Advanced Engineering Informatics 30 (2016) 26–38

Contents lists available at ScienceDirect

Advanced Engineering Informatics

journal homepage: www.elsevier.com/locate/aei

Combination of improved cosine similarity and patent attribution probability method to judge the attribution of related patents of hydrolysis substrate fabrication process ☆

Zone-Ching Lin *, De-Wei Wu, Guo-En Hong

Department of Mechanical Engineering, National Taiwan University of Science and Technology, Taipei, Taiwan

ARTICLE INFO

Article history:
Received 4 February 2015
Received in revised form 9 November 2015
Accepted 12 November 2015
Available online 11 December 2015

Keywords:
Hydrolysis substrate fabrication process
Cosine similarity
Patent
Probability
Attribution

ABSTRACT

The paper studies the attribution of patents by innovatively establishing a combination of improved cosine similarity concept and patent attribution probability method for hydrolysis substrate fabrication process in order to enhance the speed and accuracy for judgment of patent attribution. For the improved cosine similarity method innovatively established in the paper, all the vocabularies of important technical words or functional words in patent documents are regarded as a number of vector dimensions. The normalized numerical values of these vocabularies are regarded as the weights of these technical words or functional words. They are substituted in a formula of improved cosine similarity. Regarding the patent attribution probability method, it applies the normalized numerical values of the various clusters of technical or functional words as well as the formula of probability method to judge which technical category or functional category that a patent is attributed to. As the study innovatively combines improved cosine similarity with patent attribution probability method, before employing patent attribution probability method for a patent document, the study firstly uses improved cosine similarity to check with the cluster of patent group words having high relativity, and rule out the unrelated technical categories or functional categories that the patent is not attributed to, so as to decrease the number of categories for calculation and the time to calculate which technical category or functional category that a patent is attributed to when using patent attribution probability method, and to achieve a method that can more rapidly and accurately judge which technical category or functional category that a patent is attributed to.

1. Introduction

The following literatures relating to the paper‘s contents are introduced in this order: (1) The literatures concerning document classification management: Generally speaking, in the aspect of analysis of document contents, the importance of a term is usually determined by the appearance frequency of the term in documents. Since the paper has to firstly calculate the frequency of the keywords in patent documents, the related literatures are firstly introduced. (2) The literatures concerning cosine similarity: This is because the paper has to calculate the similarities of the technical categories of patent documents. (3) The literatures concerning knowledge management and knowledge extraction: Since patent is a kind of knowledge, it is related to knowledge management and knowledge extraction. (4) The literatures concerning patent analysis: This is because patent analysis and calculation of technical categories of patent have to be made. According to the above order, the related literatures are introduced below.

The literatures concerning document classification management are explained below. In the aspect of analysis of document contents, the importance of a term is usually determined by the appearance frequency of the term in documents. Singhal and Salton [1], suggested that in the aspect of analysis of document contents, the importance of a term was usually determined by the appearance frequency of the term in documents. They also explored the relationship between the appearance frequency of terms in documents and the importance of terms. Those terms with high appearance frequency but of low importance were usually the meaningless conjunctions or grammar words, which could not express the features of the document contents. Hence, during analysis of document contents, these kinds of terms had to be filtered out before retrieving keywords. The keywords with low appearance frequency

☆ Handled by C.-H. Chen.

* Corresponding author. Tel.: +886 2 27376455 (O); fax: +886 2 27376243.
E-mail address: zclin@mail.ntust.edu.tw (Z.-C. Lin).

http://dx.doi.org/10.1016/j.aei.2015.11.003
1474-0346/© 2015 Elsevier Ltd. All rights reserved.

and of low importance would be abandoned during analysis of document contents, whereas the terms with high appearance frequency and also of high importance could fully represent the features of document contents. Therefore, how to retrieve these terms is an important issue to analysis of document contents. Normally in a cluster of documents, part of the keywords appear together with other keywords. These terms usually have a certain extent of relativity. In order to save the time of reading information, some methods were developed from the past literature for automatic reading of documents and filtering information stream. After that, only related information is transmitted to the people in need. The above process is called information filter in the studies of information retrieval, and this topic has started to be studied and discussed. Besides, when extending from the definitions of the keywords themselves, other different related term can also be obtained. According to the appearance frequency of keywords in document groups, relativity of keywords can be judged, and a stock of related terms can be established. Such a stock of related terms can be applied to integration of keywords. After combination of the terms of high relativity, the set of keywords can represent the features of document contents. Lawrence and Giles [2], proposed improving the retrieval of literatures relating to techniques of general search engines and searched results of scientific information, but they did not make analysis of technical/functional words of patent documents.

