A Comparison of Patent Classifications with Clustering Analysis

Chapter · November 2015

DOI: 10.1007/978-3-319-26187-4_38

2 authors:

Mick Smith
North Carolina Agricultural and Technical Sta…

Rajeev Agrawal
Engineer Research and Development Center…

A Comparison of Patent Classifications with Clustering Analysis

Mick Smith

School of Technology
North Carolina A&T State University
1601 E. Market Street
Greensboro, NC 27411
csmith14@aggies.ncat.edu

Rajeev Agrawal, PhD

Department of Computer Systems Technology
North Carolina A&T State University
209 Price Hall
Greensboro, NC 27411
ragrawal@ncat.edu

Abstract. There is an abundance of data and knowledge within any given patent. Through the use of textual mining and machine learning clustering techniques it is possible to discover meaningful associations throughout a corpus of patents. This research demonstrates that such relationships between USPTO patents exist. Through the use of k-means and k-medians clustering, the accuracy of the USPTO classes will be assessed. It will also be demonstrated that a more refined classification process would be beneficial to other areas of analysis and forecasting.

Keywords: Clustering, K-Means, K-Medians, Patent Classification, Machine Learning, Text Mining

1 Introduction:

As the age of big data continues to evolve so does the potential for analytic opportunities within large data sets. One area in particular that may lend itself to ongoing analysis comes in the form of patent analysis. In addition to the quantitative analysis that can be done on the frequency or patterns of patents, there exists massive amounts of knowledge that can be extracted from the textual bodies of each document. It's possible that this knowledge could be used to improve on techniques used for searching and retrieving patents of various classification or content. It is important to improve on this process since the quality of the improper searching may result in unexpected overlapping into another person's intellectual property (IP).

The goal of this research is to determine which clustering method most accurately represents the classification of patents as proposed by the United States Patent and Trademark Office (USPTO). The clustering techniques examined in this paper will be k-means and k-medoids. While this is by no means an exhaustive list of clustering methods, it does offer a good starting point of investigation. Both of these methods are efficient, proven approaches for the clustering of textual documents. The establishment of a good clustering technique will provide a baseline for future research on patent clusters. More specifically, it will provide a proven methodology to be applied in a future quantitative and qualitative data mining analysis.

A lot of benefits could be realized from the development of an efficient mechanism for clustering and classifying patents. For instance, due to the sheer volume of patent data there is an increased chance of copyright infringement. Individuals who desire to submit a patent for a new idea may not be aware that the same concept already exists. Such a violation could be avoided if the patents were correctly classified. Furthermore, if the author of the patent had a mechanism to compare similarity of content to either a single patent or group of patents, then this may result in a reduction of IP infringement. Another benefit comes in the form of business value. New marketing and innovation strategies can be discovered proper classification and analysis of patents [10].

The USPTO attempts to make their current classification as specific as possible. At this time it includes 473 classes and offers even more granularity as there are over 150,000 sub-classifications. Although, such specificity may not be beneficial to the analysis of large number of patents. For instance, if someone wanted to conduct research on all "Green Energy", they could perform a key word search on the USPTO website and return a list of 643 related patents. However, this list may not be inclusive of all patents related to green energy. Another approach may be to drill down into the already defined classes that the USPTO has set forth. Though this method might only yield patents related to a specific subset of green energy. Especially since each patent may have multiple sub-classifications.

To properly encapsulate the scope of a patent search, it is necessary to consider textual and content relationships from one patent document to the next. By extension, it is suggested in this research that more meaningful relationships exist between patents than is indicated by the classification and subsequent sub-classification. It should be mentioned that this study is not fully conclusive and that this research is ongoing.

2 Background

One area in particular that exists as a byproduct of big data is the ability to autonomously retrieve textual information and associate various bodies of text. This need spans anything from a web search conducted through an internet web search engine to in-depth textual analysis and natural language processing. Recent research has used semantic and word analysis to identify hidden relationships in social media comments [12]. Additionally textual mining has been used to analyze and compare patents [2], [8, 9], [18].

Patent Clustering/Classification Techniques

Currently the USPTO suggests using a seven step searching strategy (www.uspto.gov). Their process recommends that the person performing the search, brainstorm some possible relevant terms, and using those terms recursively over multiple searching trials. However, while this method may return a wide breadth of patents related to the search, it is far from an exhaustive list. Furthermore, it is left up to the subjective aptitude of the searching party.

