J Syst Sci Syst Eng (Jun 2010) 19(2): 162-181
DOI: 10.1007/s11518-010-5134-x

ISSN: 1004-3756 (Paper) 1861-9576 (Online)
CN11-2983/N

CLUSTERING PATENTS USING NON-EXHAUSTIVE OVERLAPS

Charles V. TRAPPEY[1] **Amy J.C. TRAPPEY**[2,3] **Chun-Yi WU**[4]

[1] *Department of Management Science, National Chiao Tung University, Taiwan, China*
trappey@faculty.nctu.edu.tw

[2] *Department of Industrial Engineering & Management, National Taipei University of Technology, Taiwan, China*
trappey@ntut.edu.tw (✉)

[3] *Department of Industrial Engineering & Engineering Management, National Tsing Hua University, Taiwan, China*
trappey@ie.nthu.edu.tw

[4] *Department of Industrial Engineering & Engineering Management, National Tsing Hua University, Taiwan, China*
d9534524@oz.nthu.edu.tw

Abstract

Patent documents are unique external sources of information that reveal the core technology underlying new inventions. Patents also serve as a strategic data source that can be mined to discover state-of-the-art technical development and subsequently help guide R&D investments. This research incorporates an ontology schema to extract and represent patent concepts. A clustering algorithm with non-exhaustive overlaps is proposed to overcome deficiencies with exhaustive clustering methods used in patent mining and technology discovery. The non-exhaustive clustering approach allows for the clustering of patent documents with overlapping technical findings and claims, a feature that enables the grouping of patents that define related key innovations. Legal advisors can use this approach to study potential cases of patent infringement or devise strategies to avoid litigation. The case study demonstrates the use of non-exhaustive overlaps algorithm by clustering US and Japan radio frequency identification (RFID) patents and by analyzing the legal implications of automated discovery of patent infringement.

Keywords: Data mining, patent analysis, patent infringement, non-exhaustive overlap clustering, ontology schema

1. Introduction

Knowledge management has reshaped the competitive structure of industry and has led to dramatic changes in the way enterprises retrieve, store, and share information. Many different approaches and algorithms have been devised to discover knowledge from large sets of data. The use of clustering enables user to group data according to their characteristics when the predefined rules of categorization are not available (Berry & Linoff 1997). Clustering methods can be effectively applied to group patent documents using key phrases that are pre-defined with a domain ontology. The key phrases and the patent clusters help to describe technology trends, processes, and innovations.

A patent is a legal document that grants exclusive use and rights to the inventor or assignee. A patent records information about the technology and describes the processes related to the invention (WIPO 2009a). Effective analysis of patent documents and technology trends better enables industry to invest in research and development that does not duplicate or infringe upon the rights of others with prior claims. Patents are issued and classified into predefined categories according to international patent classification rules (WIPO 2009b) or by United States patent classification rules (USPTO 2009). The fast emergence of patents has led to discrepancies within and between the classification rule systems. Thus, researchers must search broad ranges of patent categories to document the prior art of related patents.

The specific technical terms, the legal writing style, and the length of patents documents places great demands on analysts to define emerging technologies and defend of existing claims. The cost of patent analysis is on the rise across industries and across technical domains (WIPO 2009c). Industries are especially interested in studying innovations and the overlaps of technical findings and claims. Therefore, the purpose of this research is to further automate the process of patent analysis through the use of a new clustering algorithm that allows non-exhaustive overlap clustering of patents. The paper is organized in several sections. Section 2 reviews the related patent analysis literature and clustering approaches. Section 3 develops the automated procedure and algorithms for clustering patent documents using non-exhaustive overlaps. Section 4 provides the results of an RFID patent clustering experiment and analysis. The research compares the non-exhaustive overlap approach with the K-means clustering approach using a collection of RFID patents.

2. Literature Review

The literature review provides a background on patent analysis, a comparison of different clustering algorithms, and a review of the latest developments in non-exhaustive and overlapping cluster analysis.

