Use of Pseudo Relevance Feedback for Patent Clustering with Fuzzy C-means

Noushin Fadaei
Hildesheim University
Hildesheim, Germany
fadaei@uni-hildesheim.de

Thomas Mandl
Hildesheim University
Hildesheim, Germany
mandl@uni-hildesheim.de

Abstract

Patent databases are meaningful resources for technology trend detection as they collect information on the recent key innovations; however the importance of wordings in patents and use of complex content are remarkable challenges in key word extraction in the text mining phase. Moreover patents share information by nature and depending on the criterion of classification such as materials or uses, one may belong to multiple classes. For clustering patents, this work proposes an updating fuzzy c-means clustering which employs pseudo relevance feedback originating from information retrieval in order to improve features extracted from the patent collection following feedbacks. The results show a noticeable improvement in clustering after applying pseudo relevance feedback clustering patents under topic *contact lenses*.

1 Introduction

The increasing competition in research conducted at universities and other research institutes as well as in industry, further intensified by the increasing globalization, reinforces the importance of identifying new trends at an early stage. According to a study by Thomson Reuters 70% to 90% of the information covered in patents depending on the research area is not published anywhere else [Cor07]. This quality of patents makes such databases vital resources for finding new trends in the industry. Yet patent databases are growing very fast and required to be more organized for such purposes. According to European Patent Office (EPO), there is a steady increase in this huge information resource in terms of filed patents since 2010b [Off18]. In 2016, EPO recorded the highest number of granted patents which went up to 10.1% in 2017 [Off18]. To manage the detection of subtopics and their potentiality of being a trend in such big data, further grouping of patents is inevitable.

Patent analysis approaches are either qualitative or quantitative [Hon09]. The focus of this work lies on qualitative patent analysis, i.e., it uses text-mining techniques regarding the content of patents unlike quantitative approaches which employ metadata. The process typically aims to transform documents to vectors using weighting systems (e.g., TFIDF) for key terms; Clustering methods then exploit similarity measures to examine related vectors (e.g., Euclidean distance) and group them into clusters. The efficiency of clustering is highly dependent on the selected key terms [TJC04]. While the key words extraction requires the contribution of experts in various domains, automatic methods restrict the number of selected terms by applying thresholds for term frequency in documents (TF) or document frequency (DF) [TLL07] or consider building key phrases for example by creating a co-relation matrix of high frequent terms in documents where the co-occurrence of terms in the dataset is reserved [THTL06]. This work adopts pseudo relevance feedback [MRS08b] to obtain the key terms that tend to form the core concept of clusters.

One of the most commonly used clustering algorithms in text-mining is k-means; however k-means is not the best clustering method for patent analysis. Patents are very rich documents in terms of professional information and they may cover a range of technologies, applications or use of various materials. Therefore exhaustive clustering confines patents that potentially belong to different classes of a certain crite-

rion such as technology or material [FMS+15]. Moreover k-means requires the number of clusters and cannot cope with outliers [CHPT05].

To avoid low accuracy, fuzzy non-exhaustive clustering methods are exploited for patent clustering, e.g., by [THTL06] and [DD09]. The fuzziness of the clustering method provides a likelihood of possession (or membership to cluster) instead of a rigid distinction and the overlapping characteristic of clustering allows one patent reflecting a number of claims contributes in multiple clusters. Setting a threshold for membership controls to what extent of similarity patents may show up in a cluster. Fuzzy c-means (FCM) [Bez81] is the most popular fuzzy clustering algorithms and here we use it to softly partition our patent collection on certain topics. Also, the membership matrix enables FCM to cope with the issue of the outliers as all membership values of one document to all clusters ought to add up to 1; thus, an unrelated document to all clusters receives insignificant membership values for each and every cluster and can be ignored.

