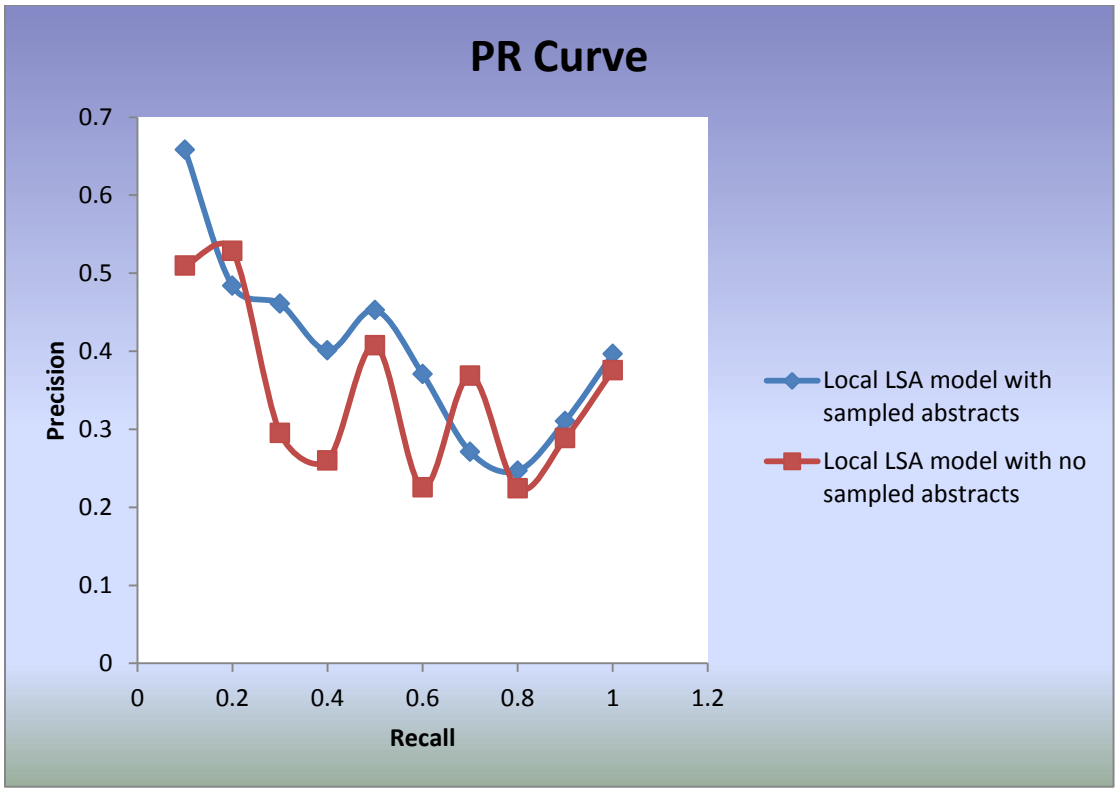

Figure 26 Averaged 11-point Precision/Recall graph plotted across selected 20 queries for Local LSA model which has representative information and local LSA model which does not have representative information.

From the above graph, the Local LSA model with sampled abstracts from its own class as well as from other classes is comparatively stable than the local LSA model with sampled representative abstracts. MAP for the Local model, where the Query based sampling of abstracts is not utilized to load the textual information is **0.20234** which is also 3% less than the proposed Local LSA model. So, Query based sampling of abstracts to define the space of local model is a positive incorporation for the model as the noise is greatly reduced and resulted in robust model. To make a comprehensive analysis about the robustness of the models, PR curve is drawn and compared for the local LSA model

with representative abstracts, local LSA model with all possible abstracts and Global model and is shown below. It is very apparent that LLSA model developed is robust compared to other two models from the Figure 27.

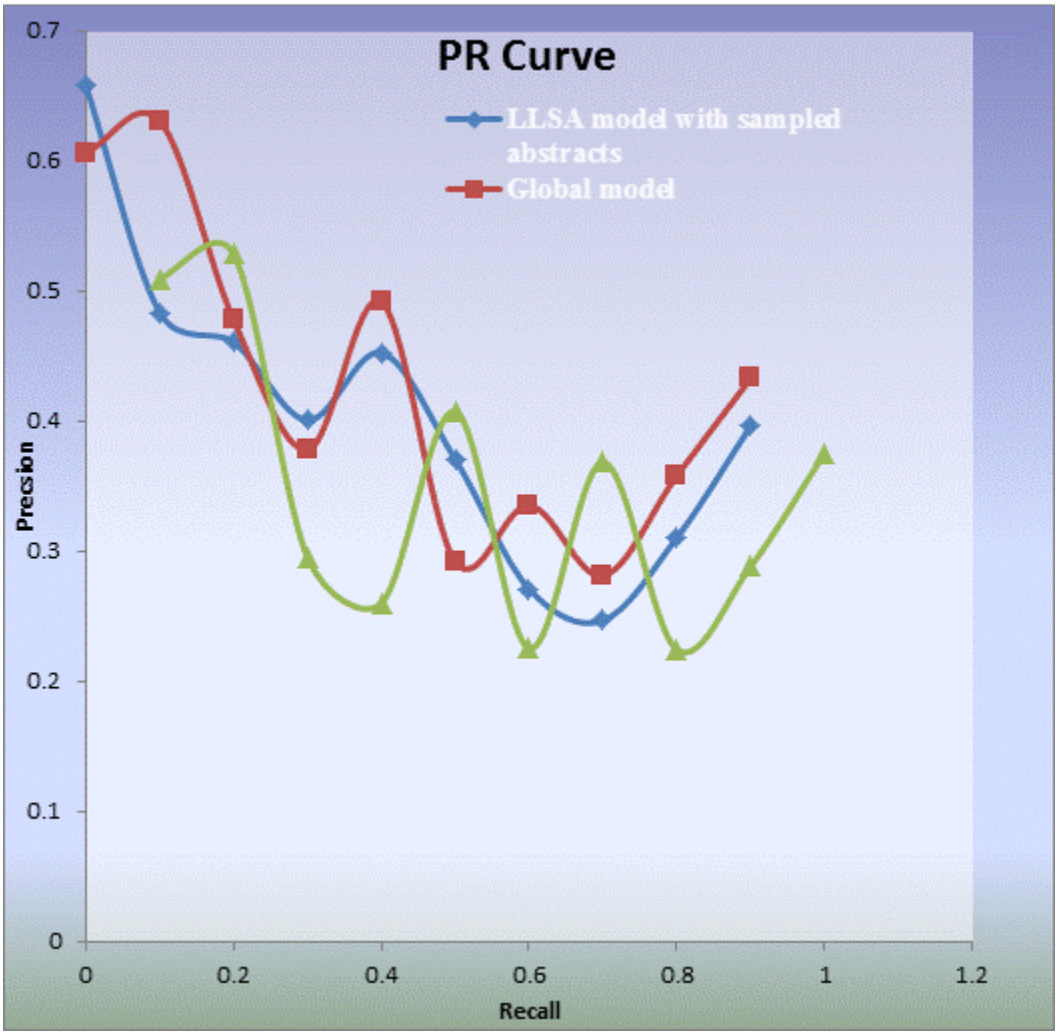

Figure 27 Averaged 11-point Precision/Recall graph plotted across selected 20 queries for LLSA model with representative information, LLSA model with all possible representative information and Global LSA model.

Performance Evaluation based on the degree of relevancy of retrieved results

As LLSA model is scalable, more features (Drugs/Chemicals in this study) can be added, as a result, evaluation of the model with queries which have voluminous relevant set would be accomplished. So, it is imperative to verify whether the model is capable of pulling high precision in the top ranked result or highly relevant set, when it is scaled. The similarity measure, cosine value in this study ranges differently for different queries, as it is data driven; hence manual thresholding to determine the highly relevant set is not feasible for every single query. So, a relevance model is developed, to automatically cluster the retrieved results into three groups like highly relevant results, moderately relevant results and low relevant results. Fuzzy c means clustering [28] is used to cluster the cosine values of retrieved drugs/chemicals for every query. The whole set of retrieved Drugs/Chemicals for every query is categorized into three groups such as highly relevant set, reasonably relevant set, and poor relevant set. The same 20 queries are used to analyze each of the relevant set.