A language based clustering approach was utilized by Kang et. al. [9] to associate patent documents. In their research, log-likelihood term frequency and data smoothing techniques are used to establish a good general information retrieval method. [11] used the k-nearest neighbor algorithm, the maximum entropy model, and support vector machines, to classify Japanese patents. Although, their grouping strategy relied heavily on the existing classifications assigned to the patents.

As proposed in this paper, similar research has been done in the field of patent classification. A back propagation neural network was used by Trappery et. al. [16] to group a specific class of patents down to the sub-class level. Chen and Chang [2] also proposed a three phase categorization method by which each patent is classified down to the same level of detail. While their method offers an incredible amount of granularity, it differs from this research in that their goal is to match patents to already existing classifications. As previously mentioned, this paper aims to demonstrate that there might be "hidden" clusters defined by textual clustering methods that offer better classification than the current USPTO system.

Clustering

The data mining technique of clustering is one that has many benefits and can be utilized in a variety of fields. According to Han & Kamber [6] examples of clustering can be seen in marketing, land use, insurance, city-planning, and earth-quake studies. Clustering is also widely used for text mining, pattern recognition, webpage analysis, and marketing analysis [17]. By grouping and subdividing a collection of documents or keyword it may be possible to discover certain trends that exist in data. This grouping is one of the benefits of using clustering techniques. As it pertains to patent mining and technology forecasting there have been many instances of use in other research [4], [7, 8], [17]. Ruffaldi et. al. [15] suggest that by using patent citations, that it is possible to understand a given technology's trajectory. The quality of a clustering result depends on both the similarity measure used by the method and its implementation. The quality of a clustering method is also measured by its ability to discover some or all of the hidden patterns.

K-Means

K-Means Clustering is a clustering technique that clusters data based on the mean vector distance between data points. It is a recursive algorithm that looks to find the minimal mean distances between pieces of data and then group them together. K-Means clustering often terminates at a local optimum. The global optimum may be found using techniques such as: deterministic annealing and genetic algorithms [6]. According to Chernoff et al. [3] there are various versions of k-means algorithms, depending on the method in which covariance matrices are estimated and the procedure of reclassifying the means. K-means has also been proven to be an effective method for classifying patent documents [2]

There are of course some limitations and weaknesses associated with the k-means clustering method. Han and Kamber [6] offer a list of such weaknesses:

- Applicable only when mean is defined, then what about categorical data?
- Need to specify k, the number of clusters, in advance
- Unable to handle noisy data and outliers
- Not suitable to discover clusters with non-convex shapes

K-Medoids

The K-Medoids clustering method is similar to the K-Means method in that both attempt to minimize the vector distance between data points. The difference between the two is that instead of finding the nearest mean the K-Medoids looks for the center points as defined by the data and works to improve that data center through the use of an arbitrary distance matrix. A common algorithm used for K-Medoids is PAM (Partitioning Around Medoids). K-Medoids clustering has a computational advantage over K-means clustering. K-Medoids clustering finds the representative objects (medoids) in clusters [8].

Other Clustering Methods

Sometimes when trying to describe a system it is necessary to use a non-discrete metric. This is even truer when attempting to classify two or more potentially similar groups. Through the use of fuzzy logic it is possible to use varying degrees of membership to measure levels of similarity. The principles of fuzzy logic can be extended to the concept of fuzzy clustering. Goswami & Shishodia [5] have shown that it is possible to use the fuzzy c-means (FCM) clustering algorithm to associate different collections of texts. In their approach they use a bag of words approach to determine the frequency of words in a document. Some important steps in their process include pre-processing the text, feature generation, and feature selection. The large dimensionality of textual collections is also challenging for classification algorithms and can drastically increase running time. Researchers apply dimensionality reduction before running the classification algorithm to alleviate these challenges [1].

Li et. al. [13] proposed the use of two different document clustering algorithms, Clustering based on Frequent Word Sequences (CFWS) and Clustering based on Frequent Word Meaning Sequences (CFWMS). Their algorithms focused on the sequence of words as opposed to the bag of words approach. The reason behind this was to cluster based on implied meanings within the document instead of word frequency. In their algorithms, each document is reduced to a compact document by keeping only the frequent words. In the compact document, they kept the sequential occurrence of words untouched to explore the frequent word sequences. By building a Generalized Suffix

Tree (GST) for all the compact documents, the frequent word sequences and the documents sharing them are found. Then, frequent word sequences are used to create clusters and describe their content for the user [13]. Ideally a comprehensive metric of text semantic similarity should be based upon the relation between the words in addition to the role played by the various entities involved in the interactions described by each of the two texts. Following this, the semantic similarity of textual components is based upon the similarity of the component words in them [12].

