European Symposium on Computer Aided Process Engineering – 15
L. Puigjaner and A. Espuña (Editors)

Knowledge discovery method for the identification of solvents for the bio-catalytic reactions

Serguei Beliaev [a], Andrzej Kraslawski [a]

[a] Dept. of Chemical Technology, Lappeenranta University of Technology
PL 20 53851 Lappeenranta, Finland

Abstract

Searching for a new solvent of a catalyst reaction is an important task in chemical reaction engineering.
Usually a literature review is the first step of the search.
However, large data sets with high dimensionality could not be effectively explored when a problem is not well defined.
Our approach to the analysis of titles, abstracts and full text of scientific articles is to use semantic analysis to identify the new solvents.
In this article, new solvent search method was developed, which is based on knowledge discovery. It has been implemented as a software tool that can considerably speed up the search for new solvents from literature sources.

Keywords: solvent, knowledge discovery, data mining, semantic analysis

1. Introduction

One of the common research tasks in process engineering is a search for new solvents for the catalytic reaction.
Usually a literature review is the first step of the search. Scientific publications are the primary instrument for the gathering and coordination of scientific knowledge.
The amount of scientific knowledge has grown a lot during the last century.
The fast advancement of information technology has resulted in accumulation of huge amounts of data that makes analysis of these data increasingly difficult.
Manual data analysis is totally useless in many scientific areas as data volumes grow exponentially. A further disadvantage of manual retrieval and analysis is that data is often of high dimensionality.
To get over those tendencies data analysis techniques should become more flexible.
Huge data sets with high dimensionality can be effectively exploited once a problem is clearly defined and the scientist knows what to look for in the data. The problem starts when the task is not well-defined as a generation of new knowledge is not a trivial task.
We believe that knowledge discovery techniques for automated data analysis play a significant role as an interface between scientists and huge data sets. Modern computers are still far from approaching human abilities in the areas of making of new knowledge.
However, automating the data filtering and reduction procedure is an important niche suitable for computers. Once data is filtered and reduced to a suitable dimension, the

scientist can proceed to analyze it using more traditional methods, like statistics or visualization.
In this article, a method for searching of new solvents was developed, which is based on knowledge discovery. It has been implemented as a software tool that can speed up the search of new solvents.

2. Knowledge discovery in databases

Knowledge Discovery in Databases (KDD) is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data (Fayyad, 1996). At the core of KDD is the application of data mining methods for pattern recognition and discovery.
Additional steps in the KDD process, such as data selection, data preparation, data transformation, data cleaning and evaluation of the results of mining, are essential to ensure that useful knowledge is extracted from the data.
Algorithms and techniques used in the data mining step are diverse and mostly have roots in the machine learning field. Data mining can use different techniques, like classification, clustering, association, regression etc. (Hand, 1981, Jain et el., 1988, Titterington et el., 1985, Weiss et el., 1991)
Each technique can use different tools, like neural network, decision trees, and induction rules.
It is important to understand, that KDD should be used as a whole process. Indiscriminate application of Data Mining methods can be a dangerous activity easily leading to discovery of meaningless patterns.

3. Strategy of search in databases

During the past two decades, Don R.Swanson (Swanson et el. 1986, 1987, 1998) developed a different approach to creation of new scientific knowledge. He proposed that combining existing, though not connected, bibliographic information results in new knowledge. His approach could be presented as a transitivity rule (Fig. 1).
The transitivity rule states: if the fact A is related with the fact B, and fact B is related with the fact C, then fact A is also related to C and we search for this relation.

$$A \rightarrow B \text{ AND } B \rightarrow C \rightarrow A \rightarrow C$$

Figure 1. Transitivity rule as the main key in searching in databases

One publication may determine the existence of a relationship between two facts A and B while another reports on the relationship between the facts B and C. If no one has reported on the correlation between A and C, this correlation can considered to be new and interesting for scientists. The main idea in this approach is that two parts of information are not related directly: there is only a hidden knowledge and could not be identified by the conventional search with the keywords.
Once those links between A and C have been found, a connection can be made and new knowledge has been created.

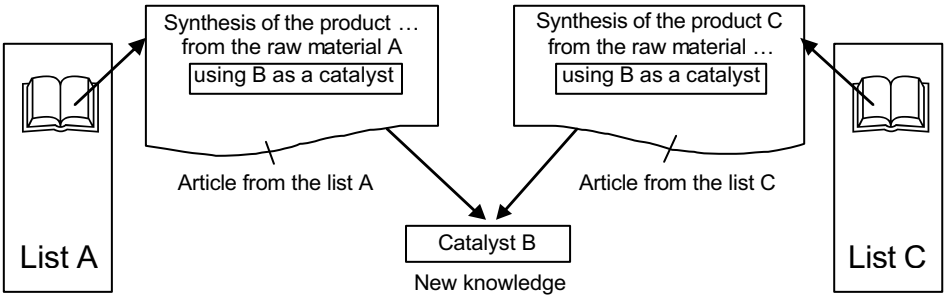

Figure 2. Searching of a new knowledge

There are two lists of articles: A, where fact A is mentioned, and C, where fact C is mentioned.