The ultimate goal of this study is to investigate the main idea of partitioning a patent dataset by means of a fuzzy clustering which allows for updating through relevance feedback. Practically this study will run with help of patent expert users however for experimental purposes and analysing the validity of the approach, we organize it with FCM and pseudo relevance feedback. For the evaluation, we developed a benchmark based on *World Intellectual Property Organization*'s patent reports on recent technology trends which were edited by experts in the domains. The reports describe technologies and trends within a domain, e.g. *contact lenses* or *robotic arms*. These reports are published at the website of the *World Intellecutal Property Organization (WIPO)*[1]. The selected reports are provided by *Gridlogics Technologies Pvt. Ltd* and were partially generated by the use of *Patent iNSIGHT Pro*. The developed benchmark can be used for experiments in classification, clustering, trend analysis or other intelligent patent processing systems.

2 Problem Statement

Given the query of the expert user which determines the scope of the topic or more generally given the International Patent Classification (IPC) [2] of the topic of interest, we retrieve the patents of the domain from the patent dataset. The set of n patents is then supposed to be represented as vectors ($X = (x_1, x_2, \ldots, x_n) 2 \leq c \leq n$). Each vector $x_k \in R^s$ is built up by s features where the features are the selected key terms from the dataset. The goal is to organize X into groups that may share vectors for further patent analysis purposes like identifying trends in the given topic [HXW+12].

2.1 Citations In Text

Citations within the text should indicate the author's last name and year[**?**]. Reference style[**?**] should follow the style that you are used to using, as long as the citation style is consistent.

3 Our Approach

3.1 Fuzzy C-Means Algorithm (FCM) [Bez81]

Fuzzy c-means algorithm (FCM) [Bez81] is based on the *fuzzy membership matrix*. The *membership* describes the likelihood of each vector (document) x_k being a part of cluster (subtopic) c_i where $1 \leq k \leq n$ and $1 \leq i \leq c$ with c being the number of clusters. The overall *membership* of each vector is normalized to 1. The algorithm starts off with initializing c vectors as centroids of the clusters. Then the *membership* (w_{ik}) is calculated through the Euclidean distances (d_{ik}) of each vector (x_k) to the centroid (P_i) of the cluster it belongs to and to the centroids of other clusters.

$$P_i = \frac{\sum_{k=1}^{n} (w_{ik})^m x_k}{\sum_{k=1}^{n} (w_{ik})^m} \tag{1}$$

$$W_{ik}^{(b)} = \sum_{j=1}^{c} \frac{1}{[(\frac{(d_{ik})^{(b)}}{(d_{jk})^{(b)}})^{\frac{2}{m-1}}]} \tag{2}$$

The FCM *membership function* is calculated as [HXW+12]:

$$\mu_{(i,j)} = [\sum_{t=1}^{c} (\frac{||x_j - v_i||_A}{||x_j - v_t||_A})^{\frac{2}{m-1}}]^{-1} \tag{3}$$

$\mu_{(i,j)}$ represents the membership value of j_{th} patent of the dataset and i_{th} topic whose centroid is v_i. $||||_A$ stands for norm function.

The *memberships* updates centroids vectors until the overall distance of the updated centroids is less than ε compared to the last set of centroids. The steps of FCM algorithm are as follows:

1. Set the number of clusters to be found (c)
2. Set an Euclidean normalization and fuzziness (m)

[1] http://www.wipo.int/patentscope/en/programs/patent_landscapes/plrdb.html

[2] According to *World Intellectual Property Organization(WIPO)*, The International Patent Classification (IPC), established by the *Strasbourg Agreement 1971*, provides for a hierarchical system of language independent symbols for the classification of patents.

3. Initialize of cluster prototype P^0, set the iterative counter (b)

4. Obtain *membership matrix* using Equation 2 above

5. Update the centroids using Equation 1 above

6. Repeat starting from step 2 until $||P^{(b)} - P^{(b-1)}|| < \epsilon$

The algorithm description and formulations are inspired by [NNB15].

