Supporting Discovery in Medicine by Association Rule Mining in Medline and UMLS

Dimitar Hristovski[a], Janez Stare[a], Borut Peterlin[b], Saso Dzeroski[c]

[a]IBMI, Medical Faculty, University of Ljubljana, Ljubljana, Slovenia
[b]Department of Human Genetics, Clinical Center Ljubljana, Ljubljana, Slovenia
[c]Institute Jozef Stefan, Ljubljana, Slovenia

Abstract

The paper presents an interactive discovery support system for the field of medicine. The intended users of the system are medical researchers. The goal of the system is: for a given starting concept of interest, discover new, potentially meaningful relations with other concepts that have not been published in the medical literature before. The known relations between the medical concepts come from the Medline bibliographic database and the UMLS. We use association rules for discovering the relationship between medical concepts. We evaluated the system by testing how successfully it predicted future discoveries (new relations between concepts). We first divided the Medline database into two segments (older and newer) using the publication date. Then we calculated how many of the new relations found by the system in the older segment become known relations in the newer segment. We found out with statistical significance that the system predicts new relations better then someone predicting randomly. The evaluation showed that our approach for supporting discovery in medicine is successful, but also that some improvements are needed, especially on limiting the number of potential discoveries the system generates.

Keywords:

Data Mining; Association Rules; Medline; UMLS

Introduction

The two main questions addressed by this paper are: 1. Is it possible to discover new, potentially meaningful relations between medical concepts by searching and analysing the documents from a bibliographic database such as Medline? and 2. To what degree can be the discovery process automated? As an attempt to deal with these issues we developed an interactive discovery support system based on association rule mining of the Medline bibliographic database. Its intended use is as a generator for research ideas that should be then investigated by traditional medical methods.

The idea of discovering new relations from a bibliographic database was introduced by Swanson [1,2] who managed to make seven medical discoveries that have been published in relevant medical journals [1,3]. Swanson's discovery support process is based on the concepts of complementary literatures and noninteractive literatures [1,2]. By combining the concepts of complementary and noninteractive literatures, Swanson developed the concept of undiscovered public knowledge.

In the beginning, Swanson performed the discovery process manually by searching the Medline database [2]. Later he added software support for some of the stages of the process. His current system is called ARROWSMITH and is described in detail in [3].

Some of the Swanson's discoveries were repeated with different methods by Gordon and Lindsay [4, 5] and by Weeber [6].

Our system is based on Swanson's ideas, but there are however, several notable differences between our approach and theirs. Instead of using title words as a representation of the Medline documents' meaning, we use the MeSH descriptors. The MeSH descriptors are assigned to documents by human experts during the document indexing stage. We believe that the MeSH descriptors represent more precisely what a particular document is about. We use association rules as a measure of relationship between medical concepts while Swanson uses word frequencies. We have built a large association rule base by pre-calculating and storing the association rules in a database management system. This allows us to build a truly interactive discovery support system with fast response.

Materials and Methods

Medline. The Medline database is a product of the US National Library of Medicine (NML). Because of its coverage and free accessibility, Medline is the most important bibliographic database in the field of biomedicine. Each citation is associated with a set of MeSH (Medical Subject Headings) terms that describe the content of the item [7].

Medical Subject Headings (MeSH). MeSH comprises NLM's controlled vocabulary and thesaurus used for indexing articles and for searching MeSH-indexed databases, including Medline [7].

Unified Medical Language System (UMLS). The Unified Medical Language System (UMLS) project that NLM began in 1986 was undertaken with the goal of providing a mechanism for linking diverse medical vocabularies as well as sources of information. UMLS consists of three components: the Metathesaurus, Specialist Lexicon and Semantic Network [7,8].

Association rules. Association rules [9] were originally developed with the purpose of market-basket analysis, where it is of interest to find patterns of the form X -> Y, e.g., beer -> peanuts, with the intuitive meaning "baskets that contain X tend to contain Y".

A basket corresponds to a single visit of a customer to a store and is called a transaction, while individual products in the basket are called items. The approach is general enough to apply to bibliographic databases, where transactions are documents and items are words or descriptors used for indexing the documents. Association rules here have the form Word1 -> Word2 or Descriptor1 -> Descriptor2 (e.g. Disease X -> Symptom Y).