Subjective analysis is done on each of the relevant set to discover whether the False Positives retrieved by the model are factually incorrect or they are not captured by the Gold Standard as relevant.

Highly Relevant Set

As it is expected, larger number of relevant results from the Gold standard is retrieved as the top ranked associated results for any selected query. As a consequence, average precision of the results for each of the query is elevated than the average precision when all retrieved results are taken into consideration and eventually MAP came up to 0.534.

The model created has higher recall in highly relevant set as well as it extracted some results which were irrelevant against PharmGKB. But when the results, drug-disease associations were checked manually they were discovered to be relevant in published literatures. As an instance, for query Alzheimer disease, Vitamin E is retrieved as highly relevant Drug/Chemical, as shown in Figure 28 which is not captured by PharmGKB. **[31]** Showed that the intake of Vitamin E exhibits the pronounced protective effect of Alzheimer disease.

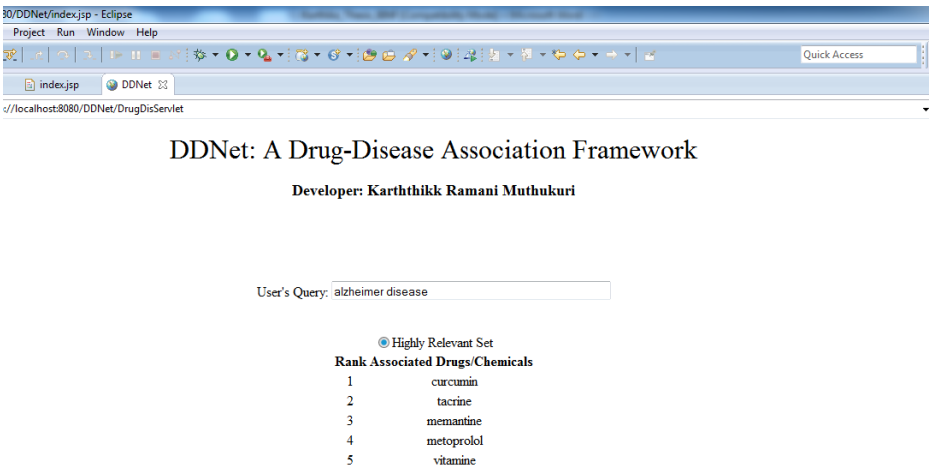

Figure 28 Screen shot showing the highly relevant set with Vitamin E at ranks 5 for the query Alzheimer disease.

Reasonably relevant set

MAP for reasonably relevant set came up to 0.17 which is significantly less than MAP for highly relevant set (0.534). This implies that comparatively, more number of

relevant results is retrieved in the top ranked set with high association values, than the number of relevant results with moderate association values. Magnesium came as reasonably relevant associated chemical from DDNet (Figure 29) which is not a relevant result for Alzheimer disease from PharamGKB. [32] Studied the effects of altered magnesium levels in mild-moderate Alzheimer disease which strongly proved that magnesium level has high impact in Alzheimer disease.

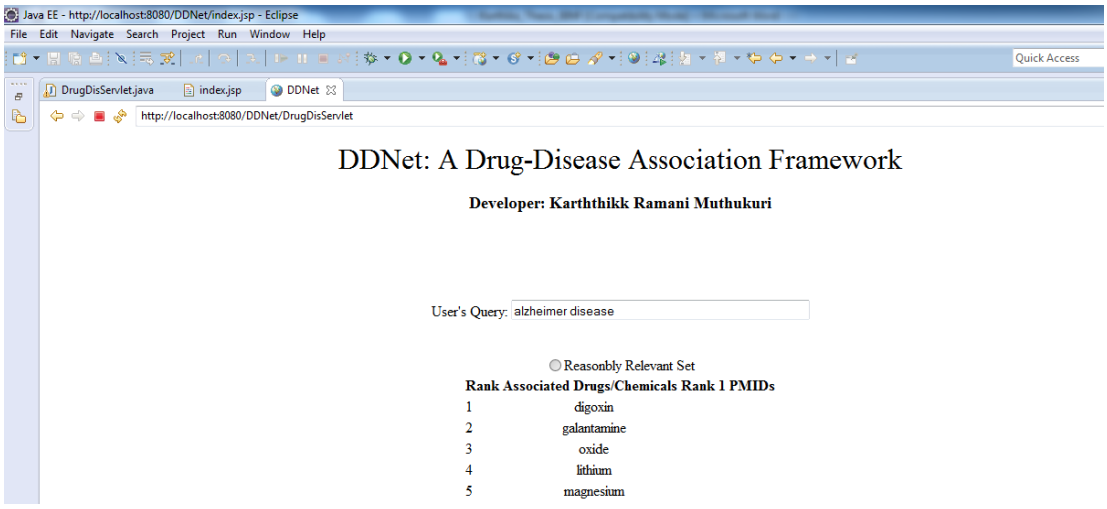

Figure 29 Screen shot showing the reasonably relevant set with magnesium at rank 5 for the query Alzheimer disease.

Poor Relevant Set

MAP for reasonably relevant set came up to 0.12 (Figure 30). [33] Suggested considering vitamin K in future investigations on the role of diet in Alzheimer's disease.

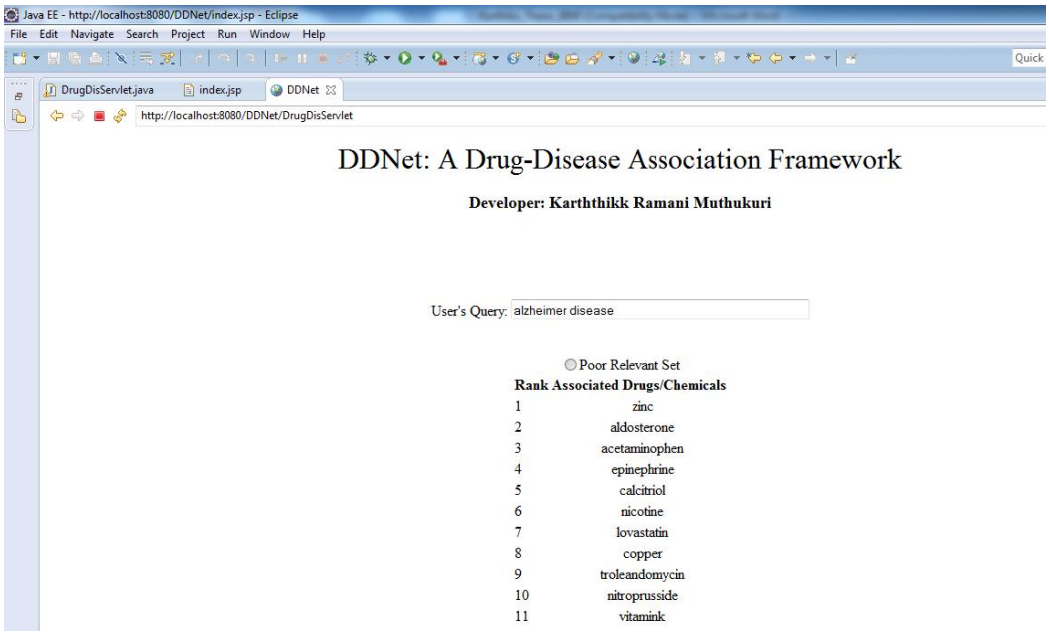

Figure 30 Screen shot showing the poor relevant set with vitamin K at rank 11 for the query Alzheimer disease.

From the MAPs obtained for the above three relevancy sets of results, it is proved that Local LSA models generated for DDNet, is capable of retrieving relevant items as the top ranked results which enable any common users to gain some fundamental knowledge about the related Drugs/Chemicals based on their information need. Also, as described in section 3.1, recall for almost all selected queries is 100% which enables the specific users like medical researchers to gain thorough knowledge about their information need. Also, some of the results which came as top ranked results but not within the highly relevant set and not published can be meticulously taken as Knowledge Discovery and be used for further research. Additionally, as the network is scalable to include lot more domain specific biological entities, precisely Drugs/Chemical here, higher Mean Average Precision can be expected for the Highly Relevant Set. Also, for

the retrieved results which are not relevant from the Gold standard, PharmGKB, above mentioned citations have proofs that they are relevant; i.e. results which are extracted as false positives are actually not false from the above mentioned citations. Even an associated Drug/Chemical for the query in the poorly relevant set seems to be fairly correct as per the citations.

