Lappeenranta
University of Technology

Robert Sitarz

IDENTIFICATION OF RESEARCH TRENDS IN THE FIELD OF SEPARATION PROCESSES.

APPLICATION OF EPIDEMIOLOGICAL MODEL, CITATION ANALYSIS, TEXT MINING, AND TECHNICAL ANALYSIS OF THE FINANCIAL MARKETS.

Thesis for the degree of Doctor of Science (Technology) to be presented with
due permission for public examination and criticism in the Auditorium 1381
at Lappeenranta University of Technology, Lappeenranta, Finland, on the
25th of October, 2013, at 14:00.

Acta Universitatis
Lappeenrantaensis 538

Supervisor	Professor Andrzej Kraslawski Department of Chemical Technology Lappeenranta University of Technology Finland
Co- Supervisor	Professor Mirosław Szukiewicz Department of Chemical and Process Engineering Rzeszów University of Technology Poland
Reviewers	Professor Jan Thullie Department of Chemistry Silesian University of Technology Poland Professor Janusz Kacprzyk Systems Research Institute Polish Academy of Sciences Poland
Opponent	Professor Janusz Kacprzyk Systems Research Institute Polish Academy of Sciences Poland
Custos	Professor Andrzej Kraslawski Department of Chemical Technology Lappeenranta University of Technology Finland

ISBN 978-952-265-483-0
ISBN 978-952-265-484-7 (PDF)
ISSN-L 1456-4491
ISSN 1456-4491
Lappeenranta University of Technology
Yliopistopaino 2013

Abstract

Robert Sitarz

Identification of research trends in the field of separation processes. Application of epidemiological model, citation analysis, text mining, and technical analysis of the financial markets.

Lappeenranta 2013
89 p.
Acta Universitasis Lappeenrantaensis 538
Diss. Lappeenranta University of Technology
ISBN: 978-952-265-483-0, ISBN 978-952-265-484-7 (PDF), ISSN-L 1456-4491, ISSN 1456-4491

Choice of industrial development options and the relevant allocation of the research funds become more and more difficult because of the increasing R&D costs and pressure for shorter development period. Forecast of the research progress is based on the analysis of the publications activity in the field of interest as well as on the dynamics of its change. Moreover, allocation of funds is hindered by exponential growth in the number of publications and patents. Thematic clusters become more and more difficult to identify, and their evolution hard to follow. The existing approaches of research field structuring and identification of its development are very limited. They do not identify the thematic clusters with adequate precision while the identified trends are often ambiguous. Therefore, there is a clear need to develop methods and tools, which are able to identify developing fields of research.

The main objective of this Thesis is to develop tools and methods helping in the identification of the promising research topics in the field of separation processes. Two structuring methods as well as three approaches for identification of the development trends have been proposed. The proposed methods have been applied to the analysis of the research on distillation and filtration.

The results show that the developed methods are universal and could be used to study of the various fields of research. The identified thematic clusters and the forecasted trends of their development have been confirmed in almost all tested cases. It proves the universality of the proposed methods. The results allow for identification of the fast-growing scientific fields as well as the topics characterized by stagnant or diminishing research activity.

Keywords: technology forecasting, text mining, knowledge management, trend analysis, distillation, filtration.

UDC 005.94:66.065/.069

Acknowledgments

First of all I would like to express my sincere gratitude to Professor Andrzej Kraslawski, Lappeenranta University of Technology, for invaluable support and substantial guidance during my PhD studies and completion of the thesis as well as for enabling the participation in scientific conferences.

I would like also to express my sincere gratitude to late Professor Jacek Jeżowski, Rzeszow University of Technology for his assistance, sound advice and meaningful guidance.

I would like also to express my sincere gratitude to Professor Mirosław Szukiewicz, Rzeszow University of Technology for invaluable advice and assistance during my work on of the dissertation.

I would also like to thank to dr Krzysztof Wałczyk for his assistance during the development of the software.

I would also like to thank the reviewers of this work, Professor Jan Thullie and Professor Janusz Kacprzyk, for their insightful comments to help make this work as good as possible.

Lappeenranta, September, 2013.
Robert Sitarz

Table of content

Chapter 1. Introduction

Some recently published papers reveal the change of the chemical and process engineering role in the public sector decision making (Clift 2006, Banares-Alcántara 2010). The engineers' role broadened from analytical problem solving to normative social and economical problem framing (Banares-Alcántara 2010). The distinction between analytical and normative role is determined by the type of decisions that are made in each case. A decision is understood as a choice from several possible ways of a problem or an issue solving (Banares-Alcántara 2010). In the analytical problem solving the criteria used to compare alternatives of the problem are available prior to the solution. Thus, the engineer shall understand the system to predict or modify its behaviour (Banares-Alcántara 2010). For example he needs to select the process or the equipment while designing the process plant, taking the economy aspects into account. Here the metrics of the decision making is known. However during the problem framing, the engineers acquire a normative role, where the norms are defined during the problem solving. In this case the criteria of success are unknown in advance and the engineers must define them within the decision making process (Banares-Alcántara 2010). Setting of the environmental standards, road planning or selection of the research field worth of investment are examples of such approach. The decisions require balancing of the techno-economic, environmental and social aspects. Including all these factors in the decision making is difficult, however such decisions are increasingly important and relate to more and more expensive high-scale projects. Therefore, engineers in general and the chemical engineers in particular must expect to be involved in this kind of decision process (Clift 2006).

The normative decision role requires a wide range of technological expertise as well as versatility in the fields like economy or management. There are many publication and patent databases being used as valuable sources of knowledge. Publications often inspire further research and lead to new discoveries. Therefore, knowledge stored in the databases ought to be well structured, organized and easily recognisable to facilitate its further growth. Identification of the valuable and desirable documents from the records gathered in the database is a very difficult task. Searching techniques based on the identification of the keywords in the documents' content are not efficient because, apart of the desired titles, they return a great number of false positives. which must be filtered out with considerable waste of time. In other words, information stored in the databases is poorly accessible. Because of the

exponential growth of the number of published articles and patents the searches become ever more laborious. Therefore new methods and tools for rapid and efficient filtration of the required information are in demand.

Consequently, the KDD techniques ("knowledge discovery in database") have been developed. The techniques manage and process the content to the more compact and useable form, which facilitate easier identification of the required information and further new knowledge discovery (Fayyad 1996a, Fayyad 1996b). The main objective of the KDD techniques is identification the potential Problem – Solution pairs, which highlight the issues worth of interest (He 1999). Such KDD process is effected by the text analysis shown in the figure 1.1.

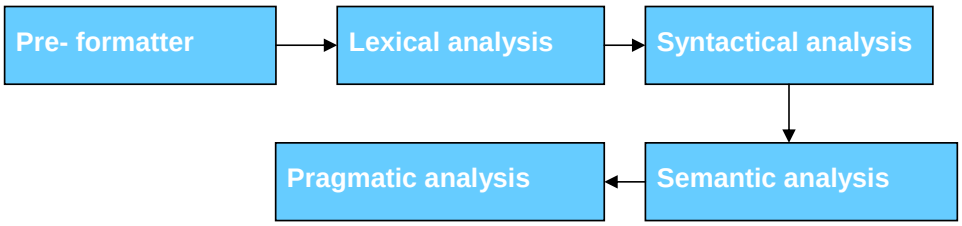


Fig. 1.1. Text analysis diagram of the KDD process.

After the text analysis process is finished, facts (informative phrases or sentences) are identified in the content of the analysed documents. The identified facts are assigned to Problem – Action – Object groups. First, the problem is identified and then are the action paths leading to a solution. The next step of KDD process discovers new Problem- Action- Object groups looking for the facts that repeat in different documents. Finally, the documents where the interesting Problem- Action- Objects groups occur, are filtered out. One of the most popular KDD tools is the ARROWSMITH (Smalheiser and Swanson 1998). Discovery of the relation between migraine and a magnesium deficiency, by means of the medical literature analysis is a sample of its efficiency.

Discovery knowledge hidden in the databases opens new fields of research, but prediction of success of such emerging research fields as well as making decision of funding such studies are very difficult. Forecasting progress on the developing research field seems to be less risky than the examination of newly discovered topics. For the known research fields it is possible to identify the research trends that allow forecasting of their future characteristics. The development trends of the research are a very important for building of a portfolio of the

research projects (Evans et al. 2009, Klapka et al. 2002, Linton et al. 2002, Rajapakse et al. 2005). The projects in the portfolio can be divided into radical and progressive (Chu et al. 1996, Ringuest et al. 2004). The radical projects offer high returns at the tradeoff of increased risk. The progressive projects bring smaller but more secure profits. The portfolio of the research projects should balance profit and risk. Increasing costs of research and the falling products' and processes' lifetime make the selection of worthy research projects and allocation of funds strategically important. Scientific research are more and more expensive because of high-priced equipment as well as highly qualified scientists, so only profitable and developing research field should be researched. New technologies provide for shortened cycle of the research to industry cycle. This requires a careful and broad analysis of the current research trends (Zapata et al. 2008, Rajapakse et al. 2005). Development of the research field is forecasted upon the analysis of past publication activity and on dynamics of its change. The number of publications and patents is a good measure of the research intensity (Fabry et al. 2006, Tsai 2011). However the change of the number of publications and patents over time must be taken into account in forecasting of the future developments. The decision of research funding is also difficult because of the exponential growth of number of publications. It makes the structuring of research and technological knowledge and analysis of its dynamics not a trivial task. It shows the need to develop methods and tools allowing on identification of potentially developing directions of research. Rapidly changing research trends entail that these tools must work quickly and efficiently. Known methods of trend detection and forecasting approximate of the time distributed number of publication by the exponential equation (Krampen et al. 2011, Lv Peng Hui et al. 2011). However, assumption that the number of publications represent the development status of the thematic field is dubious, because the exponentially growing volume of publications makes it increasingly difficult to filter those qualifying and identify the true trends. Therefore topics worth the investment, for which the development is expected are difficult to identification.

Before the trends are detected, the scanned documents must be assigned to the thematic clusters, by a structuring process, that is not effected with necessary precision. There are three main groups of structuring techniques reported in the literature:

- co- classification analysis;
- co- citation analysis;
- co- word analysis.

The first group of methods executes the clustering of papers on the basis of the journals classification, in which the analysed papers were published (Krampen et al. 2011, Lv et al. 2011). Here we get an overview of the analysed research field rather than its detailed structure. The second structuring groups of methods bases on the co-citation coefficient as a measure of the similarity between clustered documents (Barnett et al. 2011; Hung and Wang 2010; Li et al. 2007). Here, the identified clusters often are not thematically coherent and address many issues. Also the hybrid documents clustering methods based on the co-citation and co-word analysis are not precise enough (Braam et al. 1991a, Braam et al. 1991b, Jansen et al. 2006, Marshakova- Shaikevich 2005). The use of this methods result in generation of clusters containing the groups of the words belonging simultaneously to several clusters. It makes the identification of the main theme of the cluster a challenging task. The third group builds maps of knowledge, i.e. graphic representations of words shared by the analysed documents which discovers word-groups whose members are often found together. These methods offer a good representation of the structure of the analysed research field, but assignment of a paper to clusters is complicated. The number of the reference words to be found in the paper to qualify it to a cluster is not predetermined (Noyons et al. 1994, Noyons 2001, Peters and Van Raan 1993, Tseng et al. 2010, van Raan 1993). Dynamic of the research in this approach is determined by comparison the time periods on the maps of knowledge (Callon et al. 1991, Ding et al. 2001). However, this approach offers only rough differentiation between the examined time periods, and not a detailed description of development trend.

The mentioned trend identification methods as well as structuring methods are wider and more detailed described in the further chapter.

The drawbacks of the existing trend identification approaches as well as the expectations signalled by the R&D decision makers provide the rationale for the research presented in the dissertation. Especially important is the limited availability of the effective methods and tools for identification of the developing research fields. Moreover, rapidly-changing research trends as well as shortening period between generating an ideas and its industrial implementation justify the relevant tools' development attempts (Hull et al. 2001).

Chapter 2. Literature overview

Identification of the investment worthy research fields, consists of two steps. First, the thematic structure of the analysed research area is built through the identification of the particular thematic clusters (branches of research field). Then, the research trends are identified for each of the clusters.

In the first part of this chapter, the structuring processes are discussed and explained. Next, the known clustering methods, and the proximity metrics are discussed. Finally, discussed are the most popular techniques of the thematic clusters identification in the scientific content databases and the research trends detection techniques.

2.1 Structuring of the research area

The main goal of the structuring process is the identification of the specific branches of knowledge within the analysed research field. The branches of knowledge (thematic clusters) are narrow and thematic coherent research areas. The generated tree structure of the thematic clusters has its trunk representing the analysed research area while the branches represent the thematic clusters. The children offshoots represent more detailed thematic clusters. Then, we come to the leaves representing the individual structured elements. It is very important to assume in advance the required granularity i.e. how fine result structure is expected to be. In the other words, it is needed the specification of the desired layer in the tree structure, which is considered as a result of structuring. It is important, because the process shall pick both the well developed branches of knowledge, and narrow, emerging and less developed branches. Such identification of the particular branches of knowledge is achieved by clustering the similar (close) elements of the knowledge.

2.1.1 Structuring techniques

As indicated, the structuring is effected by clustering of the elements of knowledge. Before the structuring techniques are described, it seems necessary to explain the clustering process, the applied similarity metrics between the group elements and review the most popular clustering methods.

"Clustering" refers to assigning of similar objects to classes (clusters). The point is to maximise the similarity within the cluster while keeping them different from the items outside and divide the data into relatively consistent groups (Larose 2006). To cluster the elements we must evaluate the similarity between objects.

There is many of similarity metrics known in literature. The most common is Euclidean distance, defined by 2.1 (Broda 2007):

$$D_{Euclidean}(x, y) = \sqrt{\sum_i (x_i - y_i)^2} \qquad 2.1$$

where x= x= x_1, x_2, …, x_m and y= y_1, y_2, …, y_m represent values of m chosen attributes of the elements x and y.

The other often used similarity metrics are Manhattan distance and Minkowski distance defined in 2.2 and 2.3 (Larose 2006).

$$D_{Manhattan}(x, y) = \sum_i |x_i - y_i| \qquad 2.2$$

$$D_{Minkowski}(x, y) = \sqrt[q]{\sum_i |x_i - y_i|^q} \qquad 2.3$$

Despite of the popularity, the metrics do not work well in proximity evaluation between text documents, since papers vary in the number of words (Broda 2007). In the case of text documents, a vector space model is used, where each document is considered as a, so called, bag of words. In this approach a vector built of words is generated for each document. Vector space model will be described in detail in the section 2.1.1.2, where the structuring methods based on lexical analysis is discussed.

Some of the proximity metrics applied to text documents attempt to measure correspondence of the co-occurring words (Broda 2007). Other used are Dice coefficient (2.4) and Jaccard coefficient (2.5).

$$D_{Dice}(x, y) = 2\frac{\sum_i x_i y_i}{\sum_i x_i^2 \cdot \sum_i y_i^2} \qquad 2.4$$

$$D_{Jaccard}(x, y) = \frac{\sum_i x_i y_i}{\sum_i x_i^2 + \sum_i y_i^2 - \sum_i x_i y_i} \qquad 2.5$$

The value $\Sigma x_i y_i$ in the above equations, which denotes the sum of conjunction, is considered as sum corresponding attributes with the same value. So in the case of text documents value $\Sigma x_i y_i$ denotes number of co-occurred words between documents x and y. Another widely used method is the cosine of the angle between the words vectors of the compared documents (Broda 2007).

$$D_{Coscine}(x,y) = \frac{\sum_i x_i y_i}{\sqrt{\sum_i x_i^2 \cdot \sum_i y_i^2}} \qquad 2.6$$

For optimum performance, clustering algorithms require data normalization or data standardization (Larose 2006).

data normalization $X^* = \frac{X - \min(X)}{range(X)}$ 2.7

data standardization $X^* = \frac{X - average(X)}{\sigma(X)}$ 2.8

where σ - standard deviation of the variable X.

Applying any of the above mentioned metrics, we are able to evaluate the similarity between a pair of the clustered objects. However, the question arises how to evaluate similarity between groups? Basically, there are three methods reported in the literature, for evaluating the similarity between groups of elements (Larose 2006).

- single linkage approach;
- complete linkage approach;
- average linkage approach.

The first approaches consider maximum similarity between any record of the group A and any record of the group B as the similarity between groups A and B. Thus, the similarity between two clusters is defined as the proximity between two most similar elements in each group. That approach has a drawback, visible in the following example. When the elements are represented by three dimensional vectors, the dominant contribution of two dimensions leads to creation of long and thin groups, which are not perceived as similar (Larose 2006).

The second of the presented approaches is characterized by minimum proximity between any record of the group A and any record of the group B. In other words, the

proximity between two clusters is defined as the similarity between two least similar objects (one from each group). In this case, the elements arranged in a three-dimensional vector space, tend to form a compact, ball like groups (Larose 2006).

The last of the techniques was discovered to eliminate the influence of the extreme values, such as the closest and most distant elements, on the criterion of merging groups. In the case, the proximity between two groups is evaluated as the average of similarities of all elements of the group A to all elements of the group B. Such obtained clusters tend to be almost uniform variation within the groups (Larose 2006). However, the approach is significantly more time consuming, so the relevant clustering algorithms take more execution time as compared with the previous methods.

2.1.1.1. Clustering methods

In the section the clustering methods are presented. The best known methods are listed below.

- hierarchical clustering (Forster 2006, Manning et al. 2007);
- k-mean algorithm (Manning and Schutze 2001, Likas et al. 2005, Mazur 2005);
- density-based clustering (Ester et al. 1996, Sander et al. 1998, Xu et al. 1998, Ankerst et al. 1999);
- graph-theory techniques (Schenker et al. 2005);
- self-organised Kohonen's maps (Kohonen 1982, Kohonen et al. 2000);
- artificial neural networks (Haykin 1990, Osowski 1996, Ahmed 2004, Bortiz 1995).

The first two of the listed methods are the most common, therefore are discussed below. First of the presented methods refers to hierarchical clustering approach, while the second one (as shown in the example of k-means algorithm) is not hierarchical. The hierarchical method results in the dendrogram (a tree structure). The non-hierarchical approaches assign the objects into n declared clusters but there is no tree path of the clustering process.

Hierarchical clustering

The hierarchical clustering generates a tree structure (dendrogram) by the recursive division (distribution methods) or connection (agglomerative methods) of the existing groups (Larose 2006). In the agglomerative clustering, an initial group (cluster) consists of a single

element. Then, in each step of the process two most similar groups are merged. This way, the number of groups decrease by one in each step. The process continues until the proximity of the most similar clusters is lower than a threshold value (Larose 2006).

The hierarchical splitting methods start with all the records in one cluster. Then, at each step the most incoherent cluster is being split. The process terminates, when the proximity of the most dissimilar groups exceeds the threshold value (Larose 2006).

Very often, the clustering process is carried out to the end, i.e. in case of agglomeration methods to the last shared group, or in case of distribution methods, until each group consists of a single element. In this case, a complete tree structure is obtained, from which the user selects the appropriate resulting clustering level. In the case, the user does not have to declare the threshold value in advance. An example dendrogram structure has been shown in the figure 2.1. The numbers on top index the stages of the clustering process. Setting the right threshold value, or selection of the appropriate clustering layer is a major dilemma. This setting should enable identification of both large, highly developed fields of research, as well as the emerging, not yet developed ones.

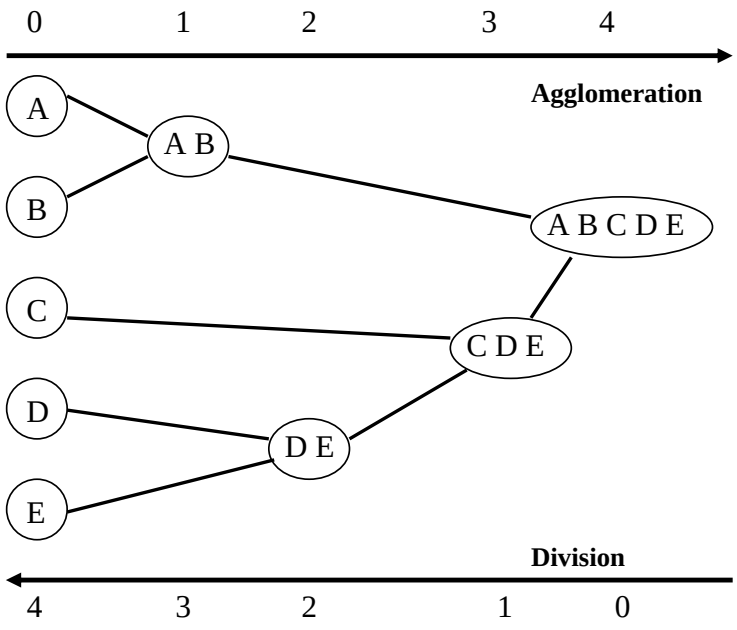


Figure 2.1. Diagram of hierarchical clustering methods.

K-means algorithm.

K-means algorithm is a simple and effective clustering technique. The advantage of the method is clustering the set of elements into k groups declared by the user. The process runs as follows: Initially k elements are selected as the initial means of groups. The choice is random, but there are extended algorithms that start with the means selected by the user. The extensions involve additional procedures, tools, or hierarchical clustering method that select the initial means' set.

Next, each of the objects is assigned to the group, whose mean is closest to this object under the metrics applied. Subsequently, the centroid of each of the k clusters is computed and the process assignment to the cluster of the closest mean is repeated. A centroid of each group is a hypothetical element placed in the cluster's "centre of gravity". The process is repeated until convergence, i.e. when the following repetitions do not change the assignment of the elements to the clusters (Larose 2006).

A modified technique is known in the literature as the k-medoid algorithm (Broda 2007). In this case (as in the original approach) k elements are selected, as the initial means clusters and the elements are assigned to the cluster having the nearest mean. But after the centroids are computed. the real elements closest to the centroids are identified. These elements are considered as medians of the clusters. In the next cycle, the elements are reassigned to the cluster having the nearest median . The process continues until convergence.

A weak point of these methods (both k-means and k-medoid) is pre-selection of the number of clusters. The chosen number may not be optimal from the analysis point of view, since some clusters should rather be merged (as would be the case with a smaller k), while others split (greater k). Finding of the optimal k is tedious and time consuming.

2.1.1.2. Structuring methods.

The structuring techniques vary by choice of the evaluated attributes and the similarity metrics applied. The rough classification goes as follows:

- co-classification methods (Amini and Goutte 2010);
- citation and co-citation methods (Cottrill et al. 2010);
- co-word methods (Ding et al. 2001, Yang et al. 2012).

The first group of methods is based on the classification of scientific papers (publications or patents) in the group, through taking into account attributes of the documents like: journal name, branch, journal profile, language, etc.. Documents are assigned to a group when they share some of the attributes.