The literatures concerning cosine similarity are explained below. Cosine similarity is a calculation way of similarity commonly used in information retrieval Salton and McGill [3]. It can be used to calculate the similarity between documents, and the similarity between vocabularies, and even the similarity between the searched phrases and documents. Before calculation of similarity between two documents, the documents have to be expressed in vector form, implying that all the important vocabularies in these documents have to be regarded as a number of vector dimensions. The weights of these vocabularies are values of the dimensions, which are combined to form a vector to represent the documents. Vallet et al. [4], used semantic search to develop a retrieval model to improve search step from a large stock of documents, and define weight calculation method and rank calculation method of classical vector space model. Based on semantic and combination optimization techniques, Gan and Chen [5], proposed improving calculation method of TF-IDF weights and applying it to vector space model (VSM). This calculation method can effectively decrease the classification of subjective factors. Wang et al. [6], proposed collaborative filtering (CF) as an effective method to solve information overload problem, and enhance the accuracy in searching the similarities of preferences among different users. Using cosine similarity and TF-IDF weights, they finished information search and filtering of retrieved information. Zhang and Odbal [7] proposed a method to automatically aim at crosscheck of Mongolian and Chinese words. Based on the statistical vector space model, keywords were retrieved from texts. Based on the contents of the combined retrieved keywords, TF-IDF method could confirm the weight for the vector of a sentence. Once the weights of terms are confirmed, the similarity between vectors of the sentence could be calculated through the cosine similarity formula. All the above literatures used cosine similarity method to make further studies of vector space model. They used TF-IDF method to calculate the weights of terms, and make improvement and studies on the TF-IDF calculation method of weight. General cosine similarity is obtained by using TF-IDF to know the appearance frequency of a certain keyword in a document and using frequency of the term to calculate the weight of the keyword. If its appearance frequency is higher, it can better know the category of the document. However, the above cosine similarity and TF-IDF are only applied to one-on-one comparison of similarity between documents, but still have not considered the effects of the total number of words or length of a patent document on weight, and have not been applied to judge which technical category or functional category that a newly added patent document is attributed to.

The literatures concerning knowledge management and knowledge extraction are explained below. Since patent is a kind of knowledge, patent analysis is related to patent management and knowledge. The paper gives a brief introduction of two research papers about knowledge management. To solve the problem of classification, information management has to be conducted in good ways. Gruber [8] defined knowledge management (KM) as "a process that helps organizations define, select, rearrange and spread important information and professional knowledge." And O'leary [9] defined KM as "the formal management of knowledge for facilitating creation, access and reuse of knowledge, typically using advanced knowledge." He further argued that KM was more than management of knowledge, and the ultimate purpose of KM was to further innovate through reuse of knowledge. Trappey et al. [10] proposed an effective analysis on patents of fundamental research and development (R&D) knowledge to decrease the development time of new products, enhance the success rate of marketing of products, and reduce the potential patent infringement. The method they proposed was to retrieve and collect patents with complete contents, and make a crucial report in the form of a tree chart. Using statistical methods to evaluate and verify the related keywords, together with a certain proportion of weight values, the accuracy of classification could be increased.

The literatures concerning patent analysis are explained below. Lin et al. [11], established an innovative integrated semantic analysis and term and word segmentation system of English patent documents. This new system method applied the part/component keywords extracted from English patent documents and the semantic analysis method of the technical words and functional words. It used the concept of appearance frequency and normalization frequency of keywords to establish word clusters of technical category or functional category. They further developed a probability analysis method and formula for judging which technical category that an LED patent document in is attributed to, and made proof accordingly. Trappey et al. [12], proposed the life spans of dental implant (DI) key technologies using patent analysis. Key patents and their frequently appearing phrases are analyzed for the construction of the DI ontology. Afterward, the life spans of DI technical clusters were defined based on the ontology schema. This research demonstrated the feasibility of using text mining and data mining techniques to extract key phrases from a set of DI patents with different patent classifications (e.g., UPC, IPC) as the basis for building a domain-specific ontology. Bermudez-Edo et al. [13], proposed a methodological approach for the definition of relationships and reasoning tasks for patent analysis by using patent ontologies, and provided a real illustration of its potential in the context of international flows of research knowledge. This declarative method was based on the formal definition of key patent analysis indicators (KPAIs). More specifically, the paper illustrated the applicability of the proposed methodology by classifying patents into the five patterns of internationalization identified by the Organization for Economic Co-operation and Development (OECD). Trappey et al. [14] proposed a novel and generic methodology combining ontology based patent analysis, and clinical meta-analysis was developed to analyze and identify the most effective patented techniques in the dental implant field. The research established and verified a computer supported analytical approach and system for the strategic prediction of medical technology development trends. As mentioned above, Lin et al. [11] developed term and word segmentation system and probability method of English patents, and once applied them to LED patent document, but they did not apply cosine similarity method to their