System Description

The system we developed is an interactive discovery support system for the field of medicine and is supposed to be used as a generator of new, potentially meaningful relations between a starting known concept of interest and other concepts.

The Medline database is used heavily by medical researchers. Traditionally it is used to check what is new in the literature on a particular topic of interest or to check if a medical discovery has already been published. In contrast to the traditional use of Medline, our system actively helps in the discovery process by generating potentially new discoveries and research ideas by analyzing the Medline database.

We used the major MeSH descriptors assigned to a Medline record as a representation of the contents of the article the record is about. Some of the MeSH descriptors are designated as major (preceded by an asterisk in the Medline record). Major descriptors are those that are the main topic of the article.

We used association rules [9] between pairs of medical concepts as a method to determine which concepts are related to a given starting concept. In our system an association rule of the form

X -> Y (confidence, support)

means that in *confidence* percent of articles containing X, Y is present and that there are *support* number such articles. In other words, we take concept co-occurrence as an indication of a relation between concepts. If X is a disease, for example, then some possible relations might be: *has-symptom*, *is-caused-by*, *is-treated-with-drug* and so on. We do not try to find out the kind of relation. This can not be done by using the MeSH descriptors assigned to an article because there is no information about the relation between the descriptors.

Algorithm. We calculated all the associations between the major MeSH descriptors. We did this regardless of the confidence and support values and for two Medline time segments: 1990-1995 and 1996-1999. The calculated associations are stored in a database management system: there are currently more than 11.000.000 associations in the rule base. The calculation of the association rules was much simplified by the use of the data contained in the UMLS, especially the co-occurrence files. Actually, for these calculations it was not necessary to access the full Medline records at all.

The large association rule base is a foundation upon which the algorithm for discovering new relations between concepts proceeds as described in Table 1. The main idea is to first find all the concepts Y related to the starting concept X (e.g. if X is a disease then Y can be pathological functions, symptoms, ...). Then all the concepts Z related to Y are found (e.g. if Y is a pathological function, Z can be a chemical regulating that function). As a last step we check if X and Z appear together in the medical literature. If they do not appear together then we have discovered a potentially new relation between X and Z. This relation should be confirmed or rejected using human judgement, laboratory methods or clinical investigations, depending on the nature of X and Z.

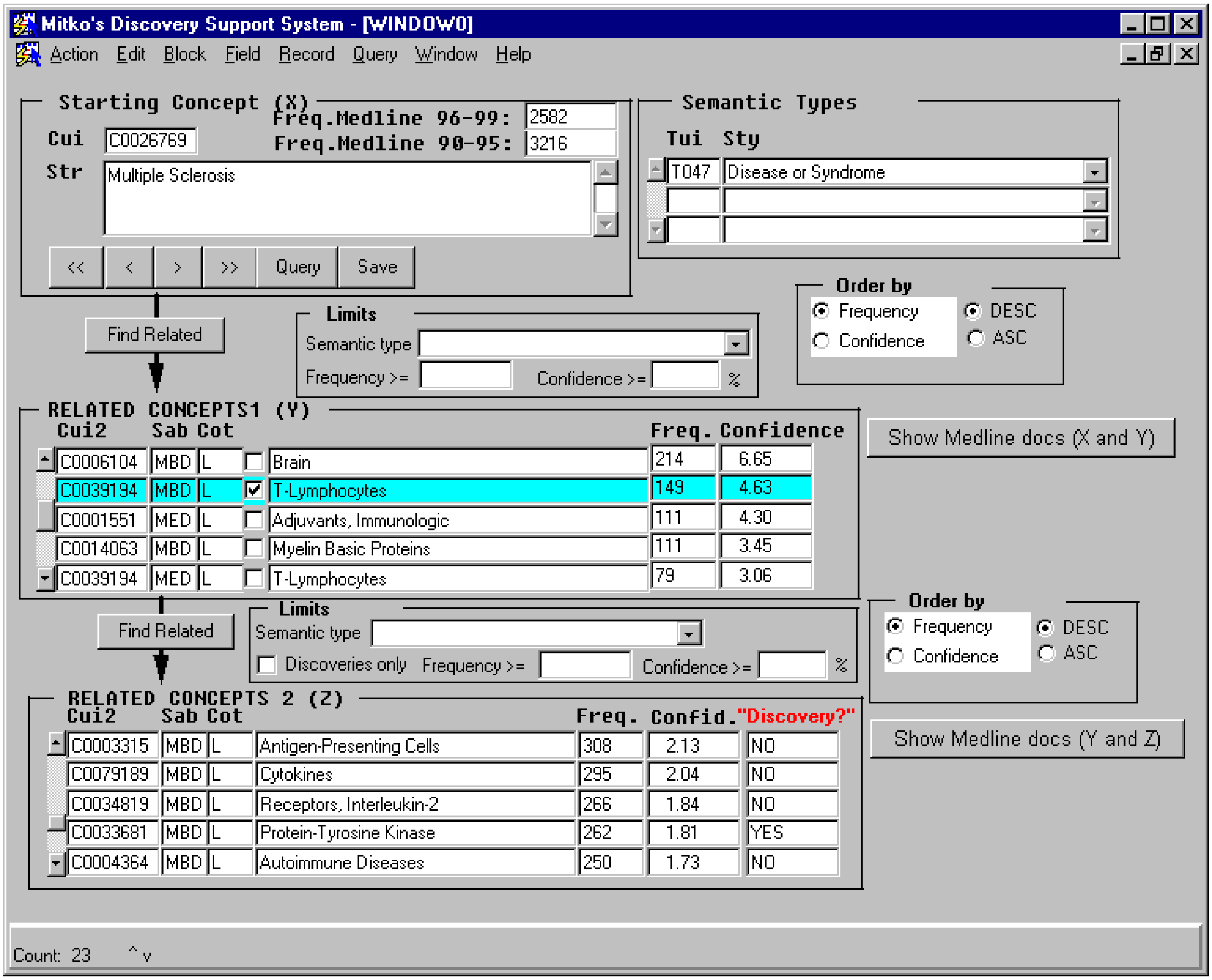
Table 1 -. The algorithm for discovering new relations between medical concepts.