3 Experimentation:

This research will utilize various data clustering methodologies to better assess the quality of patent classifications by the USPTO. To achieve the results for this research several applications were used. Figure 1 details the process used to create the desired clusters. All of the patents used in this research were obtained from the UC Berkley Fung Institute (https://github.com/funginstitute/downloads). The UC Berkley patent data was extracted from the USPTO website and converted from XML to a SQLite table structure. The database covers all US patents from 1976 to 2013 and consists of 4,823,407 patents. While it would be ideal to perform clustering on all patents, the sheer size and volume patents makes that unrealistic for a short term analysis. To address this issue a smaller sample of patents was extracted. Queries were run on the SQLite database to extract patents that meet the following characteristics:

- Granted between 1995 and 2013
- Contains the word "Technology" in either the abstract or title
- For each patent, the following information was extracted:
  - Patent title, Abstract, USPTO Class, USPTO Patent Number

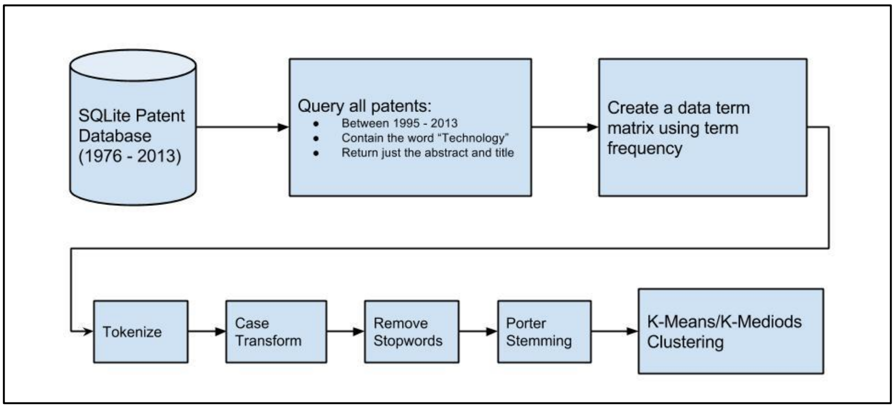

Fig. 1. – Workflow for Patent Clustering

These initial queries reduced the number of patents to 9,087. However, this was too large a sample to conduct preliminary analysis on, so further filtration was carried out.

A random sample of 449 patents was taken from the reduced set of 9,087 and it consisted of 10 USPTO classes. Table 1 contains the list of classes, descriptions, quantities of patents that are used in this experimentation. The data was saved in .csv format and the titles and abstracts were concatenated to create one "document" per row.

The .csv file was then loaded into RapidMiner so that a term frequency data term matrix could be created. RapidMiner was used to tokenize the text in each row, make all words lower case, remove stop words, and perform Porter stemming. The resulting data term matrix is then exported to another .csv file. This data term matrix that was consisted of 449 rows and 1,417 columns. R was then used to perform various clustering analyses on the data set.

USPTO Class	Classification Description	Quantity
257	Active solid-state devices (e.g., transistors, solid-state diodes)	44
340	Communications: electrical	50
359	Optical: systems and elements	51
365	Static information storage and retrieval	55
370	Multiplex communications	48
435	Chemistry: molecular biology and microbiology	8
439	Electrical connectors	54
514	Drug, bio-affecting and body treating compositions	46
705	Data processing: financial, business practice, management, or cost/price determination	50
707	Data processing: database and file management or data structures	43

Table 1. – Technology Patent Classifications, Descriptions, and Quantities

Two types of clustering was performed in R, k-means and k-medoids. The first objective of our research was to determine if each algorithm could accurately approximate the classifications assigned by the USPTO. For this reason the number of centers (k) was set to 10. Table 2 shows the clustering results for k-means while Table 3 shows the results for k-medoids.