2.1 Patent Analysis

A patent document in standard format describes the exclusive rights granted by government organizations to the inventor or assignee for a given period of time (WIPO 2009d). According to the report of the World Intellectual Property Organization (WIPO 2009e), patent documents reveal more practical insights about core technologies and innovations than academic articles. The rate of filing patents is also used as a proxy measure of national technology development and economic growth (Grilliches 1990).

Globally, product life cycles are becoming shorter which places pressure on R&D teams to provide quicker solutions to satisfy changing market demand. Patent analysis provides a means to predict product maturity and market trends (Trappey, Trappey, Hsu & Hsiao 2009). Patent analysis is employed across organizations and is a research approach frequently used by R&D engineers, academics, patent attorneys, and technology policy makers. The results of patent analysis are used to estimate trends, profitability, and performance of technology around the world (Paci et al. 1997).

Hong (2009) divides patent analysis into quantitative and qualitative approaches. Qualitative analysis focuses on the content of the patent and often uses text mining to extract the substance of patent documents. Quantitative analysis uses bibliographic information (e.g., inventors, assignees, applied date, issued date, citation, and other characteristics) to extract the patent metadata. The patent metadata is in turn used for advanced searches and information extraction about patent activity by inventors, companies and countries. The competition among companies is revealed using metadata, and careful analysis of this information often results in the discovery of new product development opportunities (Gupta 1999).

There is a time lag when a patent is first applied for and when the patent is officially issued or denied. Initial applications (or patents pending) provide useful information about early launches of commercial products and the emergence of technology trends, whereas the robustness, scope, and multiple claims of existing patents provide valuable information about strategic technology boundaries. Likewise, the number of patents within a specific domain is a valuable predictor of the technology life cycle. Granstrand (1999) proposed that successful patent analysis reduces R&D costs and strengthens the ability of a firm to develop market strategies. Patent analysis is most often applied by large technology firms while smaller firms often hire consulting companies for knowledge extraction (Mogee 2000). Patent analysis plays an important role in the effective operation of enterprises with patents being the core value of corporations' intellectual assets (Lai & Wu 2005, Kim et al. 2008).

2.2 Cluster Analysis

Supervised classification is used to group objects by a predefined criterion. Clustering, on the other hand, is the unsupervised classification of patterns into groups based on similarities of internal features or characteristics. Clustering is useful for exploratory pattern analysis, group-oriented decision making, and machine learning. The applications for clustering include targeting document retrieval, image segmentation, and pattern classification (Anderberg 1973). Clustering divides data into several groups and is applied when the criteria for classification are not defined. The clustering result can also be used as a reference for defining classification criteria (Han & Kamber 2000). Clustering methodologies have been applied across varied domains including marketing segmentation, web analysis, and computational biology (Berkhin 2002).

Cluster analysis is a sequential process consisting of data collection, measuring the similarity of data, choosing the most suitable clustering methodology, evaluating the performance of the chosen methodology, and interpreting the clustering results by domain experts. The two basic objectives of clustering are to maximize the similarity of objects within the same cluster and the diversity of objects between clusters. In order to achieve these objectives, the data sets are usually represented in terms of their relative position in space using Euclidean or Manhattan distances (MacQueen 1967).

One of the most widely used clustering algorithms is the K-means algorithm. There are several limitations cited with the K-means approach (Chen et al. 2005). First, the user must assign a number k as the expected number of clusters. Since the centroids of clusters are randomly chosen, the algorithm must repeat many times to adjust centroids and can only achieve locally optimal clustering results. Further, the algorithm is sensitive to outliers. As a response to these problems, an improved K-medoids method was developed using k objects as cluster centers instead of k calculated centers (Han & Kamber 2000). The advantages over the K-means algorithm include the reduction of the effect of outliers, the use of real objects to represent the cluster center, and the categorical clustering of non-numeric data (Hu 2006).

