Mining Generalized Association Rules on Biomedical Literature

Margherita Berardi [1] Michele Lapi [1] Pietro Leo[2] Corrado Loglisci[1]

[1]Dipartimento di Informatica – Università degli Studi di Bari
via Orabona 4 - 70126 Bari
[2]Java Technology Center - IBM SEMEA Sud
Via Tridente, 42/14 - 70125 Bari

[1]{ berardi, lapi }@di.uniba.it, { radodiadema }@libero.it
[2]{ pietro_leo }@it.ibm.com

Abstract. The discovery of new and potentially meaningful relationships between concepts in the biomedical literature has attracted the attention of a lot of researchers in text mining. The main motivation is found in the increasing availability of the biomedical literature which makes it difficult for researchers in biomedicine to keep up with research progresses without the help of automatic knowledge discovery techniques. More than 14 million abstracts of this literature are contained in the Medline collection and are available online. In this paper we present the application of an association rule mining method to Medline abstracts in order to detect associations between concepts as indication of the existence of a biomedical relation. The discovery process fully exploits the MeSH (Medical Subject Headings) taxonomy, that is, a set of hierarchically related biomedical terms which permits to express associations at different levels of abstraction (generalized association rules). We report experimental results on a collection of abstracts obtained by querying Medline on a specific disease and we show the effectiveness of some filtering and browsing techniques designed to manage the huge amount of generalized associations that may be generated on real data.

1. Introduction

In biomedicine, the decoding of the human genome has increased the number of online publications leading to information overload. Every 11 years, the number of researchers doubles [10] and Medline, the main resource of research literature, has been growing with more than 10,000 abstracts per week since 2002[1]. Therefore, it becomes more and more difficult for researchers in biomedicine to keep up with research progresses. Moreover, the data to be examined (i.e. textual data) are generally unstructured as in the case of Medline abstracts and the available resources (e.g. PubMed, the search engine interfacing Medline) do not still provide adequate

[1] http://www.nlm.nih.gov/pubs/factsheets/medline.html

mechanisms for retrieving the required information. The need to analyze this volume of unstructured data and to provide knowledge to improve retrieval effectiveness makes biomedical text mining a central bioinformatic problem and a great challenge for data mining researchers.

In this paper we present the application of association rule mining to Medline abstracts in order to detect associations between concepts as indication of the existence of a biomedical relation but without trying to find out the kind of relation. The discovery process fully exploits the MeSH (Medical Subject Headings) taxonomy, that is, a set of hierarchically related biomedical terms which permits to mine multi-level association rules (*generalized association rules*). Considering the hierarchical relations reported in the MeSH taxonomy allows the discovery algorithm to find associations at multiple levels of abstraction from one side, but generally leads to a huge amount of generalized associations from the other side. The two-fold aim of the paper is to investigate how taxonomic information can be profitably used in the task of concept relationship discovery and to evaluate the effectiveness of some filtering and browsing techniques designed to manage the huge amount of discovered associations.

The paper is organized as follows. Section 2 illustrates the background on our work and some related works on biomedical text mining. Section 3 presents the problem of mining generalized association rules and some filtering methods. In Section 4, some experimental results on a collection of abstracts obtained by querying Medline on a specific disease are reported. Finally, some conclusions are drawn and some possible directions of future work are also presented.

2. Background and Related Works

In [3], we presented a data mining engine, namely MeSH Terms Associator (MTA), that was employed in a distributed architecture to refine a generic PubMed. The idea is to support users by offering them the possibility of iteratively expanding their query on the basis of discovered correlations between their topic of interest and other terms in the MeSH taxonomy. A natural extension of this initial work is to enable an association discovery process that takes advantage of the MeSH taxonomy defined on biomedical terms. Kahng et al. [6] have already investigated an efficient algorithm for generalized association rule mining using the MeSH taxonomy. In this seminal work, no processing on Medline abstracts is performed but a MeSH-indexed representation (in Medline, to every record a set of relevant MeSH terms is manually associated as representation of the content of the document the record is about) is adopted. Moreover, the evaluation of the interestingness of mined associations with respect to the task of PubMed retrieval capabilities improvement is not an issue considered by the authors. A different perspective is taken by Srinivasan [13] and Aronson et al. [2], who state the importance of query expansion to improve retrieval effectiveness of the PubMed engine. In particular, for the indexing process they both use a MeSH-indexed representation, while for the query expansion process, Srinivasan exploits a statistical thesaurus containing correlations between MeSH terms (MeSHs) and text, and Aronson et al. use the MetaMap system to associate

UMLS (Unified Medical Language System, that is, a semantic classification of the MeSH dictionary) Metathesaurus concepts to the original query.