Chapter V

Conclusion

This study was set out to explore semantically related drug-disease associations, to expedite the application of drug repositioning, by implementing information retrieval technique, LSA, on noise free literature from PubMed. This study also sought to develop a scalable, robust, efficient and bias free framework DDNet, to rank the associations retrieved from LSA model, for user query, thereby facilitating the medical researchers to get forward in their goal. Query Based Sampling of abstracts is executed to filter the garbage from literature data and Local LSA models are created from the filtered data to ensure the scalability and robustness in the framework. LLSA incorporated on sampled textual data resulted in robust semantic model which is evident from PR curve analysis for selected twenty queries. PR curve analysis showed the robust nature of the proposed LLSA model and thereby DDNet itself. MAP was computed to be 0.2331 which is approximately 5% greater than the traditional semantic model. Even, the retrieved associations from the model which are not relevant are substantiated for its correctness through Medline citations from the subjective evaluation.

The scale of this study is limited with 53 features, but the LLSA model is generated as scalable. To achieve the complete usability of the proposed model and framework, by the medical experts and researchers, higher recall and precision is one of the chief aspects to be targeted. This can be achieved by incorporating more sampled features into the system at the local model level which will retrieve all possible relevant results.

This work has offered an accurate semantic model over traditional LSA model in the application of Drug repositioning and was conducted on specific domain of drug-disease network. The model can even be scaled with varied biological concepts as features to spread the usage; like facilitating the hot topic Personalized Medicine [34] by inclusion of genes.

References

1. Roberts RJ: **PubMed Central: The GenBank of the published literature.** *Proc Natl Acad Sci U S A* 2001, **98:**381-382.

2. Lu Z: **PubMed and beyond: a survey of web tools for searching biomedical literature.** *Database (Oxford)* 2011, **2011:**baq036.

3. Wang J, Cetindil I, Ji S, Li C, Xie X, Li G, Feng J: **Interactive and fuzzy search: a dynamic way to explore MEDLINE.** *Bioinformatics* 2010, **26:**2321-2327.

4. Plake C, Schiemann T, Pankalla M, Hakenberg J, Leser U: **AliBaba: PubMed as a graph.** *Bioinformatics* 2006, **22:**2444-2445.

5. Booth B, Zemmel R: **Prospects for productivity.** *Nat Rev Drug Discov* 2004, **3:**451-456.

6. Dudley JT, Deshpande T, Butte AJ: **Exploiting drug-disease relationships for computational drug repositioning.** *Briefings in Bioinformatics* 2010, **12:**303-311.

7. Jiao Li, Xiaoyan Zhu, mail JYC: **Building Disease-Specific Drug-Protein Connectivity Maps from Molecular Interaction Networks and PubMed Abstracts** *PLoS Comput Biol* 2008, **5**.

8. Hu G, Agarwal P: **Human disease-drug network based on genomic expression profiles.** *PLoS One* 2009, **4:**e6536.

9. Deerwester S: **Improving Information Retrieval with Latent Semantic Indexing.** *Proceedings of the 51st Annual Meeting of the American Society for Information Science* 1988, **25:**36-40.

10. Manning C, Raghavan P, Schutz H: **Introduction to Information Retrieval.** *Cambride University Press* 2008, 151-156.

11. Abedi V, Zand R, Yeasin M, Faisal FE: **An automated framework for hypotheses generation using literature.** *BioData Min*, 2012, **5:**13-25.

12. Klein TE, Altman RB: **PharmGKB: the pharmacogenetics and pharmacogenomics knowledge base.** *Pharmacogenomics J* 2004, **4:**1.

13. Shatkay H, Feldman R: **Mining the biomedical literature in the genomic era: an overview.** *J Comput Biol* 2003, **10:**821-855.

14. Tsuruoka Y, Tsujii J, Ananiadou S: **FACTA: a text search engine for finding associated biomedical concepts.** *Bioinformatics* 2008, **24:**2559-2560.

15. Becker KG, Hosack DA, Dennis G, Jr., Lempicki RA, Bright TJ, Cheadle C, Engel J: **PubMatrix: a tool for multiplex literature mining.** *BMC Bioinformatics* 2003, **4:**61-67.

16. Chen H, Sharp BM: **Content-rich biological network constructed by mining PubMed abstracts.** *BMC Bioinformatics* 2004, **5:**147-160.

17. Homayouni R, Heinrich K, Wei L, Berry MW: **Gene clustering by latent semantic indexing of MEDLINE abstracts.** *Bioinformatics* 2005, **21:**104-115.

18. States DJ, Ade AS, Wright ZC, Bookvich AV, Athey BD: **MiSearch adaptive pubMed search tool.** *Bioinformatics* 2009, **25:**974-976.

19. Altman RB: **PharmGKB: a logical home for knowledge relating genotype to drug response phenotype.** *Nat Genet* 2007, **39:**426-429.

20. Hewett M, Oliver DE, Rubin DL, Easton KL, Stuart JM, Altman RB, Klein TE: **PharmGKB: the Pharmacogenetics Knowledge Base.** *Nucleic Acids Res* 2002, **30:**163-165.

21. Zhu Q, Freimuth RR, Pathak J, Durski MJ, Chute CG: **Disambiguation of PharmGKB drug-disease relations with NDF-RT and SPL.** *J Biomed Inform* 2013,**46**:690-696.

22. Yu H, Kim T, Oh J, Ko I, Kim S, Han WS: **Enabling multi-level relevance feedback on PubMed by integrating rank learning into DBMS.** *BMC Bioinformatics*, **11 Suppl 2:**S6.

23. Fontaine JF, Barbosa-Silva A, Schaefer M, Huska MR, Muro EM, Andrade-Navarro MA: **MedlineRanker: flexible ranking of biomedical literature.** *Nucleic Acids Res* 2009, **37:**W141-146.

24. Wishart DS, Knox C, Guo AC, Shrivastava S, Hassanali M, Stothard P, Chang Z, Woolsey J: **DrugBank: a comprehensive resource for in silico drug discovery and exploration.** *Nucleic Acids Res* 2006, **34:**D668-672.

25. Davis AP, Murphy CG, Saraceni-Richards CA, Rosenstein MC, Wiegers TC, Mattingly CJ: **Comparative Toxicogenomics Database: a knowledgebase and discovery tool for chemical-gene-disease networks.** *Nucleic Acids Res* 2009, **37:**D786-792.

26. Bezdek CJ, Ehrlich R, Full W: **FCM: The fuzzy c-means clustering algorithm.** *Computers and Geosciences* 1984, **10:**191-203.

27. Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabasi AL: **The human disease network.** *Proc Natl Acad Sci U S A* 2007, **104:**8685-8690.

28. Naressi A, Couturier C, Devos JM, Janssen M, Mangeat C, de Beer R, Graveron-Demilly D: **Java-based graphical user interface for the MRUI quantitation package.** *MAGMA* 2001, **12:**141-152.

29. Li FJ, Shen L, Ji HF: **Dietary intakes of vitamin E, vitamin C, and beta-carotene and risk of Alzheimer's disease: a meta-analysis.** *J Alzheimers Dis*, **31:**253-258.

30. Barbagallo M, Belvedere M, Di Bella G, Dominguez LJ: **Altered ionized magnesium levels in mild-to-moderate Alzheimer's disease.** *Magnes Res*, **24:**S115-121.

31. Presse N, Shatenstein B, Kergoat MJ, Ferland G: **Low vitamin K intakes in community-dwelling elders at an early stage of Alzheimer's disease.** *J Am Diet Assoc* 2008, **108:**2095-2099.