From "A" list of articles, one can find some articles, where catalyst B is used in the reaction where A is a raw material. From "C" list of articles, one can find some articles, where catalyst B is used in the reaction where C is a product (as shown in Fig. 2).

Combining this knowledge, it can be assumed that solvent B could be used in the reaction A → C. The further study of the literature on raw material A and product C shows that there was no overlap at the current moment.

4. Semantic analysis

This method provides a list of overlapping facts B (in our case - catalysts). The list of B-terms is potentially very long and filtering is needed. There are two main difficulties related with this problem.

The first one is 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 is that many substances are identified by more then one word. Finding meaningful multi-word terms in 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; analysis text using advanced NLP techniques (Swanson et el. 1998, Weber et al. 2001).

Our approach to the analysis of titles, abstracts and full text of the scientific articles is based on the use of semantic analysis.

Usually, the major part of text analysis is a routine, simple search. At different stages of the discovered process, researcher uses some semantic types to filter the text. For instance, if we are looking for production of some final products there should be in the text at least one sentence containing "production" and the name of the product. In this case, any substance in such sentence with high probability could be a product.

The typical semantic structure of sentence, describing production of some products is presented in Fig. 3.

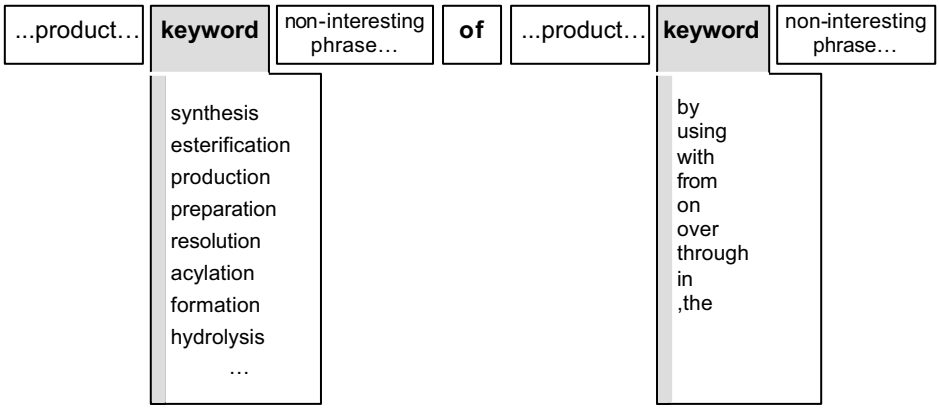

Figure 3. Typical semantic structure of the sentence, describing production of some product

The article, describing production of product C, should have at least one sentence with semantic structure from Fig.3. Only few phrases can contain a description of product, so only those could be used for comparison. Also, in the list C, at least one sentence describing the solvent should exist.

Logically, if some substance is used as a solvent, there should be at least one sentence, containing this substance and the word "solvent".

For better identification and comparison of the solvents, the sentence should be split into the small phrases using typical semantic structure in Fig. 4.

To separate those sentences into different parts, the following semantic structure could be used.

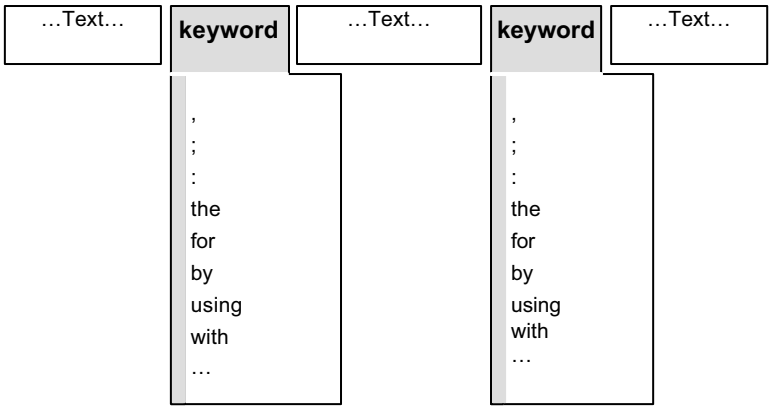

Figure 4. The semantic structure of the sentence, dividing it into small phrases using common conjunctives

Searching the raw material is a more complex task. We were not able to find the common semantic structure for identification of raw materials in the body of text. In such a situation there could be a very high probability to skip interesting substances.

All articles, containing raw material A also should have at least one sentence describing the solvent. The same applies to the list C: all sentences, containing the word "solvent"

should be split into the small phrases using typical semantic structure as in Fig. 4. for identification of solvents.
Such phrases from the list A and C, that could contain potentially interesting solvents B, should be compared between each other. In the case when any two phrases from the different lists A and C match, B becomes potentially interesting. After applying a developed method, the researcher can get a list of potentially interesting substances that could be the required solvents. The researcher can easily and quickly analyze this list.
The suggested method has been implemented as a software tool. The Elsevier database (http://www.sciencedirect.com) was used as source database.

