A Survey On Different Text Clustering Techniques For Patent Analysis

Abhilash Sharma
Assistant Professor, CSE Department
RIMT – IET, Mandi Gobindgarh, Punjab, INDIA

ABSTRACT

Patent analysis is a management tool in order to confront the management of product or service development process and organization’s technology. Patent documents contain novel ideas, inventions and important research results. The analysis of these patents can be valuable to various sectors such as industry, business, law and policy-making communities in order to assess latest technological trends and to forecast new technologies. This work has been carried out with an aim to review various text clustering techniques for effective patent analysis.

Keywords

Text Clustering; Patent Analysis; Fuzzy Approach; Bayesian Method; k-means Clustering Algorithm.

1. INTRODUCTION

A Patent is basically a type of Intellectual Property. Patent analysis is one of the ways of recognizing the advancements in technologies. But patent documents consist of large technical and legal terminology and it is difficult for non-specialists to interpret the inventions and technologies mentioned in the patents. So, some simple methods are required to deduce the valuable information from the patent documents.

Text Clustering, also referred to as Document Clustering, is closely related to the concept of data clustering. Text clustering is a more specific technique for unsupervised document organization, automatic topic extraction and fast information retrieval or filtering. This clustering technique involves the use of descriptors and descriptor extraction. Descriptors are sets of words that describe the contents within the cluster.

Patent analyses based on structured information such as filing dates, assignees, or citations have been the major approaches in practice and in the literature for many years. A typical Patent Analysis Model [7] is shown in Figure 2. But there is need of some effective techniques for analysing the patents.

Our work attempts to analyse various text clustering techniques for reviewing the patent data and that data can be beneficial for various companies to understand the present technologies, to predict the future technologies and to plan for potential competition based on new technologies.

2. METHODOLOGY

This paper illustrates the study of various text clustering methodologies for patent analysis that can help out many companies for improving their competitiveness. The main methodology used for this work was by examining the publications, journals and reviews in the field of text clustering, patent analysis and patent documents over the times.

3. RESEARCH FINDINGS

3.1 K-Means Clustering Algorithm

Young Gil Kim, et. al. [2008] proposed a new visualization method for patent analysis. In this technique, initially keywords are collected from the patent documents of a particular technology field. After that, clusters of patent documents are generated using k-means algorithm. With the clustering results, a semantic network of keywords is formed without respect of filing dates. Then, a patent map is made by rearranging each keyword node of the semantic network according to its earliest filing date and frequency in patent documents.

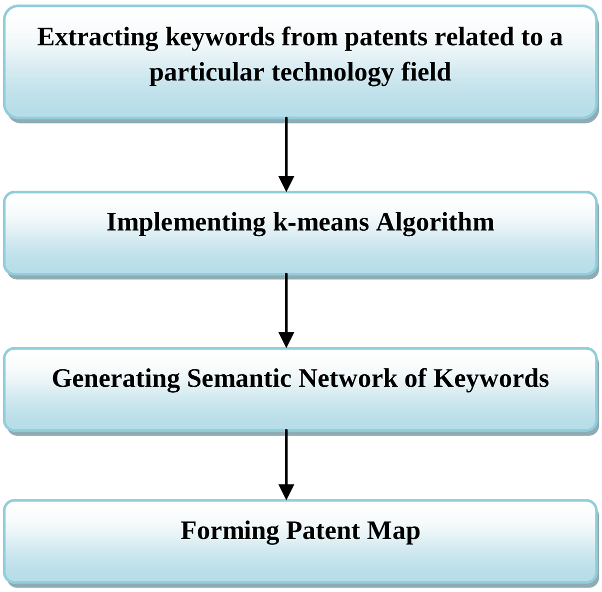
Extracting keywords from patents related to a particular technology field
Implementing k-means Algorithm
Generating Semantic Network of Keywords
Forming Patent Map

Figure 1: Flowchart of visualization method for patent analysis

Figure 1 depicts the flowchart of the proposed methodology. A patent map is the visualized expression of total patent analysis results to understand complex patent information easily and effectively. And it is generated by collecting related patent documents of a target technology field, processing, and analyzing them. In general, a patent document consists of structured and unstructured data.

3.2 Document Clustering and Time Series Analysis

In this research work, a new Patent Analysis Model is proposed for Technology Forecasting [1]. Most of the techniques which were developed earlier for Patent Analysis were based on one analytical approach such as clustering, classification and citation analyses. But, they had some limitations to predict the future state of a technology because they were dependent on only one result of a Technology Forecasting method.