For what concerns the application of association rule mining to the biomedical literature, an interesting work has been carried out by Hristovski et al. and implemented in the BITOLA system [5]. They tailor their work for the discovery of new relations involving a concept of interest, where the novelty of the relation is evaluated by matching transitive associations. Indeed, they first find all the concepts Y related to the concept of interest X, then all the concepts Z related to Y and finally, they check if X and Z appear together in the biomedical literature. If they do not appear together, the system has discovered a potentially new relation that will be evaluated by the user. The search of associations is constrained to associations involving only two terms (i.e. the concept of interest and a new related concept) and can be limited by the semantic type to which terms belong with respect to the UMLS dictionary. In particular, they exploit an association rule base gathered by the UMLS vocabulary on which the discovery of new associations will be performed. As document representation, a MeSH-indexed representation is used and no knowledge about the MeSH taxonomy is exploited.

The idea of applying the transitivity property on correlations to discover relations between concepts has been widely investigated also from a different perspective. Indeed, in [14] transitive knowledge is exploited not only for the discovery of new relations with an input topic but also for the discovery of connections between two given topics of interest that are bibliographically disjointed (e.g. two topics that have been studied independently and may belong to two different sub-areas of research). In both cases, the intermediate level of correlations is used as a transitivity level between topics in order to both discover "hidden" connections and provide the set of correlating concepts. In this work, correlations are extracted on the basis of co-occurrences computed in profiles of topics, where a profile is built in form of a vector of MeSH term vectors, that is, a vector that for each UMLS semantic type reports MeSHs weights (a measure of the conditional importance of each MeSH term). Srinivasan approach is inspired by the pioneer work of Swanson [15], who first explored potential linkages via intermediate concepts starting from two given topics. Many other works inspired to Swanson's approach mainly differs for the document processing phase. While Swanson restricted the analysis only to titles of Medline records, others consider the MeSH-indexed representation of abstracts or the whole abstracts as free-text. In this case, n-grams may be extracted and evaluated by means of different weighting schemes (e.g. TFIDF) as indexing method [9] or a UMLS-indexed representation may be obtained by applying the natural language processing capabilities of the MetaMap system [17, 11].

All these works aim at capturing connections between distinct sub-areas of biomedical literature in order to gain new knowledge on a single topic of interest or on the relation between two topics of interest. This leads to restricting the discovery to only two-term associations as in [5], which means extraction of knowledge only about co-occurrences, or to restricting the discovery to three-term associations as in the case of Swanson and works inspired by him, which means extraction of knowledge not only about co-occurrences but also about correlating terms. Moreover, in discovered associations, the topic (topics) of interest has (have) to be directly

involved in the associations. On the contrary, we are interested in mining associations involving an unknown number of terms, which should be quite certain with respect to the distribution of associations and which may directly involve the topic of interest or not. Besides, we are not interested in discovering literature connections on an unknown segment of Medline but we intend to use the topic of interest directly as a query to retrieve from Medline the segment of related abstracts and then perform an "unbiased" mining on MeSHs contained in this set of abstracts, aiming at capturing the knowledge they share.

3. The approach

In this section we present the general problem of mining association rules and the extension to the use of taxonomic knowledge on data. Moreover, some filtering techniques are discussed.

3.1 Mining association rules

Association rules are a class of regularities introduced by [1] that can be expressed by an implication:

$$X \rightarrow Y$$

where X and Y are sets of items, such that $X \cap Y = \varnothing$. The meaning of such rules is quite intuitive: Given a database D of transactions, where each transaction $T \in D$ is a set of items, $X \rightarrow Y$ expresses that whenever a transaction T contains X than T probably contains Y also. The conjunction $X \wedge Y$ is called pattern.

Two parameters are usually reported for association rules, namely the support, which estimates the probability $p(X \subseteq T \wedge Y \subseteq T)$, and the confidence, which estimates the probability $p(Y \subseteq T \mid X \subseteq T)$. The goal of association rule mining is to find all the rules with support and confidence exceeding user specified thresholds, henceforth called *minsup* and *minconf* respectively. A pattern $X \wedge Y$ is large (or *frequent*) if its support is greater than or equal to *minsup*. An association rule $X \rightarrow Y$ is *strong* if it has a large support (i.e. $X \wedge Y$ is frequent) and high confidence. Several algorithms have been presented in the literature to discover associations among items composing transactions of a database. Many of them are variations on the *Apriori* algorithm [1], which works in two phases: (1) it finds all frequent item-sets; and (2) it uses these item-sets to generate all rules whose confidence is above the *minconf* value.