5. Example

The proposed method could be illustrated by the following tasks: identify all new solvents that can be used for the synthesis of polysaccharides from sucrose, using Novozyme 435 as a catalyst., Fig.5.

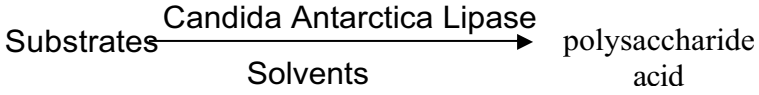

Figure 5. The initial conditions of the example

Two lists of articles are to be generated, using Elsevier database and traditional searching methods.
List A – Sucrose AND Novozyme 435
List C – polysaccharide acid AND Novozyme 435

The potentially interesting article should satisfy the following requirements:

1. There should be at least one sentence in the list C, containing semantic structure from Fig. 3 and product “polysaccharide” in the same sentence.
2. There should be at least one sentence in the list C, containing the word “catalyst”. Using the semantic structure in Fig. 4, this sentence should be split into different phrases. Those phrases could contain a name of the catalyst, so they need to be compared with the other potentially interesting phrases from the list A.
3. There should be at least one sentence from the list A, containing the word “catalyst”. Using the semantic structure in Fig. 4, this sentence should be split in different phrases and compared with the same phrases from the list C.

Researcher can look through and analyze the results of search, as shown in Fig. 6.

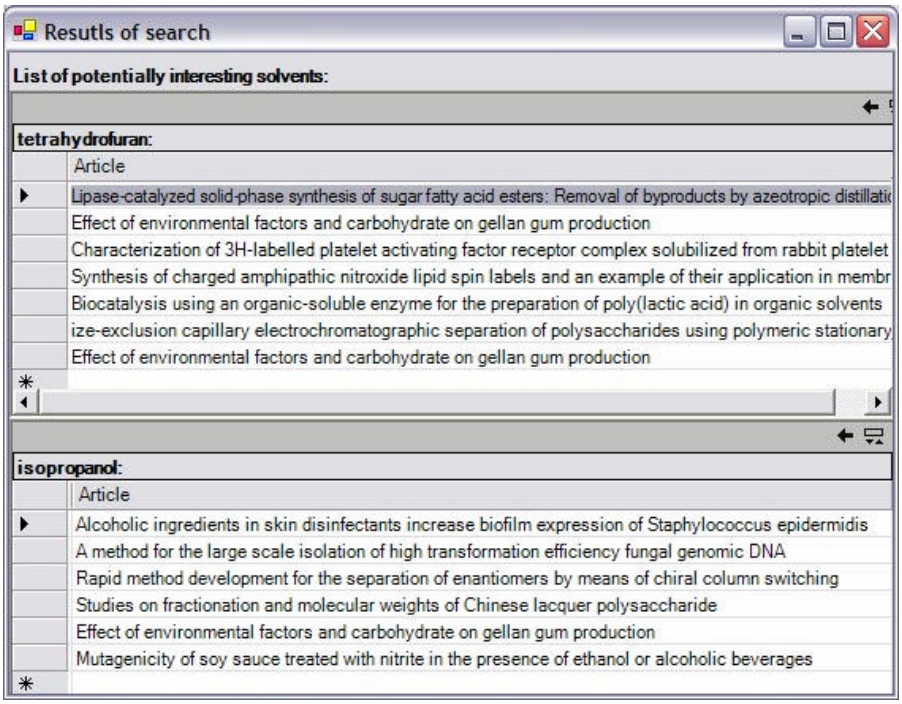

Figure 6. Screenshot with the results of search

The preliminary screening of the potentially interesting solvents allowed to identify tetrahydrofuran as a potentially interesting solvent.
In the next step, their possible applicability has to be tested experimentally.

References

Fayyad U., Piatetsky -Shapiro G., Smuth P., 1996, From Data Mining to Knowledge Discovery in Databases.

Jain, A.K., and Dubes, R.C., 1988, Algorithms for Clustering Data. Englewood Cliffs, N.J.: Prentice-Hall.

Hand, D.J.,1981, Discrimination and Classification, Chichester, U.K.: Wiley.

Swanson, D.R., 1986, Fish oil, Raynaud's syndrome, and undiscovered public knowledge, Perspectives in Biology and Medicine, 30, 7–18.

Swanson, D.R., 1987, Two medical literatures that are logically but not bibliographically connected, Journal of the American Society for Information Science, 38, 228 –233.

Smalheiser, N.R., Swanson, D.R., 1998, Using ARROWSMITH: A computer-assisted approach to formulating and assessing scientific hypotheses, Computer Methods and Programs in Biomedicine, 57, 149 –153.

Titterington, D.M., Smith, A.F.M, and Makov U.E., 1985, Statistical Analysis of Finite-Mixture Disxtibutions, Chichester, U.K.: Wiley.

Weber M., Klein H., Lolkje T.W. de Jong-van den Berg, 2001, 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):548-557

Weiss, S.I., and Kulikowski, C., 1991, Computer Systems That Learn: Classification and Prediction Methods from Statistics, Neural Networks, Machine Learning, and Expert Systems, San Francisco, Calif.: Morgan Kaufmann.