PubMedMiner: Mining and Visualizing MeSH-based Associations in PubMed

Yucan Zhang, PhD[1,2], Indra Neil Sarkar, PhD, MLIS[1,3,5], Elizabeth S. Chen, PhD[1,4,5]

[1]Department of Computer Science, [2]Plant Biology, [3]Microbiology & Molecular Genetics, [4]Medicine, [5]Center for Clinical & Translational Science, Univ. of Vermont, Burlington, VT

Abstract

The exponential growth of biomedical literature provides the opportunity to develop approaches for facilitating the identification of possible relationships between biomedical concepts. Indexing by Medical Subject Headings (MeSH) represent high-quality summaries of much of this literature that can be used to support hypothesis generation and knowledge discovery tasks using techniques such as association rule mining. Based on a survey of literature mining tools, a tool implemented using Ruby and R – PubMedMiner – was developed in this study for mining and visualizing MeSH-based associations for a set of MEDLINE articles. To demonstrate PubMedMiner's functionality, a case study was conducted that focused on identifying and comparing ~~comorbidities for asthma~~ in children and adults. Relative to the tools surveyed, the initial results suggest that PubMedMiner provides complementary functionality for summarizing and comparing topics as well as identifying potentially new knowledge.

Introduction

The MEDLINE citation database represents a major source of references to biomedical literature, containing over 20 million citations dating back to the late 1940s with 2000-4000 citations added daily[1, 2]. MEDLINE is the primary component of PubMed that provides access to over 23 million citations[1] where each record is associated with a set of metadata that includes title, abstract, and Medical Subject Headings (MeSH) descriptors that are used to index MEDLINE citations and enable searching in PubMed. With the exponential growth of published biomedical research available in resources such as PubMed (especially MEDLINE), it can be challenging for clinicians and researchers to ~~keep current with~~ findings on a topic of interest and to discover connections between seemingly unrelated concepts (e.g., as represented by MeSH descriptors) across multiple disciplines[3-5].

A number of systems have been developed that build upon the search capabilities provided by PubMed[4, 6, 7], including biomedical literature mining tools that incorporate information retrieval, entity recognition, information extraction, text mining, and data integration methods[8, 9]. As part of the literature-based knowledge or relation discovery process, a first step is the identification of biomedical concepts or entities such as diseases, drugs, and genes. While some studies have involved extracting concepts from titles and abstracts in PubMed or the full-text of articles in PubMed Central, others have focused on MeSH descriptors that are arranged in a hierarchical structure thus allowing for searching at different specificity levels (e.g., using a broader descriptor such as "Respiratory Tract Diseases" versus a more specific descriptor such as "Asthma")[7, 10-14]. These latter studies include those that have used MeSH descriptors to identify co-occurring biomedical concepts as well as those involving use of association rule mining techniques to discover putative relationships between biomedical concepts[15, 16]. There are more than 27,000 descriptors in MeSH that are continually revised and updated[17]. Since MeSH descriptors are applied to MEDLINE records by trained subject matter experts with domain knowledge, they can be seen as representing standardized and high-quality summaries of a particular publication[18]. Thus, MeSH descriptors may offer a useful window into the full text for information retrieval, extraction, and other high-level functions, and potentially allow for further automation of the discovery process.

The goal of this study was to build upon prior efforts and develop an open-source literature mining tool ("PubMedMiner") for identifying and visualizing relationships between MeSH descriptors in MEDLINE. The findings from a survey of literature mining tools that incorporate MeSH descriptors were used to guide the development of PubMedMiner to include functionality for PubMed searching, MeSH descriptor extraction, Unified Medical Language System (UMLS) semantic type filtering, basic statistical analysis and visualization, as well as association rule mining and visualization. In order to demonstrate how PubMedMiner could facilitate hypothesis generation and knowledge discovery, a case study is presented for exploring and comparing comorbidities for asthma in children and adults. A discussion of limitations of the current prototype system, challenges in its development, and planned enhancements to the system are then provided.

Background

Various literature and text mining tools have been developed to improve information retrieval and infer relationships between biomedical concepts[6, 12, 19, 20]. In this section, findings from a survey of literature mining tools including MeSH-based functions are summarized and association rule mining is described as a technique for exploring MeSH-based relationships.

Survey of Literature Mining Tools

Numerous studies have successfully revealed interesting scientific discoveries from biomedical literature in MEDLINE using a variety of knowledge discovery tools that incorporate different terminological systems (e.g., MeSH[5, 21], UniProt protein/gene names[22], or the UMLS[23]) and algorithms for discovering or predicting relations[24]. For example, literature-based discovery (LBD) approaches have been used to discover new knowledge and generate hypotheses based on findings in the literature, such as identifying connections between the beneficial effects of fish oil in treating Raynaud's syndrome and the causal effect of magnesium deficiency on migraines, both of which were later validated[25-28]. LBD studies have integrated MeSH descriptors as part of the knowledge discovery process[5, 21]. Other studies have focused on using MeSH descriptors to improve information retrieval and identify relationships between entities[25, 26, 29-31]. For example, relationships such as drug-gene, drug-effects, protein-protein, and gene-gene, have been explored using statistical co-occurrence methods[29, 31-33].

A survey of literature mining tools was conducted to identify and compare those that incorporate MeSH descriptors. Based on searching publications, Google and Google Scholar Web searching results, and specific resources (e.g., the National Network of Libraries of Medicine, National Center for Biotechnology Information, and Arrowsmith project sites), a total of 80 tools was identified[2, 4, 6, 12, 19, 34]. Of these tools, links for about one-third were found to be no longer active. For the remaining tools, those satisfying the following criteria were included: (1) is maintained and available to the public; (2) focuses on literature mining in the biomedical domain and builds upon basic PubMed functions; and, (3) incorporates MeSH descriptors in the implementation of advanced functionalities, such as basic statistical analysis and association rule mining. The 14 eligible tools after applying these criteria were found to use MeSH descriptors in their main functions besides searching and were categorized into three groups according to their most notable features: (1) filtering or clustering search results using MeSH: BibliMed[35] (2011; Private); (2) describing search results with basic statistics for MeSH: LigerCat[36] (2009; Academic), PubAnatomy[37] (2009; Academic), Anne O'Tate[38] (2008; Academic), GoPubMed[39] (2005; Private), MedSum[40] (2005; Academic), and PubMed PubReMiner[41] (2004; Academic); and, (3) exploring associations among MeSH: KNALIJ[42] (2012; Private), PubAtlas[43] (2009; Academic), AliBaba[44] (2006; Academic), BITOLA[45] (2005; Private), PubNet[46] (2005; Academic), XplorMed[47] (2001; Academic), and MEVA[48] (2001; Private).