The second group of clustering methods aggregate the documents into clusters applying a similarity metrics basing on the number of citations and co-citations. In the third group of methods, words are the attributes, on the basis of which the similarity metrics is determined. There are two approaches within this group of methods. The first clusters scientific documents taking into account the number of co-occurring words. In the second approach, words identified in the body of documents are clustered and the proximity between words is calculated on the basis of the number of documents containing them.

Co-classification methods

As it was mentioned, the group of methods base on the aggregation of the scientific documents through the classification of the journals, where the papers were published (Krampen et al. 2011, Lv et al. 2011, Amini and Goutte 2010). Publication in the approach are the clustered unit elements of the knowledge however the similarity between documents is evaluated on the basis of common attributes such as the name of the journal, the profile of the journal, language, etc. The choice of the attribute depends on the aim of the classification. If the purpose of classification is to identify the journals, where the articles are mostly published in the analyzed area of knowledge, then the magazine's name is chosen as the attribute of classification. However, if the purpose of the analysis is the identification of publication language, in which the articles are mostly published and which countries are key considered, then the language. is chosen as the classification attribute. Another one type of classification is realized in order to identify the subjects corresponding to the profiles of the journals. Subjects of such obtained clusters correspond to the journal profiles. The methods have a serious drawbacks.

Identified clusters in the group of method are not thematically coherent. In addition, the approach provides only a general overview of the analyzed research field. Obtained clusters are very often too general indicating only a wide, well-developed thematic clusters related to journal profiles. The methods do not identify newly formed research fields which have not been considered as individual branches of knowledge yet. However, the advantage of the approach is the ability to easily present the time-distribution of publications for obtained

thematic clusters, what is crucial for the identification of the research trends. Moreover such methods work very quickly, so does not required a significant computational effort.

Citation and co-citation methods

The methods of this group analyse which papers are cited by the processed publications and clusters the citing papers accordingly. The publications are being clustered quite like with the previous group of methods, only the similarity metric is different. There similarity is assessed by any of the following ways:

- bibliographic coupling,
- co-citations.

The fig. 2.2 shows the structure of bibliographic coupling. The ovals on the drawing represent the publications while the arrows represent the citation: F→A reads F cites A. Here, F and G are coupled because they cite the same five papers A–E. Similarity grows with the number of shared cited references. However, even if documents cite the same papers, their content is not necessarily similar.

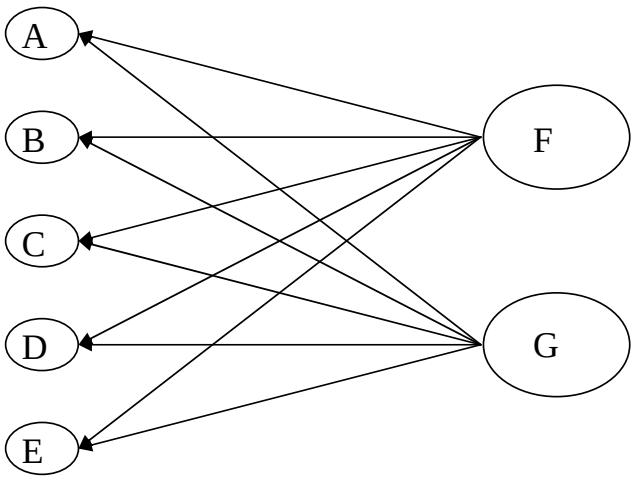


Fig. 2.2 Bibliographic coupling.

The other approach shown on the fig. 2.3 is analysis of co-citations. Here the relation is reversed. F and G are similar if there are publications where both F and G are cited. (co-citation). The similarity grows with the growing number of co-citing papers. The co-cited

documents more often deal with similar subjects than those which refer to the same documents (coupled).

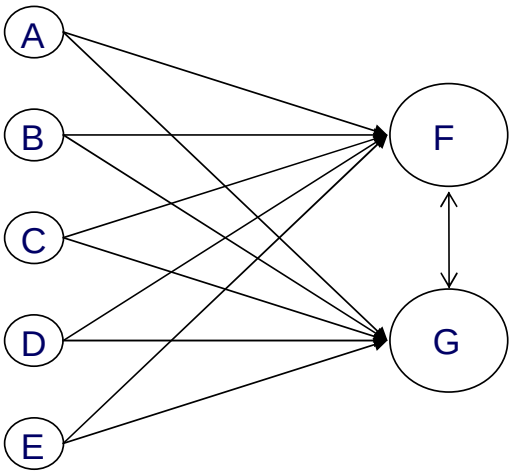
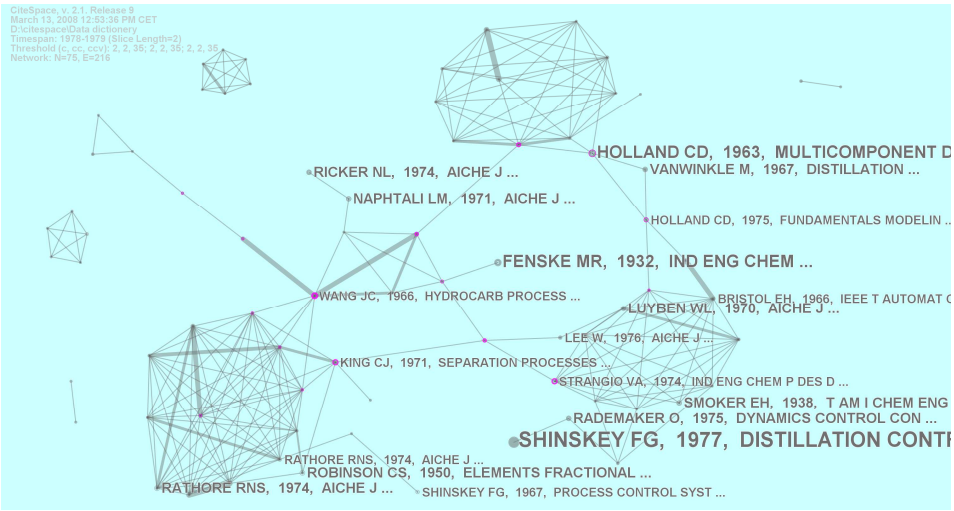


Fig 2.3 Co-citation

Having the similarity metric, we can discuss the clustering. Here the map of knowledge is built of the publications related by the similarity relation. Greater similarity is mapped with bolder lines on the map of knowledge. The more citations has the paper, the larger is the shape representing it. The fig. 2.4 shows a network built by the CiteSpace tool (Chen 2006).



Fir. 2.4 Co-citation network for the subjects of distillation

The clusters can be composed in many ways. Any clustering algorithms discussed can be applied (e.g. hierarchical or k-means). When dealing with the co-citations, immaterial connections can be cut, thus isolating the clusters (Granovetter 1973, Schildt 2006). The connection is recognised as immaterial by application of the triangle inequality (Chen 2006).

$$W_{ij} \leq \left(\sum_{k} W^{r}_{N_k N_{k-1}} \right)^{1/r} \qquad 2.9$$

W_{ij} denotes the similarity metric between items i and j, while W_{NkNk-1} is the similarity value of the items N_k and N_{k-1}, for k=1,2,…,m. Therefore it is an indirect path between the items i and j that goes through the items N_1, N_2,…, N_k, as long as these belong to the same cluster. If the W_{ij} is greater than the weight of the alternative indirect link, it violates the triangle inequality and such a link is removed (Chen 1998). There are two more parameters involved in this method: r and q. The r parameter defines the Minkowski metric space for the network, which becomes Euclidean for r=2. The q parameter limits the number of links in the triangle inequality test, where the combined weight of the path between i and j is evaluated. This parameter fits within 2 <= q <=M-1, M refers to the number of all items in the network (Chen 2003).

Citation analysis is a popular method use for structuring of various knowledge domains (Barnett et al. 2011; Hung and Wang 2010; Li et al.. 2007; Newman 2005; Price 1965; Small and Griffith 1974, Garfield 1994). Diagrams depicting the citations networks help understand how scientific and technology knowledge is organised and how it evolves in time (Wasserman 2003, Scot 2000, Yi et al. 2012). Limitations of this method derive from the perception of the document as an unit individual. Although each publication differs from others, it is rarely monothematic and usually addresses many issues (McRoberts 1996, Garfield 1985). This means that scientific papers and patents are always linked to others and cannot be perceived as unit elements of knowledge (Lee et al. 2010, Yi et al. 2012). Thus, this method also builds thematic clusters that are not coherent, quite like the co-classification methods described earlier. The resulting clusters derived from the bibliographic coupling and co-citation often address many subjects and naming of the core subject is difficult. Nonetheless, the methods basing on co-citations analysis are more precise than those relying on co-classification.

Co-word method

The third group of methods analyses how specific words co-occur (co-words). This means that a cluster is defined by the words' bundles found in the analysed documents (Courtial 1994, Yoon 2004, Yoon 2008). These methods treat each document as a ”bag of words” and assume that the words found in the scientific paper reflect its content (Larsen 2008, Cui 2008). This is the so called vector space model. Here the vectors of words found in the analysed documents are built, or matrix of such vectors. Some of these base on the preselected keywords only, so it is easy to built the vectors. Some other methods identify the words in the titles and abstracts. Eventually there are methods fishing the words in the full text. All need to select informative words that are good discriminators. These informative words vectors known as fingerprints are built in statistical processing. One of the most often used is the term frequency inverse document frequency factor (TFIDF), computed for each word with the formula 2.10 (Salton et al. 1975, Salton 1988, Manning et al. 2007):

$$tfidf_t = \sum_d^N tf_{t,d} \cdot \log \frac{N}{df_t} \text{, where} \qquad 2.10$$

$tf_{t,d}$ - number of occurrences of the word t in the document d;
N - number of document;
df_t - number of papers where the word t occurs;

This factor becomes high for words that are numerous in small number of documents. It is small when the word t is either rare or ubiquitous. Therefore the higher the TFIDF is the more discriminatory words select. Here we come to the two main subgroups of the co- words methods.

- text documents clustering;
- maps of knowledge visualisation.

The first approach clusters the text documents and the similarity is computed on the basis of the words shared (Forster 2006, Dittenbach et al. 2000, Goldszmidt and Sahami 1998, Musgrove 2003). Any of the discussed metrics is suitable (e.g. Euclidean, Jaccard's. cosine etc.). There are also hybrid methods where the thematic clusters are determined by combination of the citation method and lexical analysis (Braam et al. 1991, Jansen et al. 2006, Marshakova- Shaikevich 2005). All these clustering methods (both hybrid and lexical) suffer from severe limitations. The fingerprints of the clusters include words' groups also present in different clusters. This creates difficulties in identification of the main subject of the created

clusters. Furthermore, as it is the case with the previous structuring methods, the documents are perceived as unit elements, while in fact they are linked elements of knowledge (Lee et al. 2010, Yi et al. 2012).

The second subgroup has the knowledge maps drawn. This approach depicts the structure of the knowledge domain as a network of the co-occured words (Borner 2003, Su 2010). The words are identified with application of the discussed vector space model. The networks are built in quite the same manner as they were in the co-citations methods. The individual words are the node elements of the network while the similarity is computed by analysis of the words shared in the documents across the cluster. There were many knowledge domains depicted this way, e.g.: biochemistry (Rip and Courtial 1984), chemistry (Callon et al. 1991), neural networks (Van Raan and Tijssen 1993; Noyons and Van Raan 1998), biological safety (Cambrosio et al. 1993), optomechatronics (Noyons and van Raan 1994), adverse drug reactions (Clarke et al. 2007), software engineering (Coulter et al.1998) etc. (Lee et al. 2010). The methods do not have defined the number of words which must be shared between the document and the fingerprint in order to assign the document to the cluster (Noyons et al. 1994, Noyons 2001, Peters et al. 1993, Tseng et al. 2010, Van Raan 1993). This hampers attempts to detect the changes in the publishing activity for the clusters under scrutiny and prevents further application of this technique for the purpose research trends detection. Furthermore, the keywords marking the publication are often ambiguous, unclear and plainly incorrect (Wang et al. 2012), and certain clusters cannot be identified accurately.

2.2. Detection of the research trends

Detection of the research trends is a relatively young subject, but some methods facilitating such analyses have already been devised. They usually fit an exponential curve to the number of the papers published over time (Lv et al. 2011). Other methods apply nonlinear regression fitting polynomials or inversed exponential curves (Baskurt 2011). Other apply two years moving average (Hossain et al. 2012), or smoothed trajectory of the number of publications (Krampen et al. 2011). However, presentation of the publication activity as the change in the number of publications over time is inadequate, because this value grows exponentially. Detection of the research trends is therefore quite difficult.

In the case of the co-citation networks there are two approaches to presentation of the clusters' evolution. First approach attempts to assess differences between the networks

applicable to different time periods (Ebel 2002). This is however strenuous and ineffective, and only discovers the differences between the networks, but does not provide information about specific trends. Other approach compares the number of citations and co-citations over time within the defined clusters (Chen 2006). Comparison of these data shows time evolution of the clusters, but as with the case of the number of publications, the data volume grows over time, and therefore the method fails to depict adequately the development of the research topics. The dynamics of the development is also determined by comparison of the visualised maps of the shared words for several time periods (Callon et al. 1991, Ding et al. 2001, Hou 2006). However, like in the case of co-citation networks, this approach gives only rough insight in the time evolution, but no specific trends identified by the well defined parameters are being discovered.

2.3. Background summary

Identification of the subjects interesting as a target of investment in research has two stages: structuring of the domain, and identification of the trends within the clusters found. Because of the practicalities, combination of the structuring with the trend detection emerges as a major issue in identification of the promising research subjects, There are two kinds of solutions:

- the structuring based on the publications' clustering with trend detection based in time evolution of the number of publications;
- the structuring based on the visualisation of the maps of knowledge with trend detection based in time evolution of the maps.

The first approach provides for easy presentation of the number of publications over time for the given thematic clusters what facilitates application of the trend detection methods. However, the structuring does not bring precise results here, since the identified clusters are not coherent. Moreover, the documents are perceived as atomic elements of knowledge while in fact they are a collections of the elements of knowledge. This means that clustering of the documents does not lead to precise identification of the research subjects. A clustering based on co- classification discloses only the general thematic profile matching the already known and broad research topics. The co-citations analysis is imperfect too, since the linked publications not always deal with the similar research subjects. Despite the limitations,

these methods prove more precise than those relying on co- classification. The third type of clustering analyses the words shared by the documents and the resulting thematic clusters are more coherent than those resulting from co-citations. The methods basing on analysis of the shared words are not perfect since they do not provide for identification of the subject (topics) of the generated clusters. Furthermore, like with the case of the earlier discussed methods, the document can be assigned to one cluster only, what assumes monothematic character of each publication. Assessment of the research subject development relies on time evolution of the distribution of either citations and co-citation or the number of publications. Since these values grow exponentially in time, this approach does not adequately reflect the development of the analysed research subjects.

Shortly: this group of methods is imprecise in identification of the thematic clusters, but offers easy assessment of the time development of the clusters, thus facilitating trends detection.

The second groups of the structuring methods differ considerably. Structuring relying on visual maps of knowledge seems much more precise than any other known method. Here, individual words become the atomic elements of knowledge and this allows for generation of coherent clusters. In this case, assignment of the documents to the specific clusters becomes more difficult. It hampers attempts to determine the development dynamics of the clusters, and application of the trend detection methods based on the time distribution of the number of publications. The trends are assessed by comparison of the maps compiled for different periods. This is not satisfactory as there are no quantitative methods to identify the development trends.

Assessment of these two methods leads to the following conclusions: From the point of view of structuring, visualisation of the knowledge maps discerns the thematic clusters in the most price way. But words, adopted as the atomic elements of knowledge are sometimes ambiguous and misapplied. From the point of view of the trends detection, the first group of the combined methods proves superior since it facilitates visualisation of the clusters' dynamics and detection of the development trends. As indicated above. number of publications provides for a poor depiction of the development dynamics since these numbers are subject to exponential growth which impedes correct detection of the trends. Some compromise between these approaches seems to be the best solution.

The provided review of the known techniques indicates need of methods and tools facilitating both adequate structuring with easy assignment of the documents to the clusters, and easy detection of the present and future trends in research. These methods should perceive

the groups of the shared words as atomic elements of knowledge and allow for occurrence of the specific words in more than one thematic cluster. This would solve the problem of diverse usage of the words, a shortcoming of the knowledge mapping methods. Furthermore these methods should use techniques facilitating easy identification of the subject of the generated clusters and clear assignment of the documents to the clusters. The tools should detect the documents loosely connected to the analysed research domain and allow for assignment of the documents to more than one cluster. Inclusion of the synonym recognition would allow for even more adequate definition of clusters. Furthermore, the development of the identified thematic clusters should be represented by variables representing values different than the number of publications, to facilitate easier and more reliable detection of trends, Except for that, the methods detecting the future research trends are necessary for those deciding about the research trend funding, Also important seems to be propose a mathematical model, which allow on both trend identification and forecasting, which would not be only a empirical regression. This model shall explain time evolution of the thematic development and provide for far reaching conclusions. Also, the citation and co-citation analysis should apply more subtle metric than that based on the number of citations only. Such approach would grasp the utilisation of knowledge on top of the "knowledge production" reflected by the number of publications.

Chapter 3. Aim and objectives

The aim of the research was creation of the tools and methods supporting the decision making process in selection of the promising research topics relating to the separation processes. The designed methods and tools have been confronted with the issues specified below:

- definition of the thematic structure of the knowledge domain;
- presentation of the development history of the identified thematic clusters by means of the historical research trends;
- forecast of the future research trends for the identified thematic clusters.

The following hypotheses have been adopted:

- past to date development trends of the knowledge domain, provide data for prediction of the development trends;
- modelling of the development dynamics of a given knowledge domain is possible.

The knowledge domain where the experimental validation of the devised tools have been carried out is distillation. This subject has been chosen because it is perceived as a well developed subject whose current progress is slow, and it consists of many thematic branches. The results were also validated for filtration. Filtration is a well developed and still growing knowledge domain consisting of many branches. It was chosen to demonstrate usability of the developed techniques to separation processes other than distillation. The other important motive for choice of these subjects was practical availability of experts to discuss the results.

The structure of the dissertation is the following: Chapter 4 presents two designs of the structuring methods and three designs of trend detection.

The first proposed structuring method identifies thematic clusters without use of semantic analysis. The words are perceived as character strings an no meaning is analysed, while the thematic clusters are identified by the groups of words shared across the cluster. On the other hand, the words shared define some shared content, therefore, in certain sense, the meaning of words is taken into account. Absence of “semantic analysis” means that the synonyms will not be recognised in the process. The proposed method builds the clusters through grouping high co-occurred word sets. The sets consist of three or four words that often coexist within the thematic cluster and are relatively rare beyond. The analysis extracts

only the meaningful words from the abstracts, that fit in the fingerprints of the analysed papers.

The other structuring method is almost the same as compared with the previous method, except for some extra improvements. At first it has the ability to recognise and use the synonyms. Furthermore, the collections of words are clustered automatically by means of hierarchical agglomerative method. Moreover, the fingerprints of the analysed publications are built by application of the TFIDF, as a measure of the discriminating capabilities of the individual words.

The second part of the chapter 4 presents three methods of trend detection. The first applies the techniques used in technical analysis of the financial markets. They facilitate detection of present and forecasting of future research trends. Publishing activity of each thematic cluster is represented by the relative number of publications (RNP). This technique derives from the observation that the time distribution of the RNP oscillates. Therefore, by analogy to the financial markets where the stock quotes oscillate, the methods have been borrowed. These techniques (trend line, a simple oscillator (OSC) and a stochastic oscillator (K%D)) forecast future research trends basing on the past publishing activity in the analysed thematic clusters.

The next approach applied the NEB (novelty - exploit – burn out) mathematical model to forecast the publishing activity in the analysed thematic clusters. This procedure relies on the SIR epidemics model often applied for assessment of the disease infection progress. This technique relies on the observed analogy between epidemics and the development of a branch of knowledge.

The third approach makes use of the functions basing on citations and co-citations within the defined thematic clusters. Development of the research subject is presented by means of the analysis of "use of knowledge", while the two earlier approaches focus on the publishing activity ("production of knowledge").

Chapter 5 of the dissertation presents the results on the research trends detection as applied to distillation domain for four variants of the combined methods of structuring and trend detection. In addition, the chapter presents an assessment of the developed methods, based on the analysis of the results. Chapter 6 presents analysis of the filtration with a variant of the designed combined techniques. The dissertation ends with the summary of the results in the Chapter 7.

Chapter 4. Methods

The chapter presents the structuring methods of the thematic domains and trend detection methods.

4.1 Knowledge domain structuring by means of identification of the thematic clusters

Here the two approaches to identification of the subject clusters are shown. The section 4.1.1 deals with a simple and effective approach where no synonym recognition is involved.

The other approach in 4.1.2 shows how synonym recognition is applied to the subject cluster's identification. This technique recognises synonyms and near equivalent words for that purpose. This provides for the more precise analysis than the previous one at the expense of the computing power applied.

Each word recognised by either of the discussed methods is converted to its root form. To simplify the message, we would refer to the root form wherever the term "word" is used further on.

4.1.1 Structuring ignoring the semantic analysis

The following tools are used:

- Article_Data - MS Access database designed by the author;
- Rapidminer - open access tool used for generation of fingerprints table (http://rapid-i.com/);
- WordSet - tool designed by the author in Delphi, which allows on identification of the word sets;
- MS Excel Makro- procedure designed by the author, which allows on determination of the time distributed publication activity.

The applicable flow chart has been shown in the fig. 4.1. The elements in this chart represent the tools applied in the analysis, while the boxes show the stage of the analysis dealt with the indicated tools.

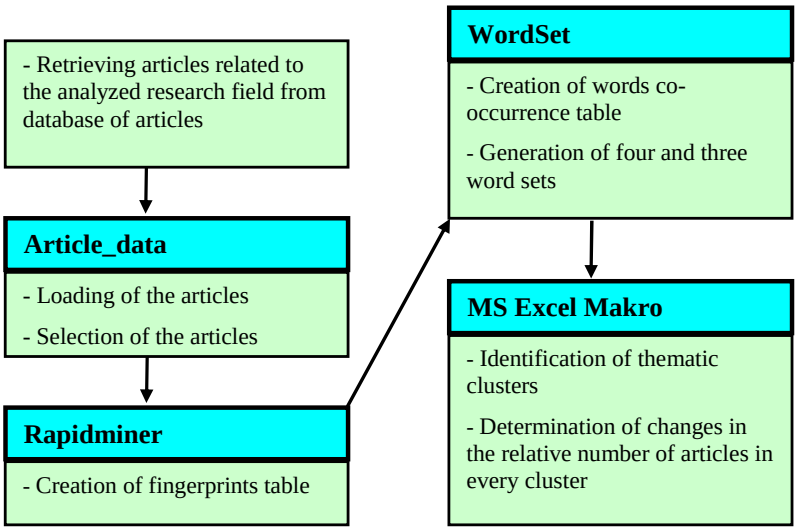


Fig. 4.1. Thematic clusters identification..