Srikant and Agrawal [12] have extended this basic mechanism in order to mine associations at the right level of a taxonomic knowledge defined on items. For this purpose, they have defined generalized association rules as association rules $X \rightarrow Y$ where no item in Y is an ancestor of any item in X in the taxonomy. The basic algorithm to mine generalized association rules first extends each transaction of the database to include each ancestor of the items contained in the transaction, then compute confidence and support for all possible association rules and finally, it prunes all the association rules that are "subsumed" by an "ancestral" rule.

3.2 Filtering association rules

Although discovered association rules are evaluated in terms of support and confidence measures, which ensure that discovered rules have enough positive evidence, the number of discovered association rules is usually high and even considering only those rules with high confidence and support it is not true that all of them are interesting. It may happen that some of them correspond to prior knowledge, refer to uninteresting items or are redundant. On the other hand, the presentation of thousands of rules can discourage users from interpreting them in order to find nuggets of knowledge. Furthermore, it is very difficult to evaluate which rules might be interesting for end users by means of some simple statistics, such as support and confidence. Therefore, an additional processing step is necessary in order to clean, order or filter interesting patterns/rules, especially when the mining is performed at different level of abstraction on items because it intrinsically introduces a degree of complexity in the amount of discovered patterns/rules.

Two different approaches can be applied to structure the set of discovered rules and filter out interesting ones, automatic and semiautomatic methods. The former allows to filter rules without using user knowledge, while the latter allows to strongly guide the exploration of the set of discovered rules on the basis of user domain knowledge. An automatic method which aims at removing redundancy in rules has been already investigated in our previous work, namely association rule covers proposed by [16]. Carrying on the work on the automatic approach, we have then investigated the effectiveness of some measures proposed by [8], which aim at evaluating the interestingness of rules from a statistical point of view different from classical support and confidence measures. In this work, the definition of interestingness of a rule is based on the following statement:

Let Π be a property of a set of association rules, $M\Pi$ the mean value, $\sigma\Pi$ the standard deviation and p a coefficient (it is often assumed to be equal to the maximum value of the statistical surprise property), two different behaviours for a rule are definable: rules behaving in a *normal* way in relation to a Π, that is rules whose value of Π is less than or equal to $M\Pi + (p*\sigma\Pi)$ and rules behaving in an *interesting* way in relation to a Π, that is rules whose value of Π is greater than $M\Pi + (p*\sigma\Pi)$.

In order to use this definition of interestingness of a rule, some statistical properties of rules have been considered. Let $X \rightarrow Y$ be an association rule to be evaluated, it is possible to define:

- the **Classical Dependency** of $X \rightarrow Y$ by estimating $P(Y \mid X) - P(Y)$
- the **Novelty** of $X \rightarrow Y$ by estimating $P(Y , X) - P(X)* P(Y)$
- the **Satisfaction** of $X \rightarrow Y$ by estimating $P(Y \mid X) - P(Y) / (1 - P(Y))$
- the **Surprise** of $X \rightarrow Y$ by estimating $P(X , Y) - P(X, \neg Y) / P(Y)$.

In particular, the first three properties are related to three different definitions of the dependency property and a definition is used rather than another on the basis of probability values of antecedent and consequent with respect to a threshold value. By

using one of these properties, the user can select interesting rules and decide to discard normal ones in dependence on the interest.

In order to augment automatic methods with user knowledge, some semiautomatic approaches have also been investigated. Indeed, in our previous work user-defined templates proposed by [7] are illustrated. An example of the template mechanism according to which the user can select and filter all the rules that satisfy and instantiate a criterion specified in a template is reported. Considering the *inclusive* template "*Analytical Diagnostics and Therapeutic Techniques and Equipment" → Mental Disorders*, only a rule that satisfies it has been selected, that is *"Analytical Diagnostics and Therapeutic Techniques and Equipment" → Mental Disorders*, while some rules instantiating it are

"Analytical Diagnostics and Therapeutic Techniques and Equipment"→Dementia
Therapeutics → Mental Disorders
Therapeutics → Dementia
Therapeutics → Alzheimer Disease.