The first group whose major feature is filtering or clustering search results using MeSH descriptors includes only one system (BibliMed) that provides a more focused search but no other advanced functions for information extraction[19]. Several systems in the other groups also have a clustering function, such as Anne O'Tate and XplorMed. Systems in the second group accept standard PubMed queries for literature search and perform frequency analysis on MeSH descriptors from retrieved records (Table 1). However, only one or two systems support search result export, statistical analysis, or visualization. LigerCat[20] was the only open-source system identified in this survey. In the third group (Table 1), systems either carry out association rule mining directly among MeSH descriptors or find connections between publications through common MeSH descriptors. Systems such as KNALIJ and AliBaba visualize associations as a network between extracted entities, whereas systems such as PubAtlas visualize associations as a two-dimensional table. While XplorMed and MEVA do not allow for direct PubMed searching, users can upload files produced from external PubMed searches that are used for subsequent analysis. Only a few of the surveyed systems allow users to adjust parameters for association rule mining, and some tools cannot visualize discovered associations.

None of the systems in the second group was found to perform statistical analysis of UMLS semantic types associated with MeSH descriptors and only two of the tools from the third group have implemented the feature of filtering MeSH descriptors by semantic type. UMLS semantic types are hierarchical subject categories for categorizing concepts in the UMLS Metathesaurus[49] and have been used to cluster or filter search results. There are 133 semantic types currently and each MeSH descriptor can be associated with one or more semantic type(s)[17, 50]. This allows users to focus on associations between MeSH descriptors of interest.

Table 1. Tools with basic statistical analysis and association rule mining functions using MeSH descriptors.

Tool	Query format	Filtering with semantic types	MeSH as entity	Search result export	MeSH frequency	Semantic type frequency	Statistics visualization	Association rule mining	Rule strength cutoff adjustable	Association visualization	Related article links	Analysis result export
LigerCat	Keywords, Gene/DNA Sequence	×	✓	×	✓	×	✓	×	×	×	✓	×
PubAnatomy	Keywords, Gene ID or Symbol	✓	✓	✓	✓	×	×	×	×	×	✓	✓
Anne O'Tate	Keywords	✓	✓	×	✓	×	×	×	×	×	✓	×
GoPubMed	Keywords	×	✓	✓	✓	×	×	×	×	×	✓	×
MedSum	Keywords	×	✓	×	✓	×	×	×	×	×	×	×
PubReMiner	Keywords, Gene	×	✓	×	✓	×	×	×	×	×	✓	✓
KNALIJ	Keywords	×	✓	×	×	×	×	✓	×	✓	✓	×
PubAtlas	Keywords, predefined terms	×	×	×	×	×	×	✓	×	✓	✓	✓
AliBaba	Keywords	×	✓	×	×	×	×	✓	✓	✓	✓	✓
BITOLA	MeSH, Gene symbols	✓	✓	×	×	×	×	✓	✓	×	✓	×
PubNet	Keywords	×	×	×	×	×	×	✓	×	✓	×	✓
XplorMed	NA	×	×	NA	✓	×	×	✓	✓	×	✓	×
MEVA	NA	✓	✓	NA	✓	×	×	✓	×	×	✓	×

Note: NA, not available.

Association Rule Mining

The concept of association rule mining was proposed by Agrawal, *et al.* in 1993[51] in the context of market basket analysis, but has since been widely adapted to other domains[30, 52-54]. An association rule is defined as an implication of the form: $X \rightarrow Y$, where $X, Y \subset I$, $X \cap Y = \varnothing$, X, Y are sets of items ("itemsets"), I is the set of total items, and $X \rightarrow Y$ means that when X occurs, Y also occurs with a certain probability. Agrawal and Srikant proposed an algorithm for mining association rules based on identifying frequent itemsets, known as the *Apriori* algorithm[55].

The strength of an association rule is typically measured using metrics referred to as support and confidence[51]. The *support* of an itemset X represents the fraction of transactions containing itemset X. The *confidence* of a rule is the probability of finding transactions containing the consequent of a given rule that also contain the antecedent of the given rule. Commonly, in order to obtain association rules with statistical significance and certain strength, minimum support and confidence values are specified[51]. Many other measures of interestingness have been proposed, such as the chi-squared (χ^2) statistic, lift measure, and Gini index[56, 57]. Of note, the χ^2 statistic has been shown as an efficient measure of the strength of a set of association rules based on co-occurrence[30, 58, 59] and found to outperform other measures such as support, confidence, lift, and conviction[60].

A number of open-source tools are available that provide a wide variety of statistical and graphical techniques for generating and visualizing association rules, such as R[61], Tanagra[62], and Weka[63, 64]. R is a widely used free software environment for statistical computing and visualization that can be extended using over 5000 packages available at the Comprehensive R Archive Network (CRAN) package repository[65, 66]. Compared with R, Tanagra lacks advanced graphical features and Weka is weaker in calculating classical statistics[64, 67, 68]. The R package "arules" implements the *Apriori* algorithm for association rule mining and can calculate various interestingness measures for generated rules[69]. The "arulesViz" R package visualizes association rules in different formats such as scatter plot, grouped matrix, and graph-based association network[70].

Materials and Methods

Overview

In this study, the combined functionality provided by the aforementioned literature mining tools was used to guide the development of an open-source literature mining tool ("PubMedMiner") to include: PubMed searching, MeSH descriptor extraction, UMLS semantic type filtering, basic statistical analysis and visualization, association rule mining and visualization, as well as results exporting. With standard searches in PubMed, MeSH descriptors extracted from retrieved literature reflect the contents of the corresponding articles. Filtering by UMLS semantic types is one way to focus on particular MeSH descriptors of interest. Basic statistical analysis such as frequency counts helps to

identify the most common concepts and semantic types to ~~inform~~ mining and visualization of associations among selected MeSH descriptors.

The general workflow of PubMedMiner involves four phases (Figure 1): (1) searching and retrieving MEDLINE records associated with a particular topic from PubMed; (2) filtering MeSH descriptors by UMLS semantic type(s); (3) generating basic statistics for MeSH descriptors and UMLS semantic types; and, (4) mining and visualizing association rules between filtered MeSH descriptors. Results associated with each phase are made available as plain text (ASCII), XML, or PDF files.