Other methods commonly used for cluster analysis include tree clustering, self-organizing maps, adaptive resonance theory, and the genetic K-means algorithm. Han, Kamber & Berkhin (2000, 2002) have organized and described these clustering techniques as classes that include partitioning methods, hierarchical methods, density-based methods, grid-based methods, model-based methods, ANN-based methods, and generic-based methods. Chen, Tai, Harrison & Yi (2005) propose a hybrid clustering approach which combines hierarchical and K-means methods. Further, Hu (2006) proposes the non-exhaustive overlapping clustering algorithm with the assumption that an object can belong to different clusters or can be an outlier without belonging to any cluster.

2.3 Non-exhaustive Overlapping Clustering

The underlying concept of non-exhaustive overlapping clustering (Hu 2006) is that an object may possess characteristics which make it acceptable for membership in multiple clusters. Since patent documents describe a broad spectrum of innovations and techniques, non-exhaustive clustering is useful for grouping patents that include multiple features or multiple claims to innovation. Further, clustering patents with non-exhaustive overlaps allows the assignment of a patent to multiple clusters helps analysts detect innovation crossovers that occur during the evolution of technology.

This algorithm computes the distance of two objects (x_i and x_j) using the Euclidean formula (1). A threshold T_s for similarity is defined to determine whether two objects (x_i and x_j) belongs to a cluster and the similarity between object x_i and x_j is expressed using Equation (2).

$$d_{ij} = \left(\sum_{d=1}^{k} \left| x_{id} - x_{jd} \right|^2 \right)^{1/2} \tag{1}$$

$$Cor_{i,j} = 1 - \min\{d_{ij}, d_{if}\} / d_{if} \tag{2}$$

where d_{if} is the top 5 percentile of distances between all object pairs.

If d_{ij} is larger than d_{if} , then the similarity ($Cor_{i,j}$) is 0. If $Cor_{i,j}$ is larger than the predefined threshold T_s , the object x_i and x_j should belong to the same cluster. In other words, the threshold T_s determines whether an object x_i belongs to the cluster while using x_j as the cluster center. Thus, a matrix can be derived from the correlation of all objects. The non-exhaustive overlap algorithm maximizes the number of objects in a cluster while maximizing the distances between pairs of cluster centers using the key values derived by the Cluster Recommending Factor (CRF) Equation (3).

$$CRF(x_i) = w1 * C_v(x_i) + w2 * Mdv(x_i)\,, \qquad (3)$$

where $C_v = nx_i / \max v$ is crowding value and $Mdv(x_i) = ndx_i / \max d$ is maximum distance value. The value n_{xi} represents the number of objects in the cluster center x_i , and ndx_i represents the distance of object x_i to the nearest center. The detailed algorithm is described by Hu (2006).

3. Headings

The methodology covers patent preprocessing, key phrase clustering, patent document clustering, cluster interpretation, and technology analysis. The procedural flow for the non-exhaustive overlap algorithm is shown in Figure 1.

The key phrase extraction is considered a pre-process prior to the execution of the non-exhaustive clustering of key phrases and patent documents. This research uses a statistic and semantic based approach to extract key phrases. The former approach is calculated using a normalized term frequency and inverse document frequency (NTF-IDF), and the latter approach builds a specific ontology using a