1. Let X be a given starting concept of interest.
2. Find all concepts Y such that X -> Y.
3. Find all concepts Z such that Y -> Z.
4. Eliminate those Z for which X -> Z already exists.
5. The remaining Z concepts are candidates for a new relation between X and Z.

Because in Medline each concept can be associated with many other concepts, the possible number of X -> Z combinations can be extremely large. In order to deal with this combinatorial problem, the algorithm incorporates *filtering (limiting)* and *ordering* capabilities. The default filtering that can not be relaxed is that only the associations between major MeSH headings are considered by the system. The related concepts can be limited by the semantic type to which they belong. Each MeSH descriptor belongs to one or more semantic types. For example, if the starting concept X is a disease (semantic type *disease or syndrome*) then the user can request that Y concepts are of semantic type *pathologic function* and that Z concepts are of semantic type *pharmacologic substance.* The last possibility for limiting the number of related concepts is by setting thresholds on the support and confidence measures of the association rules in steps 2. and 3. of the algorithm. In fact, all of the filtering options can be interactively set alone or several of them in combination.

The goal of the ordering is to present best candidates first to make human review as easy as possible. Currently the default ordering is by the decreasing association rule confidence, but it is also possible to order by support or semantic type.

Implementation. Figure 1 shows the user interface of our discovery support system. The user starts a discovery session by searching for a starting concept X, which is usually from his own research area. The concepts Y related to the starting concept are found by pressing the *Find Related* button and are presented in the *Related Concepts1* frame. Before finding related concepts the user can specify limits and the order of the related concepts. Similarly the Z concepts related to the Y concepts are found and shown in the *RELATED CONCEPTS2* frame. The frame *RELATED CONCEPTS2* contains an important additional field designated as *"Discovery?"*. This value of this field is YES if a relation (association) between the starting concept X and the current concept Z does not exist in the appropriate Medline segment and *NO* if such a relation exists. It is also possible to find and display the related Medline records in a separate window.