A second objective of this research was to determine if the clustering carried out by the USPTO was accurate. In other words, for the group of patents in this study, what is the optimal number of groupings? To address this a few different approaches were used, first a sum of squared error scree plot was created for the k-means clustering. This graph can be seen in Figure 2, where the objective is to observe an abrupt bend in the curve which would indicate a suggested number of clustering centers. From observing the plot in Figure 2, there doesn't appear to be a clear point where the curve bends.

	Clusters										
USPTO Class	1	2	3	4	5	6	7	8	9	10	
257		39	1	1	2			1			**44**
340		2	1		33			1	4	9	**50**
359		7	7		11		19	4	3		**51**
365		1		39	5			1	1	8	**55**
370					30			2	6	10	**48**
435									8		**8**
439		1	46		3			3	1		**54**
514						31			15		**46**
705	11				8				27	4	**50**
707				1	8			1	19	14	**43**
Total	11	50	55	41	100	31	19	13	84	45	**449**

Table 2. – K-means Cluster Results

	Clusters										
USPTO Class	1	2	3	4	5	6	7	8	9	10	
257	34	2	3	2	2	1					**44**
340	3	1	17	1		1	12	15			**50**
359	7	32	5	5	1				1		**51**
365			3	1	40			10		1	**55**
370		1	19	2	4		5	17			**48**
435				1	2				5		**8**
439	2	1	2	3	1	42	2		1		**54**
514			1		1		1		43		**46**
705			7		2		2	22	4	13	**50**
707			7	1	2		2	29	1	1	**43**
Total	46	37	64	16	55	44	24	93	55	15	**449**

Table 3. – K-medoids Cluster Results

Another approach to finding the right number of clusters is to use a gap statistic. Table 4 shows the gap statistic and standard error for both the k-means and k-medoids results. As the highlighted row indicates, the gap statistic suggests 14 clusters for both k-means and k-medoids. It should be noted that the gap statistic identifies the smallest value of k such that *f(k)* is not more than one standard error away from the first local minimum [14].

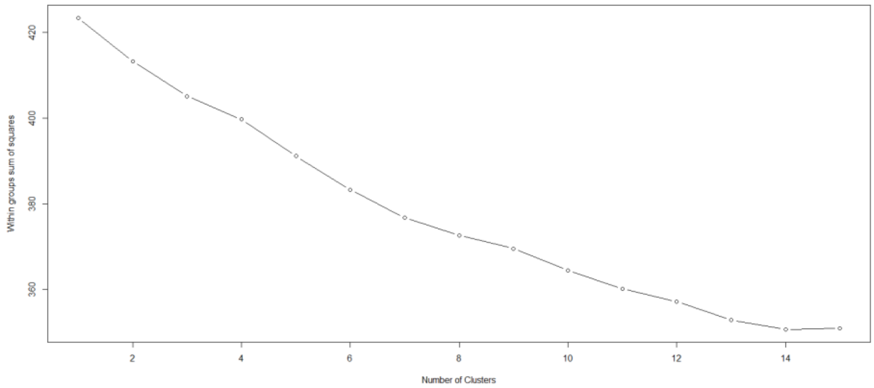

Fig. 2. – K-means SSE Scree plot

	k-means gap	k-means SE	k-medoids gap	k-medoids SE
Cluster 1	0.69245	0.00191	0.69159	0.00193
Cluster 2	0.69303	0.00171	0.69630	0.00188
Cluster 3	0.69861	0.00164	0.69984	0.00170
Cluster 4	0.70298	0.00165	0.70463	0.00197
Cluster 5	0.70837	0.00160	0.71212	0.00179
Cluster 6	0.71341	0.00166	0.71853	0.00179
Cluster 7	0.71763	0.00177	0.72389	0.00156
Cluster 8	0.72012	0.00157	0.72818	0.00161
Cluster 9	0.72537	0.00164	0.73156	0.00165
Cluster 10	0.72801	0.00151	0.73414	0.00147
Cluster 11	0.73048	0.00166	0.73712	0.00156
Cluster 12	0.73310	0.00150	0.74055	0.00182
Cluster 13	0.73534	0.00161	0.74380	0.00178
Cluster 14	0.73809	0.00165	0.74538	0.00195
Cluster 15	0.73913	0.00175	0.74710	0.00205

Table 4. – Gap Statistics

Unfortunately, the silhouette plots (figures 3 – 6) for each clustering approach don’t reveal much of an association between the generated clusters and the data points assigned to them. In the graphs, the clusters are listed in order from top to bottom with each cluster element's silhouette score displayed. The scores can range between -1 and 1, scores closer to 1 are desirable as they demonstrate a stronger association. Low or negative scores may also indicate either too many or too few clusters. However, as it has been stated, this research is in the preliminary stages, and further tweaking and adjustments of the parameters of each algorithm should produce stronger results.

