PERIODICA POLYTECHNICA SER. CHEM. ENG. VOL. 50, NO. 1, PP. 33–44 (2006)

KNOWLEDGE DISCOVERY FOR IDENTIFICATION OF ENZYME WITH A PRIORI SPECIFIED PROPERTIES

Andrzej KRASLAWSKI and Sergey BELIAEV

Department of Chemical Technology
Lappeenranta University of Technology,
P.O. Box 20, FIN-53851 Lappeenranta, Finland
ph: +358 40 591 3379, fax: +358 5 621 2199 e-mail: Andrzej.Kraslawski@lut.fi

Received: Sept. 27, 2005

Abstract

Development of new products with the given properties from the known raw materials is one of the common research tasks in process engineering. Usually the first step of research is a literature survey based on the search for the specific keywords. Nowadays there exist many vast databases of articles and patents, and the traditional, keywords-based, searching tools are not always sufficient to find the desired information. The main objective of this paper is to develop methodology for identification of new materials, based on knowledge discovery. As an example, the proposed methodology is applied for identification of new enzyme of microbial origin capable of polymerizing lactose in aqueous solution, with the number of required criteria.

Keywords: Data Mining, knowledge discovery, semantic analysis, lactose polymerization.

1. Introduction

One of the common research tasks of process engineering is a search for new products with the given properties. Usually a literature review is the first step of the search. Nowadays there exist many databases of articles on chemical technology and traditional searching tools (based on keywords) are not always sufficient to find the desired information.

The traditional computer search of the references, e.g. based on keywords, has two major drawbacks. The first one is a generation of huge, usually not-enough-specific set of the references that are impossible to be analysed. The second problem, often encountered in the scientific research, is incompletely defined search space. It means that only a part of the 'keywords' is defined. Very often there is a need to find common keywords between two facts, but preliminarily those keywords are unknown and therefore, it is impossible to find them by the traditional queries. The keywords-based approach will, of course, not allow to find the unidentified products or raw materials. The main objective of this article is the development of a methodology for identification of new multifunctional products, based on knowledge discovery.

An example of its applicability is presented for identification of a new enzyme

of microbial origin capable of polymerizing lactose in aqueous solution, with a priori specified properties.

2. Knowledge Discovery in Databases

At an abstract level, Knowledge Discovery in Databases (KDD) field is concerned with the development of methods and techniques for making sense of data. The basic problem addressed by the KDD process is one of mapping low-level data (which are typically too voluminous to understand and digest easily) into other forms that might be more compact (for example, a short report) or more useful (for example, a predictive model for estimating the value of future cases). KDD is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data [4].

At the core of KDD is the application of Data Mining methods for pattern recognition and discovery.

Data Mining is a step in the KDD consisting in applying data analysis and discovery algorithms that, under acceptable limitations of computational efficiency, produce a particular enumeration of patterns (or models) over the data [4].

Data Mining differs from the traditional techniques in that it does not recover from a collection a subset of documents which are hopefully relevant to a query, based on keyword searching. Instead, the goal is to extract from the documents (which may be in a variety of languages) salient facts about pre-specified types of entities and relationships.

Currently, KDD and Data Mining are used essentially as synonyms in many literature sources, but that is not correct. In our view, KDD refers to the overall process of discovering useful knowledge from data, and Data Mining refers to a particular step in this process. The distinction between the KDD process and the data-mining step (within the process) is a central point. The additional steps in the KDD process, such as data preparation, data selection, data cleaning, incorporation of appropriate prior knowledge, and proper interpretation of the results of mining, are essential to ensure that useful knowledge is derived from the data. Blind application of data-mining methods (rightly criticized as data dredging in the statistical literature) can be a dangerous activity, easily leading to the discovery of meaningless and invalid patterns.

Data Mining functions can be divided into two categories: supervised (directed) and unsupervised (undirected) [9].Supervised functions are used to predict a value; they require the specification of a target (known outcome). Unsupervised functions are used to find the intrinsic structure, relationships, or affinities in data. Unsupervised mining does not use a target.