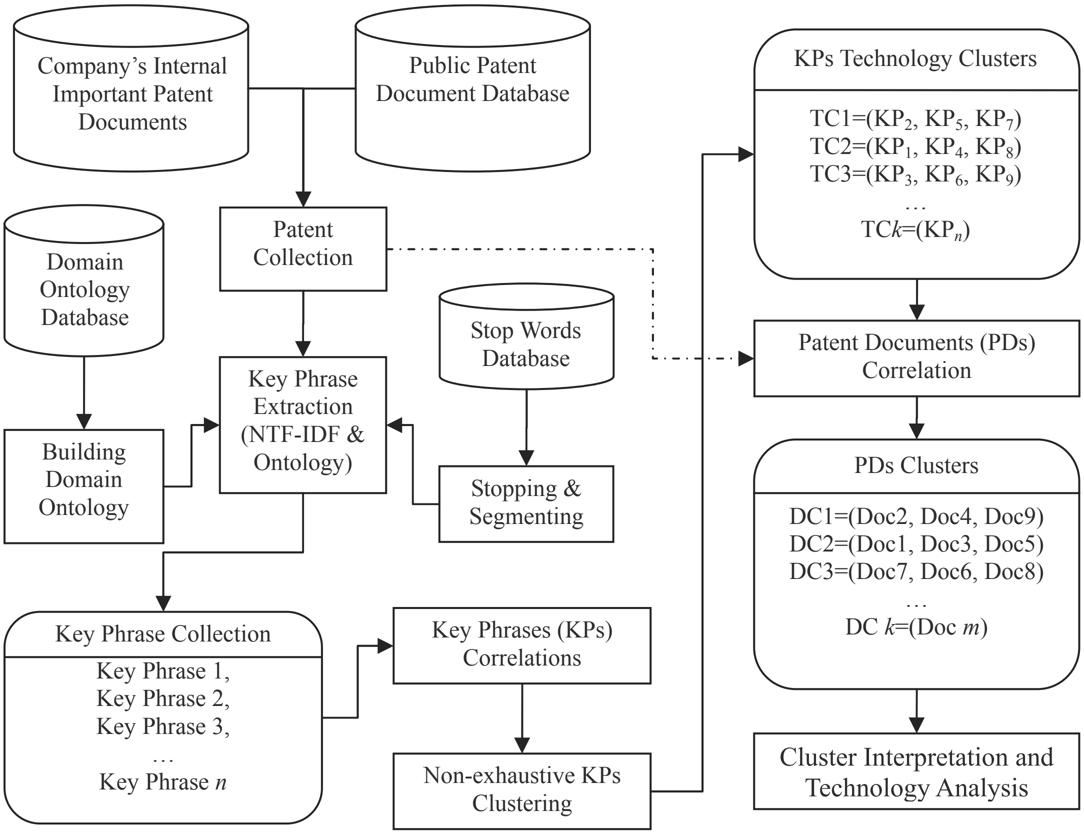
Company's Internal Important Patent Documents
Public Patent Document Database
Domain Ontology Database
Patent Collection
Stop Words Database
Key Phrase Extraction (NTF-IDF & Ontology)
Building Domain Ontology
Stopping & Segmenting
Key Phrase Collection
Key Phrase 1,
Key Phrase 2,
Key Phrase 3,
…
Key Phrase n
Key Phrases (KPs) Correlations
Non-exhaustive KPs Clustering
KPs Technology Clusters
TC1=(KP$_2$, KP$_5$, KP$_7$)
TC2=(KP$_1$, KP$_4$, KP$_8$)
TC3=(KP$_3$, KP$_6$, KP$_9$)
…
TCk=(KP$_n$)
Patent Documents (PDs) Correlation
PDs Clusters
DC1=(Doc2, Doc4, Doc9)
DC2=(Doc1, Doc3, Doc5)
DC3=(Doc7, Doc6, Doc8)
…
DC k=(Doc m)
Cluster Interpretation and Technology Analysis

Figure 1 The procedure for patent clustering and interpretation

domain expert to extract component phrases from the patent document database. The detailed methods and algorithms of the ontology based key phrase extraction are described by Trappey, Trappey & Wu (2009); and Chen & Wu (2005).