Sang Sung Park, et. al. [2012] attempted to minimize this problem by combining two analytical approaches which are patent document clustering and time series model. K-means clustering algorithm has been used as a patent clustering method and Time Series Regression (TSR) as a time series model.

K-means clustering is a clustering method for finding K clusters and assigning all points to each cluster by Euclidean distance measure. TSR is a time series method to model the function of dependent variable Y and independent variable X, where X is time and Y is the number of issued patent documents.

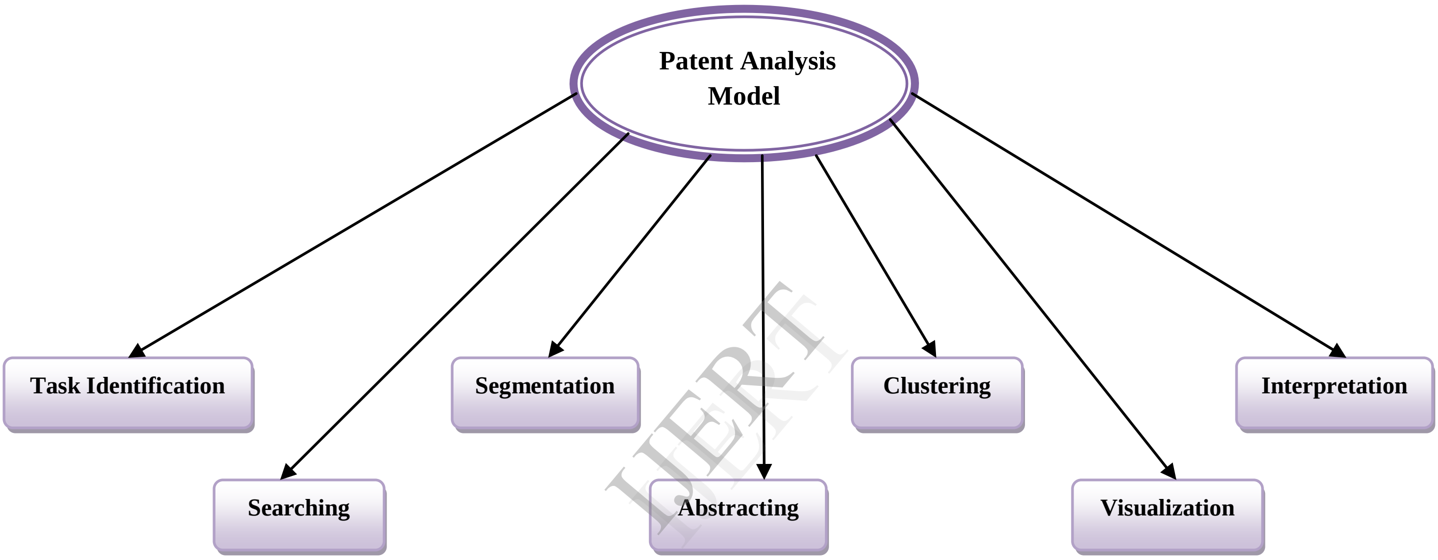

Figure 2: A Typical Patent Analysis Model

3.3 Distance Determination Approach

This research work proposed a new clustering algorithm for automatic discovery of data clusters [2]. Iterative Relocating Technique of Partitional Clustering, i.e. k-means and k-medoids have been used in this work. The algorithms have been implemented in C++ language. A partitioning based algorithm, D-M (Density Means), clustering automatically generates the clusters in this distance determination approach.

K-Means Step: The basic step of k-means clustering is to give the number of clusters k and consider first k objects from data set D as clusters & their centroid. After that, K-means algorithm's steps will be performed.

K-Medoids or Partition around Medoid (PAM) Step: The basic step of K-Medoid or PAM clustering is to give the number of clusters k and consider first k objects from data set Dn as clusters & their medoid. Then the k-medoid or PAM algorithm will perform its steps.

Table 1 and Table 2 depict the Comparison of algorithm's running time and Comparison of SSE of algorithms respectively.

Algorithm	Computational Complexity
k-means	***O(nkt)***
k-medoids	$O(k(n-k)^2)$
D-M	$O(n^i)$

Table 1: Comparison of algorithm's running time

3.4 Bayesian Approach

The patent documents include the data which is of highly dimensional structure. It is difficult to cluster the document data because of their dimensional problem. Therefore, Bayesian approach was adopted to solve this problem of dimensionality [3].

Earlier, clustering algorithms were based on similarity or distance measures, but Bayesian clustering used the probability distribution of the data.