Supervised data-mining methods:

- **Classification**: Grouping items into discrete classes and predicting to which class an item belongs [5, 3].
- **Regression**: Approximating and forecasting continuous values [3].

– **Attribute Importance**: Identifying the attributes that are most important in predicting results.

Unsupervised data-mining methods:

– **Clustering**: Finding natural groupings in the data [7, 2]
– **Association models**: Analysing 'market baskets' [1].

Each data-mining method has its own algorithms. The different algorithms serve different purposes and each algorithm has advantages and disadvantages.

- **Decision trees:** Tree-shaped structures that represent sets of decisions. Decision tree rules provide model transparency so that a chemical engineer can understand the basis of the model's predictions, and therefore, be comfortable acting on the predictions and explaining them to others.
- **Rule induction:** The extraction of useful IF-THEN rules from data, based on statistical importance.
- **Support Vector Machine Algorithm**: An algorithm with strong regularization properties, that is, the optimization procedure maximizes predictive accuracy while automatically avoiding over-fitting of the training data.
- **Orthogonal Partitioning Clustering (O-Cluster) Algorithm**: O-Cluster creates a hierarchical grid-based clustering model, that is, it creates axis-parallel (orthogonal) partitions in the input attribute space.
- **Non-Negative Matrix Factorization (NMF)**: [8].

The classification method of Data Mining and the Decision Tree algorithm will be used in this article.

3. Types of Knowledge Discovery Processes

There are two main models of knowledge discovery process: open - for generation of a hypothesis and closed – for testing of a hypothesis [15]. Usually we use first an open approach for the generation of a hypothesis and next a closed approach for testing of the generated one.

There exists also a semi-open discovery process. It is a procedure similar to open process, but performed with the specific, a priori determined, selection criteria for the hypothesis.

The traditional search, based on the keywords and data mining methods are applicable for these discovery processes.

3.1. Open Discovery Process

An open knowledge discovery process is used in generation of a hypothesis. Initially, there is only the specified problem without clear final target. For example,

the problem is associated with the set of articles containing fact A. Such set is called a literature A. In *Fig. 1* there is shown the pathway from A, proceeding through the intermediate terms B_I, that co-occur with A, to various targets C_J, each of which co-occurs with one or more B-terms. For each literature A there can be many of B-terms, and for each of B-terms there can be many of C-targets. The solid arrows indicate potentially interesting pathways of discovery, the dashed ones unsuccessful pathways.

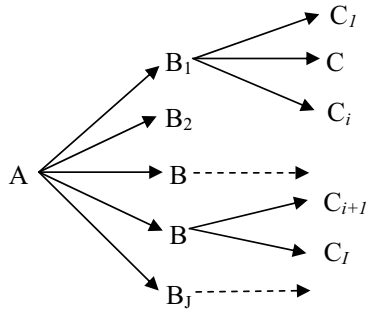

Fig. 1. Open discovery process.

3.2. Semi-Open Discovery Process

The semi-open discovery process is similar to open process, but there are specified some selection criteria for the intermediate terms B_I and targets C_J. In a classical open discovery process there exist a huge amount of possible, potentially interesting, B-terms and targets C_J. It leads to a combinatorial explosion of solutions. In order to reduce a number of possible solutions we can use some selection criteria for the intermediate terms B_I and targets C_J. For example, term B should be a separation process, that could be used for the separation of the substances A or C. The C should be a product that can be produced from the A by using a separation process B.

3.3. Closed Discovery Process

A closed discovery process is aimed at testing of the hypothesis. The search is started with the given hypothesis, possibly obtained by the open discovery route described above. This hypothesis has to be tested basing on the literature search, as shown in *Fig. 2*. For example, there is a hypothesis that two literature lists A and C are connected. We will search for any common intermediate B-terms, that connect A and C. For each member (article, patent, etc.) of the literature lists A and C there

can be many of B-terms. The solid arrows indicate potentially interesting pathways of discovery, the dashed ones unsuccessful pathways.