What may be of more interest is to notice that when the number of clusters increased to 14, as suggested by the gap statistic, the strength of the silhouette scores and plots improved. This may be initial support to one of the running arguments in this paper. The classification of patents might be done in a more efficient manner if the context of each patent is assessed analytically instead of subjectively.

To further support these findings the data was also transformed using Principal Component Analysis (PCA). One of the known problems with the clustering of vectors from a data term matrix is that there may be a lot of sparseness within the matrix. This may reduce the quality of the clustering results. To overcome this problem PCA can be applied and reduce possibility of problems due to dimensionality issues. When applied to this data the resulting gap statistic suggested 14 clusters as well.

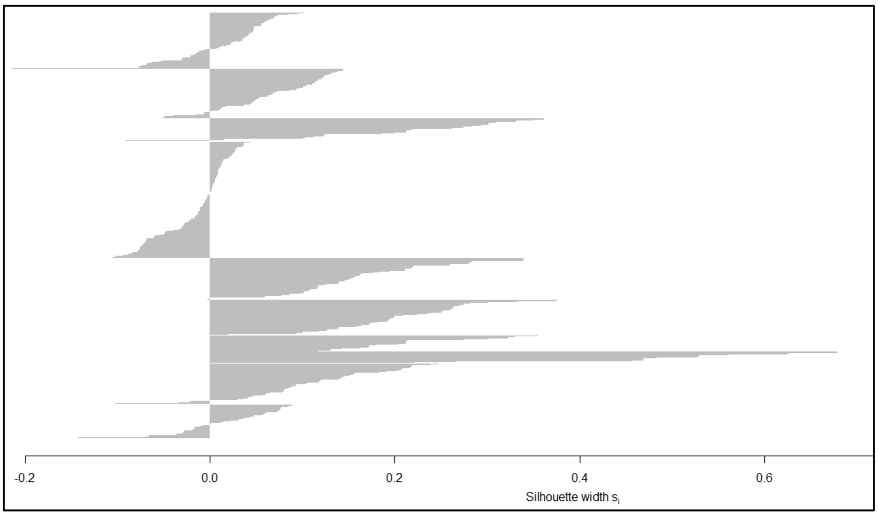

Fig. 3. – Silhouette Plot of k-means (k = 10)

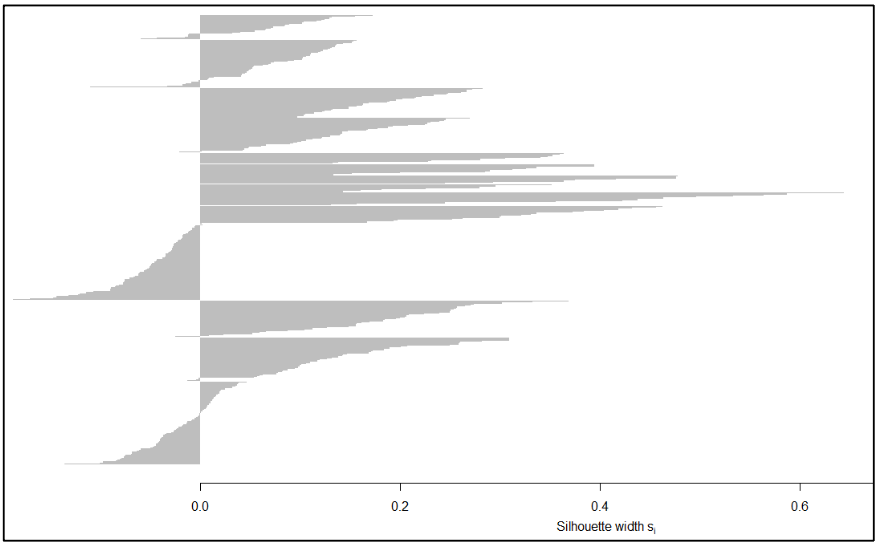

Fig. 4. – Silhouette Plot of k-means (k = 14)