3.1 Key Phrase Correlation and Clustering

Assigning key phrase weights by the term's appearance frequency (TF) in a document is a popular method in text mining (Luhn 1957). However, this method may misjudge the importance of key phrases. A common phrase (or non-unique phrase) that appears frequently in a document may not be important if it also appears in many other documents. In order to consider the uniqueness of key phrases, Salton & Buckley (1988) propose using the Term Frequency and Inverse Document Frequency (TF-IDF) approach which is modified from Jones (1972). Since this method does not account for the different number of words in each document (Sedding & Kazakov, 2004), the weights for the frequency of key phrases are normalized by the number of words in the documents. The function normalized TF-IDF or NTF approach is expressed as Equation (4).

$$NTF = tf_{ik} \times \frac{\sum_{s=1}^{n} WN_s}{n} \times \frac{1}{WN_k} \qquad (4)$$

where tf_{ik} is the number of key-phrase i in document k, WN_k is the words number of document k, and n is the total number of documents in the document set.

The correlation between key phrases is calculated using the inner product of vectors as expressed in Equation (5).

Correlation (KP_i, KP_j) =

$$\frac{KP_i \cdot KP_j}{\|KP_i\|\|KP_j\|} = \frac{\sum_{k=1}^{n} w_{ik} \times aw \times w_{jk} \times aw}{\sqrt{\sum_{k=1}^{n} w_{ik}^2 \times aw^2 \times \sum_{k=1}^{n} w_{jk}^2 \times aw^2}} \qquad (5)$$

where $KP_i = aw\left(w_{i1}, w_{i2}, \ldots, w_{in}\right)$,

$aw = \frac{\sum_{s=1}^{n} WN_s}{n \times WN_k}$ = average Word Number (WN),

$w_{ik} = tf_{ik} \times idf_i$,

df_i = the number of documents of key-phrase i in document set,

$idf_i = \log_2\left(\frac{n}{df_i}\right)$, and

n = the total number of documents in document set.

The correlations of key phrases are expressed as a symmetrical matrix used for clustering as shown in Table 1. The clustering with non-exhaustive overlaps uses the threshold T_S for similarity to determine whether two key phrases (x_i, x_j) belong to a cluster.

The similarity of object x_i and x_j is expressed as $CoR_{i,j}$. If $CoR_{i,j}$ is larger than the threshold T_S , then x_i and x_j belong to the same cluster (noted as *Y*). Otherwise, these two phrases with dissimilar features are noted as *N*. The *Y*/*N* binary matrix is obtained from the matrix of key phrase correlations as shown in Table 2. *Y* indicates a significant relationship between x_i and x_j whereas *N* denotes an insignificant relationship.

The two objectives of clustering with non-exhaustive overlaps are to maximize the number of key phrases in a cluster while maximizing the distances between each cluster center. The first objective is satisfied using the

Table 1 Key phrases correlation matrix

	KP_1	KP_2	KP_3	KP_4	…	KP_t
KP_1	$CoR_{1,1}$	$CoR_{1,2}$	$CoR_{1,3}$	$CoR_{1,4}$	…	$CoR_{1,t}$
KP_2	$CoR_{2,1}$	$CoR_{2,2}$	$CoR_{2,3}$	…	…	…
KP_3	$CoR_{3,1}$	$CoR_{3,2}$	…	…	…	…
KP_4	$CoR_{4,1}$	…	…	…	…	…
⋮						
KP_t	$CoR_{t,1}$	…	…	…	…	$CoR_{t,t}$

KP_1 to t is extracted from the collected patent documents
$CoR_{t,t}$ is defined as the correlation of object pairs

