Ian David Lockhart Bogle and Michael Fairweather (Editors), Proceedings of the 22nd European Symposium on Computer Aided Process Engineering, 17 - 20 June 2012, London.

Application of semantic and lexical analysis to technology forecasting by trend analysis - thematic clusters in separation processes

Robert Sitarz,[a,b*] Andrzej Kraslawski [a,c]

[a]*Lappeenranta University of Technology, Lappeenranta, Finland*
[b] *Rzeszów University of Technology, Rzeszów, Poland*
[d] *Technical University of Lodz, Lodz, Poland*

Abstract

The development trends in a given field of research. are very important factors influencing decisions on R&D funding. The analysis of the thematic clusters of the publications allows for the identification of the structure of given field of research. The number of the scientific publications and patents as well as their change in time are the measures of dynamics of development trends. The existing approach of structuring a research field, based on clustering of the sets of high co-occurrence words, does not take into account semantic and lexical similarities between words and requires an intervention by the experts in a given discipline.

The paper presents an improved method for identification of thematic groups of papers, based on clustering of sets of high co-occurrence words, as well as application of financial analysis techniques for prediction of development trends. The method takes into account semantic and lexical similarity factors between words as well as clustering of the sets of words with the automatic rejection of the redundant groups of terms.

The material presented in this paper is limited to the journal papers. It covers publications in the field of distillation presented in ISI Web of Science database for the period 1990-2010. The identified 26 thematic clusters of research have shown the diversified patterns of development, e.g. stagnation, revival, slow development or intensive growth.

Keywords: technology forecasting, knowledge management, semantic and lexical analysis, distillation, trend analysis;

1. Introduction

Exponential growth of the number of publications and growing specialization of research are the result of the dynamic development of university and industrial R&D activities. However, increasing costs of research as well as shortening of life time of the products and processes need very careful and profound analysis of the development in a given research subject before any decision on the allocation of the R&D funds can be made (Zapata et al. 2008). A very important factor influencing decisions on R&D funding is assessment of development trends in the given branch of technology. The dynamics of the research field may be shown by the changes in the number of scientific publications and patents, indirectly illustrating the importance and potential of specific areas of research (Fabry et al. 2006). The method of structuring a given research area,

* E-mail The Corresponding Author: robs@prz.edu.pl

based on clustering of sets of high co-occurrence words, which form the seeds of thematic clusters (Sitarz et al. 2010), looks promising. However, it has some drawbacks e.g. the method does not take into account semantic and lexical similarities between words and requires an intervention by the experts.

The requirements of R&D decision makers and the lack of the effective methods and tools for the determination of existing and emerging trends in a given research field are a major motivation for the analysis of the dynamics of knowledge flow. This paper introduces an improved method for the identification of thematic clusters and the dynamics of their change for a given research subject. It applies financial analysis techniques for prediction of development trends based on the tool presented by Sitarz et al., 2010. The main improvements in comparison to the previous method are:

- Use of a semantic and lexical similarity factor
- Exclusive use of highly discriminating words in the analysis process, characterized by a high value of the tfidf (term frequency factor)
- Automated clustering of the sets of words with the rejection of the redundant sets of words

2. Method

The method proposed in this paper is presented graphically in Fig 1.

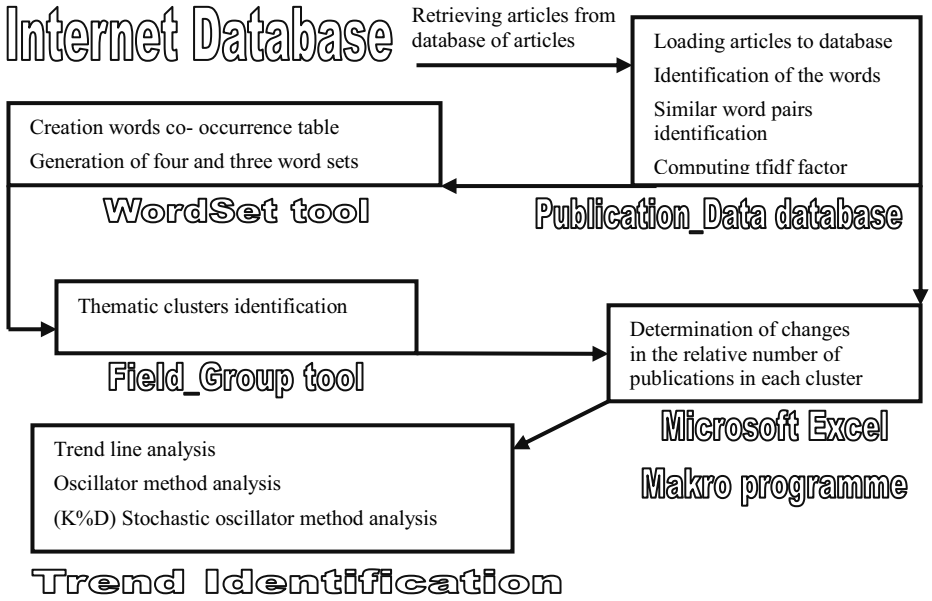

Figure. 1. Structure of the proposed method for identification of development trends in a given field of research