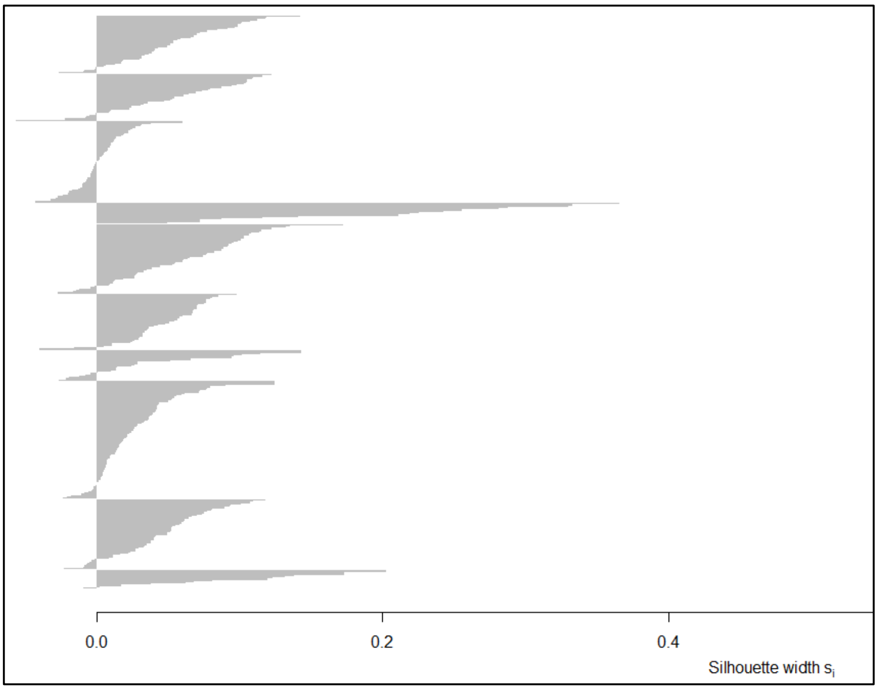

Fig. 5. – Silhouette Plot of k-medians (k = 10)

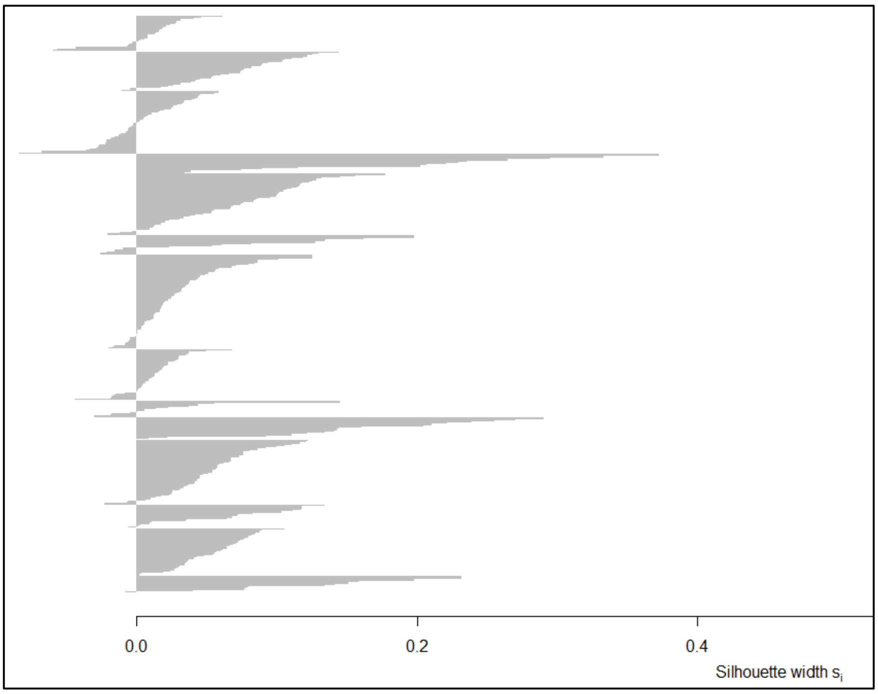

Fig. 6. – Silhouette Plot of k-medians (k = 14)

One of the challenges of this experimentation is that the cluster classification was not carried through to the final results. While this makes it impossible to construct a true confusion matrix, it allows for interpretation beyond the initial USPTO classification. Still a method for associating the generated clusters to the initial patent groups is good for visual and contextual reference. Especially in the case of 14 clusters. Observing what patents were deemed to be outside the scope of the suggested USPTO class is an important step toward the creation of an alternative classification scheme.