The process goes as follows: First, the publications addressing the analysed knowledge domain are identified in the scientific papers database. They are filtered with respect to the keywords. A paper is included in the domain when the query keyword is found in: title, keywords or an abstract of the publication. Then, the excerpts of the filtered publications are downloaded and stored. The excerpts contain the title, authors, keywords, abstract, year of publication, name of the journal, location within the issue (page number) etc.

Then, these data are loaded in the CIT table in the *Article_data* database. The next step of the analysis involves only titles, keywords and abstracts. Abstract is a mandatory feature here, as the former columns to not contain enough words to identify the clusters.

The next step applies the Rapidminer tool with the TextImput module (Mierswa et al. 2006). The tool creates a fingerprints table for all analysed documents. The array is built as a result of statistical analysis from which the immaterial words (stop list words) are removed. The fingerprints in the array store only the root forms of the words. The created table represents the meaningful words found in the documents under analysis.

The array created by the Rapidminer is fed to WordSet, which creates the word co-occurrence table. The table shows the number of documents , in which each pair of words co-occur. During further analysis the word co-occurrence table and the fingerprints table are used for identification of four words sets that denote the atomic elements of the thematic clusters.

Tests were carried out to find the optimum number of words constituting the atomic elements of the knowledge. Attempts of use of five words sets resulted in smaller number of clusters than the four words sets at the same threshold values, Furthermore, the five words collections discerned only very narrow subjects having small number of publications, Many important thematic clusters were ignored this way. On the other hand, three words sets created too broad clusters matching correspondingly broad subjects. Four words collections proved to be the best atoms for thematic clusters and returned the decently structured clusters. However, not all subjects can be identified with four words sets, so final identification of the thematic clusters was performed with four words sets and supplemented with the three words sets. In practice, the three word sets are applied to the leftovers after the four words sets searches.

The words that build a word set are characterized by a high co-occurrence between words constituting the same thematic clusters, and at the same time low co-occurrence with other words, that build other thematic clusters. Identification runs against all four word combinations and those that satisfy the condition below are recognised as atomic elements of knowledge:

$$CC_{i,j,k,l} \geq ThrA \qquad 4.1$$

$$\frac{CC_{i,j,k,l}}{\frac{1}{4}\left(D_i + D_j + D_k + D_l\right)} \geq ThrB \quad \text{, where} \qquad 4.2$$

$CC_{i,j,k,l}$ - number of publications where the words i,j,k,l occur,

D_i - number of publications where the word i occurs.

ThrA, ThrB - arbitrary threshold values.

The first formula defines the minimum number of documents in the thematic cluster. It must be low enough to facilitate discovery of new subjects with “little publicity”, as well as broad, well known subjects where a lot is being published. The second formula checks how often the candidates for a word set occur together as opposed to the company of other words. Thus, the ThrB threshold value determines the degree of significance of the words constituting a potential cluster from the discrimination point of view. The next stage of the analysis looks for three words sets not being subcollections of the approved four words sets.

The next step of the analysis is realized in MS Excel. Here the identified word sets are grouped into thematic clusters. Then comes the procedure that relates the number of

publication to time in each thematic cluster. The results are then scaled against the total number of publications for a given year for the given domain of knowledge. Distribution of the relative number of publications allows for easier recognition of the trends than the distribution of row number of publication. Also, the results are more reliable end expressive.

4.1.2 Structuring with the synonym recognition applied.

This section presents an improved variety of the procedure presented in the preceding chapter. The earlier method does not equate the synonyms while looking for the words sets and the assignment of the publications to the clusters requires a human involvement. The synonyms should be equated for the purpose of the word set building. Moreover the fingerprints building procedure should be adjusted to remove the immaterial words from the table of the fingerprints. Previously only the "stop-list" words and very rare were removed. Still too many immaterial words figure in the fingerprints table. Furthermore, the assignment of the word sets to clusters requires attention of human experts and slows the analysis down. This process should be automated.

The improved method utilises the WordNet database of English vocabulary (Fellbaum 1998, http://wordnet.princeton.edu, Miller 1995), where the words are grouped in the synsets and related by the lexical and semantic relations. The database allows for better identification of clusters after equating the synonyms in the words sets. The fingerprints are then built after computation of the TFIDF (Salton et al. 1975, Manning et al. 2007). The next improvement is application of the agglomerative clustering of the word sets (Chang et al. 2010, Voorhees 1986, Gowda et al. 1995). Cluster procedure is quick and controlled only by the two threshold values, and also provides for removal of the redundant word sets from the identified clusters.

The following tools are applied with this method: The tools not used in the previous method are presented in boldface.

- **Publication_Data**- a database developed by the author, built in MSSQL Server with functionalities of word root finding and the synonyms finding.
- WordSet - a tool developed by the author, created in Delphi, picking words sets,
- **Field_Group** - a tool developed by the author, created in Delphi, building the thematic clusters,

- MS Excel Makro- procedure developed by the author, charting the distribution of the publishing activity over time.

A diagram of an improved method of trend detection has been shown on the fig. 4.2..The elements in this chart represent the tools applied in the analysis, while the boxes show the stage of the analysis dealt with the indicated tools.

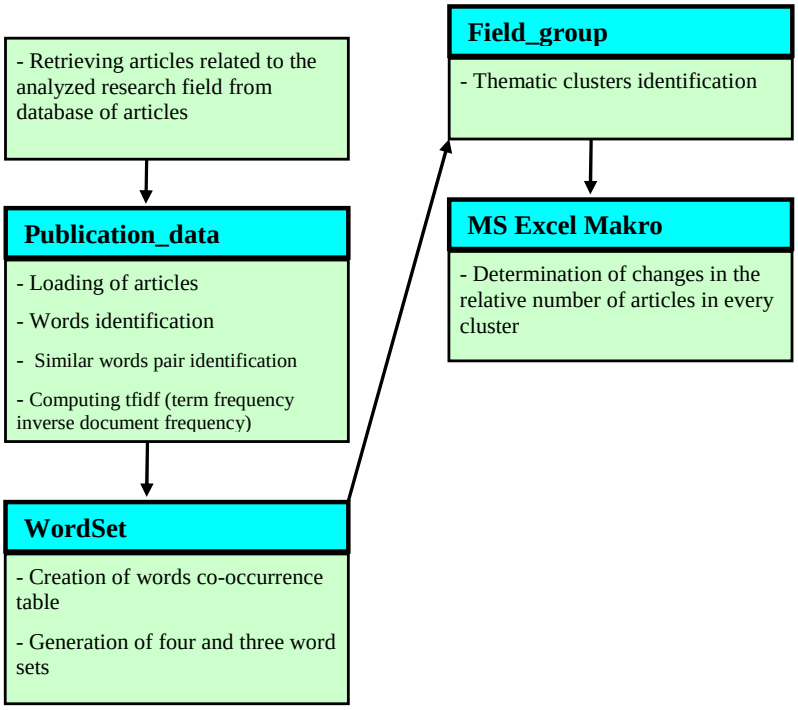


Fig. 4.2. Improved thematic clusters identification.

The process goes as follows: First, the publications addressing the analysed knowledge domain are identified in the scientific papers database. This process is an analogue to this described in the earlier subsection. Next, the found data are downloaded and fed into the CIT table in the *Publication_data* database.

The database includes the following *WordNet* database tables: „*lexlinkref*", „*linkdef*", „*semlinkref*", „*sense*", „*word*". The *WordNet* stores the words of English grouped by synonymity and related with respect to the semantic and lexical similarity relations of: hypernym, hyponym, antonym, similar, participle, derivation etc. The similarity relations are held in the „*linkdef*". table. Basing on these tables the *wordsim* table was built, where the pairs

lexically and semantically similar words are stored. The process was run automatically only for certain types of similarity in order to limit the number of the resulting pairs and select those most similar. The Publication_data database has implemented functions PorterStemmer and EnglishStopwordFilter borrowed from the Rapidminer with the TextImput module. The PorterStemmer finds the words' roots while the EnglishStopwordFilter removes immaterial words. The figure 4.3 shows the structure of the Publication_data database.

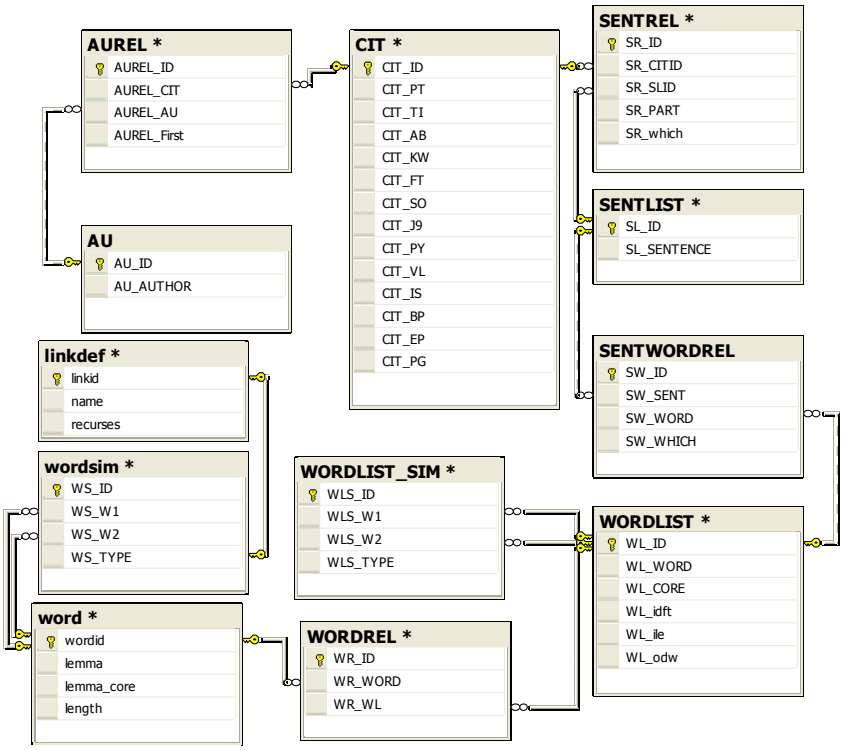


Fig. 4.3. Publication_data structure.

In the following process of thematic clusters construction the processed text is retrieved from the titles, keywords and abstracts. The sentences extracted from each of the publications are stored in the *Sentlist* table. Then, the analysis deals with the specific words. First, the words are converted with the *PorterStemmer* and then the words on the *stop-list* are removed. The resulting roots are stored in the *Wordlist* table of the *Publication_data* database. Furthermore, the *Sentwordrel* table records the (*word*, *sentence*) pairs indicating words found in the sentences. If a word has already been recorded in the *Wordlist* table, only the relation in the

Sentwordrel is recorded. Simultaneously, as the third step of Publication_data object, pairs of similar words from the *Wordlist* table are identified automatically based on the *word*, *wordrel*, and *wordsim* tables and are introduced into the *WordlistSim* table. Then the *Sentwordrel* is appended with the records of documents where the semantically similar words occur. This means that: if the words (A,B) are related in the *WordlistSim* table, and A is found in the document D, than the records equivalent to B is found in D are appended to the relevant tables The operation causes taking into account the similarity of the words during the further structuring process. The last step in the *Publication_data* database the TFIDF is computed for each word in the *Wordlist* table. Let's remind: this factor becomes high for words that are numerous in small number of documents, i.e. they discern well. Further on, the analysis deals only with the words having high TFIDF value.

Next, the *WordSet* tool is used in the same way as described in the previous subsection. It generates the words co-occurrence table and then generates four words sets and three words sets. The input data consist of two tables. *Cit-Word* records the primary key references to pairs (word, document) – this way records the containing document for each word. The records are appended to the table when the (word, document) pair satisfies the following conditions: TFIDF > A (A being an arbitrary threshold), and the number of documents where a word is found is limited by an arbitrary threshold B. *Word-Similarity*, the second input table, records primary keys of the similar words recorded in *Wordlistsim. Word- Similarity* table is extended in the *WorSet* tool to a second level of similarity explained on the example. So if word X is similar to word Y, and word Y is similar to word Z, then this situation is treated such that word X is considered similar to word Z, and the extra similarities, X, and Z are added to the *Word- Similarity* table. This operation results in elimination of the redundant word sets in the WordSet tool, which consist of synonyms. The *WordSet* builds collections in very similar way to that described in 4.1.1. There is only one extra condition in the word sets building. The limitation states that the word sets, which contain although one pairs of similar word are neglected. It causes elimination of the redundant word sets, consisted of semantically similar words. After the four words sets are built, the three word sets follow.

The next step of the analysis is realized by the Field_group tool in order to cluster the word sets into thematic clusters. This process applies an agglomerative, hierarchic clustering described in detail in 2.1.1.1 (Chang et al. 2010, Voorhees 1986, Gowda et al. 1995). The input data to the *Field_group* include the *Cit-Word* table and the table of words sets. Clustering begins with one set in each clusters. In each next step, the closest clusters

according to the adopted metric, are being merged. The loop breaks when the similarity between the closest clusters is lower than ThrC threshold. Similarity is computed for all collections' pairs, where the collections are assigned to the respective clusters. The similarity is computed as provided by the 4.3 formula. This metric of similarity resembles full linkage method described in 2.1.1.

$$MIN\left[\left(\frac{D_{Si} \cap D_{Sj}}{MIN(D_{Si}, D_{Sj})}\right), Dla \quad i = 1..I,\ j = 1..J\right], \text{ where:} \qquad 4.3$$

S_i - words set assigned to the first cluster under consideration;
S_j - words set assigned to the second cluster under consideration;
D_{Si} - number of papers where the word set S_i occurs
I- number of words sets in the first cluster;
J - number of word sets in the second cluster;
MIN- the minimum function;

Additionally, a limitation 4.4 has been introduced to avoid confusion when one of the words sets represents a narrow fields while the other words set represents a broad field overlapping with the former. When the limitation is not satisfied for the pair of words sets, this is interpreted as absence of similarity.

$$\frac{MAX(D_{Si}, D_{Sj})}{MIN(D_{Si}, D_{Sj})} \geq ThrD \qquad 4.4$$

After grouping, the clusters are checked against redundant word sets which are removed. This means that when a set of documents containing the words set A of the cluster AA is completely included in the set of documents where other words sets of AA occur, the words set A is removed. Once the clusters are defined, the output of the *Field_group* tool are fed in the *Publication_data* for further analysis.

The final result coming is a time distribution of the relative number of publications represented by each thematic cluster, what facilitates application of the trend detection procedures.

4.2 Trend detection methods

This subsection proposes three approaches to trend detection. The first borrows from the technical analysis of the financial markets. These methods are applicable because the publishing activity oscillates in time. Therefore, the trend detection can by analogy borrow

from the technical analysis of the financial markets (Murphy 1999). The main advantage of these methods is detection of the turning points of the relative number of publications (RNP), which start a new trend section. A trend is recognised as ascending when each increase in the RNP attains a new maximum, while each decrease stops at a higher level than the previous one. A descending trend is defined accordingly.

The second approach - a NEB model - utilises an adjusted SIR model of epidemics. This procedure relies on the SIR epidemics model often applied for assessment of the epidemics development progress. Its time profile shows similar shape as a time profile of the RNP.

The third approach applies newly designed citation function (CF) and co-citation function (CCF). This methodology offers alternative view on the subject development. As indicated, these previous approaches base on “knowledge production” while the proposed approach considers “use of knowledge” and therefore bases on citations and co-citations numbers. Moreover, the proposed functions allow on easy identification of the developing research areas, because its signals are more exposed than the relative number of publications signals.

4.2.1. The methods of financial markets

Trend detection discussed in this subsection applies three methods borrowed from the technical analysis of the financial markets (trend line, a simple oscillator (OSC) and stochastic oscillator (K%D)). These are the most common techniques of a technical analysis of the financial markets. Trend detection is difficult and unreliable and there are many supporting techniques improving the reliability. This approach allows for detection of periods when a given trend governed and facilitated short term forecast of future trends. The advantages of these techniques are useful for finding of the thematic clusters whose climax has not yet been reached.

4.2.1.1. Trend line method

A straight line is drawn, crossing the points of minima (maxima) of the RNP for the respectively ascending (descending) trends. The line shows both the trend direction and its velocity (angle of the ascent). When the trend line crosses the section line of the RNP, it signals reversal of the trend. The fig. 4.4 shows how to identify the ascending and descending

trends detected with this technique. The points marked with “9” on the drawing signal reversal of the trend.

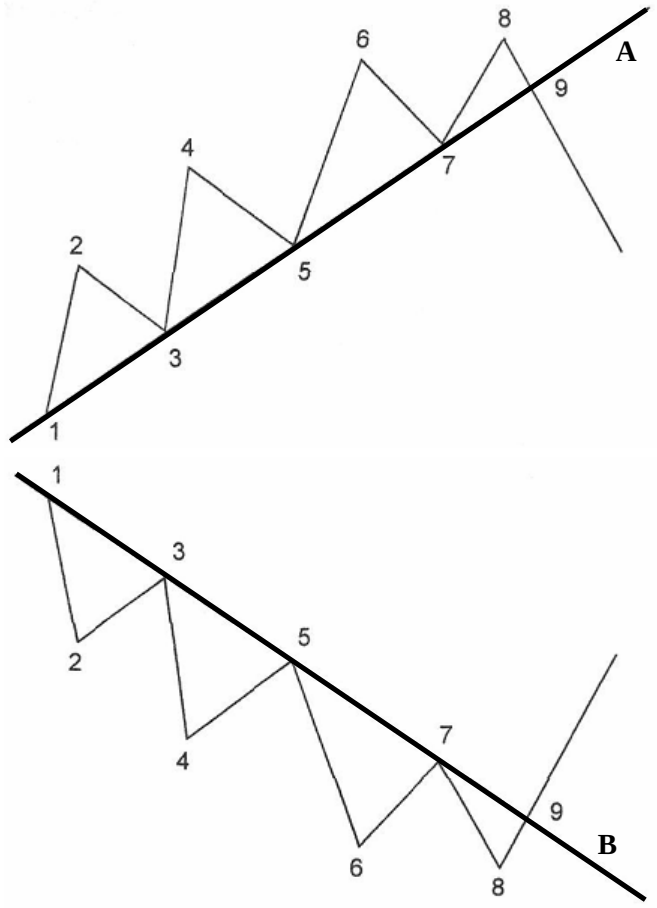


Fig. 4.4 Trend detection example for ascending (A) and descending (B) trends.

4.2.1.2. Oscillator method

This method (OSC) bases on the momentum, a value defined as a difference between the annual number of publications and the respective number of x years earlier. Positive momentum indicates an ascending trend while a negative value marks a descent. Zero value (sign change) indicates reversal of trend. Peak values (any sign) may indicate forthcoming reversal of trend. The x parameter sets the frequency of the oscillator. The lower it is, the

larger is the frequency. This means that for large values of x, the oscillator reacts slower to the trend reversal but it is less sensitive to random (noise) alterations in the RNP. Small values of x, make the oscillator to react quicker to the trend reversal but it is more sensitive to random (noise) alterations in the RNP. In practice, usually two oscillator functions are examined (small and large x) to balance the imperfections and get more reliable data. The fig. 4.5 shows an example of application of this method to various values of x. The curve illustrated in the top of the figure denotes prices, however the bottom curves illustrate momentum's lines for high and low value of parameter x. The faster momentum curve very often intersect the zero line what makes difficult trend identification. However slower momentum curve allows on easier identification of the trend but identifies trend later than faster momentum curve. Points market by arrows in the figure denote moments best for purchase

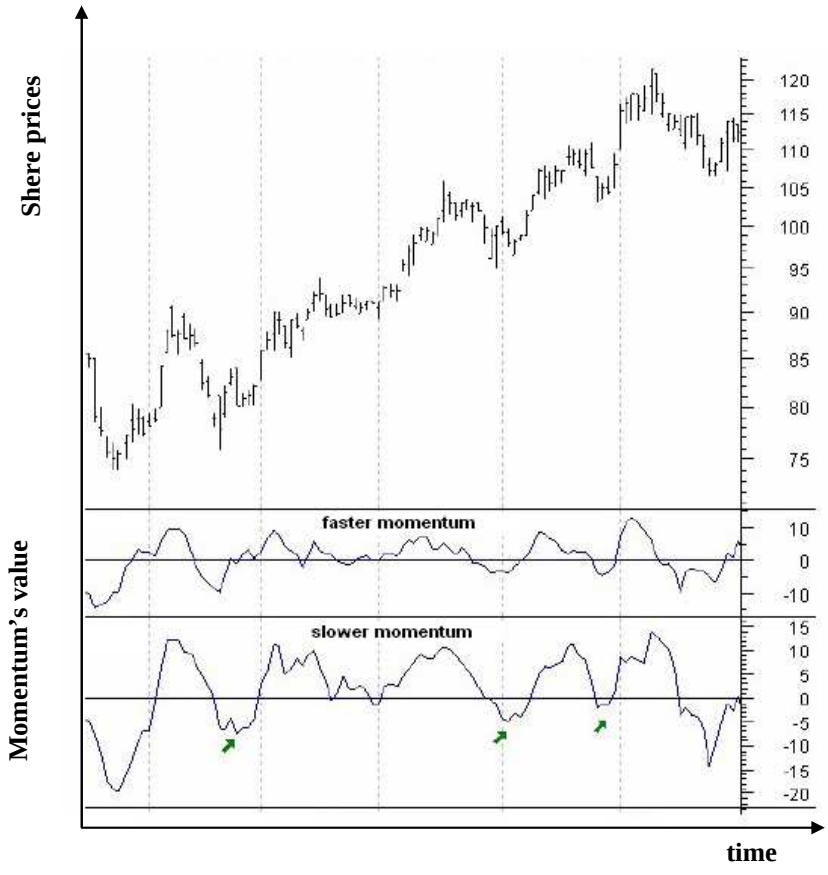


Fig. 4.5 Trend detection with OSC.

4.2.1.3. Stochastic oscillator K%D

The K%D stochastic oscillator charts two moving average lines K and D. The values of K are computed with the formula 4.5 .

$$K = 100\frac{C - L}{H - L}, \text{ where:} \qquad 4.5$$

C - current value of the RNP,

L- RNP minimum over y year back

H- RNP maximum over y year back

The line D is a three years moving average o the K line. Ascending trend is detected when the lines cross on ascent. Descending trend is detected when the lines cross on decent. An example is shown on Fig 4.6.