The distribution family of Bayesian model has Gaussian and Laplace in this model. The proposed method is a two step process, i.e. Initialization and Repetition (Clustering) for the following input and output:

Input:
(1) Given data, $X=\{x_1, x_2, \ldots, x_n\}$
(2) Prior distribution, $p(\theta)$
(3) Likelihood function, $l(x|\theta)$

Output:
(1) Posterior distribution, $p(\theta|x)$
(2) Updated parameters of hierarchical Bayesian model
(3) Dendrogram of Bayesian clustering

Algorithm's Name	Two	Four
k-means	***1053.6022***	***681.4046***
k-medoids	***1082.002***	***464.5469***
D-M	***1053.6022***	***392.6812***

Table 2: Comparison of SSE (Sum of Square Error) of Algorithms

3.5 *Fuzzy Logic Control Approach*

This research work presented a novel hierarchical clustering approach for patent analysis [4]. Keyword-based methodologies for analysis tend to be inconsistent and ineffective when partial meanings of the technical content are used for cluster analysis. Thus, a new methodology has been presented to automatically interpret and cluster knowledge documents using an ontology schema. Moreover, a fuzzy logic control approach is used to match suitable document clusters for given patents based on their derived ontological semantic webs.

Fuzzy ontological document clustering (FODC) uses the following methodology:
Initially, domain experts define the domain ontology using a knowledge ontology building and RDF editing tool called Protégé, and the words and phrases (e.g., speech, chunks, and lemmas) of the patent documents are mapped to the corresponding domain ontology concepts. The experts also create a training set of patents using a free and easy-to-use natural language processing and tagging tool. Afterwards, the probabilities of the concepts in given document chunks are computed. The concept probabilities calculated in any given patent document are then used for clustering the patents with fuzzy logic inferences. Hence, the hierarchical clustering algorithm is refined by adapting fuzzy logic to the process of ontological concept derivation. Table 3 depicts the differences between FODC and Key Phrase K-Means Clustering.

No.	FODC	Key-Phrase K-Means Clustering
1.	Extracts representative information	Extracts general phrases
2.	Ontological structure is applied to present knowledge	Key phrase sentence fragments are used to present knowledge
3.	Ontology carries meanings and relations	Key phrases are less meaningful
4.	FODC's F-Measure is high	K-Mean's F-Measure is low
5.	Documents provide various views including the main concepts and details	Documents provide a key phrase view point

Table 3: Differences between FODC and Key-Phrase K-Means Clustering

4. RESULTS

For better analysis, the methodologies proposed and inferences from each Text Clustering Technique have been shown separately. Different methodologies provide different types of analysis depending upon the user demands. Moreover, generation of patent maps and semantic networks also help in reducing analysis time.

5. CONCLUSIONS

The objective of our work is to provide a study of different text clustering techniques that can be used for efficient patent analysis. These techniques can be beneficial for various types of analysis such as visual analysis, analysis which include high dimensional data, etc. Hence, this study helps to prove effective for efficient analysis in a way that analyst can choose the appropriate Text Clustering Technique depending upon the requirements.

6. REFERENCES

[1] Sang Sung Park, et. al., "A Patent Analysis Model Combining Document Clustering and Time Series Analysis"; Brain Korea Project, 2012.

[2] Bhanu Sukhija, Sukhvir Singh, "Improved K-Means Clustering Technique using distance determination approach"; ISSN: 2277 – 9043, International Journal of Advanced Research in Computer Science and Electronics Engineering Volume 1, Issue 5, July 2012.

[3] Sunghae Jun, "A Clustering Method of Highly Dimensional Patent Data using Bayesian Approach"; IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 1, No 1, January 2012, ISSN: 1694-0814.

[4] Amy J.C. Trappey, et. al., “A Fuzzy Ontological Knowledge Document Clustering Methodology”; IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART B: CYBERNETICS, VOL. 39, NO. 3, JUNE 2009.

[5] Young Gil Kim, et. al., “Visualization of patent analysis for emerging technology”; ScienceDirect, Expert Systems with Applications 34 (2008) 1804–1812.

[6] Khaled Khelif, et. al, “Semantic Patent Clustering for Biomedical Communities”; 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology.

[7] Yuen-Hsien Tseng, et. al., “Text Mining Techniques for Patent Analysis”; ScienceDirect, Information Processing and Management 43 (2007) 1216–1247.

[8] Michele Fattori, et. al., “Text mining applied to patent mapping: a practical business case”; World Patent Information 25 (2003) 335–342.

[9] https://sites.google.com/site/analyzingpatenttrends/Home/what-is-patent-analysis

[10] http://www.patentinsightpro.com/documents/Text%20Clustering%20on%20Patents.pdf