32. Hamburg MA, Collins FS: **The path to personalized medicine.** *N Engl J Med*, **363:**301-304.

Appendices

A. Associated drugs/Chemicals for 20 selected queries for Local LSA model with sampled information (Proposed Methodology)

Alzheimer disease	galantamine, metoprolol, memantine, isoproterenol, epinephrine, nitrous oxide, tacrine, aldosterone, nitroprusside, digoxin, betamethosone, nicotine, morphine, triamcin, beclome, hydrocortisone, codeine, estrone, testosterone, budenoside, ethinyl, lithium, warfarin, levonorgestrol, pravastatin, nore, methadone, magnesium, lovastatin, tamoxifen, mifepristone, acetaminophen, zinc, dicloxacillin, copper, macrolides, prednisolone, erythromycin, amoxicillin, vitamine, curcumin, tacrolimus, sirolimus, cyclosporine, ritonavir, prednisone, troleandomycin, calcitriol, vitamink, mycophenolic, cisplatin, vincristine, idarubicin
Neurodegenerative disease	tamoxifen, cisplatin, testosterone, curcumin, memantine, estrone, galantamine, calcitriol, nicotine, mifepristone, idarubicin, sirolimus, epinephrine, isoproterenol, vincristine, nitroprusside, vitamine, tacrine, morphine, zinc, cyclosporine, magnesium, macrolides, hydrocortisone, tacrolimus, triamcin, copper, vitamink, ethinyl, prednisolone, lithium, lovastatin, aldosterone, nore, betamethosone, prednisone, pravastatin, metoprolol, troleandomycin, mycophenolic, levonorgestrol, codeine, erythromycin, beclome,

	ritonavir, dicloxacillin, budenoside, warfarin, acetaminophen
Dementia	memantine, galantamine, nicotine, nitrous oxide, tacrine, copper, epinephrine, isoproterenol, morphine, zinc, lithium, methadone, magnesium, troleandomycin, vitamine, curcumin, metoprolol, vitamink, nitroprusside, mifepristone, acetaminophen, beclome, triamcin, budenoside, cyclosporine, sirolimus, macrolides, calcitriol, hydrocortisone, betamethosone, lovastatin, tacrolimus, digoxin, erythromycin, testosterone, codeine, ritonavir, aldosterone, warfarin, ethinyl, tamoxifen, nore, pravastatin, estrone, levonorgestrol
Drug Toxicity	vincristine, warfarin, digoxin, idarubicin, prednisone, cisplatin, mycophenolic, curcumin, prednisolone, lovastatin, tacrolimus, cyclosporine, pravastatin, ritonavir, triamcin, sirolimus, acetaminophen, vitamink, isoproterenol, nitroprusside, metoprolol, nitrous oxide, macrolides, betamethosone, budenoside, codeine, vitamine, morphine, erythromycin, amoxicillin, epinephrine, dicloxacillin, beclome, calcitriol, troleandomycin, lithium, tamoxifen, aldosterone, magnesium, mifepristone, hydrocortisone, tacrine, copper, zinc, testosterone, ethinyl, memantine, nicotine, estrone, galantamine, nore, levonorgestrol

Transplantation	prednisone, mycophenolic, prednisolone, triamcin, tacrolimus, metoprolol, beclome, warfarin, cyclosporine, amoxicillin, vincristine, nitrous oxide, betamethosone, budenoside, idarubicin, digoxin, erythromycin, macrolides, dicloxacillin, ritonavir, vitamink, sirolimus, acetaminophen, aldosterone, magnesium, cisplatin, pravastatin, hydrocortisone, nitroprusside, lovastatin, lithium, zinc, troleandomycin, methadone, copper, vitamine, epinephrine, levonorgestrol, calcitriol, ethinyl, memantine, codeine, nore, isoproterenol, galantamine, tamoxifen, nicotine, mifepristone, testosterone, morphine, tacrine, estrone, curcumin
Depression	memantine, nicotine, galantamine, lithium, testosterone, zinc, isoproterenol, epinephrine, tacrine, tamoxifen, mifepristone, copper, estrone, magnesium, morphine, methadone, vitamine, nitroprusside, hydrocortisone, nore, troleandomycin, calcitriol, ethinyl, metoprolol, sirolimus, aldosterone, vitamink, curcumin, codeine, lovastatin, levonorgestrol, digoxin, acetaminophen, cyclosporine, macrolides, cisplatin, pravastatin, tacrolimus, nitrous oxide, betamethosone, triamcin, erythromycin, idarubicin, ritonavir, budenoside, beclome, mycophenolic, vincristine, warfarin, prednisolone, prednisone, dicloxacillin, amoxicillin
Venous thrombosis	warfarin, metoprolol, triamcin, prednisone, vincristine, beclome, idarubicin, prednisolone, pravastatin, mycophenolic, nitrous oxide,

	galantamine, digoxin, betamethosone, aldosterone, memantine, methadone, epinephrine, budenoside, cyclosporine, tacrolimus, cisplatin
Thrombosis embolism	warfarin, acetaminophen, curcumin, vitamine, beclome, vitamink, mycophenolic, prednisolone, budenoside, cisplatin, prednisone, lovastatin, digoxin, triamcin, pravastatin, ethinyl, levonorgestrol, amoxicillin, metoprolol, tacrolimus, lithium, betamethosone, codeine, idarubicin, cyclosporine, vincristine, nore, erythromycin, nitrous oxide, tamoxifen, macrolides, memantine, sirolimus, troleandomycin, ritonavir, methadone, magnesium, tacrine, hydrocortisone, nicotine, aldosterone, copper, nitroprusside, isoproterenol, estrone, testosterone, dicloxacillin, morphine, galantamine, mifepristone, calcitriol, epinephrine, zinc
Breast neoplasm	aldosterone, pravastatin, lovastatin, tacrine, nitroprusside, metoprolol, vitamine, isoproterenol, epinephrine, magnesium, testosterone, vitamink, morphine, tamoxifen, curcumin, hydrocortisone, troleandomycin, memantine, amoxicillin, mifepristone, codeine, levonorgestrol, nicotine, sirolimus, calcitriol, erythromycin, macrolides, digoxin, warfarin, estrone, zinc, nore, galantamine, cyclosporine, methadone, tacrolimus, lithium, idarubicin, budenoside, acetaminophen, betamethosone, ethinyl,

	beclome, prednisone, mycophenolic, prednisolone, ritonavir, cisplatin, triamcin, copper, dicloxacillin, vincristine, nitrous oxide
Hypertension	mycophenolic, nitrous oxide, prednisone, cyclosporine, lithium, tacrolimus, prednisolone, magnesium, aldosterone, vitamink, warfarin, metoprolol, nitroprusside, triamcin, beclome, ritonavir, sirolimus, acetaminophen, zinc, macrolides, hydrocortisone, idarubicin, vitamine, pravastatin, digoxin, calcitriol, budenoside, codeine, betamethosone, memantine, lovastatin, isoproterenol, epinephrine, ethinyl, troleandomycin, nore, vincristine, erythromycin, amoxicillin, levonorgestrol, methadone, tamoxifen, copper, dicloxacillin, estrone, galantamine, testosterone, morphine, mifepristone, cisplatin, nicotine, tacrine, curcumin
Nausea	codeine, morphine, acetaminophen, methadone, digoxin, cisplatin, tamoxifen, curcumin, estrone, nicotine, testosterone, ethinyl, vincristine, hydrocortisone, vitamine, nore, vitamink, metoprolol, pravastatin, mycophenolic, idarubicin, isoproterenol, warfarin, tacrine, memantine, epinephrine, copper, lithium, dicloxacillin, ritonavir, nitroprusside, amoxicillin, betamethosone, prednisone, prednisolone, troleandomycin, lovastatin, tacrolimus, galantamine, budenoside, cyclosporine, mifepristone, macrolides, erythromycin, zinc, levonorgestrol, beclome, sirolimus, nitrous oxide, magnesium,

	aldosterone, triamcin, calcitriol
Coronary artery disease	digoxin, warfarin, pravastatin, metoprolol, nitrous oxide, codeine, methadone, lovastatin, morphine, nicotine, nitroprusside, estrone, epinephrine, levonorgestrol, tamoxifen, ritonavir, erythromycin, galantamine, tacrine, acetaminophen, aldosterone, isoproterenol, budenoside, vitamine, ethinyl, triamcin, amoxicillin, nore, hydrocortisone, testosterone, betamethosone, vitamink, macrolides, lithium, sirolimus, dicloxacillin, mifepristone, curcumin, tacrolimus, prednisone, prednisolone, copper, beclome, zinc, magnesium, cyclosporine, mycophenolic, troleandomycin, calcitriol, vincristine, memantine, idarubicin, cisplatin
Myocardial infarction	metoprolol, isoproterenol, epinephrine, pravastatin, vitamine, digoxin, aldosterone, warfarin, lithium, galantamine, vitamink, nitroprusside, lovastatin, triamcin, nicotine, magnesium, hydrocortisone, zinc, memantine, testosterone, acetaminophen, tacrine, sirolimus, copper, calcitriol, nitrous oxide, tamoxifen, mifepristone, macrolides, morphine, cyclosporine, estrone, tacrolimus, troleandomycin, codeine, budenoside, betamethosone, prednisolone, methadone, beclome, curcumin, levonorgestrol, prednisone, idarubicin, ethinyl, erythromycin, cisplatin, mycophenolic, vincristine, dicloxacillin, nore, ritonavir, amoxicillin