In an open search, the sets of the articles on A and B are studied; in the closed search, the sets A and C are analysed.

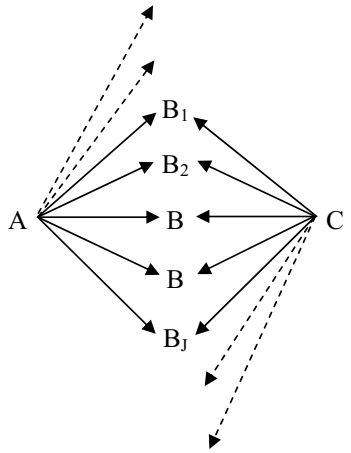

Fig. 2. Closed discovery process.

4. Development of Search Paradigm

The closed discovery process seems to be the most promising for knowledge discovery related to a search for new products to be produced from the known raw materials or the identification of new raw materials for the synthesis of the known products.

During the past two decades, SMALHEISER and SWANSON [12, 10, 14] have developed a set of interactive software and database search strategy for knowledge discovery, called ARROWSMITH. They suggested to combine existing bibliographical information for discovery of knowledge.

The approach proposed by Smalheiser and Swanson is based on the implementation of the following transitivity rule:

$$A \rightarrow B \text{ and } B \rightarrow C \Longrightarrow A \rightarrow C$$

The main idea of the transitivity rule is as follows: term A is related with term B, and term B is related with term C. Simultaneously we assume that there is a relation between A and C and we search for this relation-term B. The discoveries obtained when using the ARROWSMITH approach and software, [10, 11, 13], have been confirmed experimentally and clinically in the field of medicine.

The ARROWSMITH method is an example of the closed approach.

The idea of ARROWSMITH is presented in *Fig. 3*. At first, we search for the list of literature sources where term A is mentioned and another list of literature sources where term C is mentioned. Using semantic analysis we can identify several terms B_i, ($i = 1...n$, where n is a total number of the common terms B) that appear simultaneously in the set of literature sources dealing with A and in the other group of sources containing C. Next, in two lists we search for the articles from sources containing A and C that have common terms B_i. The group of terms B_i is composed of two sub-groups: one of the known terms B_{KNOWN} and the second of unknown to us terms, $B_{\text{UNKNOWN.}}$ The known terms B_{KNOWN} could be easily identified by the conventional search with the keywords A and C. In consequence, the papers containing the terms B_{KNOWN} are not interesting for knowledge discovery as they are explicitly given in the literature and as such they are rejected from the subsequent analysis. The further analysis in search for knowledge discovery will concentrate exclusively on the reference material containing the terms $B_{\text{UNKNOWN.}}$

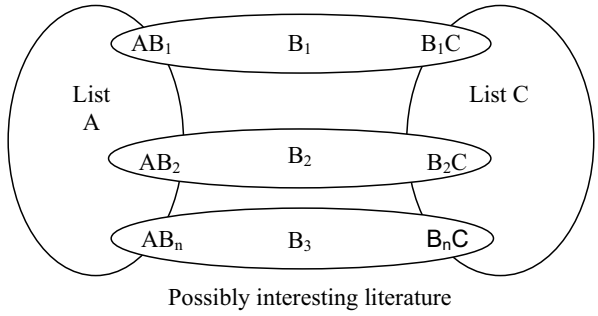

Fig. 3. Venn diagram representing the single-level approach [14]

ARROWSMITH is also an interactive software for the discovery of plausible hypotheses linking findings across specialties in biomedical literature (e.g. MEDLINE).

It contains a pre-compiled 'stop-list' of 5000 words, which are definitely non-interesting (e.g. 'the', 'for'). One of the hypotheses was that there is a biologically-

significant relationship between Raynaud's disease with fish oil and migraine with a magnesium deficiency. After Swanson published his paper bringing attention to this hypothesis, 12 papers appeared reporting experimental or clinic tests and almost all were positive.