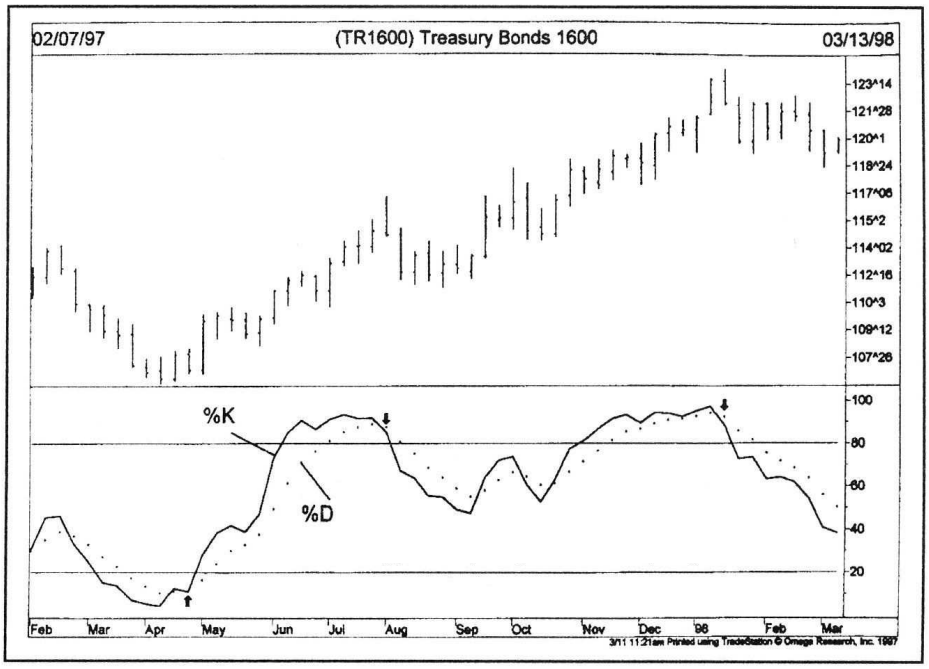


Fig. 4.6 Trend detection with K%D.

The top chart shows price change - a vertical section marks a day change. The lower part of the fig. 4.6 shows the K and D lines. Arrows indicate at the points of trend reversals.

4.2.2. Mathematical modelling as a way of trend detection

This subsection presents application of the NEB (novelty - exploit – burn out) mathematical model for visualisation of the RNP in the thematic clusters. This facilitates forecast of the future trends in research. The model has been created on a basis of the SIR epidemics model (Kermack et al. 1927, Piyawong et al. 2003). It is governed by three differential equations describing the time evolution of the three groups of members during the epidemic progress. These groups include: Susceptible – healthy – at risk of infection, Infected and Recovered. Analysis of the RNP charts shows that the RNP grows to a top point and then stabilises or decreases. This depends on the subject's level of development. Therefore the SIR model has been applied and then replaced by the NEB model, to represent the RNP distribution. The built model recognises novelty of the thematic clusters (N), changes of the publication activity (E) and their burn-out.

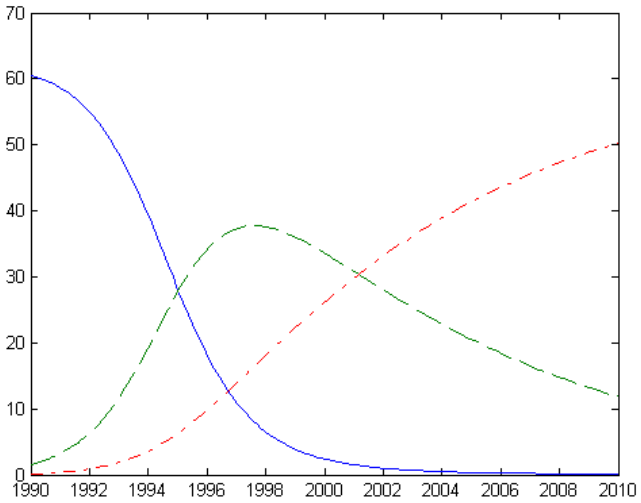
This method employs the tools created in the Matlab environment that allow fitting of the model functions to the experimental data and computation of the model parameters.

4.2.2.1. SIR Epidemics

The SIR models the evolution of the three groups during the epidemics outbreak. The model assumes that the infected (I) individuals transfer the disease to the susceptible (S) individuals, while the recovered (R) are already resistant. Therefore the speed of S individuals decrease is proportional to both their number and the number of the infecting agents (bSI). The number of infected individuals grows with the decrease of the susceptible but the recovering make the grow slower subtracting the number proportional to those infected. The speed parameters a and b denote infection and recovery rate. The described process of evolution can be represented by the following system of differential equations.

$$\begin{aligned} \frac{dS}{dt} &= -b \cdot S(t) \cdot I(t); \\ \frac{dI}{dt} &= b \cdot S(t) \cdot I(t) - a \cdot I(t); \\ \frac{dR}{dt} &= a \cdot I(t); \end{aligned} \qquad 4.6$$

An example of the time evolution of the all three groups has been shown on the fig, 4.7.



Fig, 4.7 An example of the time evolution of the Susceptible (solid blue), Infected (green dash) and Recovered (red dash point)

The epidemic begins with many susceptible and a few infected. With time, the number of the susceptible decreases and the infected grows in numbers for as long as the first part the second equation (bSI) dominates over the second part (aI), This means that there are more individuals being infected than those recovering. When these terms equal, the number of infected reaches maximum shown on Fig. 4.7. Then the number of infected individuals decreases when the number of the recovering exceeds that of being infected. The number of the recovered grows with varying speed during the epidemic.

4.2.2.2 NEB model

The proposed model NEB bases on the SIR model (He et al. 2007, Pourabbas et al. 2001, Piyawong et al. 2003). The differential equations system have been shown below.

$$\begin{aligned} \frac{dN}{dt} &= -c \cdot N(t) \cdot E(t); \\ \frac{dE}{dt} &= c \cdot N(t) \cdot E(t) - d \cdot E(t); \\ \frac{dB}{dt} &= d \cdot E(t); \\ d &= d_0 \cdot (1 + g \cdot \sin(w \cdot t)) \end{aligned} \qquad 4.7$$

The function N represents novelty of an thematic cluster, E the publishing activity in the thematic cluster, while B denote its burn out. Novelty is substantial when the analysed branch of knowledge emerges and then decreases in time with its development proportionally to novelty and publishing activity. The c parameter represents the speed of grow of the publishing activity. At the same time, the burning out equals 0 with emergence of the branch of knowledge and grows along the discoveries in proportion to the publishing activity (dE). The d parameter represents the speed of burning out. It is a harmonic time function with the amplitude equal d_0g (the force of burn out) and frequency w. The publishing activity depends on how the novelty relates to the burn out. This means that the publishing activity in a branch of knowledge depends on the novelty and the existing publications in the field. These publications inspire new discoveries in the subject. As long as the decrease in novelty (cNE) exceeds the growth in the burn out (dE), the publishing activity will increase, and it will decrease when the said relation reverses. Analogy to the SIR model is quite obvious.

Application of this model requires fitting of the E function to the experimental distribution of the RNP. This fitting will facilitate forecast of the future RNP, i.e. a trend forecast. Finding of the parameters in the NEB model for the purpose of fitting of the E function is an optimization problem.

4.2.2.3 Fitting of the model to the data

Fitting of the E function requires solution of the NEB model for the assumed model parameters. The applied optimization algorithm adjusts the assumed values according to the fit accuracy. The optimization problem is solved with the algorithm of differential evolution what allows for the parameters determination (Price et al. 1997, Price et al. 2005). The problem is solved with the least squares method, so the $\Sigma(RNP(t) - E(t))^2$ is being minimised. The process of optimisation has been presented in the fig. 4.8. Differential Evolution is a effective evolution strategy. It is simple, easy in application fast and efficient. This algorithm was used i.e. for optimization of a system of ion exchange columns (Sitarz 2007). There are two tables in the optimizer. Each contains a population of vectors (members) having real values. The members are construed here as the solutions of the optimising problem. Each member is represented by a vector of decision variables. The primary table records the current population while the secondary table stores the next generation vectors.

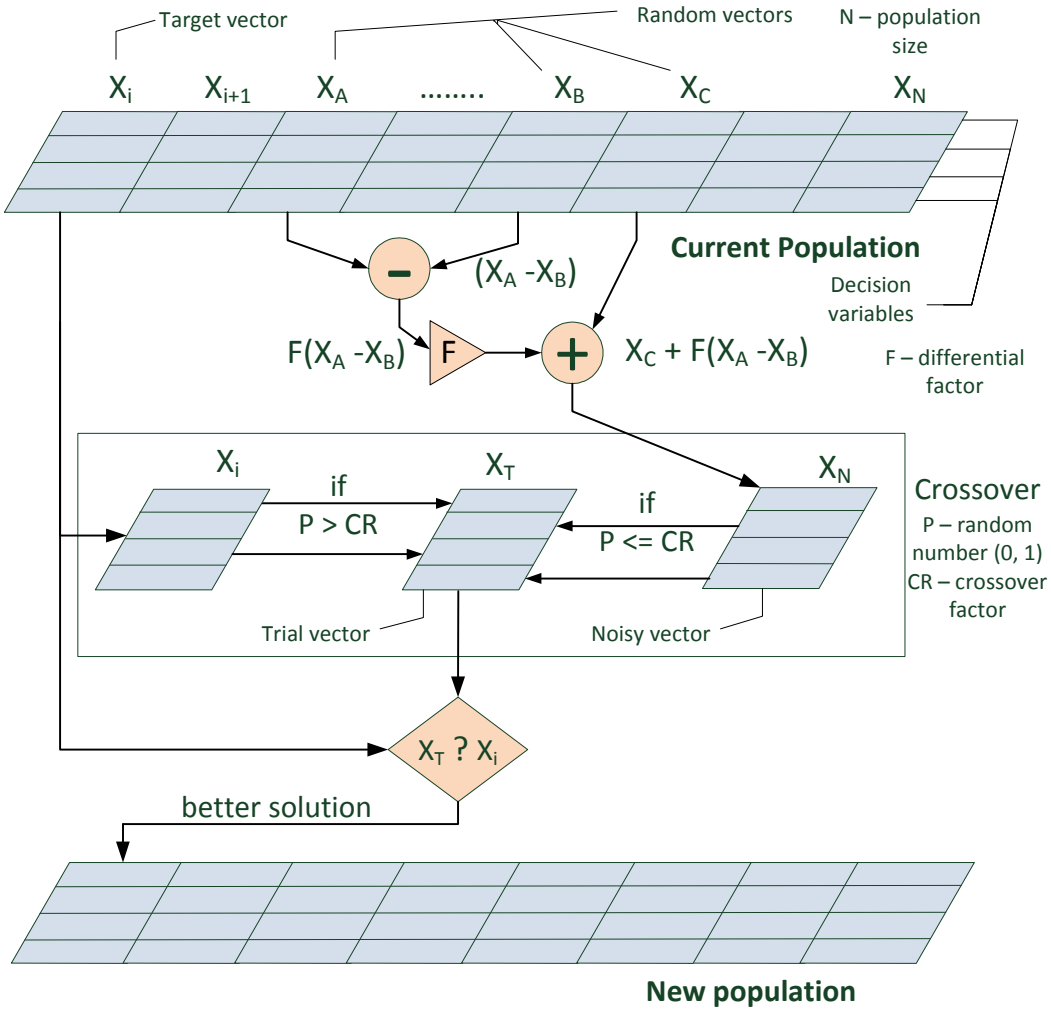


Fig. 4.8. Algorithm of the differential evolution

The process goes as follows: First, the start population is built of the assumed number of members. The members are selected by random attribution of values within the earlier defined limits for each variable. Once the optimising is launched, the first member of the start population is selected as a target vector, and then 2 vectors are drawn from among of the remaining members and subtracted. The resulting vector multiplied by the F parameter is added to a third vector selected at random from the remaining vectors in the table. This is a noise vector, which is crossovered with the target vector what results in the trial vector. During the crossover process a real number is selected in random from the section (0,1). If it

exceeds the CR parameter of the optimizer, the trial vector gets the variable value copied from the target vector, otherwise, it gets the value of the noise vector. Next, the goal function is being computed for the trial vector and compared with the goal function of the target vector. If the value of the goal function of the trial vector is lower than the goal function of the target vector, the trial vector is stored in the next generation table, otherwise the target vector is stored in the next generation. This is a cycle for a single member. Then the next vector is selected as a target vector and the cycle is repeated for all members as the target vectors (N times) and for the given number of generations (M).

The data optimized in the NEB model (decision variables) include: c, d_0, g, w, N_0 (novelty of the subject in the initial conditions), E_0 (RNP in the initial conditions). The optimization algorithm has been implemented in the Matlab script and tested on test examples. The system of NEB differential equations was solved for each iteration by the ODE15s procedure of the Matlab software, for the given decision variables.

4.2.3. Research trend detection basing on citation and co-citation analysis

The trend detection methods presented in the previous subsections based on the RNP. This means that the field of research development is perceived as the growth of the RNP, i.e. with the "knowledge production". On the other hand the new knowledge discovered and published inspires work that results in new discoveries. Therefore, if a certain knowledge domain develops, then the papers belonging to the domain should be cited often. This approach focuses on "use of knowledge" as a indicator or the domain development.

However, the number of citations an co-citations within the domain is not good measure of the development since these values grow along with the number of publications. Therefore, the number of citations and co-citations within the domain should be normalized to provide for a reliable indicator of the domain development. Furthermore, domains include key publications which are often cited and those less prominent which are often not cited at all. Also, the subsequent citation of a key paper seems more important for the subject development than citation of a marginal publication. Another issue is the period when the paper enjoyed referencing. Key publications are quoted for a long time, as opposed to the marginal ones. Another issue are the co-citations which indicate coherence of the subject matter. The number of co-citations should be normalised in a similar way as the number of citations.

Taking into account the known shortcomings of the bibliographic techniques that map the domain development on the number of citations and co-citations, some functions mapping the development have been proposed. The proposed technique solves all the shortcomings of those discussed earlier. The basic idea goes as follows.

4.2.3.1. Citation analysis method

The functions defined below (4.8 and 4.9) represent the development of the domain.

$$CF_t = \sum_{i}^{N} C_{i,t} \cdot \left(\frac{C_i}{\max C} + \frac{P_i}{\max P} \right) \qquad 4.8$$

$$CCF_t = \sum_{i}^{N} \sum_{j \neq i}^{N} CC_{ij,t} \cdot \frac{CC_{ij}}{\max CC} \cdot \sqrt{\left[C_{i,t} \cdot \left(\frac{C_i}{\max C} + \frac{P_i}{\max P} \right) \right] \cdot \left[C_{j,t} \cdot \left(\frac{C_j}{\max C} + \frac{P_j}{\max P} \right) \right]} \qquad 4.9$$

where

N - number of documents in the analysed thematic cluster.

$CC_{ij,t}$ - number of co-citations between the documents i and i within the year t.

CC_{ij} - total number of co-citations between the documents i and i.

maxCC - maximum value of the co-citations in the analysed documents.

$C_{i,t}$ - number of citations of the document i in the year t,

C_i - total number of citations of the document i,

maxC - maximum value of the citations in the analysed documents.

P_i - period when the document i has been cited,

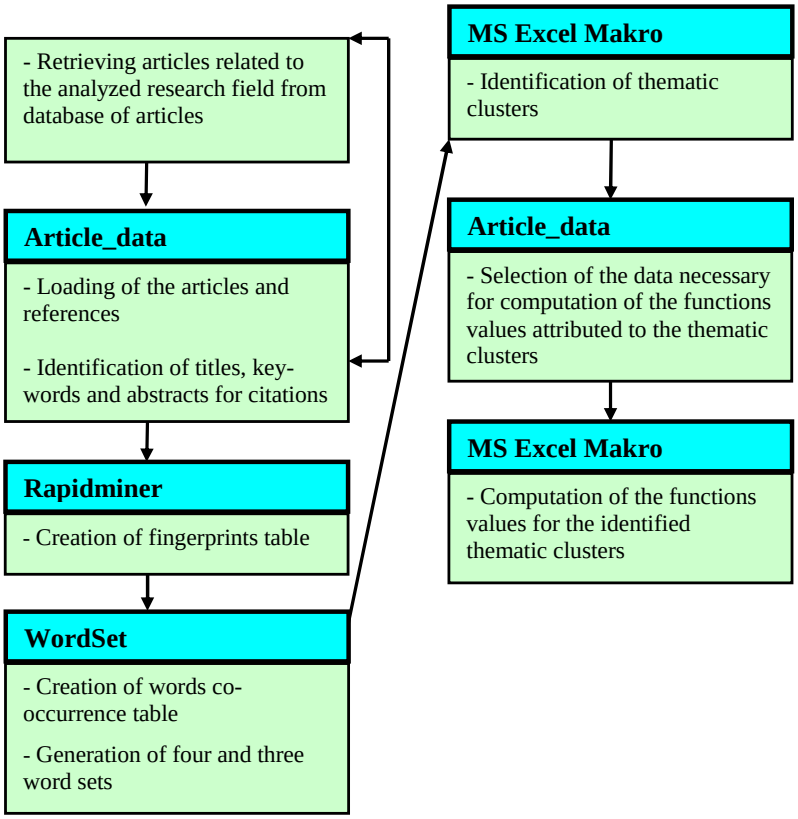
maxP - maximum period, where the document can be cited for the data under analysis.

The citation function (CF) represents the development level of a research domain and depends on the number of citations only. Co-citation function (CCF) accounts for the number of co-citations within the domain. These functions are computed annually, what allows for representation of the time progress and detection of trends. The CF value depends on the number of citations, which is normalised to the relevance of the cited documents. The normalising factor includes two elements. The (C_i/ maxC) factor weights of the citations of the document against the most cited document. This value therefore belongs to (0,1) section. The factor (P_i/ maxP) weights the period when the document under consideration against the longest observed period citations. This value therefore belongs to (0,1) section.

The CCF value depends on the number of co-citations, which is normalised to their relevance and the weights of the documents with which these are co-citation. The CC_{ij}/

maxCC factor weights of the co-citations between the i and j against the most co-cited pair of documents. This value therefore belongs to (0,1) section. The remaining terms in the CCF function represents the average of the relevance of the co-cited publications. The weights are computed analogously to those for the CF function.

The fig. 4.9 presents the process of trend detection. The elements in this chart represent the tools applied in the analysis, while the boxes show the stage of the analysis dealt with the indicated tools. The fraction of the diagram that deals with the construction of the thematic clusters matches the diagram 4.1 representing the method described in the 4.1.1.



Fog 4.9. Trend detection based on citations and co-citations

The process goes as follows: First, the publications addressing the analyses knowledge domain are identified in the scientific papers database (e.g. ISI Web of Science), with the search criteria based on the keywords associated with the analysed knowledge domain. Then,

the excerpts of the criteria matching documents are stored in records containing: title, authors, keywords, abstract, publishing date, bibliographic address etc. These data are read into the *CIT* table in the *Article_data* database and the citation data are fed into the *CR* table. Additionally, the database has relations table recording the *CIT* and *CR* indices values, and thus it is known which documents and where are cited. Citations (i.e. references to the documents included in the records) keep the following data: authors. journal name, publication year, journal issue, volume and pages in the issue. Therefore, the essential information like: title, key words, and abstract is lacking. These data need to be updated and the necessary procedure has been built.

Next step builds the table "*Powiazania*" [Relations] where the pairs of indices values of co-cited documents from the CR are stored. This process runs automatically involving another specially built procedure.

In the following step, the thematic clusters for the analysed knowledge domain are built. This process is the same like that described in the 4.1.1. The clusters can be built for the data from both the CIT or CR tables. In the case of the reported research only the CIT clusters were built because the CIT documents have been specially selected to fit in the analysed knowledge domain, while the cited documents from the CR may not concern the domain under analysis.

The last stage involves the specially designed procedure that identifies the data necessary for computation of the CF/CFF functions for each year and each thematic cluster. The data are fed to Excel and the values of the CF/CFF functions for the specific clusters are computed for the analysed period.

Chapter 5. Structure of the research area – distillation.

This section presents the results for the distillation research area in four variants:

1. Structuring without synonym recognition; trends detection in the analysis with the financial markets methods of technical analysis;
2. Structuring without synonym recognition; trends detection in the analysis with the CF and CCF functions;
3. Structuring with semantics; trends detection in the analysis with the financial markets methods of technical analysis;
4. Structuring with semantics; trends detection in the analysis with the NEB model;

5.1 Variant 1 results

The subsection shows the results for structuring with no use of synonym recognition and trends detected by the methods of financial markets technical analysis. The research was run on the documents (scientific papers) in the ISI Web of Science database. This database seems the most suitable for the analysed subject. Furthermore, the database is generally accessible, often used and facilitates batch downloads of abstracts (up to 500).

The ISI Web of Science contains 13450 papers published between 1990 and 2008, whose titles, keywords or abstracts contained the word "distillation". These data were fed into the *Article_data* database. 11402 documents that included an abstract were selected for further analysis. Next, the fingerprints table was built for 7389 identified words and 11402 documents. Words that were found in less than 5 documents only, were removed from the fingerprints. Such rare words are too detailed for further analysis. The next step created 2156 four words sets for the thresholds valued of ThrA and ThrB equal 14 and 0,09 respectively. Additionally, 993 three words sets were built, which did not include in the four words sets created earlier. The thresholds values ThrA and ThrB for the three words sets equalled 14 and 0,14 respectively. Some of the created word sets were removed since they did not refer directly to the analysed subject and dealt with the subjects like chromatography, spectroscopy, sedimentation etc.. The remaining words sets were grouped according to expert decision into 38 thematic clusters specified in the table 5.1.

Table 5.1. Thematic clusters identified for the variant 1.

No	Name of the thematic cluster	Number of Articles
1	Designing of the distillation column system	1281
2	Modelling of the distillation column	1113
3	Extractive distillation	1097
4	Modelling of the distillation column system	1055
5	Controlling of the distillation column system	809
6	Optimization and designing of the distillation model	619
7	Vapour - liquid phase model for distillation	504
8	Multiple steady states in the distillation process	481
9	Dynamic of the steady state model	319
10	Economy aspects of distillation	316
11	Membrane distillation	312
12	Modelling of the reactive distillation	299
13	Equilibrium aspect in the reactive distillation	293
14	Heat and mass transfer in the packed distillation column	271
15	Kinetics and catalysis in the reactive distillation	222
16	Designing of the reactive distillation	209
17	Distillation of alcohols	194
18	MTBE System	194
19	Azeotrope distillation	173
20	Optimization in distillation process	163
21	Distillation process for desalination	151
22	Esterification	139
23	Azeotrope residue curve map	128
24	Pressure drop aspect in the packed distillation columns	119
25	The steady state model for the system of the reactive distillation	105
26	Supercritical distillation of CO_2 and CO	104
27	Hydrophobic membranes	96
28	Energy aspects of desalination	89
29	Economical aspects in the desalination process	75
30	Equilibrium models - nrtl, and uniquac	66
31	Stefan Maxwell nonequilibrium state model	53
32	Direct contact membrane distillation	48
33	Cryogenic distillation of the hydrogen isotopes	45
34	Distillation plate	40
35	PVDF membranes	28
36	Heating boilers for distillation	28
37	Artificial neural network applications in distillation	26
38	Distillation of the essential oils	24

The examples of the trend detection for two selected thematic clusters have been presented below.