Table 2 The Y/N binary matrix

	x_1	x_2	x_3	x_4	…	x_{t-1}	x_t	CRF
x_1	Y	Y/N	Y/N	Y/N	…	Y/N	Y/N	$CRF(x_1)$
x_2	Y/N	Y	Y/N	Y/N	…	Y/N	Y/N	$CRF(x_2)$
x_3	Y/N	Y/N	Y	Y/N	…	Y/N	Y/N	$CRF(x_3)$
x_4	Y/N	Y/N	Y/N	Y	…	Y/N	Y/N	$CRF(x_4)$
⋮								
x_{t-1}	Y/N	Y/N	Y/N	Y/N	…	Y	Y/N	$CRF(x_{t-1})$
x_t	Y/N	Y/N	Y/N	Y/N	…	Y/N	Y	$CRF(x_t)$

x_1 to t represent the patent documents
Y/N determines whether the object pair belongs to the cluster
CRF(x_t) is defined as the cluster recommending factor for x_t

crowding value which maximizes the phrases in clusters. The second objective, the maximum distance value, expresses the maximal distance between cluster centers. The crowding value and the maximum distance value are combined with the weights *w*1 and *w*2 as the Cluster Recommending Factors (Equation 3). The phrase with a higher CRF value is selected as the cluster center. After all potential cluster centers are obtained, the objective function of the complementary combination of cluster centers are calculated using Equation (6).

$$Obj_{function} = w_1 \times \mathrm{Min}\left[\frac{1}{S_{\max}(x_{C1})}, \frac{1}{S_{\max}(x_{C2})}, \ldots, \frac{1}{S_{\max}(x_{Ck})}\right] + w_2 \times \frac{\sum_{i=1}^{k} Cv(x_{Ci})}{k}, \tag{6}$$

where $S_{\max}(x_{c1})$ =the largest similarity value of x_{c1}, $C_v = n_{xi} / n_{\max}$, and k (number of clusters) is predefined by the user.

A set of cluster centers is replaced to form a new set when a new combination yields a higher objective function value. This operation repeats a predefined number of times to obtain the best set of cluster centers for key phrase clustering.

3.2 Patent Document Correlation Calculation and Clustering

The key phrases are clustered into k clusters with non-exhaustive overlaps called Technology Clusters (TCs). The term weights for the key phrases of patent document j in TC_i are cumulated and noted as F_{ij}, which provides the measure to determine whether the patent document belongs to the TC groups (Table 3). Afterward, the patent documents are non-exhaustively clustered following the same procedures as key phrase clustering (Figure 2).

Table 3 Calculation of patent indices for technology clusters

	$Patent_1$	$Patent_2$	$Patent_3$	$Patent_4$	…	$Patent_n$
TC_1	F_{11}	F_{12}	F_{13}	F_{14}	…	F_{1n}
TC_2	F_{21}	F_{22}	F_{23}	F_{24}	…	F_{2n}
TC_3	F_{31}	F_{32}	F_{33}	F_{34}	…	F_{3n}
TC_4	F_{41}	F_{42}	F_{43}	F_{44}	…	F_{4n}
…	…	…	…	…	…	…
TC_m	F_{m1}	F_{m2}	F_{m3}	F_{m4}	…	F_{mn}