Nevertheless, ARROWSMITH has a range of bottlenecks:

- It can be applied to MEDLINE (the 'stop-list' has been generated especially for that area);
- B-terms are usually single words, bigrams and trigrams;
- User should create manually both lists – A and C.

In this paper, we propose the extensions of the ARROWSMITH approach, called multi-level knowledge discovery approach, combined with the developed semantic rules.

4.1. Use of Semantic Analysis in Closed Discovery Process

The list of the articles containing the B-terms, obtained in multi-level approach, is potentially very long and its filtering is desperately needed. There are two main difficulties related to this task. The first one is an identification of the searched words in the body of articles. For example, the same substances could be a solvent, a product etc. It is very difficult to recognize which one is a solvent B, and which one is a raw material A and a product C. The second difficulty lies in the fact that the searched terms are identified by more then one word, e.g. traditional names of chemical substances. Finding meaningful multi-word terms in the text is a non-trivial task.

In the last decade, many scientists suggested different approaches and concepts, like the use of an extensive stop list, a list of words such as determiners and adverbs that are considered non-relevant; analytic approach based on word frequency statistics [14, 15].

Our approach to the analysis of title, abstract and full text (articles, patents, web-sites) is based on the use of semantic analysis. We have developed the specific semantic rules to analyse the text. For instance, if we are looking for material with the specific property, there should be in the analysed text at least one phrase containing the name of this property, e.g. bacterial. In this case, any substance in such phrase, with high probability, could be a sought material. The application of the semantic analysis for the searching of products will considerably simplify and speed up the process of the text analysis.

The phrase containing the word 'enzyme' or 'polymerase' has to be identified in order to determine a name of the potentially interesting enzyme.

The following semantic structure could be used, *Fig. 4*:

It is an empirical semantic rule that has been created after 'manual' analysis of the articles by one of the authors. Usually, such rules are formulated by the experts in the given field.

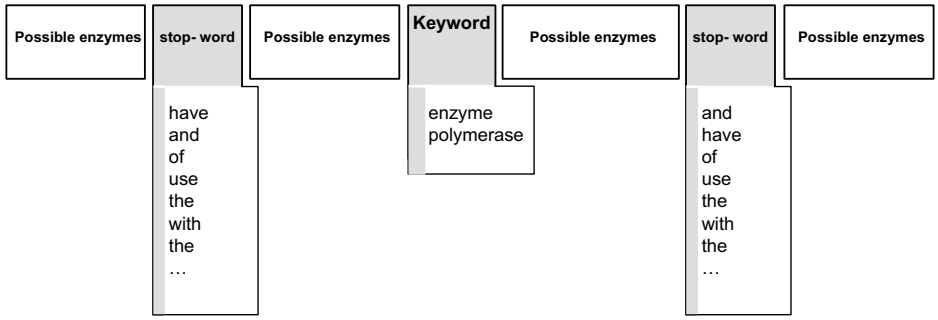

Fig. 4. The semantic structure used in the search for enzyme of microbial origin capable of polymerizing lactose.

There has been developed a web-based interactive software, designed for advanced search in internet-databases. It combines the traditional (based on keywords) search, Data Mining methods and semantic search. The software is able to analyse any documents in Adobe Acrobat (PDF), Microsoft Word (DOC) and HyperText Markup Language (HTML) format. In this paper, the Elsevier database (`www.sciencedirect.com`) has been used as a primary source, as the biggest world database. In principle, it could be any electronic database, containing the documents in the formats described above.

5. Multi-lever Approach of Knowledge Discovery Process

Very often, the researcher has to identify a new material with more then two priori specified properties. For this case, the a multi-level approach has been developed.