3.2 Updating Fuzzy C-Means using Pseudo Relevance Feedback

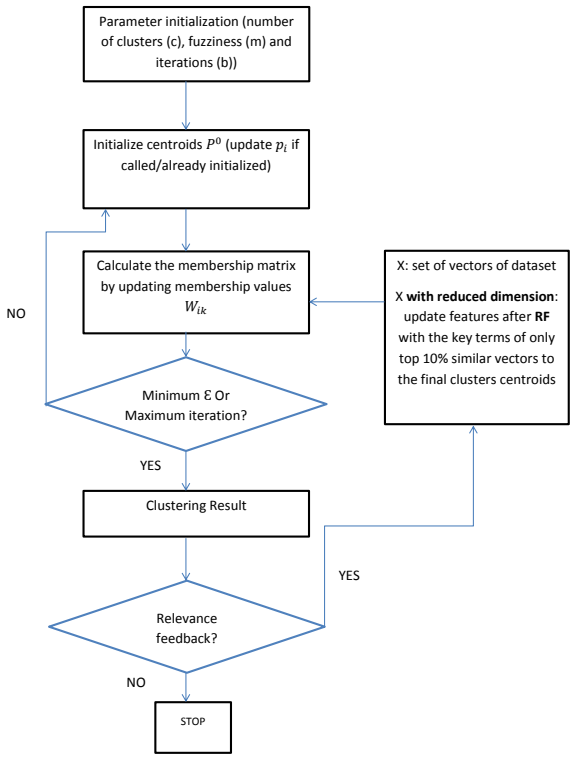

Figure 1: Updating fuzzy c-means using pseudo relevance feedback

Relevance feedback is a common concept in information retrieval that involves the user knowledge to identify and fetch more similar results to the query. Pseudo relevance feedback resembles the same concept but the entire process is carried out automatically. The goal is to identify the relevant key terms regarding the query and to expand them in order to update the query and make it in line with more relevant documents. The procedure starts with retrieving relevant documents and sorting them based on their similarity score to the query. Then the top k documents are considered as the feedback for best results. Pseudo relevance feedback considers these documents the source of relevant key terms to the query.

We make use of this method in FCM so that we find the patents that are more similar to the core concept of the cluster or centroids. The hypothesis is the key terms provided by these patents reflect the main idea of the cluster and their corresponding vectors can represent a more manageable dataset for FCM. The procedure is shown in figure 1. It generates an overlapping clustering based on the *membership matrix* of FCM. As long as a relevance feedback is requested the following procedure runs:

1. Rank patents based on their membership to the cluster

2. Obtain the top relevant patents for building the features for vectors (k% of the size of the corresponding cluster)

3. Update $X = (x_1, x_2, \ldots, x_n) 2 \leq c \leq n$ (dataset of vectors) with new feature

4. Pass X to FCM so it starts updating the membership function

5. Stop if after clustering result of FCM no relevance feedback (RF) is requested

4 Experiment

4.1 Datasets and Experiments Settings

We used freely available queries that are provided by the World Intellectual Property Organization (WIPO) to form a gold standard. The result sets were gained through the European Patent Fulltext (EPFULL) repository. For this work, we used a set, namely *contact lenses* to be clustered and another set, *robotic arms* to show an insight of the built gold standard. The code was implemented in Python using Natural Language Toolkit (NLTK) for tokenization and stemming but for removing the stop words, we used our list of stop words for patent analysis. For Fuzzy c-means clustering we adopted the code provided in Github [3]. The overlapping is controlled by allowing the patents that are at least 99% similar to a member of the cluster inside the cluster as a member. The fuzzifier is set on 1.2, the error on 0.001 and the number of iterations on 200. We run the method for c=3, c=13 and c=23 and for visualization in 2D, we have used Principal Component analysis (PCA).