The detailed results for the "Membrane distillation" thematic cluster

The words sets for the cluster have been listed below.

solut	membran	permeat	flux;
solut	membran	water	flux;
solut	membran	flux	concentr;
solut	membran	flux	feed;
membran	temperatur	permeat	flux;

membran	temperatur	water	flux;
membran	temperatur	flux	concentr;
membran	temperatur	flux	feed;
membran	permeat	pervapor	flux;
membran	permeat	water	flux;
membran	permeat	flux	concentr;
membran	permeat	flux	feed;
membran	permeat	flux	transport;
membran	permeat	flux	hydrophob;
membran	permeat	flux	contact;
membran	permeat	flux	modul;
membran	water	flux	concentr;
membran	water	flux	feed;
membran	flux	concentr	feed;
membran	flux	direct	contact;
membran	flux	transport	hydrophob;

312 documents published between 1990 and 2008 were assigned to the subject under consideration. The fig 5.1 shows the trends detected for this thematic cluster. The upper part of the chart shows the RNP segment line in the period from 1990 to 2008 and the charted trend lines (straights I- III). The middle part of 5.1 shows the momentum lines for the OSC for x=2 (orange line) and for x=4 (brown line). The lower part of the chart shows the K line (for y=4) and D line for the stochastic oscillator.

The results indicate four specific trends.

An ascending trend was present between 1990 and 1994 as the line I indicates. The signals generated by OSC or K%D during this time are inessential, since only after 1994 the OSC is fed with enough data to detect the trend, while the K%D oscillator could only be applied after 1997.

Within the 1994 - 1997 period a descending trend is present, shown by the line II. The momentum is negative what corroborated with the perception of the descent.

The 1997 - 2003 period is marked with the ascending trend with weak oscillations of the RNP. Within this period both the momentum lines are placed over zero what corroborated with the ascending trend. Crossing of K and D in 1997 corroborates with the ascending trend.

The 2003 - 2008 period is marked with the ascending trend (line III) with strong oscillations of the RNP. The crossings of K and D during this period are misleading and result from large oscillations in the number of publications. Both momenta are positive after 2004 what corroborates with the ascending trend. Negative momentum for the K%D oscillator after 2007 is misleading and results from large oscillations in the number of publications The analysis of the detected trends leads to forecast of the continuing ascent after 2008 since there are no signals of reversal.

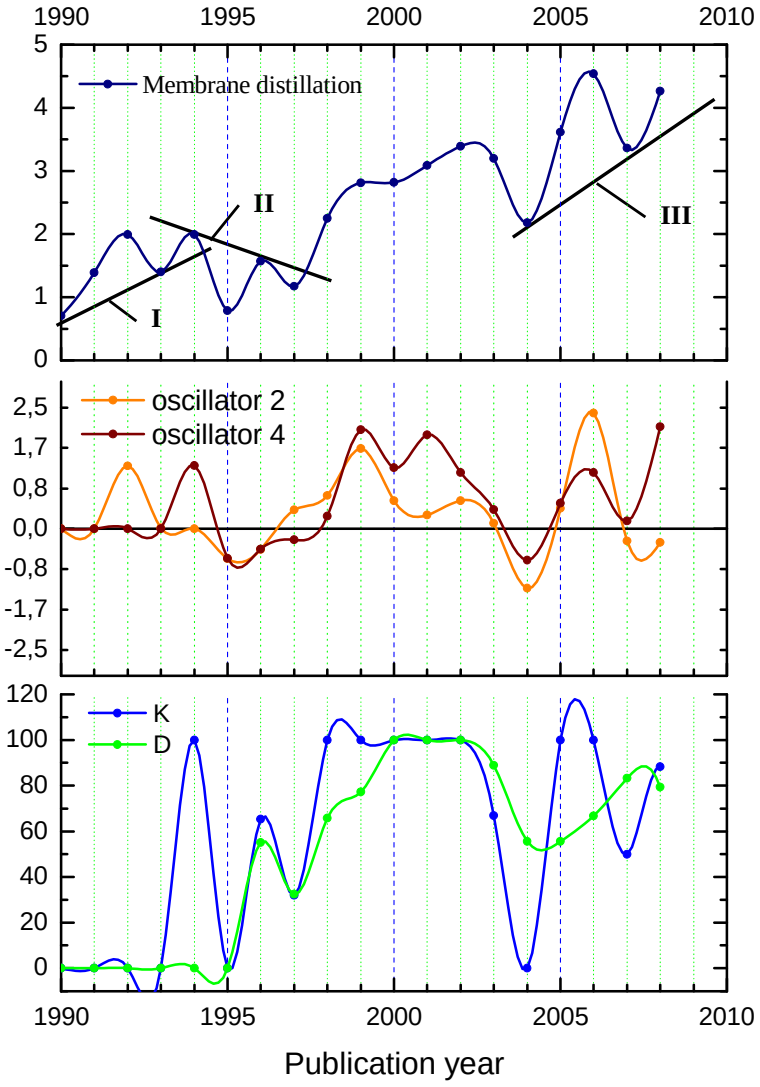


Fig 5.1 Trend detection for the "Membrane distillation" cluster.

Detailed results for the " Kinetics and catalysis in the reactive distillation " thematic cluster

222 documents published between 1990- and 2008 were assigned to the cluster under consideration. The following word collections define it.

catalyst	reaction	kinet	reactiv;
reaction	kinet	model	reactiv;
reaction	kinet	column	reactiv;

The fig 5.2 shows the trends detected for this thematic cluster. The arrangement of the lines on the chart is similar to those in the previous case.

Analysis of the trend detection results shown on 5.2 we can identify three main trends during this period.

In the beginning period from 1990 to 1995 a weak ascending trend is detected with its peak in 1994. It is represented by the line I.

Within the 1995 - 2002 period an ascending trend continues, shown by the line II. However, the growth is stronger than the recorded earlier. It is represented by greater angle of ascent for the line II. Crossing of K and D in 1998 corroborates with the ascending trend. The momenta of the OSC and the K%D oscillators are positive, and this corroborate with the ascending trend too. Crossing of the line II and the RNP line as well as crossing the K and D lines near 2002, indicate a trend reversal in 2002.

For the third period between 2002 and 2008 the lines III and IV form a descending triangle, what indicates a period of change. When the trend line III crosses the section line of the RNP, it forecasts a descending trend in 2008.. Values of the momenta were positive, but they were descending and the momentum line for the second oscillator was placed below zero between 2005 and 2007. This is a signal indicating an unstable, intermediate trend during this period. Crossing of zero by both momentums before 2008, indicates the descending trend after 2008. Furthermore, the lines K and D cross several times indicating subsequent ascents and descents.

This indicates a volatile trends during this period. However, crossing of K and D in 2007 confirms the forecast of descent after 2008.

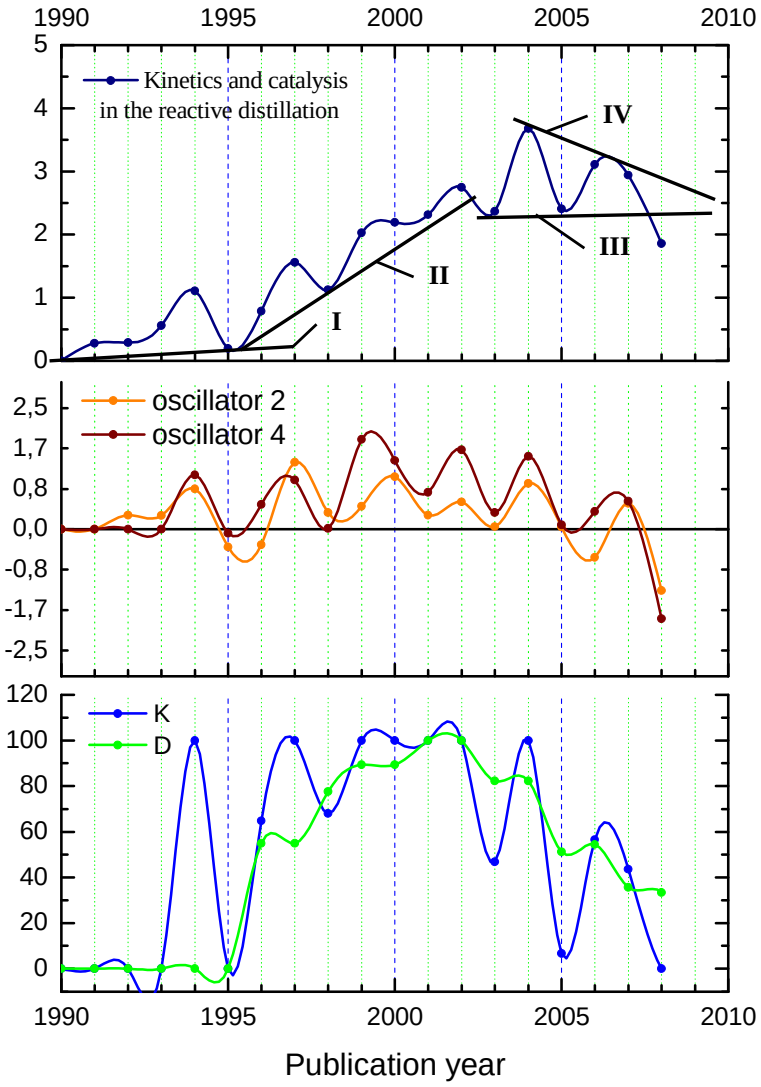


Fig 5.2 Trend detection for the " Kinetics and catalysis in the reactive distillation" cluster.

5.2 Variant 2 results

The results for structuring without the semantic analysis and the CF, CFF are as follow. This query involved 13450 papers from the ISI Web of Science published between 1990 and 2008, whose titles. keywords or abstracts contained the word "distillation". The process of the clusters construction has been described in 5.1.1. There were 38 clusters. The clusters were built with the data from the CIT table in the *.Article_data* database. There were 182050 cited papers in the 13450 found documents. Titles of 53827 cited papers were found, 46098 cited papers has their keywords identified and for 47099 abstracts were found. The co-citation table „*Powiazania*" had 1197770 records of co-cited documents. Then, the values of CF and CFF were computed for all thematic clusters as the function of time. The parameters' values were the following: maxC = 206; maxP = 18; maxCC = 68; Below, two trends detected with the CF and CFF have been presented.

The detailed results for the "Membrane distillation" thematic cluster

The fig. 5.3 shows the changes in RNP of the cluster over time (in blue at the upper chart) and the CF and CFF changes over time (red and green on the bottom chart).

The values of both functions on 5.3 grow until 2005, what indicates development of the subject over time. However, there were temporary drops of the values between 1996 and 1998, and between 2000 and 2001, understood as temporary interruptions in the development. The RNP grows until 2005 except for the temporary descent between 1994 and 1997.

After 2005 both functions fall until 2007 what signals brake of the development. But further growth after 2007 indicated continued development. Analysis of the RNP within this period indicated ascending trend with high oscillations between 2005 and 2008. The profile of both functions is similar but CCF provides for stronger and more expressive signals.

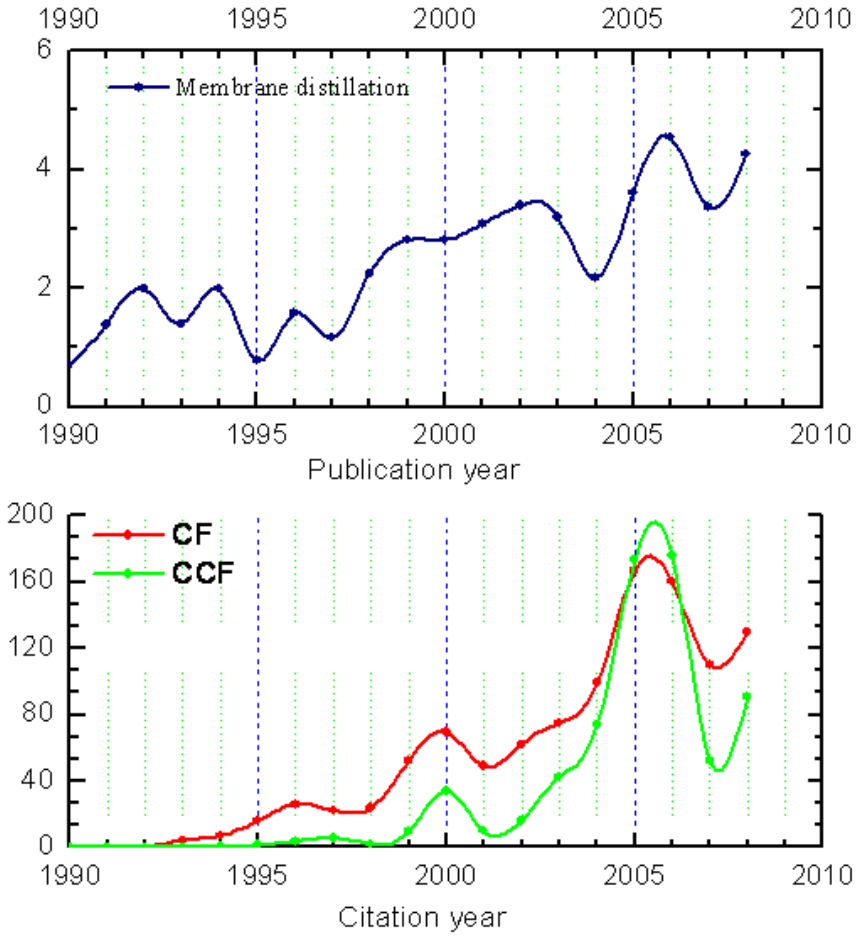


Fig. 5.3. Detection of trends with the CF and CFF functions for the "Membrane distillation" thematic cluster

Detailed results for the " Kinetics and catalysis in the reactive distillation" thematic cluster

Fig. 5.4 shows RNP over time and the CF and CFF profiles. The colours are applied as in the previous sample.

This cluster reveals two local maxima near 2000 and 2004, and - like in the previous sample - the signals of the CFF are more expressive. The proposed functions show development up to 2004 with short break between 2001 and 2003. The subject seem to expire after 2004. Similar results are coming from the RNP in this cluster, where the ascending trend is first observed until 2004 to become stagnant after that.

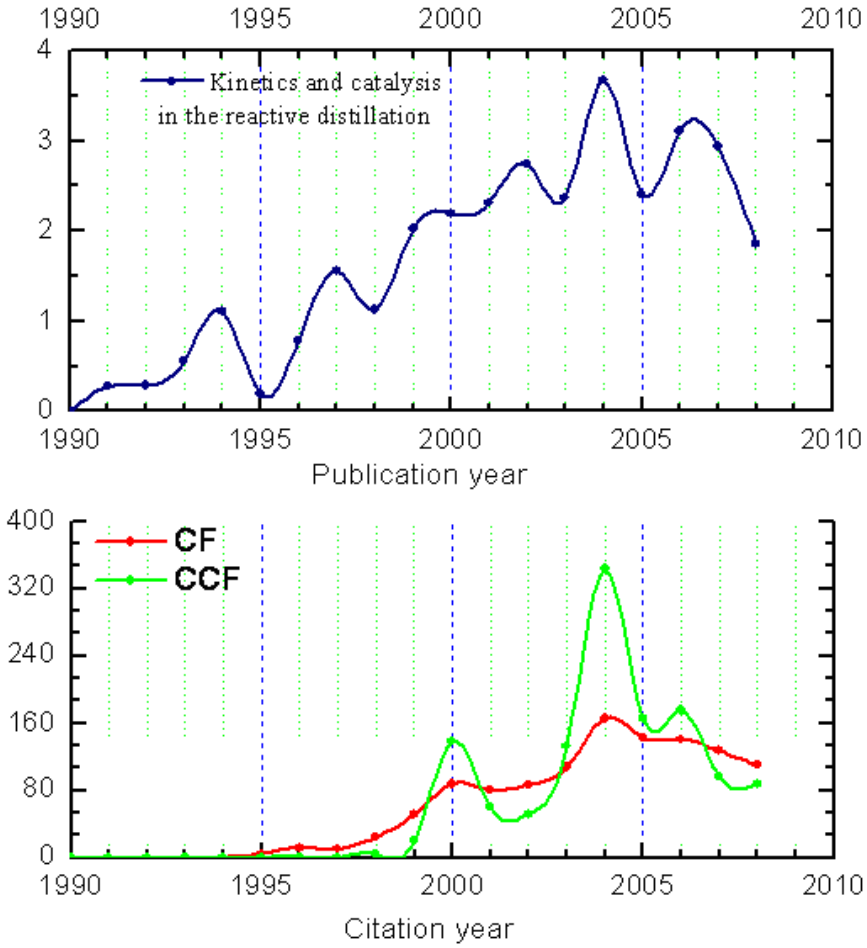


Fig. 5.4 Detection of trends with the CF and CFF functions for the " Kinetics and catalysis in the reactive distillation" thematic cluster

5.3 Variant 3 results

This subsection shows the results for structuring with use of synonyms recognition and trends detected by the methods of financial markets technical analysis. The query was run on 15576 publications in the ISI Web of Science published between 1990 and 2010. The documents that featured the word of "distillation" in the title, keywords or abstract were selected. The downloaded data were stored in the CIT table of the *Article_data* database. After the immaterial "stop-list" words were removed, 35455 words remained. Then, 4183

pairs of similar words were identified and stored in the *WordlistSim* table. Further analysis was limited to words having TFIDF greater than 100 (threshold A) and the threshold B between 14 and 30 percent of the total number of publications in the CIT table, i.e. 4673. After these limitations, the *Cit-Word* table included 1892948 records of the (word, document) pairs. *Word-Similarity*, the other input table, included records from the *WordlistSim* table and 15301 additional records of similarity. In the next step 1659 four words sets were built for the threshold values ThrA and ThrB of: 14 and 0,12 respectively. Additionally, 473 three words sets were built for the threshold values ThrA and ThrB of 14 and 0,19 respectively. Further analysis lead to 108 thematic clusters keeping the threshold values ThrC and ThrD of 0,4 and 2,5 respectively. Eventually, 27 thematic clusters were selected, strictly related to the analysed domain from the point of view of chemical and process engineering. The clusters have been listed in the table 5.2.

Table 5.2. Thematic clusters on distillation for the variant 3.

No	**Name of the thematic cluster**	**Number of Articles**
1	Modelling of the distillation column system	1558
2	Controlling of the distillation column system	1045
3	Designing of the distillation column system	948
4	Distillation process control in steady state	680
5	Vapour - liquid phase model for distillation	626
6	Extractive distillation	545
7	Mass transfer in the packed distillation columns	501
8	Equilibrium aspect in the reactive distillation	494
9	Dynamic and steady state models	493
10	Multiple steady state process	376
11	Catalysis in the reactive distillation	260
12	Membrane distillation	211
13	Supercritical distillation of CO_2 and CO	189
14	Direct contact membrane distillation	137
15	MTBE System	123
16	Ion exchange resins	115
17	Desalination process	111
18	Azeotrope residue curve map	99
19	ETBE system	93
20	Parametrical optimization of the distillation process	71
21	Distillation of alcohols	70
22	Artificial neural network application	67
23	Divinylbenzene	54
24	Equilibrium models – nrtl, and uniquac	54
25	Stefan Maxwell nonequilibrium state model	32
26	Structural optimization of the distillation process	22
27	Bio diesel transestrification	19

The examples of the trend detection for some selected thematic clusters have been presented below.

The detailed results for the "Direct contact membrane distillation" thematic cluster

137 documents published between 1990 and 2010 were assigned to the subject under consideration. The following word collections define it.

contact	direct	membran	dcmd;
contact	direct	membran	flux;
contact	membran	flux	permeat;

The fig 5.5 shows the trends detected for this thematic cluster. The upper part of the figure shows the RNP profile and the trend lines. The middle part show the OSC oscillator profiles for x =2 (oscillator 2) and x=4 (oscillator 4). The lower part of the chart shows the K line and D line for the K%D oscillator with y=4.

The subject under consideration was born about 1996 and the development was started about 2000. An ascending trend was present between 2000 and 2004 as the line I indicates. Both momenta are positive after 2001 what corroborates with the ascending trend. Crossing of the lines K and D near 2001 indicates the ascending trend in this period. Other crossings within the analysed period are not meaningful.

Within the 2004-2010 period the trend is rising (trend line II) quicker than in the previous period. Both the momentum lines lay over zero axis, what confirms the ascending trend. The K and D factors reach 100 between 2007 and 2009, and then the K and D lines cross about 2009. This indicated the descent after 2010. Other methods do not provide meaningful signals about the trend reversal, though the momenta fall after 2009, what may lead to crossing zero after 2010.

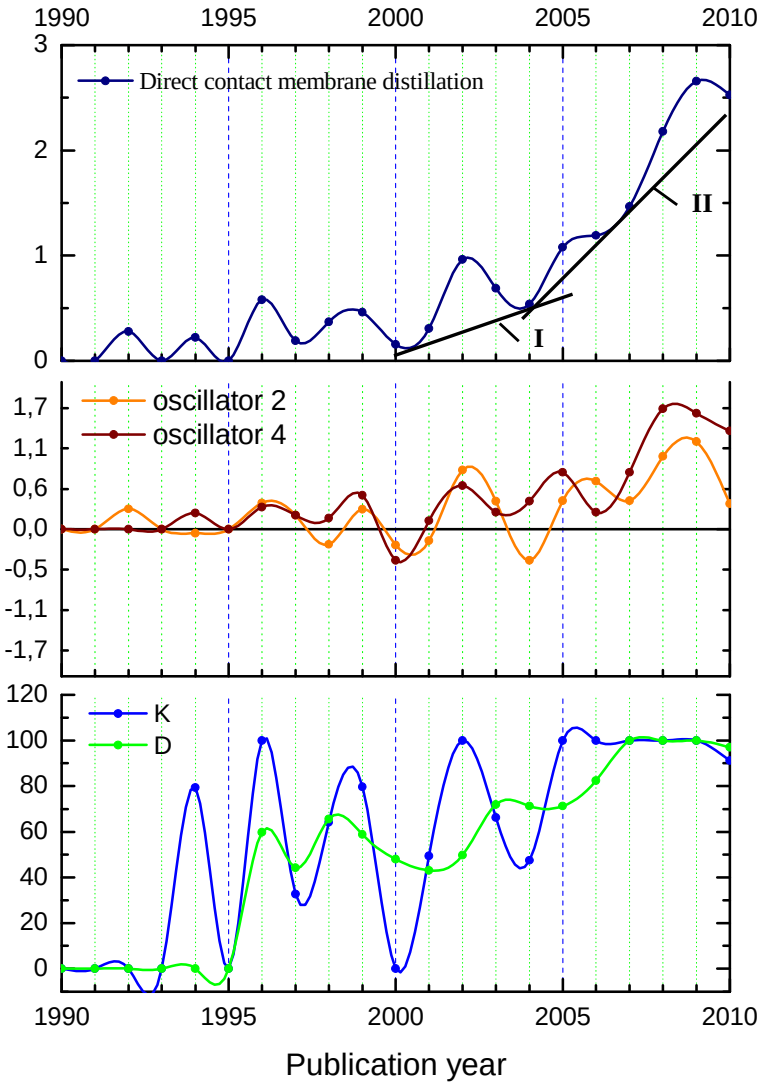


Fig 5.5 Trend detection for the "Direct contact membrane distillation" cluster.