The Ruby scripting language was used to develop the interactive, command-line version of PubMedMiner that is configurable at each phase. For the statistical analysis component, the R environment for statistical computing and graphics[71] was integrated into PubMedMiner using the RinRuby[72] Ruby gem. Functionality for basic statistics and association rule mining was enabled through the use of the "arules"[66] and "arulesViz"[66] R packages.

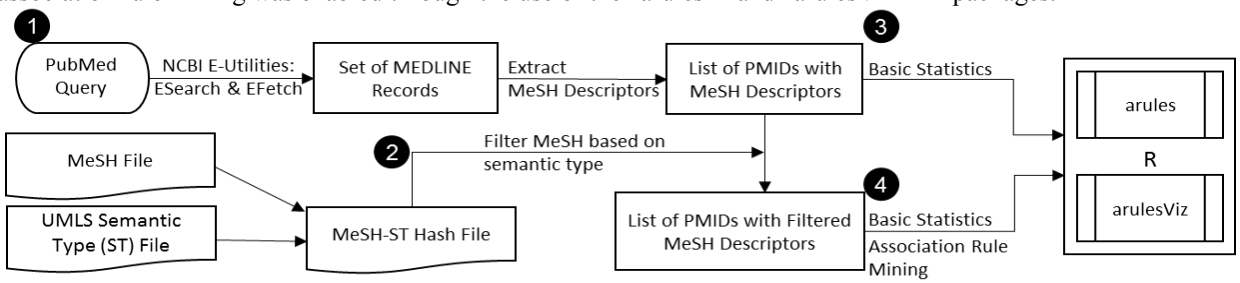

Figure 1. Overview of PubMedMiner development process.

Searching and Retrieving MEDLINE Records

PubMedMiner makes use of the National Center for Biotechnology Information Entrez Programming Utilities (NCBI E-Utilities)[73] to search and retrieve records in MEDLINE format from PubMed. Within PubMedMiner, a user can enter a PubMed query to search and retrieve MEDLINE-formatted records and is notified if no matching articles are found. At the end of a search, PubMedMiner parses each retrieved MEDLINE record and extracts the PMID and MeSH descriptors based on the corresponding PMID and MH metadata fields. Only the PMID and all associated MeSH descriptors along with subheadings are saved into a new text file for the next phase.

Filtering MeSH descriptors by UMLS Semantic Types

UMLS semantic types are used by PubMedMiner to filter MeSH descriptors for subsequent association rule mining. This step removes irrelevant associations and only focuses on association rules between MeSH descriptors corresponding to the set of semantic types specified by the user. Each MeSH descriptor and associated UMLS semantic types are extracted and stored as a hash table in a YAML (YAML Ain't Markup Language) text file, ~~which links the semantic type codes to their full names according to a downloadable list of UMLS Semantic Groups that specifies the semantic types included in each group[74].~~ For example, the MeSH descriptor "Asthma" is associated with the UMLS semantic type code "T047" for "Disease or Syndrome."

Generating Basic Statistics for MeSH Descriptors and UMLS Semantic Types

The PMID-MeSH and PMID-semantic type data are transformed into the requisite formats that enable them to be analyzed in R using the "arules" package (version 1.0-15)[66]. Both absolute counts and relative frequencies of each MeSH descriptor and UMLS semantic type are saved in a text file in a tabular format. The visualization of frequencies as histograms is also implemented using "arules." The user can specify the number of MeSH descriptors or UMLS semantic types to be viewed in the histogram graphs (e.g., top 10 or top 25). Since each MeSH descriptor may correspond to multiple UMLS semantic types, duplicates of the same semantic type can exist for a single MEDLINE record. To accommodate for this, statistics are obtained for semantic types both with and without duplicates. PubMedMiner can perform statistical analysis and visualization on MeSH descriptors and UMLS semantic types both before and after filtering.

Mining and Visualizing Association Rules

In PubMedMiner, the "arules" R package that implements the *Apriori* algorithm is used for generating association rules among selected MeSH descriptors[66, 69] and the "arulesViz" R package is used for visualizing these rules (version 0.1-7)[66, 70]. Before the rule mining step, the selected MeSH descriptors are transformed into the appropriate format for R. The user needs to specify minimum support and confidence values at the beginning of the mining process.

Association rules are generated with 17 different interestingness measures that include support, confidence, χ^2, and Gini index, which are made available in a pipe-delimited text file. In addition, links to PubMed are generated that includes the consequent and antecedent of each rule as search terms in the query: "<rule consequent>"[mh] AND "<rule antecedent>"[mh]. Rules can be visualized and saved as PDF files in five different formats: (1) scatter plot; (2) matrix-based visualization; (3) group matrix-based visualization; (4) graph-based visualization with itemsets as vertices; and, (5) graph-based visualization with items and rules as vertices. For the graph-based visualizations, users can specify the number of rules to include (e.g., all or top 20). To view larger amounts of rules (maximum 1000 rules), a GraphML formatted file is generated that can be opened and modified by visualization tools such as Gephi and Cytoscape[75, 76]. Users can choose whether or not to re-run the rule mining function with adjusted parameters (i.e., support and confidence) if previously defined values did not give meaningful results.

Results

PubMedMiner

PubMedMiner is available under a GPL2 license at: https://github.com/UVM-BIRD/pubmedminer. The PubMedMiner tool provides several modes that cover functions for searching, filtering, statistics before and after filtering, and association rule mining. Each mode represents a combination of different functions and the user can specify which set of functions to be executed by the system. There are three main modes: (1) search, filter, and rule mining (includes statistics); (2) filter and rule mining (includes statistics); and, (3) rule mining only. The first main mode covers the whole process of the pipeline while the other modes only execute part of the process. Four additional modes provide individual functionality: (1) search only; (2) filter only; (3) statistics only (before filtering); and, (4) statistics only (after filtering). A PubMedMiner log file retains a record of the timestamp, selected mode, and configuration details such as the user-specified search query, UMLS semantic types, and other parameters for the basic statistical analysis and association rule mining.