There are several lists of articles: A, where term A is mentioned; C, where term C is mentioned; D, where term D is mentioned, etc. The multi-level approach is a search performed simultaneously on the several lists A, C. D, etc, instead of the exploration of only two lists A and C as in a single approach. The objective of the search is to identify the relations-terms B between the facts A, C, D, E etc. that are implicitly given in the literature. Venn diagram representing the searching of a new knowledge for multi-level approach is shown in *Fig. 5*.

The Venn diagram from *Fig. 5* could be represented as in *Fig. 6*.

The multi-level approach may be used for the more complicated cases when the lists A, C, D and E should be composed of more than one element. For example we are interested not only in one specific solvent S but in several ones which are simultaneously applicable to the given reaction C_1, C_2 and C_3 and this same for the lists A, D and E, *Fig. 7*.

The multi-level approach is illustrated by the following task:

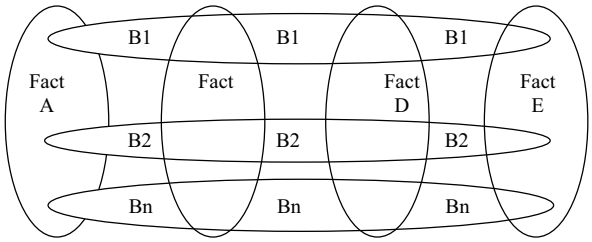

Fig. 5. Venn diagram representing the multi-level knowledge discovery

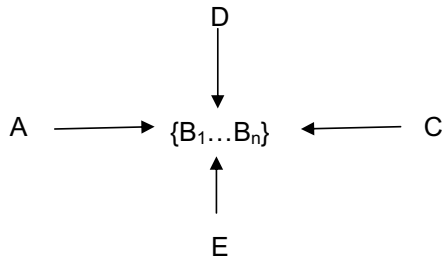

Fig. 6. Decision tree for multi-level approach

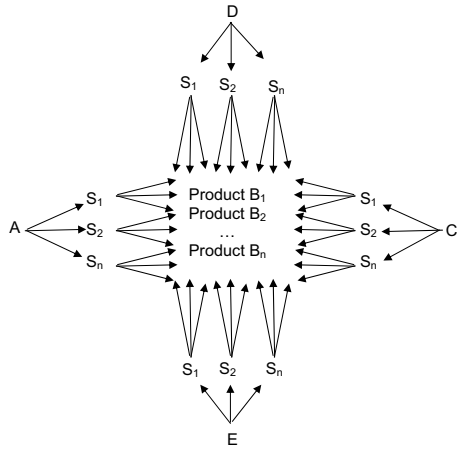

Fig. 7. Extended decision tree for multi-level approach

An enzyme of microbial origin capable of lactose polymerization in aqueous solutions without toxic co-factors and by-products is needed.

The enzyme should meet the following criteria:

At the enzyme concentration of 10 microgram/mL and lactose concentration of 10%, the enzyme should convert at least 25% of the lactose into polymeric form within 5 hours.

The degree of polymerization of the reaction product should be > 10, i.e. molecular mass of the product should exceed 3,000 Da as determined by mass spectrometry or gel permeation chromatography.

The enzyme must be extracellular, i.e. secreted into the culture medium.

Ideally, no small molecules (cofactors, mediators, etc.) should be required for the reaction. If other molecular in addition to the enzyme is needed to improve the reaction efficiency, these additives must be of food-grade and needed only in catalytic amounts with the turnover number of at least 1,000 (the concentration of the additive in the reaction should be 1,000-fold less than lactose concentration, i.e. below 1000 mg/L.)

The enzyme should be presented as 90% pure material as judged by SDS-PAGE.

At the beginning, five lists A, C, D, E and F have been generated, using Elsevier database.

Step 1. The traditional keyword search in the database of the articles containing the word 'polymerizing' AND 'lactose' – creation of **list A**.