[3] https://github.com/holtwashere/PossibilisticCMeans

4.2 Evaluation Metrics

Evaluation measures used for granulated clusters (e. g. k-means clusters) such as Dunn, Mutual Information (MI), F-measure, Rand Index and Jaccard are not quite useful for measuring overlapping clusters; for instance Dunn gives a higher score to the clustering systems that assign the data points to more distant clusters while the contents of each cluster are pretty close. Considering a criterion like uses, we know that one patent might have several usages and such measures are not revealing any required information. Purity exposes the very nature of a cluster: the degree of consistency. The higher is the purity the less is random clustering. This is one important characteristic of clusters in the patent clustering task, nevertheless we cannot neglect the drawbacks of Purity: it is highly dependent on the number of clusters. Like Purity, MI is also influenced by the number of clusters, while Normalized Mutual Information (NMI) enables us to compare the clusters with each other (it ranges between 0 and 1) and it is not affected greatly by the inaccurate number of clusters. For assessing the quality of this clustering, we have used Normalized Mutual Information (NMI) by Fred and Jain [FJ03] and Purity [MRS08a].

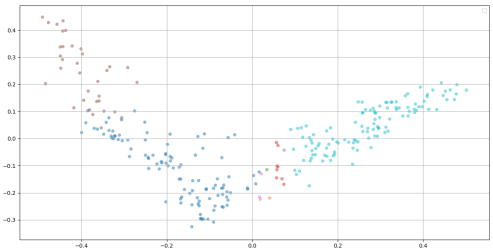

Figure 2: FCM results on 13 clusters under topic Contact Lenses

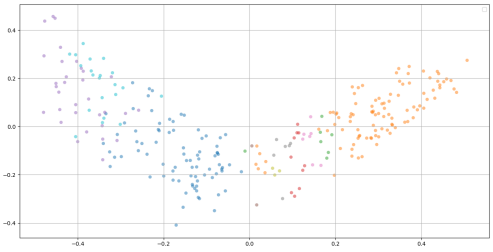

Figure 3: FCM and RF results on 13 clusters under topic Contact Lenses

4.3 Gold Standard based on World Intellectual Property Organization (WIPO) Reports

There are a number of technological reports available by WIPO under the universal title of *public health/life science*. These reports usually cover two types of queries; one results in the main class (usually shares the title with the topic that is reported) and one is breaking the main query into pieces; thus produces subclasses. Some reports categorize the foregoing subclasses and provide fewer yet more general subclasses. The following example depicts search strings provided by WIPO report which results to a main class, namely *robotic arms* followed by a further query that along with the main query leads into a subclass of *robotic arms*:

1. **Query to the main class *robotic arm*:**

   (FT=(robot* or (artificial w/2 intelligence) or android or cyborg or humanoid*)) or (TAC= (manipulator* or manipulater* or actuator* or actuater* or drives or joint or joints or actuation or ("end effector" or "end effecter") or ((pneumatic* or air) w/2 muscle*))) and ((IC= B25J9/02 or B25J9/04 or B25J9/06 or B25J13/02 or B25J13/08 or B25J17 or B25J18) or (UC=901/2 or 901/14 or 901/19 or 901/27 or 901/31 or 901/39 or 700/245 or 700/248 or 700/261))

2. **Query to its subclass *Anthropomorphic Robot*:**

   (TAC) contains (humanoid or android or anthropomorphic* or anthropomorfic*)

   Query guide: FT-Full Text, TAC- Title Abstract Claim, IC- International Class, UC- US Class. w/2 shows the maximum number of intervening unmatched positions doesn't exceed 2.

Depending on the authors of the reports, queries are described in different languages and formats. Therefore the queries had to be adjusted to match Json data type. Using Elasticsearch's API, we collected data through the EPFULL database and obtained the main class and subsequently the corresponding subclasses. For instance under the topic *robotic arms*, 519 patents were retrieved, out of which 511 patents were covered by subclasses from which 293 hits belong to the subclasses of the *types* criterion, 474 hits fit in the subclasses of *applications* criterion and the

Table 1: Evaluation results of clustering of the topic Contact Lenses

#Clusters	NMI		Purity	
	FCM	FCM&RF	FCM	FCM&RF
3 clusters	0.18	0.22	0.76	0.77
13 clusters	0.05	0.16	0.78	0.77
23 clusters	0.09	0.13	0.76	0.70

subclasses of *parts* criterion cover 701 hits. Patents may share different subclasses and some available patents in EPFULL may not be covered by any subclass. The results of the proposed clustering method have been examined against these retrieved subclasses.

