Search automation of the generalized method of device operational characteristics improvement

I Yu Petrova[1], A A Puchkova[2], V M Zaripova[1]

[1] Astrakhan State University of Architecture and Civil Engineering, 18, Tatisheva Str., Astrakhan, 414040, Russia
[2] Astrakhan State University, 20, Tatisheva Str., Astrakhan, 414040, Russia

E-mail: irapet1949@gmail.com, aa.puchkova@gmail.com, vtempus2@gmail.com

Abstract. The article presents brief results of analysis of existing search methods of the closest patents, which can be applied to determine generalized methods of device operational characteristics improvement. There were observed the most widespread clustering algorithms and metrics for determining the proximity degree between two documents. The article proposes the technique of generalized methods determination; it has two implementation variants and consists of 7 steps. This technique has been implemented in the "Patents search" subsystem of the "Intellect" system. Also the article gives an example of the use of the proposed technique.

1. Introduction

By developing new technical solutions, it is extremely important to define the most perspective directions for further researches. In particular, it is possible by detemination of the methods of operational characteristics improvement, which were already used in existing analogues of the developing device, and their further join in generalized methods. The method is a set of changes in construction, composition or technology of device manufacturing, which allows achieving a positive effect compared to the prior art. The determination of the generalized method and its furter using with the another prototype often results in developing a new, more advanced technical solution. However, the vast number of existing today technical solutions (in the form of registered patents) practically excludes possibility of manual produce of this operation. Thus, it is advisable to develop an information system to support this process. It is proposed to cluster scientific and technical documents, found on the Internet, with the further annotation of resulting clusters. Each cluster will thus correspond to one generalized method, and its annotation will include a text description of this method.

In addition, to determine the research degree of specific knowledge areas and to detect the most perspective one in terms of development of new patentable technical solutions, it is necessary to create a method of forecasts forming and trend determination, what will define the current status of each direction (being actively developed, the activity level is falling, is poorly studied, etc.).

The majority of currently existing patent search services (Google Patents, USPTO, Expacenet, etc.) receives the text query from the user and furter produces full-text search according to the entered query. They can promptly produce search in the vast patent databases, but they do not allow detetmining an improvement method used in one or another document and do not support the function of the search of the nearest documents to original and trend forming.

The research team, headed by G. Osipov, developed an "Exactus Expert" system, in which the user can search documents by a text query or documents closest to the specified one and also form trends. Among the documents storing in the database there are papers, dissertation abstracts and patents. However, this service in terms of solving problems has several drawbacks.

1. This system is universal and allows producing the search on the different branches of knowledge: technical, medical, agricultural, humanitarian and other sciences. Consequently it does not have features specific for area of new technical solutions development, namely method definition, the objective of the research determination, etc.

2. Analogue search in the system presupposes the presence of the source text of the prototype patent or formed text query, which is impossible to create in short time for several tens of synthesized technical solutions to define the most patentable variant.

3. Functions of trend forming and scientific directions analysis are available only by keywords which can be selected from the list existing in the system, but not all that may interest the user. The system on the basis of entered keywords classifies documents stored in the database. In other words, the system does not allow determining existing research directions in specific subject subregion, the user must independently apply the closest keywords (from the existing in the list) to the description of each direction and produce the search of the according documents.

The problem of patent trends forming is greatly similar with the problem of generalized method determination. Many researchers studied its solution. So, P. Bronwyn [1] proposed the use of the patent citation graph, but this method has low precision because the fact of citation does not allow judging the degree of two patents' proximity.

K. Fritch and P. Neuhäusler [2] classified patents on different grounds: IPC (internetional patent classification) classes, rightholder country, rightholder company, etc., but this approach cannot define trends within the same class or define a trend common for two classes.

B. Yoon [3] instead of the citation graph used the relationship graph with its subsequent clustering. Each patent has a vector of keywords which are selected from the text of the patent using TF-IDF metric after excluding stop words. The disadvantages of this method include the use of certain keywords instead of phrases, because a single word often bears no semanthic load. Another disadvantage of this approach is the need to determine a large cluster size to achieve required construction accuracy, resulting in the definition of the most general trends.