Case Study: Exploring and Comparing Comorbidities for Asthma in Children and Adults

To demonstrate the use of PubMedMiner, a case study is presented here for how the tool can be used to facilitate exploration and comparison of comorbidities for asthma in either a pediatric or adult population. The respective queries (performed on February 13th, 2014) for these patient populations were:

- Pediatric Asthma: "asthma"[mh] AND ("infant"[mh] OR "child"[mh] OR "adolescent"[mh]) NOT "adult"[mh]
- Adult Asthma: "asthma"[mh] AND "adult"[mh] NOT ("infant"[mh] OR "child"[mh] OR "adolescent"[mh])

The first main mode, as described above, was used by PubMedMiner for both queries where the tool performed the PubMed search, filtered and analyzed extracted MeSH descriptors, and generated association rules. The pediatric asthma search returned 23,448 results containing 6,524 unique MeSH descriptors with 123 semantic types; whereas, the adult asthma search returned 22,839 articles containing 8,688 unique MeSH descriptors with 127 semantic types. Absolute counts and relative frequencies for all MeSH descriptors and UMLS semantic types were calculated, and the top 25 MeSH descriptors and top ten semantic types ordered by frequency were chosen to be displayed as histograms in separate PDF files. MeSH descriptors were subsequently filtered by the following three UMLS semantic types in the UMLS Semantic Network: (1) "Disease or Syndrome," (2) "Mental or Behavioral Dysfunction," and (3) "Neoplastic Process," where the latter two are children of the first[50]. After the filtering step, 669 MeSH descriptors were left for pediatric asthma and 948 MeSH descriptors remained for adult asthma. The tool carried out basic statistical analysis again on these filtered MeSH descriptors and further generated association rules using user-specified parameters. To maximize the rules generated by PubMedMiner, a minimum support of 0.01, minimum confidence of 0.01, and maximum rule length of 3 were specified as the settings for both example runs in this study.

According to the statistical analysis of the extracted UMLS semantic types, most semantic types in both sets of articles had similar frequencies (< 2 fold difference). For example, both sets of articles included the same top five UMLS semantic types ordered by absolute count without duplicates: "Age Group," "Population Group," "Human," "Disease or Syndrome," and "Organism Attribute." Several semantic types had frequencies with greater than a two-fold difference. For example, the frequency of the semantic types "Family Group," "Environmental Effect of Humans," "Conceptual Entity," "Organization," and "Regulation or Law" were higher in publications related to pediatric asthma; whereas "Anatomical Abnormality," "Organophosphorus Compound," "Body Substance," "Cell Function," and "Neoplastic Process" appeared more often in articles for adult asthma.

Similarly, the frequency patterns of MeSH descriptors after filtering differed between the two sets of articles. Table 2 presents a selection of MeSH descriptors that had relative frequencies greater than 2%. The fold differences of relative

frequencies were calculated for MeSH descriptors that were higher than 0.5% between the two sets, and fold differences greater than two are highlighted in Table 3, which shows MeSH descriptors with fold difference greater than five. Several MeSH descriptors were more frequently mentioned for pediatric asthma such as "Attention Deficit Disorder with Hyperactivity," "Virus Diseases," and "Respiratory Syncytial Virus Infections," and "Bronchiolitis, Viral" was only found in publications related to pediatric asthma. Some descriptors were more frequent for adult asthma such as "Asthma, Occupational," "Pulmonary Disease, Chronic Obstructive," "Hypertension," "Lung Diseases, Obstructive," and "Churg-Strauss Syndrome" (Table 3). This type of analysis enables one to explore potential asthma-related diseases and reveals those that are more studied relative to the two patient populations (children and adults).

Table 2. MeSH descriptors after filtering by UMLS semantic types.

Pediatric Asthma			Adult Asthma		
MeSH Descriptor	**Absolute Count**	**Relative Frequency**	**MeSH Descriptor**	**Absolute Count**	**Relative Frequency**
Asthma	9761	97.61%	Asthma	9692	97.20%
Dermatitis, Atopic	390	3.90%	Pulmonary Disease, Chronic Obstructive	738	7.40%
Rhinitis, Allergic, Perennial	389	3.89%	Occupational Diseases	667	6.69%
Acute Disease	357	3.57%	Bronchial Hyperreactivity	521	5.23%
Rhinitis, Allergic, Seasonal	329	3.29%	Chronic Disease	408	4.09%
Bronchial Hyperreactivity	311	3.11%	Rhinitis	310	3.11%
Respiratory Tract Infections	299	2.99%	Acute Disease	300	3.01%
Chronic Disease	292	2.92%	Rhinitis, Allergic, Perennial	270	2.71%
Rhinitis	275	2.75%	Rhinitis, Allergic, Seasonal	246	2.47%
Eczema	259	2.59%	Eosinophilia	214	2.15%

Table 3. Comparison of MeSH descriptor frequencies for pediatric asthma and adult asthma.

Pediatric Asthma		Adult Asthma	
MeSH Descriptor	**Fold Difference**	**MeSH Descriptor**	**Fold Difference**
Attention Deficit Disorder with Hyperactivity	53.00	Asthma, Occupational	83.00
Virus Diseases	18.80	Occupational Diseases	47.64
Respiratory Syncytial Virus Infections	14.33	Pulmonary Disease, Chronic Obstructive	29.52
Bronchiolitis	8.00	Hypertension	13.57
Eczema	7.40	Lung Diseases, Obstructive	11.67
Dermatitis, Atopic	6.61	Churg-Strauss Syndrome	10.50
Respiratory Tract Infections	5.34	Lung Neoplasms	8.43
Bronchiolitis, Viral	*	Drug Hypersensitivity	7.05
		Aspergillosis, Allergic Bronchopulmonary	5.20

* MeSH descriptor "Bronchiolitis, Viral" is only associated with pediatric asthma.

With a minimum support value of 0.01 and minimum confidence value of 0.01, 35 and 34 association rules were generated for articles related to asthma in children and adults respectively. The resulting text file included the rules along with the different interestingness measures sorted by χ^2 value in descending order and links to PubMed (e.g., http://www.ncbi.nlm.nih.gov/pubmed/?term="Dermatitis, Atopic"[mh] and "Asthma"[mh] for the rule "Dermatitis, Atopic => Asthma"). The five types of graphs were generated along with the GraphML file for visualizing the rules. Figure 2 shows the grouped matrix-based visualization of all the association rules where rules with the top *k* (user-specified) consequent (Right-Hand Side [RHS]) MeSH descriptors sharing common antecedents (Left-Hand Side [LHS]) are grouped together. If the LHS contains more than one MeSH descriptor, only the most frequent MeSH descriptor is shown in the column label and the number of other descriptors is shown as "+*N*" where *N* represents the number. Size and color represent support and χ^2 value respectively. The columns and rows are ordered according to the χ^2 value that is decreasing from top down and from left to right so that the group of most interesting rules according to χ^2 value are shown in the top-left corner of the plot. Numbers below LHS represent the number of rules containing the corresponding LHS MeSH descriptors.