Step 2. Search for the potentially interesting sub-phrases that could be enzymes, using the developed semantic rule. It means there must be at least one phrase with enzyme from semantic structure in *Fig. 4*. All sub-phrases should be saved into the databases for the further analysis. There were 1100 articles in the database.

The phrase 'Nonetheless, Arg residues have recently emerged as general bases in several enzymes: IMP dehydrogenase, pectate/pectin lyases, fumarate reductase, and L-aspartate oxidase.' [6] from the *Fig. 8* is an example, showing how the semantic rule works.

After analysis, the following sub-phrases have been selected and saved to the database: 'Nonetheless, Arg residues', 'recently emerged as general bases in several', 'IMP dehydrogenase, pectate/pectin lyases, fumarate reductase' and 'L-aspartate oxidase'. The last two phrases are really containing the possible enzymes. As a result, only those sub-phrases will be used for the further search as the facts A, instead of the whole sentence.

Step 3. The traditional keyword search in the database of the articles containing the word 'procariotic' OR 'bacterial' – creation of **list C**.

Step 4. Search for the potentially interesting sub-phrases that could be enzymes, using the developed semantic rule. It means there must be at least one phrase with enzyme from semantic structure in *Fig. 4*. All sub-phrases should be saved into the databases for the further analysis. There were more than 10 000 articles in the database.

Step 5. The traditional keyword search in the database of the articles containing the word 'extracellular' OR 'secretion' – creation of **list D**.

Step 6. Search for the potentially interesting sub-phrases that could be enzymes, using the developed semantic rule. It means there must be at least one phrase

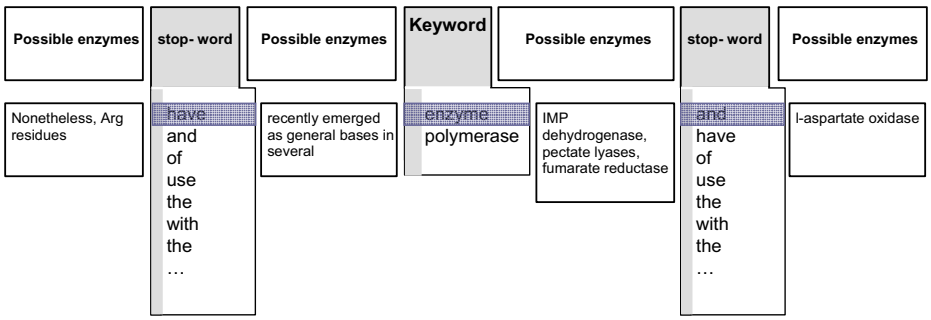

Fig. 8. The example of the phrase that satisfies the semantic structure from *Fig. 4*

with enzyme from semantic structure in *Fig. 4*. All sub-phrases should be saved into the databases for the further analysis. There were more than 10 000 articles in the database.

Ste 7. The traditional keyword search in the database of the articles containing the word 'Cofactor' AND 'Free' – creation of **list E**.

Step 8. Search for the potentially interesting sub-phrases that could be enzymes, using the developed semantic rule. It means there must be at least one phrase with enzyme from semantic structure in *Fig. 4*. All sub-phrases should be saved into the databases for the further analysis. There were more than 6 000 articles in the database.

Step 9. The traditional keyword search in the database of the articles containing the word 'SDS-PAGE' AND 'purity' – creation of **list F**.

Step 10. Search for the potentially interesting sub-phrases that could be enzymes, using the developed semantic rule. It means there must be at least one phrase with enzyme from semantic structure in Fig. 4. All sub-phrases should be saved into the databases for the further analysis. There were more than 2 000 articles in the database.

Step 11. Application of the classification method and decision tree algorithm, shown in *Fig. 6*, to the saved sub-phrases in the list A, C, D and E to identify all possible products $B_i., i = 1...N$.

A range of potentially interesting products, common for the lists A, C, D, E, F were found e.g: RNA polymerase, ribose polymerase, lactose polymerase, arginyl aminopeptidase, etc.