Thus, it is reasonable to develop a new technique to determine general improvement methods, as well as a new technique of trend forming. It is proposed to cluster documents with subsequent annotating the resulting clusters. Each cluster will thus match one generalized method, and its abstract will contain description of this method. The trend forming process can be perfomed using cluster analysis of existing patents. In this case, the result of clustering will contain lesser number of larger clusters (clusters of similar methods can be joined according to the principle of status: being actively developed, is poorly studied, etc.). Each cluster will thus present a separate direction of research. Chronological classification of the content of each cluster will allow determining the current status of this direction.

The purpose of the research is to create a technique of automatic determination of the generalized method for technical solution presented in the form of fthe ull sourse text of the patent, as well as the technique of forecasts and trends forming.

2. Materials and methods

In the world, there are many methods of text documents clustering, each of which has its own advantages and disadvantages and, as a consequence, different application areas. During the research, there was made the analysis of the most widespread clustering methods to select the best one to solve the problem of automatic generalized method determination.

1. Split or flat methods (K-means [4], spherical K-means [5], etc.) produce the partition of elements into N clusters. Their advantage is a high cluster speed, the shortcomings are the need for a priory selection of the number of clusters and sensitivity to "emission" elements.

2. Hierarchical methods (Single Link, Complete Link [4], etc.) produce the building of the cluster tree by establishing a system of partitions of nested elements. The advantages of these methods are high precision and the partition structure; the shortcomings are the need for maximum cluster size definition and high algorithmic complexity.

3. Semanthic methods. Clusters are nodes of a suffix tree formed from suffix trees of the input documents (trees containing all suffixes of a string). An original STC [6] method has a great contextual dependence and low accuracy, so it has developed its DIG [7] modification, which precision is about 70%. Its shortcoming is a high price of the tree or graph building in the case of receiving documents by network [8]. The advantages of these methods are the high work speed and the ansence of necessity to specify the number of clusters or threshold.

4. Kohonen self organising maps are a variant of the neural network; the result of their work is a distribution map of vectors, which are initial documents [9]. The advantage of this approach is learning without a teacher, and its main shortcoming is dependence on the random initial values of neuron weights and on the training set size.

During clustering, it is extremely important to select the principle of documents comparing and determination of the distance between them. Almoust always, the document is a point in N-dimensional vector space, and to determine the proximity degree between two points (x and x'), a specific metric can be used. Below, there are given the most widespread metrics [10].

1. Euclidean distance $$p(x,x') = \sqrt{\sum_{i}^{N}(x_i - x_i')^2}$$
2. Spearman distance $$p(x,x') = \sum_{i}^{N}(x_i - x_i')^2$$
3. Manhattan distance $$p(x,x') = \sum_{i}^{N}|x_i - x_i'|$$
4. Chebyshev distance $$p(x,x') = \max(|x_i - x_i'|)$$

The main difficulty of this task is to select the attributes vector. For each concrete situation this choice should be special. So, for instance, to select the most interesting one for particular researcher scientific papers [11], there were formulated the following attributes: the authors, keywords, abstract. To determine duplicate web pages [12], the basic words mechanism was proposed .

However, to solve the problem of generalized methods determination, it is necessary to take into account a number of its features. In particular, it should be implemented in the "Patent search" subsystem of the "Intellect" system, which is the automated system for support of the conceptual design stage of scientific and technical creativity [13]. In the heart of this system there is the energy-information model of chains (EIMC). Patents in this case are stored in the subsystem database in the form of their passports and are divided into two categories: added automatically or with assistance of an expert. An expert, by adding the patent into the database, fills in several fields of its passport; one of the most important is the list of physical and technical effects (PTE), used in this patent. By automatic adding the patent, the subsystem itself on the basis of available data determines relevance of this patent to all PTEs storing in the database. The relevance of the patent to particular PTE must also be considered for generalized method determination. Consequently, it is necessary to develop a new clustering method taking into account above-mentioned features.