In the graph-based visualization in Figure 3, MeSH descriptors and rules are represented as vertices, where the size of the rule vertex represents the support value of the rule and color denotes the χ^2 value (e.g., darker shades corresponds to higher values). Bi-directional association rules are identified between pairs of associated MeSH descriptors; the graph contains two rules for each pair with the same strengths in opposite directions. This type of graph emphasizes the composition of the rules and shows which MeSH descriptors share the same rule. Examination of disease-disease

associations from articles for pediatric asthma and adult asthma further highlights several diseases as possible common comorbidities of asthma in both populations such as "Bronchitis," "Bronchial Hyperreactivity," and "Rhinitis," whereas several diseases may be more common in one population than the other, such as "Eczema," "Dermatitis, Atopic," and "Food Hypersensitivity" in children, and "Airway Obstruction," "Churg-Strauss Syndrome," and "Eosinophilia" in adults. In addition to these pairwise relationships, the visualizations for pediatric asthma reveal relationships between "Rhinitis, Allergic, Perennial" and "Dermatitis, Atopic" as well as triple associations such as between "Asthma," "Rhinitis, Allergic, Seasonal," and "Rhinitis, Allergic, Perennial".

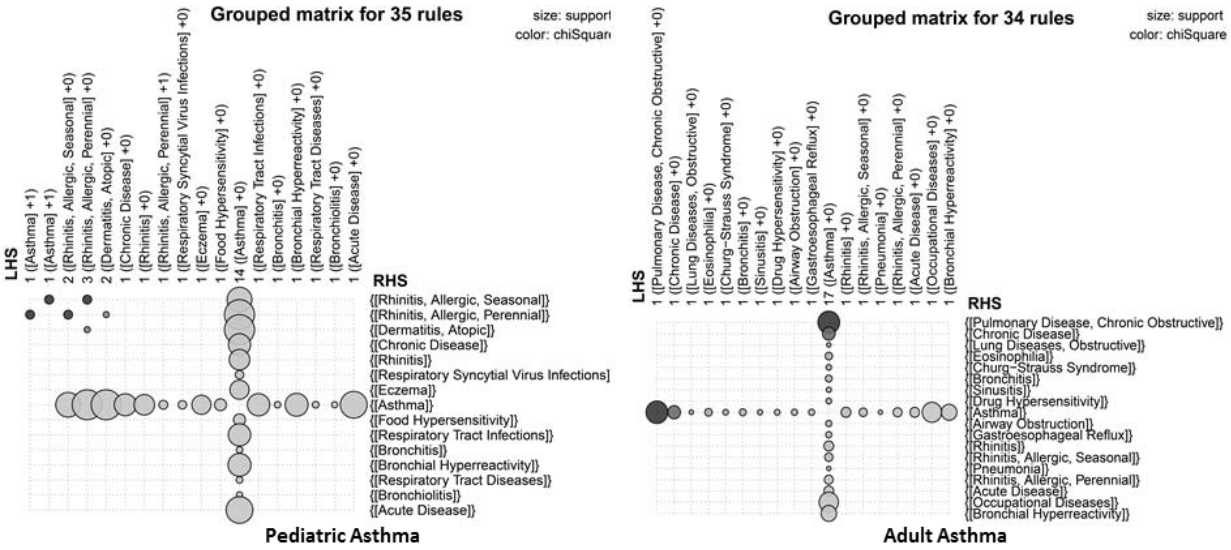

Figure 2. Grouped matrix-based visualization of all association rules for pediatric and adult asthma.

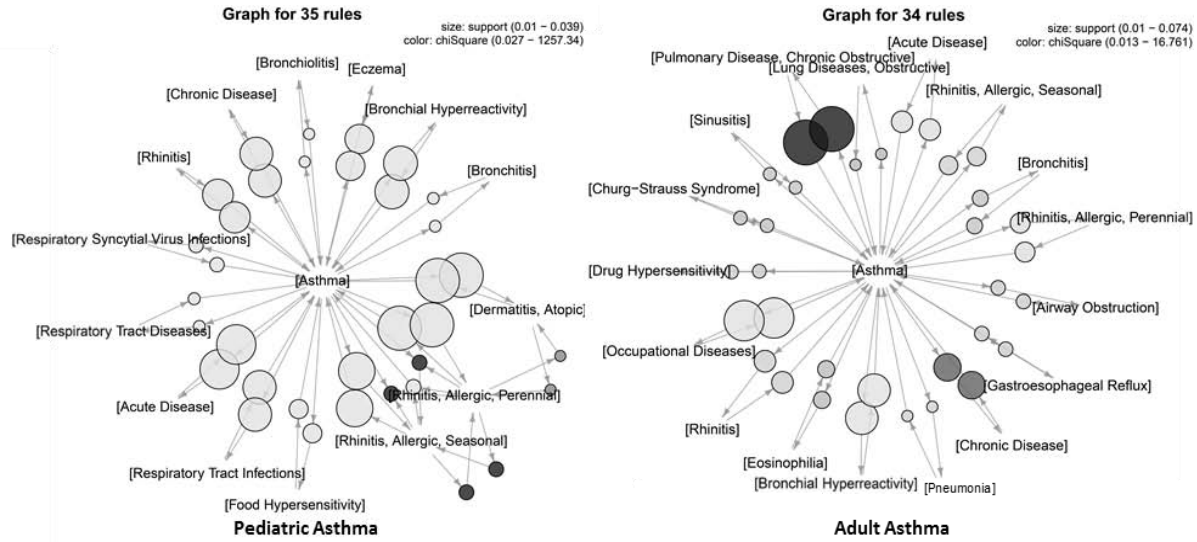

Figure 3. Graph-based visualization with items (i.e., MeSH descriptors) and rules as vertices of all association rules.