After analysing the identified articles, RNA polymerase was selected as a potentially interesting product.

6. Summary

A new method of knowledge discovery have been developed and applied in the research of the biochemical reactions. The Data Mining techniques and semantic rules allow a huge number of literature sources to be treated and reduce the amount of routine work in finding new knowledge. Therefore, it can significantly reduce the development time for the design of a new product.

As example, it was applied for the search for a new enzyme of microbial origin capable of polymerizing lactose in aqueous solution, with the number of required criteria. The obtained results are the subject of the experimental verification.

References

[1] AGRAWAL, R. – MANNILA, H. – SRIKANT, R. – TOIVONEN, H. – VERKAMO, I., *Fast Discovery of Association Rules*, Advances in Knowledge Discovery and Data Mining, eds. U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, (1996) pp. 307–328. Menlo Park, Calif.: AAAI Press.

[2] BARZILAY R. AND LEE L., Learning to Paraphrase: An Unsupervised Approach Using Multiple-sequence Alignment,. *LIn Proceedings of HLT-NAACL.*, (2003) pp. 16–23.

[3] BREIMAN, L. – FRIEDMAN, J. H. – OLSHEN, R. A. – STONE, C. J., *Classification and Regression Trees.* Belmont, Calif.: Wadsworth (1984).

[4] FAYYAD U. – PIATETSKY-SHAPIRO G.– SMUTH P., *From Data Mining to Knowledge Discovery in Databases*, American Association for Artificial Intelligence, (1996), pp. 37–54.

[5] FURNKRANZ J., Exploiting Structural Information for Text Classification on the www. In *Proceedings of the Third International Symposium on Advances in Intelligent Data Analysis*, (1999). pp. 487–498. Springer-Verlag.

[6] GUILLÉN SCHLIPPE Y.V. – HEDSTROM L., A Twisted Base? The Role of Arginine in Enzyme-catalyzed Proton Abstractions, *Archives of Biochemistry and Biophysics*, **433/1**, (2004). pp. 266–278.

[7] JAIN, A.K. – DUBES, R.C., *Algorithms for Clustering Data*. Englewood Cliffs, N.J.: Prentice-Hall, 1988.

[8] LEE D. D.– SEUNG H. S., Learning the Parts of Objects with Nonnegative Matrix Factorization, *Nature* **401**, (1999), p. 788.

[9] STEPHENS, S. – TAMAYO, P. Supervised and Unsupervised Data Mining Techniques for Life Sciences, Curr Drug Disc, 2003.

[10] SMALHEISER, N. R. – SWANSON, D. R., Using ARROWSMITH: A Computer-Assisted Approach to Formulating and Assessing Scientific Hypotheses, *Computer Methods and Programs in Biomedicine*, **57**, 1998, pp. 149–153.

[11] SWANSON, D. R., Fish Oil, Raynaud's Syndrome, and Undiscovered Public Knowledge, *Perspectives in Biology and Medicine*, **30**, 1986, pp. 7–18.

[12] SWANSON, D. R., Two Medical Literatures that are Logically but not Bibliographically Connected, *Journal of the American Society for Information Science*, **38**,(1987) pp. 228–233.

[13] SWANSON, D. R., Migraine and Magnesium: Eleven Neglected Connections, *Perspectives in Biology and Medicine*, **31**, 1988, pp. 526–557.

[14] SWANSON, D. R., A Second Example of Mutually Isolated Medical Literatures Related by Implicit, Unnoticed Connections, *Journal of the American Society for Information Science*, **40**, 1998, pp. 432–435.

[15] WEBER, M. – KLEIN, H. – LOLKJE, T. W. – DE JONG-VAN DEN BERG, Using Concepts in Literature-based Discovery: Simulating Swanson's Raynauld-Fish Oil and Migrane-Magnesium Discoveries. *Journal of the American Society for Information Science and Technology*, **52/7**, 2001, pp. 548–557.