5 Results

The results of clustering under the topic *contact lenses* shows using pseudo relevance feedback can definitely help patents picking the better cluster. According to table 1, while the Purity results remain more or less the same the normalized mutual information (NMI) improved 4% to 11% for all three number of clusters. Purity is significantly dependent on the number of clusters as when the data themes are noticeably fewer than the number of clusters the chance of having more consistent clusters would raise. However The results of Purity under the topic *Contact Lenses* do not change diversely for different number of clusters and it remain relatively high. This reflects the acceptable ability of FCM in this clustering. Moreover, visualizing clustering at 13 clusters with FCM (figure 2) and FCM with influence of RF (figure 3) using PCA, we observe the data points are less scattered after using RF.

6 Future work

The main purpose of this study is to test whether the use of relevance feedback in clustering can play an important role in grouping patents. In future we would use explicit relevance feedback of the expert users to modify the dimension of the vectors with more useful features. We would also make use of lexical resources on top of the feedbacks. In future we would also survey the position of feedback vectors as they are selected by users and may not be necessary appear around the centroid.

References

[Bez81] James C. Bezdek. *Pattern Recognition with Fuzzy Objective Function Algorithms*. Kluwer Academic Publishers, Norwell, MA, USA, 1981.

[CHPT05] Bernard Chen, R. Harrison, Yi Pan, and Phang C. Tai. Novel hybrid hierarchical-k-means clustering method (h-k-means) for microarray analysis. In *Proceedings of the 2005 IEEE Computational Systems Bioinformatics Conference - Workshops*, CSBW '05, pages 105–108, Washington, DC, USA, 2005. IEEE Computer Society.

[Cor07] The Thomson Corporation. Global patent sources: An overview of international patents. 2007.

[DD09] Türkay Dereli and Alptekin Durmuşoğlu. Classifying technology patents to identify trends: Applying a fuzzy-based clustering approach in the turkish textile industry. *Technology in Society*, 31(3):263 – 272, 2009.

[FJ03] Ana L N Fred and Anil K. Jain. Robust data clustering. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, 2003.

[FMS+15] Noushin Fadaei, Thomas Mandl, Michael Schwantner, Mustafa Sofean, Julia M. Struß, Katrin Werner, and Christa Womser-Hacker. Patent analysis and patent clustering for technology trend mining. In *HIER workshop*, Hildesheim, Germany, July 2015.

[Hon09] Soonwoo Hong. The magic of patent information. *World Intellectual Property Organization (WIPO). Available via DIALOG.*, December 2009.

[HXW+12] Ming Huang, Zhixun Xia, Hongbo Wang, Qinghua Zeng, and Qian Wang. The range of the value for the fuzzifier of the fuzzy c-means algorithm. *Pattern Recogn. Lett.*, 33(16):2280–2284, December 2012.

[MRS08a] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008.

[MRS08b] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, Cambridge, UK, 2008.

[NNB15] Janmenjoy Nayak, Bighnaraj Naik, and HS Behera. *Computational intelligence in*

data mining. Springer, New Delhi, India, 2015.

[Off18] European Patent Office. Epo quality report 2017. 2018.

[THTL06] Amy J. C. Trappey, Fu-Chiang Hsu, Charles V. Trappey, and Chia-I Lin. Development of a patent document classification and search platform using a back-propagation network. *Expert Syst. Appl.*, 31:755–765, 2006.

[TJC04] Y. H. Tseng, D. W. Juang, and S. H. Chen. Global and local term expansion for text retrieval. In *Proceedings of the fourth NTCIR workshop on evaluation of information retrieval, automatic text summarization and question answering*, Tokyo, Japan, June 2004.

[TLL07] Yuen-Hsien Tseng, Chi-Jen Lin, and Yu-I Lin. Text mining techniques for patent analysis. *Inf. Process. Manage.*, 43(5):1216–1247, September 2007.