In the first step, the articles related to a given research field, are downloaded from ISI Web of Science database, and loaded into the "Cit" table. The database contains "wordsim" table generated from the following tables extracted from the WordNet database (Miller 1995):"lexlinkref", "linkdef", "semlinkref", "sense", "word". WordNet is a lexical database of English words in which the parts of speech are grouped into the sets of synonyms, and additionally interlinked by means of semantic and lexical relations for various types of similarity. The "Publication Data" database has a "PorterStemmer", and "EnglishStopwordFilter" functions used by the dedicated tool

Rapidminer (Mierswa et al. 2006) with a TextImput operator. The first of the above-mentioned functions identifies the core of the word, cutting the ending of the word, and the second one eliminates non-specific "stop list" words like: "do", "be", "the" etc.
In the second step, in Publication_data block, the base form of each identified word by the "PorterStemmer" function, is introduced into the "Wordlist" table as well as the relation is introduced into the "Sentrel" table denoting occurrence of a given word in a given article.
Simultaneously, in the third step, the pairs of similar words from the "Wordlist" table are generated automatically and they are introduced into the "WordlistSim". Moreover, the relations, that the words similar to the considered one exist in the article, are added to the "Sentrel" table for each considered word in the "Wordlist" table.
In the fourth step, aimed to identify "informative" words, the term-inverse document frequency factor (tfidft) is calculated for each word from the "Wordlist" table (Salton et al. 1975) from the following equation:

$$tfidf_t = \sum_{d}^{N} tf_{t,d} \cdot \log \frac{N}{df_t} \text{,where} \quad (1)$$

$tf_{t,d}$ – frequency of the word t in the document d;
N – number of the articles;
df_t – number of the articles, where the word t existed;

This factor has high values for the words appearing frequently in a small number of articles; thus, such words are good discriminators. The next block is realized using a "WordSet" tool in order to build the word co-occurrence table and finally to create highly co-occurring four- and three-word sets. The input data for this tool are the words from the "Sentrel" table satisfying two limitations: the tfidf factor of these words is greater than the A threshold value (words which are good discriminators), and the number of publications in which a considered word exists is within the range specified by the B threshold (the discriminating words should exist in a specific number of publications). The second input table is "Wordlistsim". Four-word sets are generated checking all possible combinations of the four word sets. The sets are identified in the similar way as presented Sitarz el al. 2010. In the next step, thematic clusters identification, the generated four- and three-word sets are grouped into thematic clusters based on an agglomerative clustering method (Voorhees 1986) using the "Field_Group" tool developed by the authors. At the beginning of the clustering process, each thematic cluster consists of one set of words. In each step of the clustering process, the most similar thematic cluster are merged until the most similar cluster pair factor is smaller than the E threshold value. In the next block a Microsoft Excel Makro programme developed by the authors is used to determine the yearly distribution of the relative number of papers belonging to a given thematic cluster (Sitarz et al. 2010).
The last block, trend identification, is composed of three methods used for technical analysis of financial markets (Murphy 1999). The application of the methods (trend line method, oscillator method and K%D stochastic oscillator method) was described by Sitarz et al. 2010.

3. Results

The keyword "distillation" was identified as a topic (title, keywords, abstract) in 15576 articles published between 1990 and 2010 in ISI Web of Science database, and then introduced into the "CIT" table in the "Publication_Data". In the next step, there were identified 35455 base forms of the words. For the further analysis, there were used 4664 words from the "Wordlist" table with a tfidf factor greater than 100 (A threshold) and a B threshold between 14% and 30% of the annual number of publications. There were identified 1659 four- word sets for the C =14% and D thresholds =12%. And 473 three-word sets for the thresholds 14% and 19% respectively. The "Field_Group" tool identified 108 thematic clusters for an E threshold equal to 40%, but finally 26 clusters directly related to the distillation process were chosen by the experts as directly related to chemical and process engineering.

The example of trend identification for the thematic clusters "MTBE system" is given below. The following word sets were generated for this thematic cluster, consisting of 123 articles published between 1990 and 2010:

Methyl, butyl, etherify, tert;
Methyl, etherify, tert, mtbe;
Butyl, etherify, tert, mtbe;
Methyl, butyl, etherify, mtbe;
Butyl, oxidizable, etherify, mtbe;

Fig. 2 presents the change in the relative number of papers belonging to this cluster (solid blue line) using the trend line method (solid lines I and II). It shows the trend tendency in the given periods of time, and two oscillator methods, for the parameter x equal 4 (oscillator 4) and 2 (oscillator 2).