3. Results

3.1. Technique of generalized method determination

As the result of analysis of advantages and disadvantages of existing documents clustering mechanisms, the following comprehensive technique to determine generalized methods was developed. There are two variants of its application available: fast and full. In each variant as settings,

there can be set the operation precision, which raise will increase the operation duration, but at the same time, to the same cluster less number of the closest patents will be assigned .

The fast variant of this technique consists of several steps (full variant differs by the absence of the flat clustering step and hierarchical clustering of the entire sample instead of training one):

- Choise of the PTE list, for which the generalized methods determination will be produced.
- Selection from the database of patents, in which using these PTEs is guaranteed. It can be achieved by selecting only among the patents that have been added with the assistance of an expert.
- Formation of the attributes vector.
- Hierarchical clustering of the formed sample.
- Flat clustering of the remaining patents, which relevance to the selected PTEs exceeds a certain threshold value.
- Removal of the “emission” elements.
- Annotation of resulting clusters.

Let us consider these steps in detail.

1. Selecting PTEs to determine methods. To reduce the set of processed documents, the clipping of elements, obviously belonging to another generalized method groups, is produced.

2. Formation of the training sample. For generalized methods detection, it is impossible to determine in advance the number of clusters, therefore flat clustering algorithms are inapplicable in this case. On the other hand, methods of other groups available to independently determine the number of clusters have either low accuracy or high durability. Because of these facts, it was proposed to combine both these methods; thus at the first stage, a hierachhical approach will be used to determine the number of clusters and their initial centers of mass. During the second stage, we will be working out the flat algorithm.

3. Formation of the attributes vector. As attributes it is proposed to use keywords taken from the previously formed thesaurus. At the beginning, one produces the analysis of the source text of all patents from the training sample and counting of the frequency of occurences of each attribute. After that, we produce the truncation of the space dimension by excluding trivial and unique attributes (presence percent of which is approaching 0 or 100 with an accuracy of up to 0.01). Synonyms are thus recorded as one attribute in a vector. Furthermore, one more attribute is added, which value is defined by each comparison of two patents: if the first one refers to another one. The weight of this attribute is proposed to be taken equal to 0.1.

4. Hierarchical clustering of the reference sample. At this stage, ascending hierarchical clustering of the formed training sample is proposed. As a proximity measure between two clusters we proposed taking the principle of distance between centers of mass, because this variant is the fastest (in the full variant instead of it, we take an unweighted pair distance as the most accurate one). As to the metric we chose Spearman distance as the most sensitive to emissions. The attribute of the stopping of the process will be a limit value of proximity between clusters on Hamming metric (it is determing from the precision paramether, the default value is 60% of space dimension).

5. Clustering of all patents. From all automatically added to the database patents, the selection of elements most relevant to PTEs determined on the first step (the limit relevance is determined from the precision paramether and the default value equal to 0.1) is produced. Then using a k-means algorithm, we produce the clustering of selected patents.

As the closure sign, we offer to take full repetition of partition of the one of the previous k-means steps or changing of clusters content not more than some threshold (it is also dependent on precision, the default value is 5%).

6. Removal of emissions. Since there is a possibility of use of a generalized method, different from methods used in the training sample, in automatically added patents, a k-means algorithm can include such patents to the wrong cluster, because the appopriate cluster will be absent. Therefore, it is necessary to analyze the formed clusters to find elements belonging to a particular cluster with a low probability. These patents should be excluded from clusters and placed in a special secondary cluster.

Such emissions can be clustered later after replenishment the training sample by an expert.

7. Annotation of clusters. At the final stage, formed clusters should be annotated. Received annotations will be the basis for formulation of the determined generalized method. To define keywords, it is necessary to determine the average distance between all pairs of patents included in the cluster by the Hamming metric. All patents attributes will point to keywords for this method, from these words the subsystem will be able to form the annotation of cluster. If necessary, the expert can produce the more accurate formulation of the generalized method independently.

4. Discussions

The described technique is realized in the "Patent search" subsystem of the "Intellect" system. It is a web application developed in the Visual Studio 2015 environment with the use of such technologies as ASP.NET, Entity Framework, AJAX.