Discussion

PubMedMiner was developed as a prototype literature mining tool for MeSH descriptor extraction, basic statistical analysis, and association rule mining of biomedical literature for a given topic. There are several currently available literature mining tools that are capable of advanced literature retrieval and literature-based discovery. Based on the survey and comparison of existing tools including functionality for MeSH descriptors, a command line version of PubMedMiner was developed, which incorporates PubMed searching, MeSH descriptor extraction and filtering, basic statistical analysis, association rule mining, and visualization.

PubMedMiner can be used to explore MeSH descriptors and UMLS semantic types of interest for enhancing focused literature search as well as to validate existing and potentially discover new clinical knowledge. This latter functionality could be valuable for assisting clinicians and researchers keep up-to-date with changes in disease knowledge (e.g., disease-disease and disease-drug) over time. The case study demonstrated a more advanced use of this tool in which it helped highlight the similarities and differences between two PubMed searches for asthma in children and asthma in adults. This example shows how PubMedMiner may be used to facilitate a comorbidity analysis of asthma in different populations. While these searches excluded articles that include MeSH descriptors for both children and adults, further analysis could involve comparing comorbidities based on asthma articles for both children and adults (e.g., "asthma"[mh] AND "adult"[mh] AND ("infant"[mh] OR "child"[mh] OR "adolescent"[mh])). In addition, future work may include formally evaluating the results from such analyses with domain experts (e.g., clinicians and biomedical researchers) to confirm if the associations represent known or unknown knowledge, exploring how this tool may be beneficial for providing highly summarized domain knowledge for researchers and clinicians, guiding studies on particular conditions such as asthma in different populations, and validating results from clinical data mining studies (e.g., comorbidities based on data from the electronic health record).

For a given topic, there could be thousands of MeSH descriptors associated with a set of retrieved MEDLINE citations. PubMedMiner can carry out association rule mining among all MeSH descriptors or only filtered ones according to user-specified UMLS semantic types. Continued development of PubMedMiner will include advanced filtering functions such as incorporating hierarchical information (e.g., enabling the selection of MeSH descriptors for sub-categories of a specified UMLS semantic type) and considering other groupings of MeSH descriptors (e.g., leveraging the MeSH hierarchies or UMLS Semantic Groups). Additionally, the filtering functionality of PubMedMiner may be further enhanced by allowing users to filter out particular MeSH descriptor(s) before generating the statistics and rules in order to highlight relationships between other descriptors (e.g., excluding the original search descriptor "Asthma" and its descendants to highlight other co-morbidities). Other statistical approaches like TF-IDF could also be used to filter out frequently occurring MeSH descriptors (e.g., check tags such as "Humans," "Adult," and "Child").

Limitations include exclusion of newer citations due to the delay in assigning MeSH descriptors[77] and potential bias of selected MeSH descriptors due to consistency issues between different indexers[78, 79]. Besides MeSH descriptors, there are many other metadata fields in MEDLINE records, such as authors, publication types, titles, and abstracts. A number of tools have incorporated such metadata into their functionalities such as PubReMiner that calculates basic statistics on year, author, journal, and substances[19], and KNALIJ that generates and visualizes associations between authors, journals, and latest citations[42]. Future extensions to PubMedMiner may include combining basic statistics with association rule mining using other metadata for a given topic, and enhancing the current association rule mining functionalities by exploring biomedical concepts in titles and abstracts.

Using open-source technologies reduced the costs in terms of labor and time for implementing PubMedMiner. Association rule mining and visualization have been implemented in the "arules" and "arulesViz" R packages. In order to exploit the functions of these packages, R was integrated into Ruby (using the RinRuby Ruby gem) in the implementation of PubMedMiner and several limitations were encountered. Since the generated graphs are not self-descriptive, one motivation for creating the PubMedMiner log file was to record details about each use of the tool, including the user-specified parameters for creating all the graphs. Graphs have to be manually labeled when one needs to compare the same type of graphs generated with different queries. In addition, incomplete labels might be shown on graphs due to the size limitation when visualizing relatively large amounts of information. Twenty or fewer rules give the best resolution of node labels in the network graphs saved in PDF format and network graphs saved in the GraphML file format can be used to visualize up to 1000 rules. Future improvement of this tool may include enhancing interactive features by allowing users to set the font of node labels as well as the title of each graph. Furthermore, current rule metrics include 17 different interestingness measures that can assist users in selecting the most significant association rules. However, visualization of generated rules is implemented with only support, confidence, and χ^2 values. In future development, the tool could allow users to specify what set of measure(s) to use in visualization and incorporate a preview function based on user needs.

For the asthma case study, the frequency of MeSH descriptors was compared manually by calculating the fold difference of the same MeSH descriptors in two different queries. Allowing users to compare results from different queries is an additional functionality that may be included as a future enhancement in PubMedMiner. Although the command line version of PubMedMiner may satisfy many of the needs of studies such as described here, a graphical user interface (GUI) would likely be more desirable for the interactive process. The prototype system developed in this study has provided the foundation for implementation of a web-based GUI version of PubMedMiner (e.g., using MySQL and Ruby on Rails) that is publicly accessible and includes enhanced functionality.

Conclusion

MeSH descriptors can be seen as high quality summaries of biomedical literature and have been used as a window to literature indexed in MEDLINE. Guided by a survey of currently available MeSH-based literature mining tools, a command line prototype system (PubMedMiner) was developed that includes functionality for PubMed searching, MeSH descriptor extraction, UMLS semantic type filtering, basic descriptive statistics, association rule mining, and visualization. Based on a case study of comparing pediatric and adult asthma literature, PubMedMiner was able to identify association rules that may represent characteristic comorbidities.

Acknowledgements

The work was supported in part by the National Library of Medicine of the National Institutes of Health under award number R01LM011364. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Reference