Mitko's Discovery Support System - [WINDOW0]
Action Edit Block Field Record Query Window Help
<< < > >> Query Save
Find Related
Show Medline docs (X and Y)
Find Related
Show Medline docs (Y and Z)

Starting Concept (X)

Cui C0026769

Freq.Medline 96-99: 2582

Freq.Medline 90-95: 3216

Str Multiple Sclerosis

Semantic Types

Tui	Sty
T047	Disease or Syndrome

Order by

● Frequency ○ Confidence

● DESC ○ ASC

Limits

Semantic type

Frequency >= Confidence >= %

RELATED CONCEPTS1 (Y)

Cui2	Sab	Cot			Freq.	Confidence
C0006104	MBD	L	□	Brain	214	6.65
C0039194	MBD	L	☑	T-Lymphocytes	149	4.63
C0001551	MED	L	□	Adjuvants, Immunologic	111	4.30
C0014063	MBD	L	□	Myelin Basic Proteins	111	3.45
C0039194	MED	L	□	T-Lymphocytes	79	3.06

Limits

Semantic type

□ Discoveries only Frequency >= Confidence >= %

Order by

● Frequency ○ Confidence

● DESC ○ ASC

RELATED CONCEPTS 2 (Z)

Cui2	Sab	Cot		Freq.	Confid.	"Discovery?"
C0003315	MBD	L	Antigen-Presenting Cells	308	2.13	NO
C0079189	MBD	L	Cytokines	295	2.04	NO
C0034819	MBD	L	Receptors, Interleukin-2	266	1.84	NO
C0033681	MBD	L	Protein-Tyrosine Kinase	262	1.81	YES
C0004364	MBD	L	Autoimmune Diseases	250	1.73	NO

Count: 23

Figure 1 - The user interface of the interactive discovery support system.

*Table 2 - The results of the prediction of new relationships between medical concepts in the newer Medline segment (1996-1999) based on the older segment (1990-1995) using the system. The column names ending with **1** are for the AVGS constraint and those ending with **2** for the 2*AVGS constraint. The columns have the following meaning: **n** - all the relationships that can be predicted; **k** - new relationships in the newer segment that were not present in the older segment; **m** - predicted relationships based on the older segment; **l** - successfully predicted relationships; **p** - probability of achieving **l** or more successfully predicted relations by chance; **r** - the number of successfully predicted relations by chance alone.*

Disease	***n***	***k***	***m1***	***l1***	***p1***	***r1***	***m2***	***l2***	***p2***	***r2***
Multiple Sclerosis	15965	635	6848	521	0	272	3151	366	0	125
Temporal Arteritis	17190	187	4735	148	0	52	1157	72	0	16
Melanoma	15336	692	6272	560	0	283	2812	392	0	127
Parkinson Disease	15966	594	5995	477	0	223	2322	309	0	86
Incontinentia Pigmenti	17504	44	3435	37	0	9	873	23	0	2
Chondrodysplasia Punctata	17422	18	2864	15	0	3	1046	9	0.00000016	1
Charcot-Marie-Tooth Disease	17355	131	3150	105	0	24	1019	66	0	8
Focal Dermal Hypoplasia	17527	23	1511	14	0	2	610	8	0.00000037	1
Noonan Syndrome	17384	68	3015	59	0	12	536	23	0	2
Ectodermal Dysplasia	17322	124	3301	96	0	24	967	45	0	7

Results

The ultimate proof of the system would be to (help) discover medical discoveries that could be published in relevant medical journals. However, we have managed to do a statistical evaluation so far. The goal of this evaluation was to see how many of the potential discoveries made by the system at some point of time become realised at a later time. For us, a potential discovery is a relationship between two concepts proposed by our system, but not present in Medline at some point of time. We consider the potential discovery realised if the two concepts later appear together in a document in the Medline database. In other words, the goal of the evaluation was to see how good our system was in predicting what discoveries would be made in the future.

We approached this goal by first dividing the Medline database and the corresponding association rules into two segments according to the publication date of the documents stored: the older segment is from 1990 to 1995 and the newer segment is from 1996 to 1999. We then analysed ten diseases, which are listed in Table 2.

Here we will give a discussion of the analysis of *Multiple sclerosis (MS).* MS appears in 2582 documents in the older segment. It is related to 1610 distinct concepts. When analysing the old segment, the system proposed 15617 concepts as potential discoveries. MS is related to 662 new concepts in the new segment that it was not related to in the old segment. Our system successfully predicted 95.5% (632 out of 662) realised discoveries in the new segment. However, only 4% (632 out of 15617) of the proposed potential discoveries got realised. It should be stressed that MS was not related to 15965 out of 17575 distinct concepts appearing in the older segment. The system proposed 97.8% of the concepts MS was not yet related to as potential discoveries. The conclusion is that without using limits on the strength of relationship the system is very successful at predicting future discoveries, but proposes far too many potential discoveries. Then we repeated the evaluation with two values for thresholds on the support level of the association rules. In one case the threshold was set to the average support of the associations between one concept and the others (AVGS) and in the other case it was set to 2*AVGS. Only associations with support greater or equal to the threshold were taken into account. The number of proposed potential discoveries dropped from 15617 without thresholds to 6848 for AVGS and to 3151 for the 2*AVGS threshold. The percent of successfully predicted realised discoveries dropped from 95.5% (632 of 662) without thresholds to 78.7% (521 of 662) for AVGS and to 55.2% (366 of 662) for 2*AVGS. However, the ratio of realised to proposed potential discoveries improved from 4% (632 out of