Detailed results for the " Equilibrium aspect in the reactive distillation" thematic cluster

Next analysed thematic cluster contains 494 documents and is defined by the following words collections.

equilibrium	react	sensitiv;
reactiv	react	dynam;
equilibrium	reactiv	react;

The fig 5.6 shows the trends detected for this thematic cluster in a similar way like the previous cluster. The lines colours and styles apply respectively. Three trends can be detected on this chart.

Line I shows a rising trend until 2001. Both momenta are positive after 2001 what corroborates with this trend.

The period 2001-2003 reveals stagnation marked with crossing of the K and D lines about 2000. However, the next crossing in 2003 ends the stagnant period. The momentum of the oscillator 2 is negative while the momentum of the oscillator 4 positive, what translates into an ambiguous signal. This is typical for a temporary stagnant period.

The period between 2004 and 2010 shows the descending trend - line II. Crossing of the lines K and D near 2005 also indicates descent. Momentum of the oscillator 2 indicated possible fall after 2005 what was corroborated with the momentum of the oscillator 4 in 2007. Very low values of the K and D factors after 2009 may signal the rising trend or stagnation after 2010, but the negative value of the momentum after 2009 and the line II indicate continuation of the descent. Forecasting of the trend after 2010 is risky but stagnation seems to be more probable.

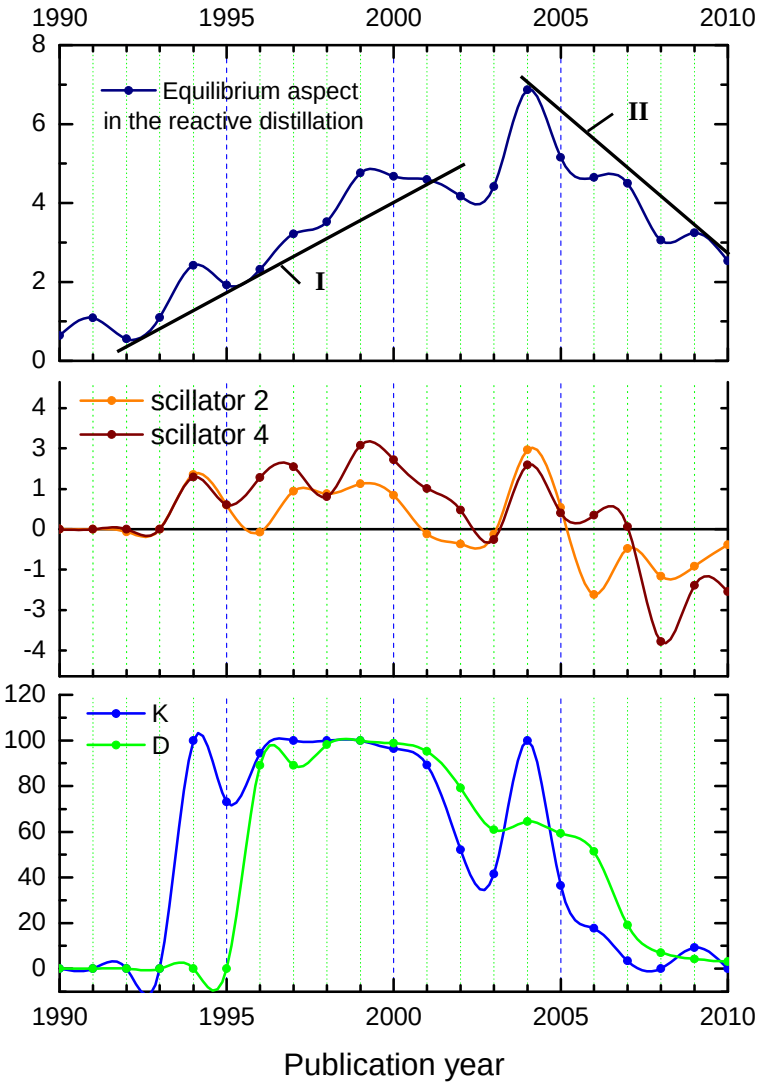


Fig 5.6 Trend detection for the " Equilibrium aspect in the reactive distillation" cluster.

5.4 Variant 4 results

This subsection shows the results for structuring with use of synonyms recognition and trends detected by application of the NEB model. Structuring of the data for this variant have been discussed in subsection 5.3 (variant 3). The trends were detected in the NEB model, and required fitting of the E curve to the RNP. The fitting of E curve was done to the RNP data in the period 1990- 2008. However the RNP data for period 2009 -2010 were used for testing the predictive possibilities of the NEB model. Extrapolation of the E curve until 2012 was meant to forecast the future values of RNP (tend forecast). The E curve was optimised by application of the differential evolution for N=35 members (selected randomly) and M=3000 generations. During the optimisation process the F parameter changes as a linear function of the generation number from 0.8 to 1.3, while the CR parameter was constant CR=0.5. The parameter values were selected by experiment during initial queries. The results of modelling for some selected thematic clusters have been presented below.

The detailed results for the " Modelling of the distillation column system " cluster.

This cluster included 1558 papers published between 1990 and 2010. The following word collections define it.

model	system	column	dynamic;
model	simulate	column	dynamic;
dynamic	system	model;	
dynamic	simulate	system	column;
dynamic	model	system	column;
model	simulate	system	column;
dynamic	model	simulate	column;

The best fit of the E curve of the NEB model until 2008 has been shown on fig. 5.7. The red line on the figure denotes E model curve. The blue stars on the figure denotes RNP data for 1990-2008 period of time. The circles on the chart show the RNP for 2009 and 2010. The variables of the model and optimised values of the variables have been shown in table 5.3.

Table 5.3 - E curve regression results for the NEB model for the " Modelling of the distillation column system" cluster.

	C	d_0	g	w	N_0	E_0
Bottom limit	0	0	0	0	0	0
Top limit	1	1	50	8	17	10
Optimal values of the model	0.01945	0.01862	29.77	3.207	11.48	5.187

The model derived curve on the 5.7 fits very well to the RNP between 1990 and 2008. The ascending trend is visible till 2003, and then the trend descends. Forecasting of the trend by the model E curve for 2009 and 2010 has been confirmed by real RNP values. Further descent is forecasted for this cluster after 2010.

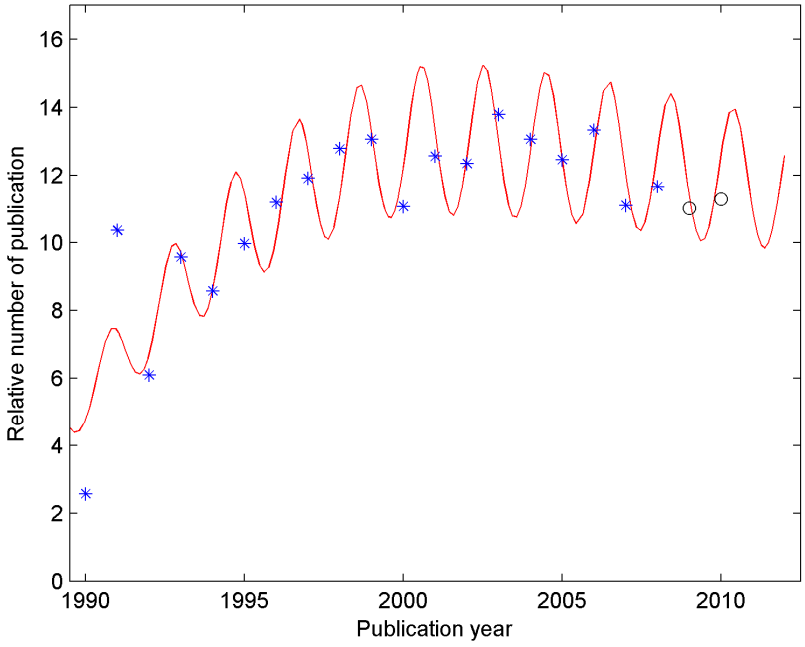


Fig 5.7 Trend detection for the " Modelling of the distillation column system " cluster.

The detailed results for the "Membrane distillation" thematic cluster

The words collections for this cluster of 211 papers have been shown below.

solut	membran	flux	permeat;
membran	porou	flux	permeat;
membran	feed	flux	permeat;
modul	membran	flux	permeat;

The best fit of the E curve of the NEB model until 2008 has been shown on fig. 5.8. The chart also shows the RNP values after 2008 to show the accuracy of prediction. The colours and marks are applied as in the previous sample.

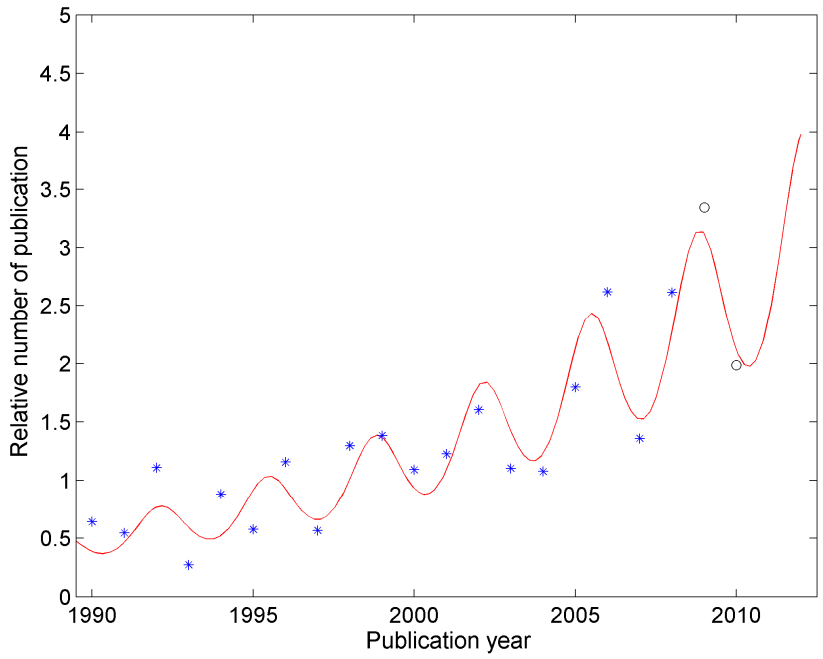


Fig 5.8 Trend detection for the "Membrane distillation" cluster.

The values of the NEB model for the best found solution have been shown in the table 5.4

Table 5.4 - E curve regression results for the NEB model for the "Membrane distillation" cluster.

	C	**d_0**	**g**	**w**	**N_0**	**E_0**
Bottom limit	0	0	0	0	0	0
Top limit	1	1	70	8	20	10
Optimal values of the model	0.006108	0.009398	59.69	1.88	16.15	0.5697

In this case the NEB curve shows a rising trend with increasing oscillations of the RNP. High oscillations at the end of the period may be construed as short term ascending and descending trends. The curve fits well to the RNP data, particularly at the end of the period. Forecasting of the trend by the model E curve for 2009 and 2010 has been confirmed by real RNP values. After 2010 we should expect an increase in the RNP, as compared with 2010.

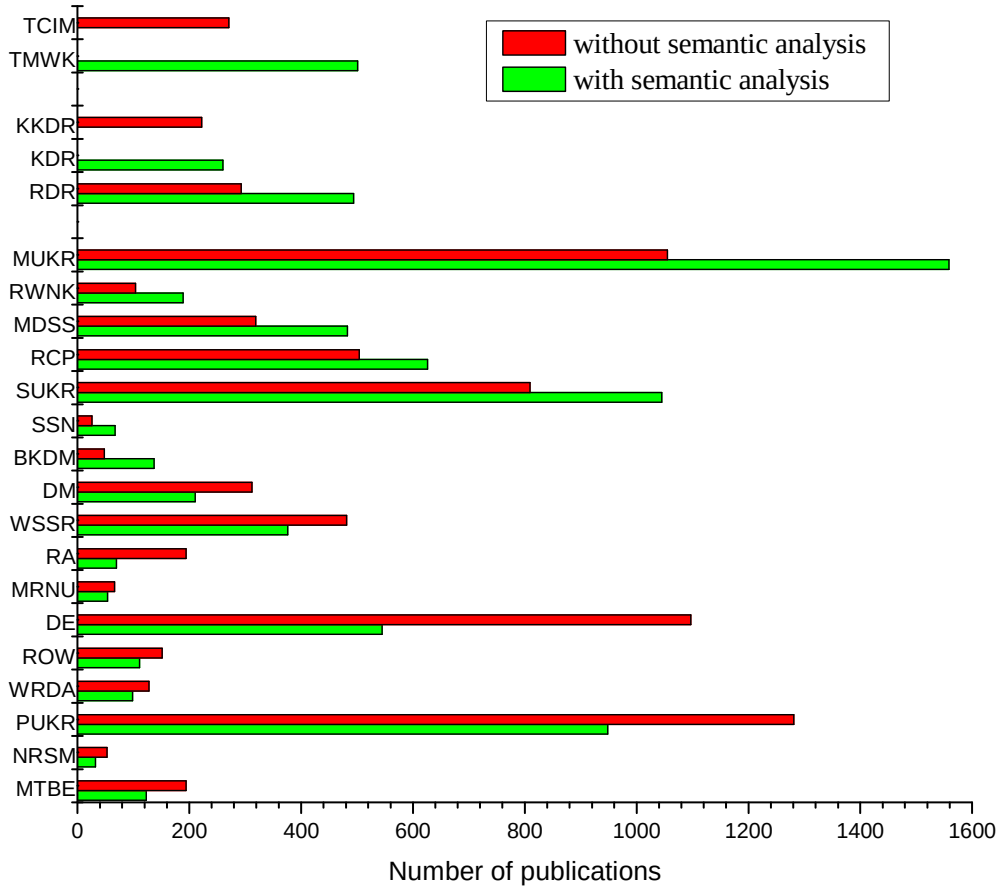
5.5. Analysis of the results for the subjects relating to distillation

This subsection will compare the results found for the designed structuring techniques and the trend detection for the "distillation" domain. First, the structuring in the two cases – with and without semantic analysis – shall be analysed. Then, the subsection will compare the trend detection results that were found for the four variants of combined structuring and trend detection.

5.5.1. Comparison of the Structuring methods

Figure 5.9 shows the size of 20 clusters that have been identified for both structuring analysis i.e. using the simplified and advanced methods. Under the figure, a legend is presented which explains abbreviations of used clusters names. Red bars on the figure denote the size of clusters obtained by simplified structuring method. However green bars denote the size of clusters obtained by advanced structuring method.

As a results of an advanced analysis 27 clusters were found - i.e. less than with the simplified method (38). Total numbers of the clusters found by both methods amounted to 45. This means that 20 clusters were identified by both methods. The shared clusters amount to 44% of all built and 74% of the clusters constructed with the advanced technique. The difference can be explained by a few factors. First, application of the synonyms recognition caused merging similar clusters that were earlier discerned. Second, the advanced method was applied to the higher threshold value ThrB of 0,12 than the simple one (ThrB = 0,09), and also automatic cluster construction of the found words collections was applied. This means that the analyses applied different levels of structuring, and therefore some clusters built with the simplified method were merged when the improved method was applied. In the case of the advanced method, higher threshold value ThrB proved necessary. Application for the advanced technique of the same or lower threshold than in the simplified one lead to construction of too many word sets that would hinder further construction of the thematic clusters.



TCIM Heat and mass transfer in the packed distillation column
TMWK Mass transfer in the packed distillation columns
KKDR Kinetics and catalysis in the reactive distillation
KDR Catalysis in the reactive distillation
RDR Equilibrium aspect in the reactive distillation
MUKR Modelling of the distillation column system
RWNK Supercritical distillation of CO2 and CO
MDSS Dynamic of the steady state model
RCP Vapour - liquid phase model for distillation
SUKR Controlling of the distillation column system
SSN Artificial neural network application
BKDM Direct contact membrane distillation
DM Membrane distillation
WSSR Multiple steady state process
RA Distillation of alcohols
MRNU Equilibrium models – nrtl, and uniquac
DE Extraction distilling
ROW Desalination process
WRDA Azeotrope residue curve map
PUKR Designing of the distillation column system
NRSM Stefan Maxwell nonequilibrium state model
MTBE MTBE System

Fig 5.9. Comparison of the numbers of the thematic clusters constructed with the two structuring methods.

The following observations corroborate with the assumption that the simplified method clusters were merged when the advanced structuring method was applied. Advanced method structured one cluster of water desalination: " Desalination process ", while the simplified method found additional two relating to economy and energy aspects of the process. Other example, a "Membrane distillation", built with the advanced method, reveals also "PVDF membranes" and "Hydrophobic membranes" with the simplified method. Another example was “Modelling of the distillation column system” found with the semantics sensitive method while the simplified revealed also "Modelling of the distillation column", etc.. Therefore, we can conclude, that the clusters built with the advanced method are more coherent, and the similar clusters merged, while the simplified method still discerns them. The advanced method therefore facilitates a more precise construction of subjects and assignment of the documents to the clusters. It is therefore better and more accurate. Further advantage of the semantics sensitive method is ability to find clusters that are not identified by the simplified technique. There are 7 such clusters found, i.e. 16% of all clusters, among them: i.e. "ETBE System" or "Distillation process control in steady state".

There are clusters built by both methods where the detailed names of the cluster differ. There were five such clusters found. These are first shown on the fig 5.9. First two clusters refer to the same subject, but the simple method found "Heat and mass transfer in the packed distillation column" while the advanced method found " Mass transfer in the packed distillation columns". The following three clusters on this drawing, refer to the reactive distillation, with the provision that although the cluster referring to the equilibrium of this process has been found by both methods, the remaining clusters differ. In the case of the simplified method, the cluster " Kinetics and catalysis in the reactive distillation" was found while the advanced method found " Catalysis in the reactive distillation" The differences are caused by different structuring methods applied.

Furthermore, comparison of the numbers of the matching clusters indicates, that the first 10 subjects on the fig. 5.9 are more numerous when the synonym sensitive method is applied. This means, that the synonym sensitive methods assign more publications to certain clusters, because identification of the synonyms leads to greater number of documents being assigned to a cluster. Such subjects make about a half of all clusters built in structuring, but some of those found with the advanced method are smaller. This is explained by the fact, that the words belonging to the fingerprints precisely discern the subjects and therefore smaller number of documents are assigned there. Thus, only the documents closely related to the

subject are assigned to a respective cluster. Therefore the generalisation, that one of these methods leads to more numerous clusters is not justified.

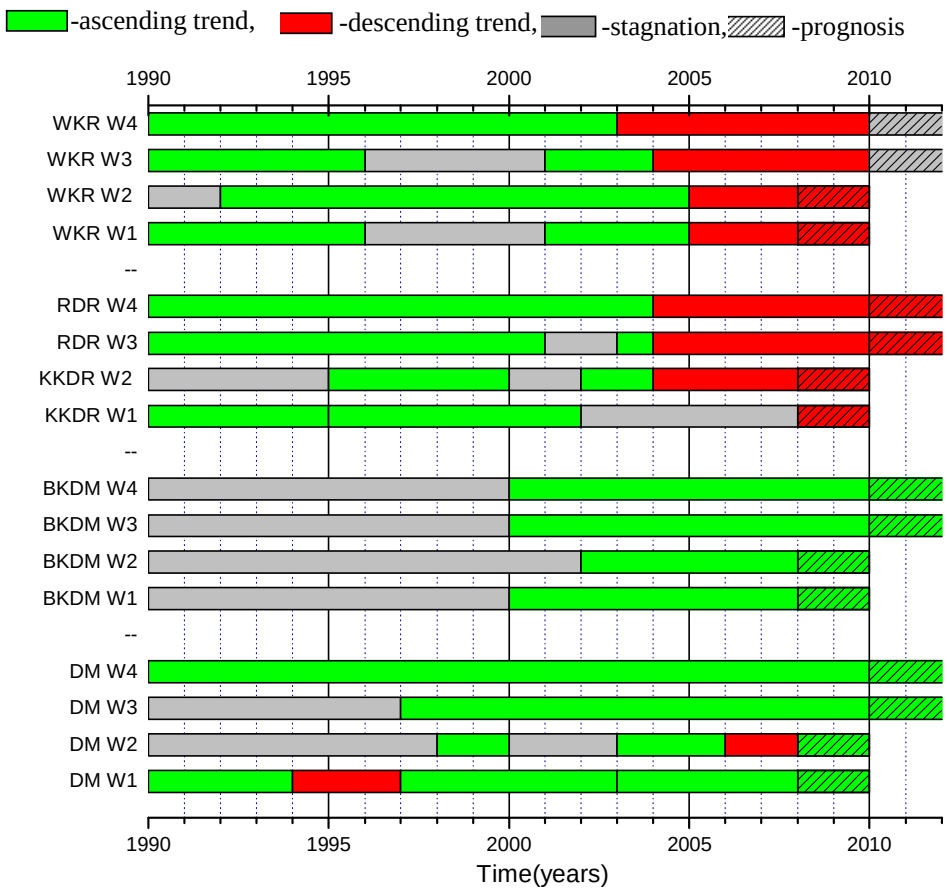
5.5.2. Comparison of the trend detection methods

The presented results of the trend detection for the four variants of combined structuring and trend detection allow for comparison of the results. Figure 5.10 presents comparison of the trend detection and forecasting results for four example of the thematic clusters. The chart uses multicolour bars to mark the trends detected for the specific periods. The abbreviations list has been included below the chart.

The first of the analysed clusters refers to "Azeotrope residue curve map". The trend detection results refers to:

- variant 1 (structuring with the simplified method and trend detection with the methods of financial markets);
- variant 2 (structuring with the simplified method and trend detection with the CF and CFF co-citation analysis);
- variant 3 (structuring with the advanced method and trend detection with the methods of financial markets);
- variant 4 (structuring with the advanced method and trend detection with the NEB model).

It is easy to observe that the trends detected with the methods 1, 3 and 4 are practically identical. An ascending trend was detected for the period up to 2004, with a stagnant period between 1996 and 2001, and the descending trend afterwards. This means, that the structuring method had little impact on the detection of trends. However, there are differences between the variant 1 and 2. Both variants see a descending trend after 2005, but while the variant 2 detects grow from 1992 to 2005, the variant 1 sees stagnation during this period. This is not a significant difference and generally, the types of the detected trends are matching. This means, that the trend detection method was not essential here. In general, it could be concluded, that the ascending trend with a temporary stagnation dominated until 2004, while the period after 2005 is marked with a decent. Both variants 1 and 2 forecast the same trend after 2008. Most important, the forecast has been confirmed with the real fall between 2008 and 2010, with the analysis of the variant 3 and the variant 4. This shows, that the presented methods are useful tools of trend forecast.