1. U.S. National Library of Medicine Fact Sheets. Available from: http://www.nlm.nih.gov/pubs/factsheets/
2. Arrowsmith. Available from: http://arrowsmith.psych.uic.edu/arrowsmith_uic/index.html
3. Srinivasan P. Text mining: generating hypotheses from MEDLINE. Journal of the American Society for Information Science and Technology. 2004;**55**(5):396-413.
4. Jensen LJ, Saric J, Bork P. Literature mining for the biologist: from information retrieval to biological discovery. Nature reviews genetics. 2006;**7**(2):119-29.
5. Agarwal P, Searls DB. Literature mining in support of drug discovery. Briefings in bioinformatics. 2008;**9**(6):479-92.
6. Weeber M, Kors JA, Mons B. Online tools to support literature-based discovery in the life sciences. Briefings in bioinformatics. 2005;**6**(3):277-86.
7. Cohen AM, Hersh WR. A survey of current work in biomedical text mining. Briefings in bioinformatics. 2005;**6**(1):57-71.
8. De Bruijn B, Martin J. Getting to the (c) ore of knowledge: mining biomedical literature. International journal of medical informatics. 2002;**67**(1):7-18.
9. Hearst M. What is text mining. Retrieved February. 2003;**7**:2011.
10. Yu H, Agichtein E. Extracting synonymous gene and protein terms from biological literature. Bioinformatics (Oxford, England). 2003;**19**(suppl 1):i340-i9.
11. Cohen AM, Hersh WR, Dubay C, Spackman K. Using co-occurrence network structure to extract synonymous gene and protein names from MEDLINE abstracts. BMC bioinformatics. 2005;**6**(1):103.
12. Lu Z. PubMed and beyond: a survey of web tools for searching biomedical literature. Database: the journal of biological databases and curation. 2011.
13. Cohen KB, Johnson H, Verspoor K, Roeder C, Hunter L. The structural and content aspects of abstracts versus bodies of full text journal articles are different. BMC bioinformatics. 2010;**11**(1):492.
14. Lin J. Is searching full text more effective than searching abstracts? BMC bioinformatics. 2009;**10**(1):46.
15. Srinivasan P, Hristovski D. Distilling conceptual connections from MeSH co-occurrences. Medinfo. 2004;**11**(Pt 2):808-12.
16. Hristovski D, Stare J, Peterlin B, Dzeroski S. Supporting discovery in medicine by association rule mining in Medline and UMLS. Studies in health technology and informatics. 2001(2):1344-8.
17. Medical Subject Headings. Available from: http://www.nlm.nih.gov/mesh/
18. Bhattacharya S, Ha V, Srinivasan P. MeSH: a window into full text for document summarization. Bioinformatics (Oxford, England). 2011;**27**(13):i120-i8.
19. NCBI literature search tool archive. Available from: http://www.ncbi.nlm.nih.gov/CBBresearch/Lu/search/
20. NCBI text mining tools. Available from: http://www.ncbi.nlm.nih.gov/CBBresearch/Lu/Demo/tmTools/
21. Srinivasan P, Libbus B. Mining MEDLINE for implicit links between dietary substances and diseases. Bioinformatics (Oxford, England). 2004;**20**(suppl 1):i290-i6.
22. Rebholz-Schuhmann D, Kirsch H, Arregui M, Gaudan S, Riethoven M, Stoehr P. EBIMed—text crunching to gather facts for proteins from Medline. Bioinformatics (Oxford, England). 2007 January 15, 2007;**23**(2):e237-e44.
23. Kilicoglu H, Fiszman M, Rodriguez A, Shin D, Ripple A, Rindflesch TC. Semantic MEDLINE: a web application for managing the results of PubMed Searches. Proceedings of the third international symposium for semantic mining in biomedicine; 2008; 2008. p. 69-76.
24. Chen H, Sharp B. Content-rich biological network constructed by mining PubMed abstracts. BMC bioinformatics. 2004;**5**(1):147.
25. Swanson DR. Fish oil, Raynaud's syndrome, and undiscovered public knowledge. Perspectives in biology and medicine. 1986;**30**(1):7.
26. Swanson DR. Migraine and magnesium: eleven neglected connections. Perspectives in biology and medicine. 1987;**31**(4):526-57.
27. DiGiacomo RA, Kremer JM, Shah DM. Fish-oil dietary supplementation in patients with Raynaud's phenomenon: a double-blind, controlled, prospective study. The American journal of medicine. 1989;**86**(2):158-64.
28. Schiapparelli P, Allais G, Gabellari IC, Rolando S, Terzi MG, Benedetto C. Non-pharmacological approach to migraine prophylaxis: part II. Neurological Sciences. 2010;**31**(1):137-9.
29. Xu R, Wang Q. Toward creation of a cancer drug toxicity knowledge base: automatically extracting cancer drug–side effect relationships from the literature. Journal of the American Medical Informatics Association. 2013.
30. Chen ES, Hripcsak G, Xu H, Markatou M, Friedman C. Automated Acquisition of Disease–Drug Knowledge from Biomedical and Clinical Documents: An Initial Study. Journal of the American Medical Informatics Association. 2008 1//;**15**(1):87-98.
31. Xiang Z, Qin T, Qin Z, He Y. A genome-wide MeSH-based literature mining system predicts implicit gene-to-gene relationships and networks. BMC Systems Biology. 2013;**7**(Suppl 3):S9.
32. Xu R, Wang Q. A knowledge-driven conditional approach to extract pharmacogenomics specific drug–gene relationships from free text. Journal of biomedical informatics. 2012;**45**(5):827-34.
33. Blaschke C, Andrade MA, Ouzounis CA, Valencia A. Automatic extraction of biological information from scientific text: protein-protein interactions. Ismb; 1999; 1999. p. 60-7.
34. Medicine NNoLo. PubMed Online and App Resources. Available from: http://nnlm.gov/training/resources/pubmedalt.html
35. BibliMed. Available from: http://www.biblimed.com/appli/index.php