Nevertheless, templates seem to be a quite dispersive method because it is useful to select all the rules satisfying a certain criterion but in this way, a large number of rules in any case could be proposed to the user. For this reason, we have also provided to the user a browsing functionality which allows to look at the set of discovered rules as a set of subspaces of rules, where for each subspace a representative rule is identifiable.

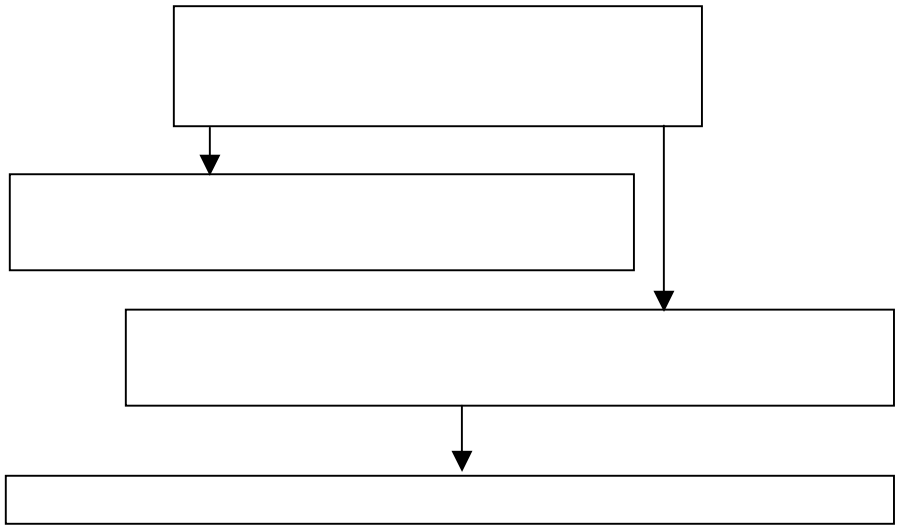

Fig. 1 Exploration of a set of rules by means of subspaces of rules. Rules that are representative of each subspace are reported in boldface. The exploration is based on the enhancement of one of the side of the representative rule.

Then, the user can visit the space of rules following his/her interest and moving towards more and more specific subspaces. An example of the exploration of a rule set by means of subspaces is shown in Fig. 1. In particular, on the basis of the users'

interest (e.g. the set of rules involving the *Tauopathies* MeSH term), he/she can explore subspaces of rules at different level of specialization by selecting which side of the rule should be enhanced.

4. Experimental results

In this section, we intend to compare results on *generalized* and *flat* association rule discovery on datasets generated by means of PubMed queries formulated by experts in the biomedical sector. An example of PubMed query formulated by biomedical researchers may ask for discovering the factors related to the reactions to Diabetes treatments (i.e. "Diabetes Drugs Response").

Submitting the query to PubMed, a set of retrieved abstracts is found out and initially annotated by the BioTeKS Text Analysis Engine (TAE) provided within the IBM UIM Architecture [5], by using a local MeSH terms dictionary. For each query, a single table of a relational database is created and fed with MeSHs occurring in the corresponding set of retrieved abstracts. In particular, each transaction of a single table is associated to an individual abstract and is described in terms of items that correspond to MeSHs. The simplest representation, namely the boolean representation, is adopted in order to represent the occurrence of a MeSH term in an abstract. More precisely, we consider only the most frequent MeSHs (about 50) with respect to the set of retrieved abstracts and we use the "canonical" form of each MeSH term, which is available in the MeSH dictionary. This allows to introduce a light control on redundancy in the data, since many MeSHs may occur referring to the same canonical term. The MeSH taxonomy is organized in 15 distinct hierarchies structured in a tree form that is about 11 levels deep.

In this study, two segments of Medline have been considered, that is the sets of abstracts related to two queries, namely "*Hypertension Adverse Reaction Drugs"* and "*Alzheimer Drug Treatment Response"*. By submitting the former, 130 abstracts have been found, while 653 abstracts for the latter. For each set of abstracts, the contingency table has been created. Depending on the set of MeSHs occurring in a set of abstracts, a different part of the MeSH taxonomy should be considered. Indeed, for the "*Hypertension Adverse Reaction Drugs"* query five hierarchies (*Diseases, Biological Science, "Chemicals and Drugs", "Psychiatry and Psychology", "Analytical Diagnostics and Therapeutic Techniques and Equipment"*) have been used; while for the "*Alzheimer Drug Treatment Response"* query six hierarchies (*Diseases, Biological Science, "Chemicals and Drugs", "Psychiatry and Psychology", "Analytical Diagnostics and Therapeutic Techniques and Equipment", Anatomy*) have been used.