WKR W4	Azeotrope residue curve map variant 4
WKR W3	Azeotrope residue curve map variant 3
WKR W2	Azeotrope residue curve map variant 2
WKR W1	Azeotrope residue curve map variant 1
RDR W4	Equilibrium aspect in the reactive distillation variant 4
RDR W3	Equilibrium aspect in the reactive distillation variant 3
KKDR W2	Kinetics and catalysis in the reactive distillation variant 2
KKDR W1	Kinetics and catalysis in the reactive distillation variant 1
BKDM W4	Direct contact membrane distillation variant 4
BKDM W3	Direct contact membrane distillation variant 3
BKDM W2	Direct contact membrane distillation variant 2
BKDM W1	Direct contact membrane distillation variant 1
DM W4	Membrane distillation variant 4
DM W3	Membrane distillation variant 3
DM W2	Membrane distillation variant 2
DM W1	Membrane distillation variant 1

Fig 5.10 Comparison of the detected trends of similar clusters by the methods applied

The second presented cluster concerns reactive distillation. The fig. 5.10 shows the trends of this subject, but in the clusters distinguishing between various aspects of this process. These clusters show ascending trend until 2004 and then the trend descends. Despite that the clusters varied slightly, the detected trend are consistent for all applied analyses. It is worth noticing that the variant 1 and the variant 2 forecasted consistently descending trend after 2008, and this descent was confirmed with the variant 3 and the variant 4, what confirms forecasting ability of the proposed methods.

The next cluster under consideration was the "Direct contact membrane distillation". Almost identical trends were detected by every variant of analyses here. The subject began developing at about 2000 and grew for the entire analysed period. Variants 1 and 2 forecasted the ascending trend after 2008 what was confirmed with the real growth shown with the variants 3 and 4 in the period between 2008 and 2010.

The "Membrane distillation" is the last subject analysed on the fig. 5.10 and shows the ascending trend for the entire period. However, there are temporary slowdowns in this cluster between 1994 and 1997 and between 2006 and 2008, which were detected by all methods (except for the NEB) and recognised as a fall or stagnation. The NEB model shows a long term rising trend with the increasing oscillations of the RNP. However, the oscillations can be interpreted as short term ascending and descending trends, what would corroborate with the signals of slowdowns detected by other methods. In this case the variant 1 and the variant 2 consistently forecast an ascending trend after 2008. The forecast has been confirmed by the analysis marked as variant 3 and 4 when real ascending trend was detected after 2008.

The following conclusions can be drawn after the analysis of the trend detection. For certain subjects (e.g. "Direct contact membrane distillation") the choice of the trend detection method or the structuring method was insignificant, and all techniques used returned the same trends. In other cases (e.g. " Azeotrope residue curve map ") slight and insignificant variations were seen, but the main trends were easy to detect. I most cases the main trends were detected and forecasted by all the methods applied even if the clusters were not identical. However, there were only few cases (e.g. the subjects concerning optimising), where the applied techniques revealed differing trends for the same time periods. This means that although the probability of correct trend detection is very big, there remains some margin of uncertainty. To minimise the risk of prediction, a number of independent tools of trend analysis is important. The techniques applied for the purpose of this dissertation utilise varying methods of trend detection. These methods based on the methods of financial markets technical analysis applied to RNP should detect similar trends. The NEB model detects only long term

trends while individual short term trends are represented by oscillations. This fact leads to differences between the trends detected by different methods. The NEB model advantage as compared with other methods, is a long term prediction of trends. The CF, CFF method detects trends basing on the co-citations number, and therefore differs considerably from other techniques.

It is noteworthy, that almost all discussed cases returned consistently similar trends after 2008. Moreover, these forecasts were indeed confirmed afterwards. This means the trend forecasts resulting from the discussed techniques are highly probable.

The trend detection and forecast is uncertain and difficult, what leads to a conclusion that high reliability of the forecast would require application of all designed methods of trend detection. Use of only one of these techniques increases probability of error.

5.6 Main trends of distillation

The clusters built for the distillation domain reveal the following development profiles: Some of them (e.g.: " Optimization and designing of the distillation model ": or "Modelling of the distillation column system", found with the simplified method) revel an ascending trend in the beginning, than stagnation and the descent afterwards. Other clusters (e.g. "Membrane distillation" or "Direct contact membrane distillation") develop strongly after a period of stagnation. The clusters found in the improved structuring process also reveal varying development dynamics. Some clusters (e.g. :"Azeotrope residue curve map" or "Equilibrium aspect in the reactive distillation") are currently descending after an earlier period of strong development. Other clusters, (e.g. "Direct contact membrane distillation" or "Membrane distillation") still show an ascending trend. There were also new clusters found, that feature a strong growth ("Bio diesel transestrification ").

The research reveals that distillation is a well developed and slowly developing domain. There are 8 developing clusters specified below:

Extractive distillation
Esterification
Equilibrium models – nrtl, and uniquac
Membrane distillation
Direct contact membrane distillation

Economy aspects of distillation

Distillation of alcohols

Bio diesel transestrification

Chapter 6. Structure of the research area - filtration

The proposed methods fascinated structuring with trend detection and forecasting for the clusters relating to distillation. To verify efficiency of the designed methods, queries concerning filtration were performed. The research was run in the same way as variant 3 of distillation – structuring with synonyms recognition and trend detection with the methods of financial markets.

The *ISI Web of Science* contains 72256 papers published between 1990 and 2010, whose titles. keywords or abstracts contained the word "filtration". The data were read into the *CIT* table in the *Publication_data* database. During the initial analysis the papers published in journals not related to the chemical and process engineering (e.g. medical, biological, veterinary) were removed. Since there are uses of the term "filtration" that do not refer the process known in the chemical and process engineering, this operation was necessary to speed up the analysis and limit the number of clusters not related to the subject under consideration. After that, there were 25703 publications remaining in *CIT*. Further analysis lead to identification of 52802 significant words and 3020 synonym pairs, Further analysis was limited to the well discerning words: having the TFIDF >100, and satisfying the threshold B limitation: 14<= presence in publications <=7711 (upper limit equals 30% of the number of the *CIT* records). After the selection there were 7032 words remaining. The *Cit-Word* input table of the *WordSet* program had 2768769 records, while the other input table contained the records of the *Wordlistsim* table plus 2014 additional word similarity records. Further analysis lead to 3061 four words sets for the threshold values of 14 and 0,14 respectively. Further analysis lead to 1393 three words sets, with the threshold values of 14 and 0,16 respectively. Last stage of structuring lead to 328 thematic clusters in the "filtration" domain with the threshold values ThrC and ThrD of 0,4 and 2,5 respectively. Eventually, 23 clusters strictly connected to filtration were selected. They are listed in the table 6.1.

The examples of the trend detection for two selected thematic clusters have been presented below.

Table 6.1 Clusters concerning "filtration".

No	Name of the thematic cluster	Number of Articles
1	Ultrafiltration	3423
2	Filtration in wastewater treatment	3021
3	Membrane fouling	2747
4	Operation of membrane filtration	2531
5	Modeling of the membrane filtration	2502
6	Efficiency of ultrafiltration	2353
7	Modification of membrane material	2094
8	Determination of flux in filtration process	1793
9	Permeate properties	1655
10	Model of media transport in pores	1518
11	Membrane porosity	1331
12	Efficiency of membrane filtration	1121
13	Cake resistance in microfiltration	834
14	Cake properties	785
15	Reverse osmosis	750
16	Resistance in the microfiltration	746
17	Microfiltration of suspensions	653
18	Sludge filtration	638
19	Nanofiltration	531
20	Application of filtration in the desalination process	122
21	Electrospinning in membrane fibres manufacturing	68
22	Artificial neural networks in modelling of the membrane filtration	60
23	Van der Waals forces in membrane processes	23

The detailed results for the "Efficiency of membrane filtration" thematic cluster

This cluster included 1121 papers published between 1990 and 2010. The words collections representing the subject have been listed below:

flux	increase	permeat;	
membran	flux	increase	microfiltration;
ultrafiltration	membran	flux	increase;
membran	foul	flux	increase;

The fig 6.1 shows the trends detected for this thematic cluster. The upper part charts the RNP profile and the trend lines. The chart in the middle shows the momenta for the OSC for x=2 (oscillator 2) and x=4 (oscillator 4). The lower part of the chart shows the K line and D line for the K%D oscillator with y=4. There are 4 main trends found in the analysed period.

The first period until 1998 shows a rising trend (I). This is confirmed with the positive momentum value.

The next period between 1999 and 2003 is stagnant (trend line II). The momenta values are close to 0, what confirms the stagnant trend during this period. The lines K and D of the K%D cross a number of times during this period what also confirms the stagnant trend.

The K and D crossing in 2003 and positive momentum after 2004 mark the ascending trend between 2003 and 2006. Crossing of the lines K and D near 2006 indicates the reversal and descent in the fourth period. The descent between 2006 and 2010 is confirmed with the trend line III, and the analysis of the results leads to forecast of further descent after 2010.

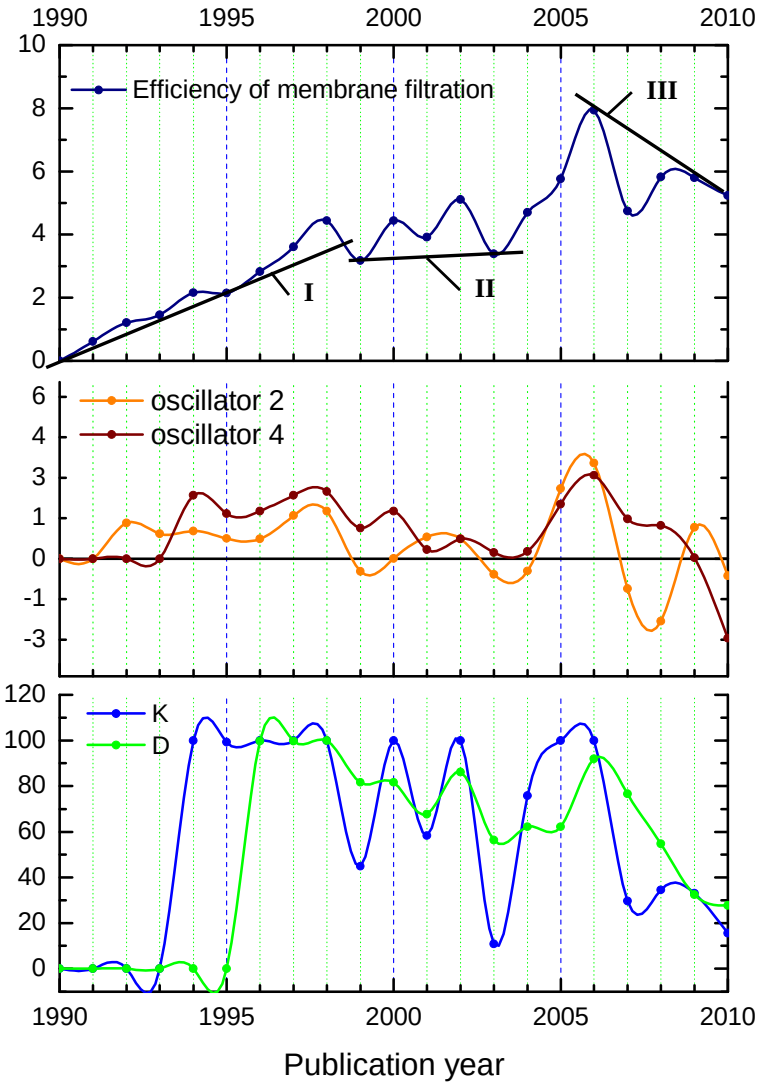


Fig. 6.1 "Efficiency of membrane filtration" cluster – RNP, the trend line method and the OSC method.

The detailed results for the "Nanofiltration" thematic cluster

The nanofiltration cluster has 531 papers published between 1990 and 2000. The following word collections define it.

reject	nanofiltration	nf;	
membran	reverse	nanofiltration	osmosi;
reverse	reject	nanofiltration	osmosi;
uf	nanofiltration	nf;	

The fig 6.2 shows the trends detected for this thematic cluster. The results are depicted in the same manner as in the previous example.

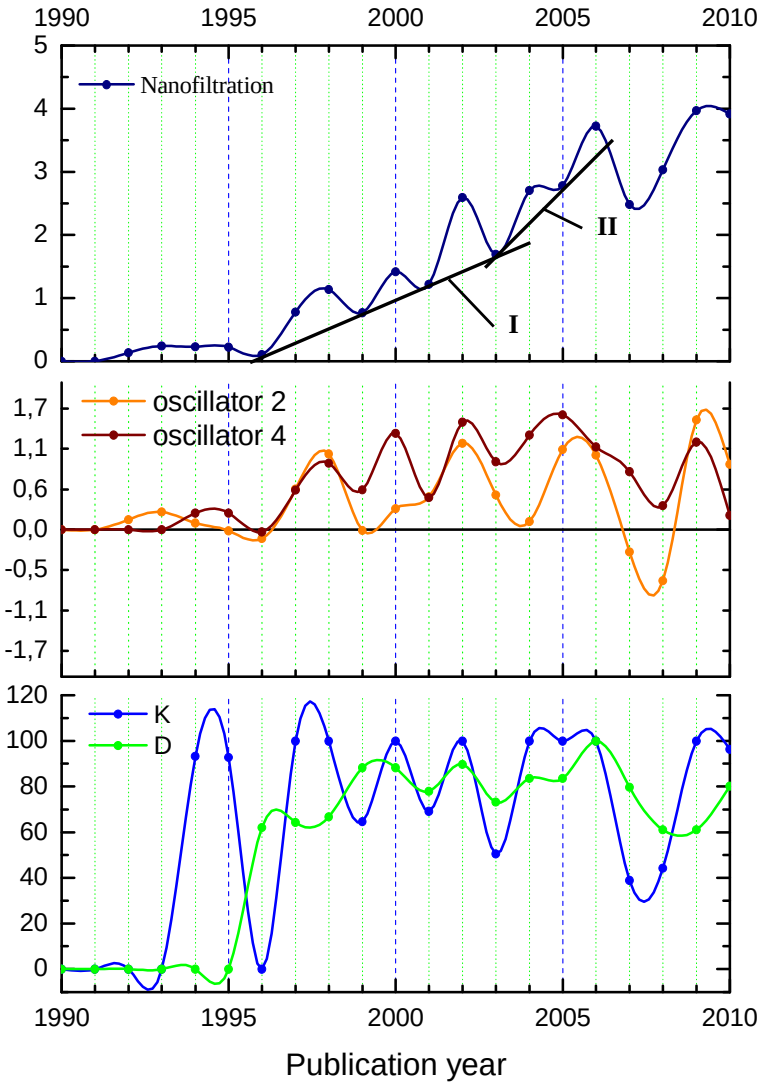


Fig. 6.2 "Nanofiltration" cluster – RNP, the trend line method and the OSC method.

Analysis of the fig. 6.2 indicates into a rising trend until 2006 (trend line I) The trend reverses between 2007 and 2008 and continues to rise until 2010. The OSC show the ascending trend along the entire period. Crossing of the F and D lines around 2006 signals a descent, but next crossing in 2008 returns to the ascending trend after 2008. Currently, there are no strong signals about trend reversal after 2010.

Trends in filtration

Like distillation, filtration is a well developed domain. There is a strong subdomain related to the membrane processes. There are two types of trend changes observed. Certain clusters (e.g. "Efficiency of membrane filtration") are currently falling after previous development. Other clusters (e.g. " Determination of flux in filtration process " or " Operation of membrane filtration") reveal an ascending trend after a period of stagnation. Some clusters (e.g. "Nanofiltration" or "Ultrafiltration") reveal a continuous rising trend. Eventually, "Permeate properties" or "Modeling of the membrane filtration" are stagnant after a period of development. The domain reveals the following developing clusters:

Efficiency of ultrfiltration
Determination of flux in filtration process
Operation of membrane filtration
Modification of membrane material
Nanofiltration
Reverse osmosis
Ultrafiltration
Resistance in the microfiltration
Model of media transport in pores

Chapter 7. Conclusions

Both finding of the relevant and desirable publications and recognition of the promising research subjects justifying investment and further study meet considerable difficulties because of the huge number of publications that is still growing exponentially. Therefore it is important to find the developing and new research subjects where further development might be expected. This information is very useful for the decision making about research funding and for the scientists interested in intensifying the research in the most quickly developing areas. The main goal of this dissertation has been a construction of the tool and methods supporting the decision making process about funds allocation in the specific research areas.

The conducted research lead to construction of two methods of the knowledge domain structuring, combined with three trend detection techniques. The first proposed method of cluster construction involved structuring without the synonyms recognition. This simplified method is fast and efficient, and the resulting thematic clusters cover the entire knowledge domain under analysis. Because the words sets are assumed as unit elements of knowledge, this method solves the problem of word ambiguity, that affects the more advanced methods of visual maps of knowledge, where individual words are treated as the atomic elements of knowledge. Furthermore, each selected word can represent more than one cluster, and each document can be assigned to more than one cluster, what provided for flexibility of the technique and allows for precise construction of the thematic clusters. The proposed structuring technique is also universal and can be applied to many domains of knowledge, what can be shown by the results found for various research areas (distillation, filtration), which show varying levels and speeds of development.

The second approach to structuring involved synonyms recognition. The open WordNet database was used for synonyms finding. This allows for more precise construction of the thematic clusters because the synonyms are recognised and the spelling differences ignored. Further improvement involved more accurate selection of words for the fingerprints of the thematic clusters, by application of the TFIDF.

The technique allows for both manual and automatic composing of the words sets in the thematic clusters. Manual operation is more accurate but time consuming and it requires expertise in the subject under consideration, while automatic clustering is fast and objective, though less accurate.

As confirmed by the examples presented in the subsection 5.5. simplified structuring returns more thematic clusters while application of the synonyms recognition is more precise and merges similar clusters. For instance desalination has three simplified clusters and only one when the synonyms are recognised. This means that synonym recognition leads to more accurate results, however on the other hand the simplified method is faster. Therefore, synonyms recognition should be applied to underdeveloped subjects where the number of publications is relatively low. Simplified method should be applied to analysis of the large domains where the number of publications is big. In this case, precision in construction of the thematic clusters is less important while the simplified method saves time.

The designed methods provide reliable results what is confirmed by the clusters returned in the analysis of the distillation domain. However, there are ways to improve the proposed techniques. First, the proposed methods require considerable computing power and the structuring takes quite a long time, so it is desirable to make the tools faster. One of the possible improvements is optimization of the software code or construction of the procedures preselecting the analysed documents, what would remove the documents insignificant for the analyses of the knowledge domain. Also synonyms recognition seems to need improvements, since the current method relies on the WordNet database. Unfortunately, this method recognises as synonymous also words that have only remote resemblance. Therefore, a method of selection of the meaningful synsets seems necessary. Another improvement may be a better criterion for fingerprints finding, that would be more restrictive than the currently applied TFIDF. Yet another improvement would involve more effective method of assignment of the word sets to the thematic clusters, or definition of another similarity measure for the aggregated objects.

The proposed structuring and trend detection techniques can be combined, what sometimes is not easy for the methods known in literature. The level of development of the subject is represented by the RNP what provides a better measure than the row number of publications, which grows every year and hampers trend detection.

The applied methods borrowed from technical analysis of the financial markets accurately detect the current and forecasted trends. Note, that these techniques do not require any computing power, therefore the trends are detected quickly and easily. Nonetheless, the discussed methods must be fed with the appropriate parameter values for current operation. Despite this, there are cases when one of the techniques detects a trend not confirmed by the others. This means that the techniques are not 100% accurate and should be used together to make the trend detection and forecast more reliable.

The other approach of trend detection analyses the citations and co-citations numbers and is very important for the task in question. This method perceives the development of the cluster as "use of knowledge" while other techniques rely on the RNP, i.e. "production of knowledge". The curves representing the forecasting CF and CFF functions are smooth and there are no significant oscillations, what leads to a transparent and easy trend detection. The trends detected with this technique are usually consistent with the trends detected by other methods dealing with the RNP. The technical analysis and the citation functions return very similar trends for the matching thematic clusters, but certain results differ. Therefore, application of the both methods allows for more reliable conclusions than use of the one only. The method basing on the forecasting functions requires a laborious preparation of the "Powiazania" table and the computations of the functions values. Therefore this method is relatively slow,

The third method of trend detection applied the NEB mathematical model for the purpose of finding of time evolution of the RNP. This method detects trends very well and by fitting of the E curve to the RNP we are able to find what is the development stage of the cluster under consideration and predict future RNPs. The proposed NEB model very accurately depicts changes of the RNP in time, and allows for long term trend forecast. Forecasting ability of the proposed model has been confirmed with the queries. As shown on the sample results for the distillation domain, the forecasted trends were indeed confirmed in the measured RNP. Furthermore, this method detects mostly the long terms trends, while certain oscillations reflect the short term trends. This technique is therefore very reliable, but it requires considerable computing power and is the slowest of all proposed trend detection techniques. The applied Matlab procedure (ODE15s) limits the time of solving by optimising the parameters that offer the best fit. It can be speed up with an alternative method for solution of the differential equations or provision of another model, which would quickly return the detected and forecasted trends.

The trend detection is a very difficult task, and therefore it is advisable to use all of the proposed methods to forecast the trends. The methods are independent and the use of each next, provides for more reliable results.

Analysis of the results presented herein, indicates that the detected and forecasted trends were consistent across the applied methods. The most important is however, that the forecasted trends for after 2008 were confirmed by the real values of RNP. This confirms verifiability of the trend forecast by the proposed methods. The proposed methods facilitate

structuring of the domain of knowledge and detection of the current and future research trends, while the results confirm that they are useful, applicable, effective and universal.

The research reported herein indicates the developing areas for the selected research domains, and detects the stagnant subjects and those enjoying a diminishing interest of the researches. This information is important for the researches and for those who construct portfolios of the R&D projects. There were 8 developing clusters relating to distillation:

Extractive distillation
Esterification
Equilibrium models – nrtl, and uniquac
Membrane distillation
Direct contact membrane distillation
Economy aspects of distillation
Distillation of alcohols
Bio diesel transestrification

Other subjects e.g. relating to distillation with chemical reaction or "Azeotrope residue curve map" show descending or stagnant trends.

There were 9 developing clusters relating to filtration:

Efficiency of ultrfiltration
Determination of flux in filtration process
Operation of membrane filtration
Modification of membrane material
Nanofiltration
Reverse osmosis
Ultrafiltration
Resistance in the microfiltration
Model of media transport in pores

Other subject in this domain are descending or stagnant. The examples include: "Efficiency of membrane filtration" or "Modelling of the membrane filtration".