Below there is an example of identifying a cluster of calorimetric biosensors on the basis of the pyroelectric effect. The "Intellect" system is constantly expanding, and currently it creates a new module for the synthesis of biosensors [14]. Biosensors consist of two parts: a bioreceptor and a transducer. The "Patent search" subsystem received the following search conditions: the use of pyroelectric effect in the patent and the presence of the words "biosensor" or "pyroelectric" in the text of the patent. The search was conducted in documents in English and Russian languages. As a result, there were found 9 patents: US20100028969 A1, RU 2266959, US 5108576, US4551425, US20110182776 A1, US20130052632 A1, US4829003 A, US 20050196322 A1, WO 1990013017 A1. After clustering these documents, three generalized methods were determined:

1. The use of materials in the construction of biosensors, characterized by low thermal conductivity and high heat capacity, which provides a minimal loss of a thermal signal arising at reaction of the enzyme substrate. As a result, the sensitivity of the device is increased.
2. The pyro-optical detection: irradiation by light pulses having a certain wavelength of pyroelectric film (PVDF), covered by film electrodes with the immobilized reagent (antibody) deposited on the surface. The reagent can bind to the analyzing material by irradiation with release of additional heat, which is converted by pyroelectric into an electrical sign. As a result, the signal/noise ratio of the biosensor is increased.
3. The differential circuit of connection of two thin film pyroelectric detectors is used in the construction of the biosensor, one of which has covered with the special epoxy film, to which proteins (ferments and antibodies) can be connected with the help of photochemical reagents.

5. Conclusion

During the studies, a technique of automatic determination of generalized methods for technical solutions, provided in the form of patents, was created. This technique was implemented in the "Patent search" subsystem of the "Intellect" system. As a result of trial operation, we made a conclusion about its effectiveness and necessity of further researches in this area. In subsequent studies, it is proposed to develop and implement a technique for trend forming. Also, in the future, we plan to define the optimal threshold values for various clustering steps. In addition, it is necessary to improve the mechanism of cluster annotation and registration as an evidence of not only keywords, but also the key phrases that will help to improve the accuracy of clustering.

6. Acknowledgments

This research was partially supported by the Russian Fund of Basic Research (grants No. 116-37-00258\16).

References

[1] Bronwyn P 2001 *NBER Working Paper* **8498** 60-77
[2] Neuhäusler P, Rothengatter O, Frietsch R and Feidenheimer A 2015 *Patent Applications: Structures, Trends and Recent Developments 2014* (Berlin: Expertenkommission Forschung und

Innovation (EFI)).
[3] Yoon B and Park Y 2004 *The Journal of High Technology Management Research* **15** 37-50
[4] Jain A and Dubs R 1988 *Algorithms for clustering data* (Upper Saddle River: Prentice-Hall Inc)
[5] Dhillon I S and Modha D S 2001 *Machine learning* **42(1-2)** 143-175
[6] Zamir O and Etzioni O 1998 *Proc. Int. Conf. on research and development in information retrieval (New York)* (New York: ACM Press) pp 46-54
[7] Hammouda K M and Kamel M S 2004 IEEE *Transactions on knowledge and data engineering* **16(10)** 1279-1296
[8] Andrews N O and Fox E A 2007 *Recent developments in document clustering* (Blacksburg: Virginia Tech)
[9] Kohonen T 2001 *Self-Organizing Maps* (Berlin: Springer)
[10] Komarova A S 2006 *SPIIRAN Works* **3** 288-299
[11] Barahnin V B, Nehaeva V A and Fedotov A M 2008 *Vestnik of NSU* **6** 3-9
[12] Ilyinsky S, Kuzmin M, Melkov A and Segalovich I 2002 *Proc. 11th Int. Conf. on World Wide Web*
[13] Zaripova V and Petrova I 2014 *Proc. 11th Joint Conf. JCKBSE (Volgograd)* (Zheneva: Springer) pp 521-532
[14] Petrova I, Zaripova V, Lezhnina Yu, Sokolskiy V and Mitchenko I 2015 *Vestnik of ASTU. Series management, computer science and informatics* (Astrakhan: ASTU publisher) **3** 35-48