In Fig. 2, the number of discovered associations is drawn varying both *minsup* and *minconf* values. The great difference in the number of generalized association rules compared with the number of flat association rules is a quite obvious observation considering that generalized rules include flat rules since flat association rule discovery corresponds with generalized association rule discovery restricted to leaves of the taxonomy.

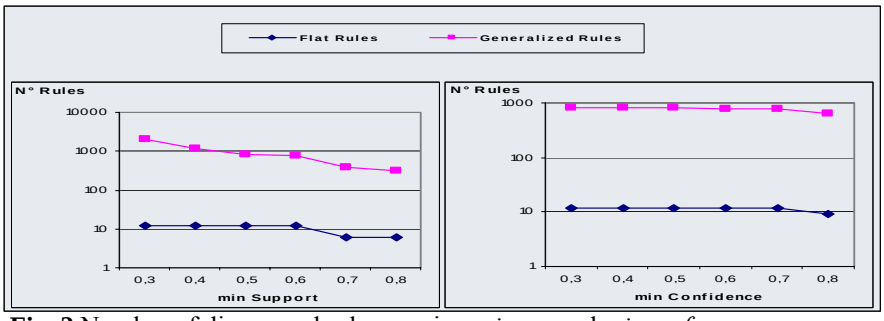

Fig. 2 Number of discovered rules varying *minsup* and *minconf.*

Generally, association rules with low support express only a casual information since it is knowledge not probabilistically justified. Indeed, flat rules generated with low *minsup* often express this kind of knowledge. In contrast, generalized association rules generated with low *minsup* may represent knowledge with a probabilistic evidence as well. An example comes from considering the following two rules that have been both discovered with low (0.4) as well as with high (0.8) value of *minsup* by means of generalized association rule mining.

Tauopathies → *Alzheimer Disease, Delirium, Dementia, Amnesic Cognitive Disorders*
0.807 support, 1 confidence

Neurodegenerative Diseases → *Mental Disorders, Brain Diseases*
0.807 support, 1 confidence

Moreover, by generalized association rule discovery it is possible to check whether a rule is a specific case of a more general one. Thus, though a certain rule has been discovered with low *minsup* value, in any case it is probabilistically justified, because its related generalized rule is probabilistically justified too. For instance, if we consider the flat rule

Therapeutics → *Alzheimer Disease* 0.621 support, 0.813 confidence

and the corresponding one discovered by generalized association discovery

Therapeutics → *Alzheimer Disease* 0.621 support, 0.813 confidence

we can explore the following ancestor rules and verify that the most general one has in any way enough probabilistic evidence.

Therapeutics → *Dementia* 0.663 support, 0.868 confidence
Therapeutics → *Mental Disorders* 0.669 support, 0.876 confidence
"Analytical Diagnostics and Therapeutic Techniques and Equipment" → *Mental Disorders* 0.717 support, 0.925 confidence

Moreover, we discover association rules from datasets which contain only the 50 most frequent MeSHs. Therefore, it is possible that an association rule that has low

support in these datasets may correspond with a pattern that is strongly supported in the datasets containing all the MeSHs.

When we compare results on flat association rules and generalized rules on the same dataset, an interesting observation can be done about some rules that are generally considered "trivial" except if the knowledge about ancestor rules is provided. Indeed, the MeSH taxonomy sometimes presents nodes that are duplicate in different part of the hierarchies. It aims to represent a different perspective of the same term. For instance, it may happen that discovered rules capture associations like X → X, where X is a MeSH that belongs to two different hierarchies in the MeSH structure. In the case of flat rules they should be discarded, while in the case of generalized rules, by exploring their ancestor rules the user may justify this kind of rules.