The dissertation proves that the research trends for a given research branch are predictable upon the past profiles, what has been confirmed by the results of trend detection. Furthermore, it has been demonstrated that mathematical modelling of the development dynamics of a research branch is possible, what in turn allows for trend forecasting. These statements prove, that the research hypotheses in the dissertation were sound. It has also been proved, that the tools and methods designed allow for selection of the developing subjects of research. The proposed methodology facilitates therefore: definition of the thematic structure of the knowledge domain, depiction of the history of development of the built thematic clusters and forecast of future research trends. The results presented in the dissertation are essential for the persons involved in the financial decision making concerning the research funding and for those balancing the portfolios of the R&D projects. Furthermore the results are essential for the researchers, who contemplate choice of the research area, and plan the research projects. It shall be noted, that this research contributes in the development of Science, by means of identification of the significant scientific information and the developing branches of knowledge.

References

Ahmed, S. R. (2004). Effectiveness of neural network types for prediction of business failure. Information Technology: Coding and Computing, 2, 455–459.

Amini, M. R., Goutte, C. (2010). A co-classification approach to learning from multilingual corpora. Machine Learning, 79(1–2), 105–121.

Ankerst, M., Breunig, M. M., Kriegel, H.-P., Sander, J. (1999)Optics: ordering points to identify the clustering structure. SIGMOD Rec., vol. 28(2), pp. 49–60.

Banares-Alcántara R.(2010). Perspectives on the potential roles of engineers in the formulation, implementation and enforcement of policies. Computers and Chemical Engineering 34, 267–276.

Barnett, G. A., Huh, C., Kim, Y., Park, H. W. (2011). Citations among communication journals and other disciplines: a network analysis. Scientometrics, 88(2), 449–469.

Baskurt O.K. (2011). Time series analysis of publication counts of a university: what are the implications? Scientometrics, 86, 645- 656.

Borner, K., Chen, C., & Boyack, K. W. (2003). Visualizing knowledge domains. Annual Review of Information Science and Technology, 37(1), 179–255.

Bortiz, J. E., & Kennedy, D. B. (1995). Effectiveness of neural network types for prediction of business failure. Expert Systems with Applications, 9, 503–512.

Braam R.R., Moed H.F., Van Raan A.F.J. (1991a). Mapping of science by combined cocitation and word analysis .1. Structural aspects. Journal of the American Society for Information Science, 42, 233-251.

Braam R.R., Moed H.F., Van Raan A.F.J. (1991b). Mapping of science by combined cocitation and word analysis .2. Dynamic aspects. Journal of the American Society for Information Science, 42, 252-266.

Broda B. Mechanizmy grupowania dokumentów w automatycznej ekstrakcji sieci semantycznych dla języka polskiego. Master Degree Thesis. Warsav University of Technology 2007.

Callon M., Courtial J. P., Laville F. (1991). Co-word analysis as a tool for describing the network of interactions between basic and technological research: The case of polymer chemistry. Scientometrics, 22, 153–205.

Cambrosio, A., Limoges, C., Courtial, J. P., & Laville, F. (1993). Historical scientometrics? Mapping over 70 years of biological safety research with coword analysis. Scientometrics, 27(2), 119–143.

Chang C.T., Lai J.Z.C., Jeng M.D. (2010). Fast agglomerative clustering using information of k-nearest neighbors. Pattern Recognition, 43, 3958-3968.

Chen, C. Generalised Similarity Analysis and Pathfinder Network Scaling. Interacting with Computers, 10, 2 (1998), 107-128.

Chen, C. (2003) Mapping Scientific Frontiers: The Quest for Knowledge Visualization (Springer, London).

Chen C. CiteSpace II: Detecting and Visualizing Emerging Trends and Transient Patterns in Scientific Literature. Journal of the American Society for Information Science and Technology, 57(3):359–377, 2006

Chu P.Y.V., Hsu Y.L., Fehling M. (1996). A decision support system for project portfolio selection. *Computers In Industry*, 32, 141-149.

Clarke, A., Gatineau, M., Thorogood, M., Wyn-Roberts, N. (2007). Health promotion research literature in Europe 1995–2005. The European Journal of Public Health, 17(Supplement 1), 24–28.

Clift, R. (2006). Sustainable development and its implications for chemical engineering. Chemical Engineering Science, 61, 4179–4187.
Cottrill, C. A., Rogers, E. M., Mills, T. (2010). Co-citation analysis of the scientific literature of innovation research traditions diffusion of innovations and technology transfer. Journal of Information Science, 36(3), 383–400.
Coulter, N., Monarch, I., Konda, S. (1998). Software engineering as seen through its research literature: A study in co-word analysis. Journal of the American Society for Information Science, 49(13), 1206–1223.
Courtial, J. P. (1994). A coword analysis of scientometrics. Scientometrics, 31(3), 251–260.
Cui, L., Liu, W., & Yan, L. (2008). Development of a text mining system based on the co-occurrence of bibliographic items in literature databases. New Technology of Library and Information Service, 24(8), 70–75.
Ding Y., Chowdhury G.G., Foo S. (2001). Bibliometric cartography of information retrieval research by using co-word analysis. Information Processing and Management, 37, 817 – 842.
Dittenbach, M., Merkl, D., Rauber, A. Using Growing Hierarchical Self-Organizing Maps for Document Classification. W Proc of the European Symposium on Artificial Neural Networks (ESANN 2000), pp. 7–12, D-Facto Publications, Bruges, Belgium, 2000.
Ebel, H., Davidsen, J., & Bornholdt, S. (2002). Dynamics of social networks. Complexity, 8(2), 24–27.
Ester, M., Kriegel, H.-P., Sander, J., Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. W Second International Conference on Knowledge Discovery and Data Mining, (E. Simoudis, J. Han, U. Fayyad, eds.), pp. 226–231, AAAI Press, Portland, Oregon, 1996.
Evans R., Hinds S., Hammock D. (2009). Portfolio analysis and R&D decision making. *Nature Reviews Drug Discovery* , 8, 189-190.
Fabry B., Ernst H., Langholz J., Köster M. (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, 215-225.
Fayyad, U. M., Piatetsky-Shapiro, G., & Smyth, P. (1996a). From data mining to knowledge discovery: an overview. In: Advances in Knowledge Discovery and Data Mining (pp. 1–34). California: American Association for Artificial Intelligence.
Fayyad, U. M., Piatetsky-Shapiro, G., & Smyth, P. (1996b). The KDD process for extracting useful knowledge from volumes of data. Communications of the ACM, 39(11), 27–34.
Fellbaum C. (1998). WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press. Princeton University "About WordNet." WordNet. Princeton University. 2010.
Forster, R. Document Clustering in Large German Corpora Using Natural Language Processing. PhD Thesis, University of Zurich, 2006.
Garfield E., History of citation indexes for chemistry—a brie review, JCICS 25 (3) (1985) 170–174.
Garfield, E. (1994). Scientography: Mapping the tracks of science. Current Contents: Social & Behavioural Sciences, 7(45), 5–10.
Goldszmidt, M., Sahami, M. A probabilistic approach to full–text document clustering. Raport tech., SRI International, 1998.
Gowda K.C., Ravi T.V. (1995). Agglomerative clustering of symbolic objects using the concepts of both similarity and dissimilarity. Pattern Recognition Letters, 16(6), 647-652.
Granovetter, M. (1973). Strength of weak ties. American Journal of Sociology, 78(6), 1360–1380.

Haykin S. Neural Networks: A comprahensive foundation, Prentice Hall, Upper Saddle River, NJ 1990
He, Q. (1999). Knowledge discovery through co-word analysis. Library Trends, 48(1), 133–159.
He D., Earn D.J.D., (2007). Epidemiological effects of seasonal oscillations in birth rates. Theoretical Population Biology, 72, 274–291.
Hossain M.D., Moon J., Kang H.G., Lee S.C., Choe Y.C. (2012). Mapping the dynamics of knowledge base of innovations of R&D in Bangladesh: triple helix perspective. Scientometrics, 90, 57-83.
Hou, H., Liu, Z., & Chen, Y. (2006). Mapping of science studies: The trend of research fronts. Science Research Management, 27(3), 90–96.
http://rapid-i.com/ Accessed: 21.04.2012.
http://wordnet.princeton.edu. Accessed: 15.10.2011.
Hull D., Aït-Mokhtar S., Chuat M., Eisele A., Gaussier E., Grefenstette G., Isabelle P., Samuelsson C., Segond F. (2001). Language technologies and patent search and classification. World Patent Information, 23, 265-268.
Hung, S. W., Wang, A. P. (2010). Examining the small world phenomenon in the patent citation network: a case study of the radio frequency identification (RFID) network. Scientometrics, 82(1), 121–134.
Janssens F., Leta J., Glanzel W., De Moor B. (2006). Towards mapping library and information science. Information Processing and Management, 42, 1614–1642.
Kermack W.O., McKendrick A.G. (1927). A Contribution to the Mathematical Theory of Epidemics. Proceedings of the Royal Society, London, 115, 700-721.
Klapka J., Pinos P. (2002). Decision support system for multicriterial R&D and information systems projects selection Source: *European Journal Of Operational Research*, 140, 434-446.
Kohonen, T. Self-organized formation of topologically correct feature maps, Biological Cybernetics, 43, 59-69. 1982
Kohonen, T., Kaski, S., Lagus, K., Salojrvi, J., Honkela, J., Paatero, V., Saarela, A. Self organization of a massive document collection. IEEE Transactions on Neural Networks„ vol. 11, pp. 574–585, 2000.
Krampen G., von Eye A., Schui G. (2011). Forecasting trends of development of psychology from a bibliometric perspective. Scientometrics, 87, 687-694.
Larose D.T. Discovering Knowledge in Data. Introduction to DATA MINING. John Wiley and Sons, Inc. 2006
Larsen, P. (2008). The state of the art in publication counting. Scientometrics, 77(2), 235–251.
Lee, P.-C., Su, H.-N., Chan, T.-Y. (2010). Assessment of ontology-based knowledge network formation by Vector-Space Model. Scientometrics, 85(3), 689–703.
Li, X., Chen, H., Huang, Z., Roco, M. C. (2007). Patent citation network in nanotechnology (1976–2004). Journal of Nanoparticle Research, 9, 337–352.
Likas, A., Vlassis, N., Verbeek, J. The global k-means clustering algorithm. Raport tech., Computer Science Institute, University of Amsterdam, The Netherlands, 2001.
Linton J.D., Walsh S.T., Morabito J. (2002). Analysis, ranking and selection of R&D projects in a portfolio. *R & D Management*, 32, 139-148.
Lv P.H., Wang G.F., Wan Y., Jia L., Qing L., Fei-Cheng M. (2011). Bibliometric trend analysis on global graphene research. Scientometrics, 88, 399-419.
Manning, C. D., Sch¨utze, H. Foundations of Statistical Natural Language Processing. The MIT Press, 2001.

Manning, C. D., Raghavan, P., Sch¨utze, H. Introduction to Information Retrieval. Cambridge University Press, 2007.
Marshakova-Shaikevich I. (2005). Bibliometric maps of field of science. Information Processing and Management, 41, Page 1534–1547.
Mazur, D. Metody grupowania i ich implementacja do eksploracji danych postaci symbolicznej. PhD Thesis, Silesian University of Technology, 2005.
MacRoberts M., MacRoberts B., Problems of citation analysis, Scientometrics 36 (1996) 3.
Mierswa I., Wurst M., Klinkenberg R., Scholz M., Euler T. (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).
Miller G.A. (1995). WordNet: A Lexical Database for English. Communications of the ACM, 38, (11), 39-41.
Murphy J. J. (1999). Technical Analysis of the Financial Markets, New York, Institute of Finance
Musgrove, P. B., Binns, R., Page-Kennedy, T., & Thelwall, M. (2003). A method for identifying clusters in sets of interlinking web spaces. Scientometrics, 58(3), 657–672.
Newman, M. E. J. (2005). Power laws, Pareto distributions and Zipf's law. Contemporary Physics, 46(5), 323–351.
Noyons, E. (2001). Bibliometric mapping of science in a science policy context. Scientometrics, 50, 83–98.
Noyons E.C.M., Van Raan A.F.J. (1994). Bibliometric cartography of scientific and technological developments of an research-and-development field - the case of optomechatronics. Scientometrics, 30, 157–173.
Noyons, E. C. M., Van Raan, A. F. J. (1998). Monitoring scientific developments from a dynamic perspective: Self-organized structuring to map neural network research. Journal of the American Society for Information Science, 49(1), 68–81.
Osowski, S. Sieci neuronowe w ujeciu algorytmicznym. WNT, Warsaw, 1996.
Peters H. P. F., Van Raan A. F. J. (1993). Co-word-based science maps of chemical-engineering. Part2. Combined clustering and multidimensional scaling. Research Policy, 22, 47–71.
Piyawong W., TwizellE.H., Gumel A.B.(2003). An unconditionally convergent finite-difference scheme for the SIR model. Applied Mathematics and Computation, 146, 611-625.
Pourabbas E., d'Onofrio A., Rafanelli M. (2001). A method to estimate the incidence of communicable diseases under seasonal fluctuations with application to cholera. Applied Mathematics and Computation, 118, 161-174.
Price, D. J. D. (1965). Networks of scientific papers. Science, 149(3683), 510–515.
Price K., Storn R. (1997). Differential Evolution: Numerical Optimization Made Easy. Dr. Dobb's Journal, 18 – 24.
Price K., Storn R., Lampinen J. (2005). Differential Evolution - A Practical Approach to Global Optimization, Springer, Berlin.
Rajapakse A., Titchener-Hooker N. J., Farid S. S. (2005). Modelling of the biopharmaceutical drug development pathway and portfolio management. *Computers & Chemical Engineering*, 29, 1357-1368.
Ringuest J.L., Graves S.B., Case R.H. (2004). Mean-Gini analysis in R&D portfolio selection. European Journal Of Operational Research, 154, 157-169.
Rip, A., Courtial, J.P. (1984). Co-word maps of biotechnology: An example of cognitive scientometrics. Scientometrics, 6(6), 381–400.

Salton G.,Wong A., Yang C.S. (1975). A vector space model for automatic indexing. Communications of the ACM, 18, 613–620.
Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval* 1. Information Processing & Management, 24(5), 513–523.
Sander, J., Ester, M., Kriegel, H.-P., Xu, X. Density-based clustering in spatial databases: The algorithm gdbscan and its applications. Data Min. Knowl. Discov., vol. 2(2), pp. 169–194, 1998. ISSN 1384-5810.
Schenker, A., Bunke, H., Last, M., Kandel, A. Graph-Theoretic Techniques for Web Content Mining. World Scientific Publishing Co. Pte. Ltd., 2005.
Schildt, H. A., & Mattsson, J. T. (2006). A dense network sub-grouping algorithm for co-citation analysis and its implementation in the software tool Sitkis. Scientometrics, 67(1), 143–163.
Scott, J. (2000). Social network analysis: A handbook. London: Sage Publications Ltd.
Smalheiser, N. R., Swanson, D. R., (1998) Using ARROWSMITH: A computer-assisted approach to formulating and assessing ccientic hypotheses, Computer Methods and Programs in Biomedicine, 57, 149-153.
Small, H.G., & Griffith, B.C. (1974). The structure of scientific literatures: I. Identifying and graphing specialties. Science Studies, 4, 17–40.
Sitarz R. Projektowanie optymalnych ciągów kolumn jonowymiennych przy wykorzystaniu metod ewolucyjnych. Master Degree Thesis. Rzeszow University of Technology 2007.
Su, H.-N., & Lee, P.-C. (2010). Mapping knowledge structure by keyword co-occurrence: a first look at journal papers in technology foresight. Scientometrics, 85(1), 65–79.
Tsai H.H. (2011). Research trends analysis by comparing data mining and customer relationship management through bibliometric methodology. *Scientometrics,* 87, 425–450.
Tseng Y.H., Chang C. Y., Rundgren S. N. C., Rundgren C. J. (2010). Mining concept maps from news stories for measuring civic scientific literacy in media. Computers & Education, 55, 165-177.
Van Raan, A. F. J. (1993). Advanced bibliometric methods to assess research performance and scientific development: basic principles and recent partial applications. Research Evaluation, 3, 151–166.
Van Raan, A. F. J., Tijssen, R. J. W. (1993). The neural net of neural network research: An exercise In bibliometric mapping. Scientometrics, 26(1), 169–192.
Voorhees E.M. (1986). Implementing agglomerative hierarchic clustering algorithms for use in document retrieval. Information Processing & Management, 22(6), 465-476.
Wang Z.Y., Li G., Li C.Y., Li A. (2012). Research on the semantic-based co-word analysis. Scientometrics, 90, 855-875.
Wasserman, S., & Faust, K. (2003). Social network analysis: Methods and applications. Cambridge University Press.
Xu, X., Ester, M., Kriegel, H.-P., Sander, J. A distribution-based clustering algorithm for mining in large spatial databases. W ICDE ’98: Proceedings of the Fourteenth International Conference on Data Engineering, pp. 324–331, IEEE Computer Society, Washington, DC, USA, 1998. ISBN 0-8186-8289-2.
Yang Y., Wu M., Cui L. (2012). Integration of three visualization methods based on co-word analysis. Scientometrics, 90, 659-673.
Yi S., Choi J. (2012). The organization of scientific knowledge: the structural characteristics of keyword networks. Scientometrics, 90, 1015-1026.
Yoon, B., & Park, Y. (2004). A text-mining-based patent network: Analytical tool for high-technology trend. Journal of High Technology Management Research, 15(1), 37–50.

Yoon, B., Lee, S., & Lee, G. (2008). Keyword-based knowledge map of academic research. Presented at the third European conference on management of technology, Nice, France.

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

ACTA UNIVERSITATIS LAPPEENRANTAENSIS

496. VAUTERIN, JOHANNA JULIA. The demand for global student talent: Capitalizing on the value of university-industry collaboration. 2012. Diss.

497. RILLA, MARKO. Design of salient pole PM synchronous machines for a vehicle traction application. 2012. Diss.

498. FEDOROVA, ELENA. Interdependence of emerging Eastern European stock markets. 2012. Diss.

499. SHAH, SRUJAL. Analysis and validation of space averaged drag model for numerical simulations of gas-solid flows in fluidized beds. 2012. Diss.

500. WANG, YONGBO. Novel methods for error modeling and parameter identification of redundant hybrid serial-parallel robot. 2012. Diss.

501. MAXIMOV, ALEXANDER. Theoretical analysis and numerical simulation of spectral radiative properties of combustion gases in oxy/air-fired combustion systems. 2012. Diss.

502. KUTVONEN, ANTERO. Strategic external deployment of intellectual assets. 2012. Diss.

503. VÄISÄNEN, VESA. Performance and scalability of isolated DC-DC converter topologies in low voltage, high current applications. 2012. Diss.

504. IKONEN, MIKA. Power cycling lifetime estimation of IGBT power modules based on chip temperature modeling. 2012. Diss.

505. LEIVO, TIMO. Pricing anomalies in the Finnish stock market. 2012. Diss.

506. NISKANEN, ANTTI. Landfill gas management as engineered landfills – Estimation and mitigation of environmental aspects. 2012. Diss.

507. QIU, FENG. Surface transformation hardening of carbon steel with high power fiber laser. 2012. Diss.

508. SMIRNOV, ALEXANDER. AMB system for high-speed motors using automatic commissioning. 2012. Diss.

509. ESKELINEN, HARRI, ed. Advanced approaches to analytical and systematic DFMA analysis. 2013.

510. RYYNÄNEN, HARRI. From network pictures to network insight in solution business – the role of internal communication. 2013. Diss.

511. JÄRVI, KATI. Ecosystem architecture design: endogenous and exogenous structural properties. 2013. Diss.

512. PIILI, HEIDI. Characterisation of laser beam and paper material interaction. 2013. Diss.

513. MONTO, SARI. Towards inter-organizational working capital management. 2013. Diss.

514. PIRINEN, MARKKU. The effects of welding heat input usability of high strength steels in welded structures. 2013. Diss.

515. SARKKINEN, MINNA. Strategic innovation management based on three dimensions diagnosing innovation development needs in a peripheral region. 2013. Diss.

516. MAGLYAS, ANDREY. Overcoming the complexity of software product management. 2013. Diss.

517. MOISIO, SAMI. A soft contact collision method for real-time simulation of triangularized geometries in
multibody dynamics. 2013. Diss.

518. IMMONEN, PAULA. Energy efficiency of a diesel-electric mobile working machine. 2013. Diss.

519. ELORANTA, LEENA. Innovation in a non-formal adult education organisation – multi-case study in four education centres. 2013. Diss.

520. ZAKHARCHUK, IVAN. Manifestation of the pairing symmetry in the vortex core structure in iron-based superconductors. 2013. Diss.

521. KÄÄRIÄINEN, MARJA-LEENA. Atomic layer deposited titanium and zinc oxides; structure and doping effects on their photoactivity, photocatalytic activity and bioactivity. 2013. Diss.

522. KURONEN, JUHANI. Jatkuvan äänitehojakautuman algoritmi pitkien käytävien äänikenttien mallintamiseen. 2013. Diss.

523. HÄMÄLÄINEN, HENRY. Identification of some additional loss components in high-power low-voltage permanent magnet generators. 2013. Diss.

524. SÄRKKÄ, HEIKKI. Electro-oxidation treatment of pulp and paper mill circulating waters and wastewaters. 2013. Diss.

525. HEIKKINEN, JANI. Virtual technology and haptic interface solutions for design and control of mobile working machines. 2013. Diss.

526. SOININEN, JUHA. Entrepreneurial orientation in small and medium-sized enterprises during economic crisis. 2013. Diss.

527. JÄPPINEN, EERO. The effects of location, feedstock availability, and supply-chain logistics on the greenhouse gas emissions of forest-biomass energy utilization in Finland. 2013. Diss.

528. SÖDERHOLM, KRISTIINA. Licensing model development for small modular reactors (SMRs) – focusing on the Finnish regulatory framework. 2013. Diss.

529. LAISI, MILLA. Deregulation's impact on the railway freight transport sector's future in the Baltic Sea region. 2013. Diss.

530. VORONIN, SERGEY. Price spike forecasting in a competitive day-ahead energy market. 2013. Diss.

531. PONOMAREV, PAVEL. Tooth-coil permanent magnet synchronous machine design for special applications. 2013. Diss.

532. HIETANEN, TOMI. Magnesium hydroxide-based peroxide bleaching of high-brightness mechanical pulps. 2013. Diss.

533. TYKKÄLÄ, TOMMI M. Real-time image-based RGB-D camera motion tracking and environment mapping. 2013. Diss.

534. PEKKOLA, SANNA. Performance measurement and management in a collaborative network. 2013. Diss.

535. PANOREL, IRIS CHERRY. Pulsed corona discharge as an advanced oxidation process for the degradation of organic compounds in water. 2013. Diss.

536. TORKKELI, LASSE. The influence of network competence of internationalization of SMEs. 2013. Diss.

537. MOLANDER, SOLE. Productivity and services – safety telephone services for the elderly. 2013. Diss.