The assignment of USPTO classes back to the created clusters starts with the transformation of the cluster results table to indicate the quantity of patents in each cluster as a percentage. This is shown in Table 5. To assign a cluster to a Classification, the steps outlined with the pseudo code in Figure 7 were followed. The highlighted cells in Table 5 indicate the assigned class.

After associating a USPTO class to a cluster, the seven most frequent terms from each cluster were extracted. As it can be seen in Table 6, the frequent terms do seem to coincide with the USPTO classification descriptions. The terms listed are in order of occurrence rank within the cluster. Also, clusters 1, 3, 7, and 8 were not assigned a class and may warrant a new classification.

```
While column j in classification matrix M has no asso-
ciated row
  For each row i in classification matrix M
    x = max(mi1, mi2,… mij,)
    class assignment is given to the column j associ-
ated with x
    if column j is already assigned to a row then
      Next
    else
      assign class value for column j to row i
    end if
  Next
End While
```

Fig. 7. – Pseudo Code for Assigning USPTO Classes to Clusters.

Cluster	257	340	359	365	370	435	439	514	705	707
1	0.14	0.10	0.67	0	0	0	0.07	0	0	0
2	0.80	0.03	0.12	0	0	0	0.03	0	0	0
3	0.22	0.26	0.08	0.08	0.12	0	0.14	0	0.02	0.06
4	0.13	0.06	0.26	0.06	0.13	0.06	0.2	0	0	0.06
5	0.04	0.02	0	0.77	0.04	0.04	0.02	0.02	0.02	0.02
6	0	0.73	0.06	0	0	0	0.06	0.06	0.06	0
7	0	0.10	0.02	0.13	0.24	0	0	0	0.20	0.28
8	0	0.27	0	0	0.35	0	0.02	0	0.22	0.12
9	0	0.08	0	0.23	0.38	0.07	0	0	0.23	0
10	0	0	0.94	0	0	0	0.06	0	0	0
11	0	0	0.02	0	0	0.07	0	0.86	0.03	0
12	0	0	0	0	0.06	0	0	0	0.33	0.61
13	0	0	0	0	0	0	1	0	0	0
14	0	0	0	0	0	0	0	0	0.92	0.08

Table 5. – K-Medoids Cluster Results by Percentage (k =14)

Cluster	USPTO Class	Initial Quantity	Cluster Quantity	USPTO Classification Description	Terms
1	n/a	n/a	28	n/a	optical, light, surface, includes, system, member, plane
2	257	44	31	Active solid-state devices (e.g., transistors, solid-state diodes)	layer, substrate, region, formed, device, semiconductor, dielectric
3	n/a	n/a	49	n/a	device, semiconductor, devices, system, signal, control, network
4	435	8	15	Chemistry: molecular biology and microbiology	sub, signal, surface, conductive, lens, film, light
5	365	55	48	Static information storage and retrieval	memory, voltage, cell, line, device, bit, cells
6	340	50	15	Communications: electrical	battery, circuit, rfid, communication, power, message, electrical
7	n/a	n/a	74	n/a	data, system, method, memory, includes, plurality, device
8	n/a	n/a	40	n/a	information, network, user, method, system, based, data
9	370	48	13	Multiplex communications	signal, sequence, reference, using, method, value, assets
10	359	51	17	Optical: systems and elements	lens, group, side, power, positive, sub, zoom
11	514	46	51	Drug, bio-affecting and body treating compositions	methods, invention, thereof, treatment, use, pharmaceutical, relates
12	707	43	18	Data processing: database and file management or data structures	search, content, query, results, method, web, user
13	439	54	37	Electrical connectors	connector, includes, portion, housing, terminal, body, contact
14	705	50	13	Data processing: financial, business practice, management, or cost/price determination	order, system, orders, trading, method, deposit, received

Table 6. – Patent Cluster Classifications based on USPTO classes

4 Conclusion:

In this paper a subset of technology patents were clustered using k-means and k-medians clustering. The goal was to determine how accurate these methods could be and if the classifications suggested by the USPTO were sufficient for further patent analysis. The quality of the classification scheme is extremely important to those looking to delve deeper into trends and themes that are available in a patent corpus. It is almost a certainty that the quality of more advanced textual mining and natural language processing of patents would be improved with a less subjective grouping.