Pain	codeine, nitrous oxide, methadone, metoprolol, warfarin, morphine, digoxin, prednisone, beclome, prednisolone, acetaminophen, nicotine, mycophenolic, pravastatin, betamethosone, budenoside, memantine, triamcin, hydrocortisone, nore, aldosterone, dicloxacillin, levonorgestrol, amoxicillin, lovastatin, epinephrine, nitroprusside, ritonavir, galantamine, ethinyl, estrone, lithium, idarubicin, tacrolimus, testosterone, cyclosporine, tamoxifen, vincristine, vitamink, vitamine, isoproterenol, erythromycin, troleandomycin, macrolides, mifepristone, magnesium, zinc, sirolimus, tacrine, cisplatin, copper, calcitriol, curcumin
Leukemia	erythromycin, vitamink, idarubicin, amoxicillin, zinc, ritonavir, cisplatin, lithium, morphine, macrolides, dicloxacillin, copper, triamcin, prednisolone, tacrolimus, vincristine, prednisone, cyclosporine, betamethosone, magnesium, mifepristone, codeine, acetaminophen, methadone, nitrous oxide, lovastatin, warfarin, tamoxifen, digoxin, calcitriol, sirolimus, mycophenolic, vitamine, levonorgestrol, nicotine, metoprolol, beclome, budenoside, testosterone, hydrocortisone, pravastatin, troleandomycin, epinephrine, memantine, curcumin, nore, nitroprusside, ethinyl, estrone, aldosterone, isoproterenol, galantamine, tacrine
Pulmonary	warfarin, acetaminophen, curcumin, vitamine, beclome, vitamink,

embolism	mycophenolic, prednisolone, budenoside, cisplatin, prednisone, lovastatin, digoxin, triamcin, pravastatin, ethinyl, levonorgestrol, amoxicillin, metoprolol, tacrolimus, lithium, betamethosone, codeine, idarubicin, cyclosporine, vincristine, nore, erythromycin, nitrous oxide, tamoxifen, macrolides, memantine, sirolimus, troleandomycin, ritonavir, methadone, magnesium, tacrine, hydrocortisone, nicotine, aldosterone, copper, nitroprusside, isoproterenol, estrone, testosterone, dicloxacillin, morphine, galantamine, mifepristone, calcitriol, epinephrine, zinc
Warfarin	warfarin, pravastatin, prednisone, vincristine, lovastatin, amoxicillin, methadone, erythromycin, idarubicin, dicloxacillin, prednisolone, ritonavir, metoprolol, vitamink, digoxin, nitrous oxide, codeine, mycophenolic, macrolides, sirolimus, tacrolimus, triamcin, vitamine, acetaminophen, morphine, tamoxifen, budenoside, betamethosone, cisplatin, cyclosporine, nicotine, beclome, galantamine, nitroprusside, calcitriol, hydrocortisone, estrone, aldosterone, tacrine, epinephrine, testosterone, magnesium, isoproterenol, troleandomycin, lithium, memantine, zinc, levonorgestrol, curcumin, copper, ethinyl estradiol
Osteosarcoma	vincristine, idarubicin, cisplatin, prednisone, mycophenolic, curcumin, tacrolimus, cyclosporine, macrolides, sirolimus,

	tamoxifen, erythromycin, amoxicillin, vitamink, lovastatin, vitamine, dicloxacillin
Neutropenia	vincristine, idarubicin, cisplatin, prednisone, mycophenolic, nore, morphine, ethinyl, levonorgestrol, nitrous oxide, dicloxacillin, estrone, memantine, digoxin, aldosterone, galantamine, tacrolimus, cyclosporine, amoxicillin, epinephrine, ritonavir, nitroprusside, budenoside, nicotine, prednisolone, codeine, testosterone, tacrine, erythromycin, hydrocortisone, methadone, acetaminophen, tamoxifen, macrolides, beclome, magnesium, betamethosone, copper, lithium, mifepristone, zinc, pravastatin, lovastatin, isoproterenol, vitamin E
Schizophrenia	curcumin, nitroprusside, memantine, tamoxifen, isoproterenol, testosterone, codeine, epinephrine, copper, mifepristone, zinc, morphine, vitamine, nicotine, tacrine, nitrous oxide, magnesium, estrone, galantamine, lithium, cisplatin, ethinyl, acetaminophen, calcitriol, methadone, sirolimus, hydrocortisone, cyclosporine, troleandomycin, vitamink, nore, macrolides, betamethosone, lovastatin, tacrolimus, idarubicin, pravastatin, aldosterone, triamcin, metoprolol, vincristine, ritonavir, digoxin, budenoside, warfarin, mycophenolic, erythromycin, beclome, prednisolone, levonorgestrol, dicloxacillin, prednisone, amoxicillin

B. Associated drugs/Chemicals for 20 selected queries for Global LSA model

Alzheimer disease	Curcumin, tacrine, memantine, metoprolol, vitamine
Neurodegenerative disease	Tamoxifen, cisplatin, testosterone, curcumin, memantine
Dementia	Memantine, galantamine, nicotine, nitrous oxide, tacrine
Drug Toxicity	Vincristine, warfarin, digoxin, idarubicin, prednisone
Transplantation	Prednisone, mycophenolic acid, prednisolone, triamcin, tacrolimus
Depression	Memantine, galantamine, nicotine, nitrous oxide, tacrine
Venous thrombosis	Warfarin, metoprolol, triamcinolone, prednisone, vincristine
Thrombosis embolism	Curcumin, lovastatin, cisplatin, morphine, nicotine
Breast neoplasm	Aldosterone, pravastatin, lovastatin, tacrine, nitroprusside
Hypertension	mycophenolic, oxide, prednisone, cyclosporine, lithium
Nausea	codeine, morphine, acetaminophen, methadone, digoxin
Coronary artery disease	digoxin, warfarin, pravastatin, metoprolol, oxide
Myocardial infarction	prednisone, metoprolol, digoxin, mycophenolic, vincristine
Pain	codeine, oxide, methadone, metoprolol, warfarin
Leukemia	erythromycin, vitamink, idarubicin, amoxicillin, zinc

Pulmonary embolism	warfarin, acetaminophen, curcumin, vitamine, become
Warfarin	nitrous oxide, aldosterone, tacrine, galantamine, morphine
Osteosarcoma	vincristine, idarubicin, cisplatin, prednisone, mycophenolic
Neutropenia	vincristine, idarubicin, cisplatin, prednisone, mycophenolic
Schizophrenia	curcumin, nitroprusside, memantine, tamoxifen, isoproterenol

C. Associated drugs/Chemicals for 20 selected queries for Local LSA model without sampled information

Alzheimer disease	Tacrine, Galantamine, Digoxin, MycophenolicAcid, Pravastatin, CopperSulfate, Tacrolimus, Lithium, VitaminE, Cyclosporine, Lovastatin, Sirolimus, Nicotine, Metoprolol, Zinc, Magnesium, Curcumin, Macrolides, Nitroprusside, Aldosterone, Isoproterenol, Acetaminophen, Epinephrine, VitaminK, Prednisone, Memantine, Prednisolone, Beclomethasone, Budesonide, Hydrocortisone, Triamcinolone, Erythromycin, Betamethasone, Warfarin, Cisplatin, Amoxicillin, Calcitriol, Troleandomycin, Testosterone, Tamoxifen, Ritonavir, Estrone, EthinylEstradiol, Morphine, Idarubicin, Mifepristone, NitrousOxide, Vincristine, Norethindrone
Neurodegenerative disease	Curcumin, CopperSulfate, Tacrolimus, Cyclosporine, Lovastatin, MycophenolicAcid, Sirolimus, Tamoxifen, Memantine, Tacrine, Macrolides, Calcitriol, Lithium, Ritonavir, Pravastatin, Cisplatin, Zinc, Triamcinolone, Magnesium, Galantamine, Acetaminophen, Estrone, Nitroprusside, VitaminE, Digoxin, Prednisolone, EthinylEstradiol, Mifepristone, Norethindrone, VitaminK, Prednisone, Erythromycin, Betamethasone, Hydrocortisone, Troleandomycin, Aldosterone, Nicotine, Testosterone, Levonorgestrel, Metoprolol, Isoproterenol, Vincristine,