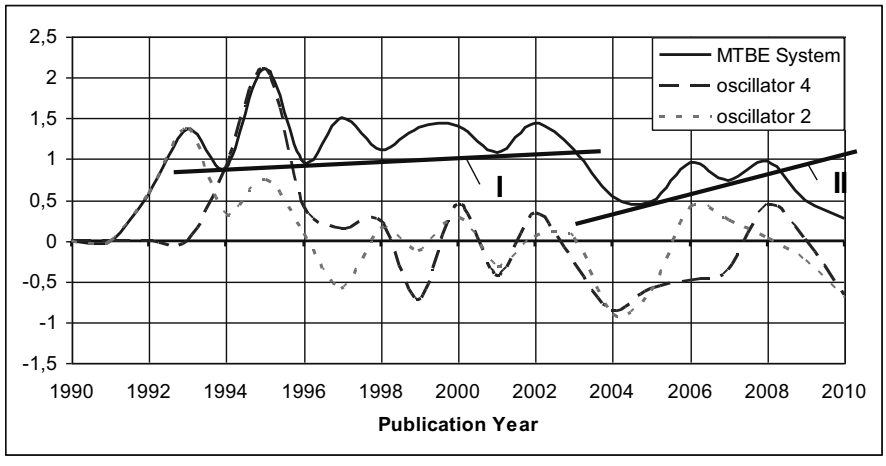

Figure 2. Cluster, "MTBE system" – relative number of publications, trend line method, and oscillator method.

The results of K%D method are shown in Fig. 3, where the solid line presents the %K for the y parameter equal to 4, and the dotted line presents %D. The analysed time period was divided into three periods. The first one, 1990-2002, is characterized by very slowly increasing trend, practically a stagnation. An additional indicator of stagnation is intersection of the zero line by the momentum lines for both oscillators. The K%D method has not provided the distinctive signals for this period. The second period, between 2002 and 2005, is characterized by a decreasing trend, illustrated as intersection of the %K and %D lines in 2002, and a second intersection in 2005. The

negative values indicate a decreasing trend. The period, after 2005, shows an increasing tendency with a signal of a future trend reversal.

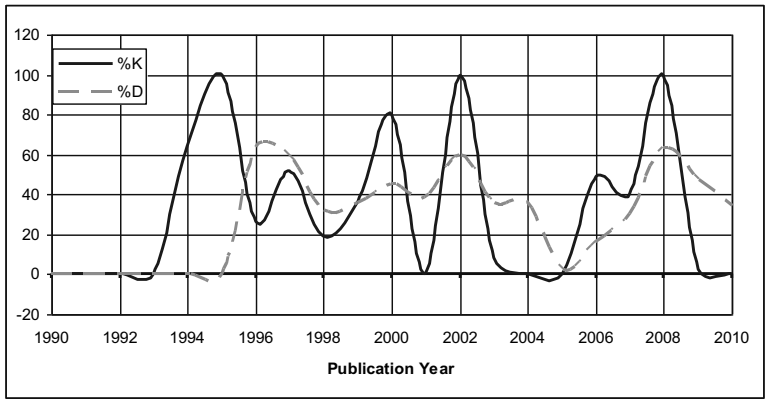

Figure 3. The cluster "MTBE system" - K%D oscillator method

4. Summary

The goal of the proposed method is improvement of identification of thematic clusters as well as their development trends The method is illustrated by the analysis of the research activity in the field of distillation. The applied semantic and lexical analysis allows for objective determination of the thematic clusters. It is realised by taking into account the synonyms of the words and in consequence, allowing for more precise allocation of the articles to the corresponding clusters. The clustering step is preformed automatically by removing the redundant word sets.

Due to the lack of space only one cluster is presented in this paper. It is worth to mention that the identified clusters showed diversified patterns of development. Some clusters, e.g. "Azeotrop curve map" or "Equilibrium in reactive distillation," show signs of decline after a revival period. The other clusters, e.g. "Membrane distillation," show an increasing trend. The examples of emerging, new fields of research are also found, e.g. "Bio diesel transestrification".

References

B. Fabry, H. Ernst, J. Langholz, M. Köster, 2006, Patent portfolio analysis as a useful tool for identifying R&D and business opportunities—an empirical application in the nutrition and health industry, World Patent Information, 28, 3, 215-225

I. Mierswa, M. Wurst, R. Klinkenberg, M. Scholz, T. Euler, 2006, YALE: Rapid Prototyping for Complex Data Mining Tasks, in Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-06)

G.A. Miller, 1995, WordNet: A Lexical Database for English, Communications of the ACM, 38, 39-41

J.J. Murphy, 1999, Technical Analysis of the Financial Markets, New York Institute of Finance

G. Salton, A. Wong, C.S. Yang, 1975, A vector space model for automatic indexing, Communications of the ACM, 18, 613–620

R. Sitarz, A. Kraslawski, J. Jezowski, 2010, Dynamics of Knowledge Flow in Research on Distillation, Computer Aided Chemical Engineering, 28, 583-588

E.M. Voorhees, 1986, Implementing agglomerative hierarchic clustering algorithms for use in document retrieval, Information Processing & Management, 22, 465-476

J.C. Zapata, V.A. Varma, G.V. Reklaitis, 2008, Impact of tactical and operational policies in the selection of a new product portfolio, Computers & Chemical Engineering, 32, 307-319