5. Conclusion and Future Work

In this paper the application of generalized association rule mining to biomedical literature has been presented. Given a biomedical topic of interest as input query to PubMed, the set of related abstracts in Medline is retrieved and a MeSH-based representation of them is produced by means of the annotation capabilities of BioTeKS TAE. Associations are generated on a single set of abstracts with the aim of discovering potentially meaningful knowledge in form of relations among MeSHs. We assume that discovered associations play the role of relevant knowledge shared by the set of abstracts under study and that can be profitably used to expand the query on the topic of interest. Some browsing and filtering techniques have also been used to support the user in the complex task of evaluating the huge amount of discovered associations. Nevertheless, a number of improvements on this work are worth to be explored. In particular, further work on the document processing phase is necessary to evaluate how the document representation model affects the quality of discovered rules. First, we intend to remove the threshold on the number of MeSHs to consider in the contingency table and to employ a better feature selection method. For instance, instead of representing the simple boolean occurrence of a MeSH term, the occurrence frequency can be used as well as it could be interesting to consider some form of "context" in which a term occurs. A solution is to use n-grams rather than single term in combination with a weighting schema to evaluate the relevance of the n-gram with respect to the set of documents. Another solution is to use natural language processing to extract information from the sentence in which the term appears. By using these techniques, we can also aspire to gain information about the kind of relation among co-occurrent MeSHs and investigate the application of multi-relational approaches to association rule mining. Finally, the most important step to improve our work is to evaluate the quality of rules by submitting results to the judgement of a biomedical expert in order to perform the whole query expansion process.

References

1. R. Agrawal, and R. Srikant: “Fast Algorithms for Mining Association Rules”, Proceedings of the Twentieth Int.Conf. on Very Large Databases, Santiago, Cile, 1994.
2. A. R. Aronson and T. C. Rindflesch: “Query expansion using the UMLS Metathesaurus” Proceedings of AMIA, Annual American Medical Informatics Association Conference, Nashville, TN, October 25-29, 1997, pp. 485-489.
3. M. Berardi, M. Lapi, P. Leo, D. Malerba, C. Marinelli, and G. Scioscia: “A data mining approach to PubMed query refinement”, 2nd International Workshop on Biological Data Management in conjunction with DEXA 2004, Zaragoza, Spain, September 2, 2004.
4. D. Ferrucci, and A. Lally: “UIMA: An Architectural Approach to Unstructured Information Processing in the Corporate Research Environment”, Natural Language Engineering, September 2004, 10(3-4), pp. 327-348.
5. D. Hristovski, J. Stare, B. Peterlin, & S. Dzeroski: “Supporting discovery in medicine by association rule mining in Medline and UMLS”, Proceedings of MedInfo Conference, London, England, September 2-5, 2001, 10(2), pp 1344-1348.
6. J. Kahng, W.-H. K. Liao, D. McLeod: “Mining Generalized Term Associations: Count Propagation Algorithm”, KDD, 1997, pp 203-206.
7. M. Klemettinen, H. Mannila, P. Ronkainen, H. Toivonen, and A. I. Verkamo: “Finding Interesting Rules from Large Sets of Discovered Association Rules”. Proc. of the 3rd Int'l Conf. on Conference on Information and Knowledge Management, Gaithersburg, Maryland, November 1994, pp. 401-407.
8. Y. Kodratoff, J. Azé: “Rating the Interest of Rules Induced from Data within Texts”, Proceedings of Twelfth Int. Conf. On Database and export Systems Application, DEXA 2001, Munich, Germany.
9. R. K Lindsay, & M.D. Gordon: “Literature-based discovery by lexical statistics”, Journal of the American Society for Information Science, 50(7), pp 574-587.
10. M. F. Perutz: “Will biomedicine outgrow support?”, Nature 399, 1999, pp 299-301.
11. W. Pratt, M. Yetisgen-Yildiz: “LitLinker: Capturing Connections across the Biomedical Literature”, Proceedings of the International Conference on Knowledge Capture (K-Cap'03), Florida, October 2003.
12. R. Srikant and R. Agrawal: “Mining Generalized Association Rules”, Proc. of the 21st Int'l Conf. on Very Large Databases, Zurich, Switzerland, Sep. 1995.
13. P. Srinivasan: “Query expansion and MEDLINE”, Information Processing and Management, 1996; 32(4), pp 431-443.
14. P. Srinivasan: “Text Mining: Generating Hypotheses from Medline”, Journal of the American Society for Information Science, 2004, 55 (4), pp. 396-413.
15. DR. Swanson: “Fish oil, Raynaud’s syndrome, and undiscovered public knowledge”, Perspectives in Biology and Medicine, 30, pp. 7-18.
16. H. Toivonen, M. Klemetinen, P. Ronkainen, K. Hatonen, and H. Mannila: “Pruning and grouping discovered association rules”, Mlnet Workshop on Statistics, Machine Learning, and Discovery in Databases, 1995, pp. 47-52.
17. M. Weeber, H. Klein, L. T. W. de Jong-van den Berg, R. Vos: “Using concepts in literature-based discovery: simulating Swanson's Raynaud-fish-oil and migraine-magnesium discoveries”, Journal of the American Society for Information Science, 2001, 52(7), pp. 548-557.