	Epinephrine, Idarubicin, Warfarin, Budesonide, Amoxicillin, Morphine, Methadone, NitrousOxide, Codeine, Beclomethasone
Dementia	Tacrine, Memantine, Lithium, Digoxin, MycophenolicAcid, Galantamine, Tacrolimus, Methadone, Metoprolol, Acetaminophen, Cyclosporine, Triamcinolone, Tamoxifen, Ritonavir, Prednisone, Curcumin, Sirolimus, Warfarin, Prednisolone, NitrousOxide, Calcitriol, Morphine, Codeine, Magnesium, Pravastatin, Macrolides, Isoproterenol, Lovastatin, Betamethasone, Aldosterone, Nitroprusside, Zinc, VitaminK
Drug	Digoxin, Troleandomycin, Codeine, Idarubicin, Metoprolol, Morphine, Prednisone, Acetaminophen, MycophenolicAcid, Methadone, Warfarin, Ritonavir, Cyclosporine, Prednisolone, Amoxicillin, Erythromycin, Lithium, Pravastatin, Cisplatin, Tacrine, Vincristine, Tacrolimus, Dicloxacillin, Nicotine, Curcumin, Macrolides, NitrousOxide, Sirolimus, Lovastatin, Beclomethasone
Transplantation	Ritonavir, Tamoxifen, Memantine, Tacrine, Tacrolimus, MycophenolicAcid, Cyclosporine, Lovastatin, Prednisone, Warfarin, Triamcinolone, Sirolimus, Calcitriol, Betamethasone, Acetaminophen, Cisplatin, Pravastatin, Digoxin, Methadone,

	Galantamine, Mifepristone, Prednisolone, NitrousOxide, Codeine, Hydrocortisone, Norethindrone, Vincristine, Metoprolol, Aldosterone, Macrolides, EthinylEstradiol, Isoproterenol, Estrone, Levonorgestrel, Lithium, Nitroprusside, Idarubicin
Depression	Tacrine, Memantine, Lithium, Digoxin, MycophenolicAcid, Galantamine, Tacrolimus, Methadone, Metoprolol, Acetaminophen, Cyclosporine, Triamcinolone, Tamoxifen, Ritonavir, Prednisone, Curcumin, Sirolimus, Warfarin, Prednisolone, NitrousOxide, Calcitriol, Morphine, Codeine, Magnesium, Pravastatin, Macrolides, Isoproterenol, Lovastatin, Betamethasone, Aldosterone, Nitroprusside, Zinc, VitaminK
Venous thrombosis	Warfarin, VitaminK, Levonorgestrel, Norethindrone, Lovastatin, Pravastatin, EthinylEstradiol, Tamoxifen, Calcitriol, Prednisolone, Prednisone, Estrone, Cisplatin, Sirolimus, Triamcinolone, Mifepristone, Testosterone, Troleandomycin, Cyclosporine, VitaminE, Vincristine, Macrolides, Tacrolimus, MycophenolicAcid, Hydrocortisone, Ritonavir, Idarubicin, Betamethasone, Dicloxacillin, Magnesium, Digoxin, Zinc, Aldosterone, Erythromycin, CopperSulfate, Nitroprusside, Acetaminophen, Beclomethasone, Metoprolol, Curcumin, Epinephrine, Isoproterenol, Budesonide, Lithium, Tacrine,

	Amoxicillin, NitrousOxide, Galantamine, Nicotine, Memantine
Thrombosis embolism	Warfarin, Ritonavir, Lovastatin, Pravastatin, Prednisone, Vincristine, Dicloxacillin, Levonorgestrel, Idarubicin, Prednisolone, Cyclosporine, Calcitriol, Norethindrone, Tamoxifen, Cisplatin, Mifepristone, EthinylEstradiol, Troleandomycin, MycophenolicAcid, Sirolimus, Erythromycin, Tacrolimus, Testosterone, VitaminK, Amoxicillin, Estrone, Macrolides, Triamcinolone, Digoxin, Betamethasone, Budesonide, Aldosterone, Hydrocortisone, Nitroprusside, Beclomethasone, Curcumin, Metoprolol, Isoproterenol
Breast neoplasm	Tamoxifen, Norethindrone, Estrone, EthinylEstradiol, Levonorgestrel, Testosterone, Cisplatin, Mifepristone, Vincristine, Idarubicin, Prednisone, Hydrocortisone, Calcitriol, Prednisolone, Ritonavir, Triamcinolone, Cyclosporine, Sirolimus, Galantamine, Lithium, MycophenolicAcid, Tacrolimus, Betamethasone, Memantine, Magnesium, Epinephrine, Zinc, Aldosterone, Nicotine, VitaminE, Curcumin, Macrolides, Isoproterenol, Digoxin, Lovastatin, Warfarin, Morphine, CopperSulfate, Nitroprusside, Tacrine, Budesonide, Pravastatin, VitaminK, Troleandomycin, Beclomethasone, Acetaminophen, Methadone, Codeine, Erythromycin, Amoxicillin, Metoprolol, Dicloxacillin

Hypertension	Budesonide, Mifepristone, Prednisolone, Testosterone, Levonorgestrel, Lovastatin, Pravastatin, Betamethasone, Prednisone, Beclomethasone, Troleandomycin, Hydrocortisone, Norethindrone, EthinylEstradiol, Calcitriol, Estrone, Cisplatin, Cyclosporine, Warfarin, Vincristine, Aldosterone, Sirolimus, Triamcinolone, Macrolides, Ritonavir, Tacrolimus, MycophenolicAcid, Nitroprusside, Zinc, Dicloxacillin, Epinephrine, Isoproterenol, VitaminE, VitaminK, Tamoxifen, Magnesium, Erythromycin, Idarubicin, Nicotine, CopperSulfate, Digoxin, Amoxicillin, Acetaminophen, Lithium, Metoprolol, Curcumin, Morphine, Codeine, NitrousOxide, Galantamine, Methadone, Tacrine, Memantine
Nausea	Nill
Coronary artery disease	Pravastatin, Lovastatin, Digoxin, Warfarin, Sirolimus, Aldosterone, Cyclosporine, Troleandomycin, Nitroprusside, Macrolides, Prednisolone, Metoprolol, MycophenolicAcid, Tacrolimus, Magnesium, Ritonavir, VitaminE, Curcumin, Zinc, Calcitriol, Nicotine, Levonorgestrel, Prednisone, Testosterone, VitaminK, EthinylEstradiol, Isoproterenol, CopperSulfate, Lithium, Mifepristone, Epinephrine, Budesonide, Triamcinolone,

	Norethindrone, Estrone, Erythromycin, Hydrocortisone, Cisplatin, Betamethasone, Beclomethasone, Dicloxacillin, Idarubicin, Vincristine, Tamoxifen, Acetaminophen, Amoxicillin, Galantamine, Memantine, Tacrine, NitrousOxide, Morphine, Methadone, Codeine
Myocardial infarction	Troleandomycin, Budesonide, Prednisolone, Beclomethasone, Betamethasone, Mifepristone, Levonorgestrel, Pravastatin, Lovastatin, Prednisone, Testosterone, Hydrocortisone, Warfarin, EthinylEstradiol, Cisplatin, Calcitriol, Norethindrone, Cyclosporine, Macrolides, Sirolimus, Aldosterone, Estrone, Triamcinolone, Tacrolimus, Nitroprusside, Zinc, MycophenolicAcid, Vincristine, Magnesium, Erythromycin, Isoproterenol, Epinephrine, Digoxin, VitaminK, Dicloxacillin, VitaminE, Nicotine, Ritonavir, Tamoxifen, Amoxicillin, Acetaminophen, Metoprolol, CopperSulfate, Lithium, Idarubicin, Curcumin, Morphine, Codeine, NitrousOxide, Galantamine, Tacrine, Methadone, Memantine
Pain	nill
Leukemia	Vincristine, Prednisone, Idarubicin, Prednisolone, Cisplatin, Ritonavir, Cyclosporine, Budesonide, Troleandomycin, Testosterone, Calcitriol, Levonorgestrel, Mifepristone, Macrolides,