36. Sarkar IN, Schenk R, Miller H, Norton CN. LigerCat: using "MeSH clouds" from journal, article, or gene citations to facilitate the identification of relevant biomedical literature. AMIA Annual Symposium Proceedings 2009 American Medical Informatics Association Available from: http://ligercat.ubio.org/
37. Xuan W, Dai M, Buckner J, Mirel B, Song J, Athey B, et al. Cross-domain neurobiology data integration and exploration. BMC genomics, 2010 11(Suppl 3): p S6 PubAnatomy Available from: http://brainarray.mbni.med.umich.edu/Brainarray/prototype/PubAnatomy/
38. Smalheiser NR, Zhou W, Torvik VI. Anne O'Tate: A tool to support user-driven summarization, drill-down and browsing of PubMed search results. Journal of biomedical discovery and collaboration, 2008 3(1): p 2 Anne O'Tate Available from: http://arrowsmith.psych.uic.edu/cgi-bin/arrowsmith_uic/AnneOTate.cgi
39. Doms A, Schroeder M. GoPubMed: exploring PubMed with the gene ontology. Nucleic acids research, 2005 33(suppl 2): p W783-W786 Available from: http://gopubmed.com/web/gopubmed/
40. MEDSUM, developed by Galsworthy, M.J., Hosted by the Institute of Biomedical Informatics (IBMI), Faculty of Medicine, University of Ljubljana, Slovenia. Available from: http://webtools.mf.uni-lj.si/public/medsum.html
41. PubMed PubReMiner, developed by Jan Koster, Academisch Medisch Centrum, Universiteitvan Amsterdam. Available from: http://bioinfo.amc.uva.nl/human-genetics/pubreminer/
42. KNALIJ, developed by Steve Melnikoff at iWakari. Available from: http://knalij.com/
43. Parker D, Chu W, Sabb F, Toga A, Bilder R. Literature Mapping with PubAtlas—extending PubMed with a 'BLASTing interface'. Summit on translational bioinformatics, 2009 2009: p 90 PubAtlas Available from: http://www.pubatlas.org/
44. Plake C, Schiemann T, Pankalla M, Hakenberg J, Leser U. AliBaba: PubMed as a graph. Bioinformatics, 2006 22(19): p 2444-2445 AliBaba Available from: http://alibaba.informatik.hu-berlin.de/
45. Hristovski D, Peterlin B, Mitchell JA, Humphrey SM. Using literature-based discovery to identify disease candidate genes. International journal of medical informatics, 2005 74(2): p 289-298 BITOLA Available from: http://ibmi3.mf.uni-lj.si/bitola/
46. Douglas SM, Montelione GT, Gerstein M. PubNet: a flexible system for visualizing literature derived networks. Genome biology, 2005 6(9): p R80 PubNet Available from: http://pubnet.gersteinlab.org/
47. Perez-Iratxeta C, Bork P, Andrade MA. XplorMed: a tool for exploring MEDLINE abstracts. Trends in biochemical sciences, 2001 26(9): p 573-575 XplorMed Available from: http://xplormed.ogic.ca/
48. Tenner H, Thurmayr GR, Thurmayr R. Data mining with Meva in MEDLINE. Medical Data Analysis 2003, Springer p 39-46 Meva Available from: http://www.med-ai.com/meva/index.html
49. MEDLINE Fact Sheet. Available from: http://www.nlm.nih.gov/pubs/factsheets/medline.html
50. UMLS - Current Semantic Types. Available from: http://www.nlm.nih.gov/research/umls/META3_current_semantic_types.html
51. Agrawal R, Imieliński T, Swami A. Mining association rules between sets of items in large databases. ACM SIGMOD Record; 1993: ACM; 1993. p. 207-16.
52. Lin W, Alvarez S, Ruiz C. Efficient Adaptive-Support Association Rule Mining for Recommender Systems. Data Mining and Knowledge Discovery. 2002 2002/01/01;**6**(1):83-105.
53. Antonie M-L, Zaiane OR, Coman A. Application of Data Mining Techniques for Medical Image Classification. MDM/KDD. 2001;**2001**:94-101.
54. Srivastava J, Cooley R, Deshpande M, Tan P-N. Web usage mining: discovery and applications of usage patterns from Web data. SIGKDD Explor Newsl. 2000;**1**(2):12-23.
55. Agrawal R, Srikant R. Fast algorithms for mining association rules. Proc 20th Int Conf Very Large Data Bases, VLDB; 1994; 1994. p. 487-99.
56. Geng L, Hamilton HJ. Interestingness measures for data mining: A survey. ACM Computing Surveys (CSUR). 2006;**38**(3):9.
57. Tan P-N, Kumar V, Srivastava J. Selecting the right interestingness measure for association patterns. Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining; 2002: ACM; 2002. p. 32-41.
58. Diaconis P, Efron B. Testing for independence in a two-way table: new interpretations of the chi-square statistic. The Annals of Statistics. 1985;**13**(3):845-74.
59. Cao H, Hripcsak G, Markatou M. A statistical methodology for analyzing co-occurrence data from a large sample. Journal of biomedical informatics. 2007;**40**(3):343-52.
60. Wright A, Chen ES, Maloney FL. An automated technique for identifying associations between medications, laboratory results and problems. Journal of biomedical informatics. 2010;**43**(6):891-901.
61. Team RC. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2012: ISBN 3-900051-07-0; 2012.
62. Rakotomalala R. TANAGRA: a free software for research and academic purposes. Proceedings of EGC; 2005; 2005. p. 697-702.
63. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software: an update. ACM SIGKDD explorations newsletter. 2009;**11**(1):10-8.
64. Zupan B, Demsar J. Open-source tools for data mining. Clinics in laboratory medicine. 2008;**28**(1):37-54.
65. Zhao Y. R and Data Mining: Examples and Case Studies: Access Online via Elsevier; 2012.
66. CRAN. Available from: http://cran.us.r-project.org/
67. Rakotomalala R. TANAGRA: une plate-forme d'expérimentation pour la fouille de données. Revue MODULAD. 2005;**32**:70-85.
68. Witten IH, Frank E, Trigg LE, Hall MA, Holmes G, Cunningham SJ. Weka: Practical machine learning tools and techniques with Java implementations. 1999.
69. Hahsler M, Grün B, Hornik K, Buchta C. Introduction to arules–A computational environment for mining association rules and frequent item sets. The Comprehensive R Archive Network. 2009.
70. Hahsler M, Chelluboina S. Visualizing Association Rules: Introduction to the R-extension Package arulesViz. R project module. 2011.
71. The R Project for Statistical Computing. Available from: http://www.r-project.org/
72. Dahl DB, Crawford S. Rinruby: accessing the r interpreter from pure ruby. J Stat Softw. 2008;**29**(4):1-18.
73. Entrez Programming Utilities Help [Internet]. Available from: http://www.ncbi.nlm.nih.gov/books/NBK25501/
74. Semantic Groups. Available from: http://semanticnetwork.nlm.nih.gov/SemGroups/
75. Gephi. Available from: https://gephi.org/
76. Cytoscape. Available from: http://www.cytoscape.org/
77. Huang M, Névéol A, Lu Z. Recommending MeSH terms for annotating biomedical articles. Journal of the American Medical Informatics Association. 2011 May 25, 2011.
78. Funk ME, Reid CA. Indexing consistency in MEDLINE. Bulletin of the Medical Library Association. 1983;**71**(2):176.
79. Gay CW, Kayaalp M, Aronson AR. Semi-automatic indexing of full text biomedical articles. AMIA Annual Symposium Proceedings; 2005: American Medical Informatics Association; 2005. p. 271.