This research demonstrated that it was possible to create clusters of patents based on the frequency of terms within each patent. The initial assessments certainly suggest that it might be possible to improve on the accuracy of the clustering methods by adjusting and testing different parameters. However, since many variations have already been checked, the enhancements may be minimal. In line with the second goal of this research, the results showed that it is possible to generate new clusters and classifications from existing ones. It may be possible to categorize each new cluster better with Topic Modeling techniques such as Latent Dirichlet allocation. This will be the focus of future related research.

5 References:

1. Blake, C.: Text mining, *Annual Review of Information Science and Technology*, 45(1), 121 – 155, (2011)
2. Chen, Y-L, Chang, Y-C: A Three-Phase Method for Patent Classification, *Information Processing and Management*, 48, 1017 – 1030 (2012)
3. Chernoff, H., Gillick, L.S., Hartigan, J.A.: k-Means Algorithms, *Encyclopedia of Statistical Sciences*, Volume 6, 3858 – 3859 (2006)
4. Chou, L-Y: Knowledge Discovery through Bibiometrics and Data Mining: An Example on Marketing Ethics, *International Journal of Organizational Innovation*, 3, 106 – 139 (2011)
5. Goswami, S., Shishodia, M.S.: A Fuzzy Based Approach to Text Mining and Document Clustering, *International Journal of Data Mining & Knowledge Management Process*, 3(3), 43 – 52 (2013)
6. Han, J., Kamber, M., Pei, J.: Data mining: Concepts and Techniques, Elsevier, Waltham, MA (2012)
7. Hsu, C.C., Huang, Y.-P., Chang, K.-W.: Extended Naïve Bayes Classifier for Mixed Data, *Expert Systems with Applications*, 35, 1080 – 1083 (2008)
8. Jun, S., Park, S.S., Jang, D.S.: Technology Forecasting Using Matrix Mapping and Patent Clustering, *Industrial Management & Data Systems*, 112, 786 – 807 (2011)
9. Kang, I-S, Na, S-H, Kim, J., Lee, J-H.: Cluster Based Patent Retrieval, *Information Processing and Management*, 43, 1173 – 1182 (2007)
10. Kasravi, K., Risov, M.: Patent Mining - Discovery of Business Value from Patent Repositories, Proceedings of the Fortieth Annual Hawaii International Conference on System Sciences, Waikoloa, Hawaii, USA (2007)
11. Kim, J-H, Choi, K-S: Patent Document Categorization Based on Semantic Structural Information, *Information Processing and Management*, 43, 1200 – 1215 (2007)
12. Karmakar, S., Zhu, Y.: Mining Collaboration through Textual Semantic Interpretation, *2011 11th International Conference on Hybrid Intelligent Systems (HIS)*, 728 – 733 (2011)
13. Li, Y., Chung, S.M., Holt, J.D.: Text Document Clustering Based on Frequent Word Meaning Sequences, *Data & Knowledge Engineering*, 64, 381 – 404 (2008)

14. Maechler, M.: "Finding Groups in Data": Cluster Analysis Extended Rousseeuw et al, Package "Cluster" (R Documentation), July 21, 2015, retrieved from: https://cran.r-project.org/web/packages/cluster/cluster.pdf (2015)
15. Ruffaldi, E., Sani, E., Bergamasco, M.: Visualizing Perspectives and Trends in Robotics based on Patent Mining, *2010 IEEE International Conference on Robotics and Automation*, May 3-8, 2010, Anchorage, Alaska, USA (2010)
16. Trappery, A.J.C., Hsu, F-C, Trappery, C.V., Lin, C-I: Development of a Patent Document Classification and Search Platform Using a Back-Propagation Network, Expert Systems with Applications, 31, 755 – 765 (2006)
17. Trappey, C.V, Wu, H-Y, Taghaboni-Dutta, F., Trappey, A.J.C.: Using patent data for technology forecasting: China RFID patent analysis, *Advanced Engineering Informatics*, 25, 53 – 64 (2011)
18. Tseng, Y.H., Lin, C.J., Lin, Y.I.: Text Mining Techniques for Patent Analysis, *Information Processing and Management*, 43, 1216 – 1247 (2007)