	EthinylEstradiol, Estrone, Norethindrone, Lovastatin, Sirolimus, Tacrolimus, MycophenolicAcid, Pravastatin, Triamcinolone, Betamethasone, Hydrocortisone, Zinc, Curcumin, Dicloxacillin, Warfarin, Beclomethasone, Magnesium, Erythromycin, VitaminK, VitaminE, Aldosterone, Tamoxifen, Nitroprusside, Amoxicillin, Digoxin, CopperSulfate, Isoproterenol, Lithium, Epinephrine, Acetaminophen, Nicotine, Metoprolol, Morphine, NitrousOxide, Codeine, Methadone, Memantine, Galantamine, Tacrine
Pulmonary embolism	Warfarin, VitaminK, Levonorgestrel, Norethindrone, EthinylEstradiol, Tamoxifen, Pravastatin, Lovastatin, Prednisolone, Calcitriol, Prednisone, Cisplatin, Troleandomycin, Estrone, Mifepristone, Sirolimus, Cyclosporine, Digoxin, Triamcinolone, VitaminE, Acetaminophen, Hydrocortisone, Testosterone, Betamethasone, Ritonavir, Zinc, Macrolides, Magnesium, Metoprolol, Tacrolimus, Vincristine, Nitroprusside, Aldosterone, MycophenolicAcid, Dicloxacillin, Isoproterenol, CopperSulfate, Beclomethasone, Epinephrine, Lithium, Erythromycin, Idarubicin, Curcumin, Budesonide, Tacrine, Amoxicillin, Nicotine, NitrousOxide, Memantine, Galantamine, Morphine, Codeine, Methadone
Warfarin	Nicotine, Morphine, Metoprolol, NitrousOxide, Epinephrine,

	Isoproterenol, Beclomethasone, Nitroprusside, Codeine, Digoxin, Troleandomycin, Lithium, Magnesium, Acetaminophen, Aldosterone, Budesonide, Zinc, Methadone, CopperSulfate, Galantamine, Tacrine, Hydrocortisone, VitaminE, Betamethasone, Testosterone, Memantine, Mifepristone, Erythromycin, VitaminK, Macrolides, Curcumin, Amoxicillin, Dicloxacillin, Estrone, EthinylEstradiol, Prednisolone, Cisplatin, Levonorgestrel, Norethindrone, Triamcinolone, Warfarin, Sirolimus, Cyclosporine, Pravastatin, Calcitriol
Osteosarcoma	NitrousOxide, Methadone, Ritonavir, Tacrine, Memantine, Metoprolol, Isoproterenol, Aldosterone, Hydrocortisone, Morphine, Nitroprusside, Digoxin, Epinephrine, Acetaminophen, Levonorgestrel, EthinylEstradiol, Warfarin, Lithium, Norethindrone, Galantamine, Estrone, Mifepristone, Testosterone, Tamoxifen, Lovastatin, Betamethasone, Prednisone, Vincristine, Codeine, Budesonide, Triamcinolone, Cyclosporine, Prednisolone, Calcitriol, Nicotine, Magnesium, Pravastatin, Beclomethasone, Tacrolimus, Zinc, Idarubicin, Sirolimus, VitaminK, Troleandomycin, Macrolides, Cisplatin, MycophenolicAcid, Curcumin, VitaminE, Dicloxacillin, Erythromycin, CopperSulfate
Neutropenia	Dicloxacillin

Schizophrenia	Tacrine, Nicotine, Lithium, Galantamine, Morphine, Methadone, NitrousOxide, Acetaminophen, Memantine, Epinephrine, EthinylEstradiol, Levonorgestrel, Zinc, Magnesium, Codeine, Norethindrone, Estrone, Metoprolol, Hydrocortisone, Digoxin, VitaminE, Nitroprusside, VitaminK, Isoproterenol, Mifepristone, CopperSulfate, Tamoxifen, Testosterone, Aldosterone, Troleandomycin, Betamethasone, Warfarin, Beclomethasone, Curcumin, Calcitriol, Budesonide, Cisplatin, Triamcinolone, Macrolides, Prednisolone, Pravastatin, Cyclosporine, Ritonavir, Erythromycin, Lovastatin, Sirolimus, Tacrolimus, Prednisone, Amoxicillin, Dicloxacillin, Vincristine, Idarubicin, MycophenolicAcid

D. Cosine values for the associated drugs/chemicals for 20 selected queries for the proposed LLSA model in the retrieved order

Alzheimer disease	0.0903 0.0683 0.0667 0.0565 0.0565 0.0563 0.0541 0.049 0.0297 0.0291 0.0229 0.0221 0.0186 0.0156 0.0143 0.0135 0.0122 0.011 0.009 0.0084 0.0071
Neurodegenerative disease	0.1368 0.1326 0.1295 0.1141 0.1135 0.1112 0.1076 0.1027 0.1019 0.0983 0.098 0.0977 0.0953 0.0915 0.0875 0.0863 0.0828 0.0799 0.0681 0.0677 0.0669 0.0657 0.0652 0.0649 0.0633 0.0613 0.0588 0.0588 0.0531 0.0529 0.0496 0.048 0.0473 0.0471 0.047 0.0464 0.046 0.0414 0.0398 0.0278 0.0257 0.0228 0.0141 0.0078 0.0045 0.0041 0.0033 0.0026 0.001
Dementia	0.1853 0.1607 0.1379 0.1355 0.1289 0.1185 0.1185 0.1124 0.1114 0.1102 0.1092 0.103 0.0898 0.0786 0.0782 0.0764 0.0713 0.0684 0.0679 0.0665 0.0651 0.0563 0.0552 0.0542 0.054 0.0538 0.0518 0.0517 0.0503 0.0484 0.0472 0.0419 0.0375 0.0292 0.0287 0.0266 0.025 0.0244 0.0231 0.023 0.0192 0.0176 0.011 0.0088 0.0069
Drug Toxicity	0.1726 0.1684 0.1651 0.1622 0.1587 0.1539 0.1341 0.1306 0.1305 0.1264 0.1208 0.1154 0.1153 0.1149 0.114 0.1104 0.1101 0.1079 0.105 0.1027 0.1008 0.0997 0.0988 0.0981 0.0962 0.0937 0.0906 0.0904 0.0893 0.084 0.0833 0.0797 0.0754 0.0753 0.075 0.0715

	0.0687 0.0685 0.0644 0.0629 0.0554 0.0533 0.0429 0.0391 0.0386 0.0359 0.0351 0.0321 0.0275 0.0226 0.0223 0.0175
Transplantation	0.3722 0.3688 0.3491 0.3378 0.332 0.3257 0.3199 0.3182 0.3173 0.3128 0.304 0.3034 0.3018 0.2997 0.2976 0.297 0.2962 0.2793 0.265 0.2609 0.2576 0.2551 0.255 0.2512 0.2395 0.2384 0.2379 0.2335 0.2312 0.2291 0.2257 0.2252 0.2252 0.2215 0.2163 0.2148 0.2137 0.2127 0.2121 0.2118 0.2101 0.2028 0.2009 0.1997 0.1986 0.1833 0.1799 0.1608 0.1592 0.1544 0.1501 0.1478 0.1409
Depression	0.1853 0.1607 0.1379 0.1355 0.1289 0.1185 0.1185 0.1124 0.1114 0.1102 0.1092 0.103 0.0898 0.0786 0.0782 0.0764 0.0713 0.0684 0.0679 0.0665 0.0651 0.0563 0.0552 0.0542 0.054 0.0538 0.0518 0.0517 0.0503 0.0484 0.0472 0.0419 0.0375 0.0292 0.0287 0.0266 0.025 0.0244 0.0231 0.023 0.0192 0.0176 0.011 0.0088 0.0069
Venous thrombosis	0.1156 0.1023 0.0766 0.0746 0.0491 0.0479 0.0435 0.0386 0.0357 0.0318 0.0293 0.0289 0.0261 0.025 0.0235 0.0216 0.0149 0.0146 0.0101 0.0089 0.0088 0.0028
Thrombosis embolism	0.3225 0.3224 0.3006 0.2943 0.2918 0.291 0.2879 0.2847 0.2826 0.2815 0.2774 0.273 0.2723 0.2723 0.2679 0.2673 0.2655 0.2648 0.2571 0.2562 0.256 0.2559 0.2549 0.2544 0.2537 0.2535 0.2522