15617) without thresholds to 7.6% (521 of 6848) for AVGS and to 11.6% (366 of 3151) for 2*AVGS. We conclude from this that with the use of proper thresholds the usability of the system is much better because a smaller number of better potential discoveries are generated.

Results of the statistical evaluation for the ten selected diseases are in Table 2. The values obtained by our system were tested against the null hypotheses of random hits. *p* values in the table are then probabilities of our results (or results better than ours) if predictions were based on chance alone. The *p* values were obtained using the hypergeometric distribution using the respective *n, m, k* and *l* values.

Values in the columns *r1* and *r2* tell us how many realized relations would be correctly predicted if predictions were based on chance alone and *m1* and *m2* predictions were made respectively. Zeros in the *p* value column actually mean that the probability is less than 10^{-16}.

Discussion and Further Work

The paper presented an interactive discovery support system for the field of medicine. For a given starting medical concept it discovers new, potentially meaningful relations with other concepts that have not been published in the medical literature before. The proposed relations should be evaluated and verified by a qualified medical professional.

As a measure of the relation between concepts we use association rules calculated from the Medline bibliographic database. In the statistical evaluation, the system proved to be successful at predicting future discoveries. However, this came at the expense of generating a large number of potential discoveries that have to be judged and verified by the user of the system.

We have several ideas for improving the system. One of them is to develop a Web version of the system that will increase the number of users. More users means better chance for real medical discoveries. The preliminary evaluation showed that properly set thresholds are crucial for successful use of the system. Thus, we plan to work on setting good default values for the thresholds that can be changed by the user if necessary. Another important way to improve the system is to include additional information sources, such as molecular biology sequence databases.

References

[1] Swanson, D.R.: Fish oil, Raynaud's syndrome, and undiscovered public knowledge. Perspect Biol Med. 1986 Autumn;30(1):7-18.

[2] Swanson, D.R.: Online search for logically-related noninteractive medical literatures: a systematic trial-and-error strategy. J Am Soc Inf Sci. 1989 Sep;40(5):356-8.

[3] Swanson, D.R., Smalheiser, N.R.: An interactive system for finding complementary literatures: a stimulus to scientific discovery. Artif. Intell. 91 (1997) 183-203.

[4] Gordon MD, Lindsay RK. Toward discovery support systems: A replication, re-examination, and extension of Swanson's work on literature-based discovery of a connection between Raynaud's and fish oil. J Am Soc Inf Sci 1996; 47(2):116-128.

[5] Lindsay RK, Gordon MD. Literature-based discovery by lexical statistics. J Am Soc Inf Sci 1999; 50(7):574-587.

[6] Weeber M, Klein H, Aronson AR, Mork JG, Jong-Van Den Berg L, Vos R. Text-based discovery in biomedicine: the architecture of the DAD-system. Proc AMIA Symp. 2000;(20 Suppl):903-7.

[7] U.S. National Library of Medicine. http://www.nlm.nih.gov/<30.04.2000>

[8] Humphreys, B.L., Lindberg, D.A.B., Schoolman, H.M., Barnett, G.O.: The Unified Medical Language System: an informatics research collaboration. JAMIA 1998;5(1):1-11.

[9] Agrawal, R. et al: Fast discovery of association rules. In U. Fayyad et al, editors, Advances in Knowledge Discovery and Data Mining. MIT Press, Cambridge, MA. (1996)

Address for correspondence

Dimitar Hristovski

Institute of Biomedical Informatics, Medical Faculty, University of Ljubljana; Vrazov trg 2/2, 1105 Ljubljana, Slovenia. e-mail: dimitar.hristovski@mf.uni-lj.si