	0.2517 0.2498 0.2494 0.2488 0.2487 0.2477 0.2469 0.2461 0.2452
	0.2451 0.2444 0.2444 0.2439 0.2385 0.2371 0.2349 0.2332 0.232
	0.2317 0.2282 0.2255 0.2246 0.2153 0.2145 0.2044 0.1876
Breast neoplasm	0.2294 0.2262 0.2236 0.2138 0.2119 0.2067 0.2007 0.1967 0.1942
	0.1923 0.1923 0.1875 0.1866 0.1826 0.1809 0.1809 0.1807 0.1804
	0.1781 0.1778 0.1768 0.176 0.1753 0.1723 0.1719 0.1709 0.17
	0.168 0.1665 0.1654 0.1635 0.1633 0.1631 0.1613 0.1598 0.1585
	0.1575 0.1571 0.1565 0.1557 0.1537 0.1531 0.1506 0.1496 0.1481
	0.148 0.1476 0.1454 0.1447 0.1429 0.133 0.1301 0.1108
Hypertension	0.4949 0.4747 0.467 0.4669 0.4639 0.4591 0.4589 0.4546 0.4517
	0.4489 0.448 0.446 0.4423 0.442 0.4391 0.4383 0.4366 0.4346
	0.432 0.4272 0.4264 0.4237 0.4226 0.4225 0.4187 0.4163 0.4143
	0.4122 0.4117 0.4103 0.4083 0.4039 0.4028 0.3999 0.3998 0.3986
	0.3983 0.3975 0.3944 0.394 0.3908 0.3823 0.3801 0.3785 0.3721
	0.3716 0.3704 0.3646 0.3626 0.3595 0.3395 0.3311 0.3194
Nausea	0.4106 0.3823 0.3481 0.3451 0.3377 0.3354 0.3312 0.3259 0.3255
	0.3215 0.3184 0.3172 0.3148 0.3133 0.3129 0.3121 0.3066 0.3058
	0.3056 0.3054 0.3046 0.3024 0.3023 0.3023 0.3022 0.3018 0.3011
	0.2994 0.2985 0.2974 0.2963 0.2958 0.2946 0.293 0.2923 0.2909
	0.2896 0.2892 0.2883 0.2883 0.2869 0.2843 0.2833 0.2808 0.2795

	0.2787 0.2755 0.2744 0.2738 0.2694 0.2647 0.2578 0.2523
Coronary artery disease	0.1696 0.1555 0.1487 0.1388 0.1386 0.1256 0.1211 0.1208 0.1168 0.1081 0.1026 0.0919 0.0901 0.0886 0.0861 0.0854 0.0843 0.0813 0.0786 0.0783 0.0776 0.0769 0.0766 0.0749 0.0719 0.0713 0.071 0.07 0.0687 0.0684 0.0649 0.064 0.0628 0.0618 0.0601 0.0503 0.0495 0.0488 0.0484 0.0473 0.0472 0.0454 0.0442 0.0439 0.0438 0.0362 0.0359 0.0318 0.022 0.0212 0.0187 0.0143 0.003
Myocardial infarction	0.3417 0.3055 0.2976 0.2949 0.2937 0.2922 0.2903 0.2878 0.2826 0.2812 0.2772 0.2754 0.2751 0.2691 0.267 0.2618 0.2548 0.2541 0.2529 0.2514 0.2511 0.2502 0.2482 0.2365 0.2333 0.2309 0.2271 0.2264 0.2241 0.2229 0.2211 0.2204 0.2195 0.2176 0.2166 0.2159 0.2155 0.2137 0.2127 0.209 0.2034 0.2016 0.1995 0.1962 0.1933 0.193 0.1919 0.1916 0.1869 0.1786 0.1734 0.1706 0.1695
Pain	0.3418 0.3118 0.3067 0.2973 0.2914 0.2891 0.2873 0.2622 0.2589 0.2569 0.256 0.2555 0.2533 0.2489 0.2424 0.24 0.234 0.233 0.2311 0.2303 0.2241 0.2238 0.2234 0.2209 0.2196 0.219 0.2154 0.2154 0.2127 0.2103 0.2091 0.2077 0.2048 0.1998 0.1975 0.1951 0.1942 0.1882 0.1875 0.1835 0.1834 0.1803 0.1729 0.1615 0.1608 0.1598 0.153 0.151 0.1484 0.1479 0.1394 0.1293 0.0992
Leukemia	0.1227 0.1134 0.1133 0.1125 0.1054 0.1016 0.0939 0.0919 0.0918 0.0896 0.088 0.0853 0.0841 0.0831 0.083 0.0825 0.0819 0.0795

	0.0775 0.0764 0.0735 0.0724 0.0719 0.0718 0.0704 0.0699 0.0689 0.0689 0.0674 0.0661 0.0657 0.0653 0.0645 0.0586 0.0583 0.0573 0.0565 0.0546 0.0545 0.0542 0.0529 0.0506 0.0504 0.0472 0.0431 0.04 0.0399 0.0385 0.0373 0.0339 0.0164 0.0132 0.011
Pulmonary embolism	0.1821 0.1543 0.1493 0.1486 0.1412 0.1348 0.1331 0.1303 0.1284 0.123 0.1226 0.12 0.1193 0.1161 0.116 0.1158 0.1141 0.1098 0.1096 0.1096 0.1072 0.1069 0.1062 0.1051 0.1022 0.0988 0.0967 0.0959 0.095 0.0944 0.0932 0.0918 0.0891 0.0865 0.0843 0.0836 0.0812 0.081 0.0809 0.0784 0.0757 0.0752 0.0717 0.0717 0.0707 0.0706 0.0701 0.0689 0.061 0.0597 0.0596 0.0553 0.0517
Warfarin	0.0812 0.0633 0.0606 0.0538 0.0495 0.0456 0.0421 0.0412 0.0399 0.0383 0.0372 0.0359 0.0345 0.0319 0.0282 0.0252 0.0229 0.0227 0.0221 0.0203 0.0196 0.0193 0.0164 0.0161 0.0152 0.0146 0.0144 0.0141 0.0124 0.0114 0.011 0.0106 0.0098 0.007 0.0048 0.0047 0.0029 0.0028 0.001
Osteosarcoma	0.1078 0.103 0.0944 0.0478 0.0326 0.0325 0.0286 0.0273 0.025 0.0225 0.0208 0.0159 0.0077 0.0024 0.0024 7.0E-4 3.0E-4
Neutropenia	0.0879 0.0819 0.0796 0.0682 0.0679 0.0657 0.0651 0.0616 0.0579

	0.053 0.0529 0.0517 0.0488 0.0484 0.0469 0.0466 0.0458 0.0444 0.0442 0.0412 0.0406 0.0405 0.0402 0.0362 0.0354 0.0342 0.0327 0.0325 0.0312 0.03 0.0289 0.0253 0.0219 0.0214 0.0208 0.0204 0.0202 0.0185 0.0142 0.0093 0.0078 0.0071 0.0058 0.0011 0.001
Schizophrenia	0.2344 0.2157 0.2119 0.2019 0.1959 0.1939 0.1934 0.1914 0.1904 0.1904 0.1888 0.1883 0.1869 0.1853 0.1853 0.185 0.1828 0.1825 0.1788 0.1699 0.1643 0.1565 0.1564 0.1551 0.1518 0.1507 0.1506 0.1452 0.1444 0.1434 0.1433 0.1421 0.1292 0.1292 0.1256 0.1237 0.117 0.1166 0.1102 0.1087 0.1056 0.1041 0.1036 0.1033 0.1025 0.1008 0.095 0.0904 0.0853 0.0797 0.0678 0.0671 0.0519