TC_m represents the technology cluster m

F_{mn} represents the total NTF-IDF of patent n's key phrases belonging to TC_m

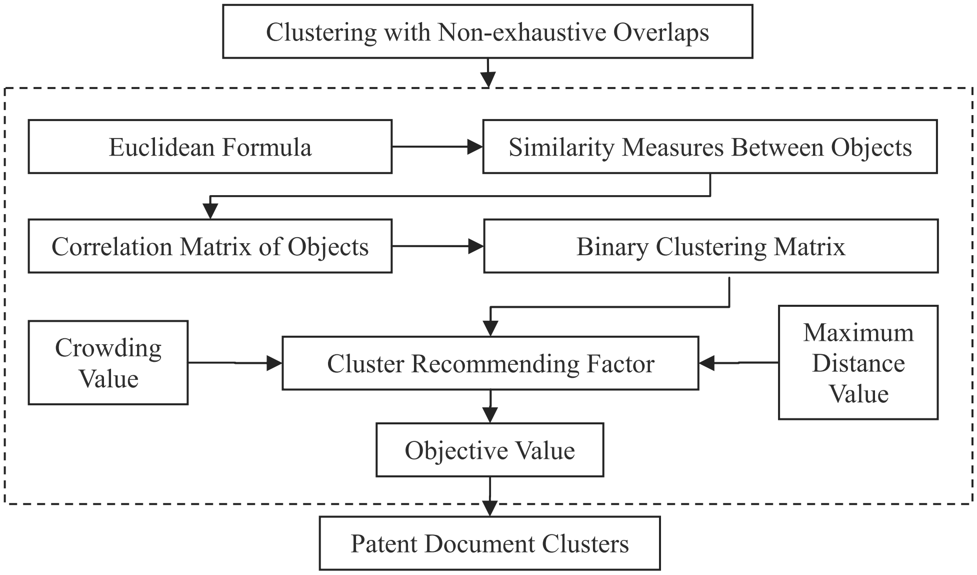
Clustering with Non-exhaustive Overlaps
Euclidean Formula
Similarity Measures Between Objects
Correlation Matrix of Objects
Binary Clustering Matrix
Crowding Value
Cluster Recommending Factor
Maximum Distance Value
Objective Value
Patent Document Clusters

Figure 2 The procedure for patent clustering and interpretation

3.3 Technology Analysis

After creating a database of representative patents in a given domain, this research derives the key phrases and determines the number of patent clusters. Each cluster contains related patent documents with overlapping claims and innovations. These patent characteristics are used to analyze the trend of technology development trends. Based on the time sequence and frequency of patents filed, analysts can track the growth and forecast the maturity of patents. The life cycle of specific techniques can also be evaluated to measure the linkages between technology development and market growth.

4. Implementation and Evaluation

This section discusses the procedures using a predefined ontology schema, key phrase extraction, and non-exhaustive overlaps. In order to evaluate the patent document clustering on an actual dataset, this research collects 160 RFID patents issued by the United States Patent Trade office (USPTO 2009) and the Japanese patent Office (JPO 2009). These patents include RFID antennas, readers, tags and systems and where collected from the online database using the International Patent Classification system definition for RFID. After clustering the patent documents, the proposed clustering algorithm is compared to the centroid based K-means method.

4.1 Key Phrase Extraction and Clustering

The proposed system in key phrase extraction requires the development of a domain ontology by an expert prior to patent document analysis. Each node of the patent technology ontology tree represents specific domain concepts which correspond with key phrases in the patent document. Thus the system collects key phrases according to the ontology schema and records the term frequencies. Figure 3 illustrates the tree structure of the RFID ontology which is constructed hierarchically using a formal ontology engineering methodology (Trappey et al. 2008). Ontological engineering is the process of developing an ontology model, which involves iterative steps in the entire lifecycle of the ontology design, including domain questions, formal competency questions, existent reference information, search phrase methods, and proposed hierarchical methods. After the construction of RFID ontology (Trappey et al. 2009), the preprocessing of patent document extracts the key phrases using the ontology schema and the normalized TF-IDF method.

Table 4 provides an example of the NTF-IDF of the key phrases (KPs). After extraction from patent document number JP 2008-204234. Table 5 shows the partial NTF-IDF matrix for KPs extracted from 80 training patent documents. After extracting the KPs from the patent documents, the system calculates the NTF-IDF values, ranks the top KPs and creates a KP correlation matrix.

For the system implementation and test, 80 RFID patents are used as training documents and another 80 patents are used as test documents. The number of technology clusters (TCs) is set at 5, the non-exhaustive threshold value is set at 0.9, and the weights of the crowding value and maximum distance is set to 0.5. After computing the key phrase correlation matrix, the Cluster Recommending Factor (CRF) values, the crowding values (CV), and the maximum distance values (MDV), the cluster centers are computed as shown in Table 6. Moreover, Figure 4 illustrates the results of Table 6 with overlapping key phrase clustering. As shown in Figure 4, only cluster TC1 is exclusive and the other clusters are mutually overlapping with common key phrases.

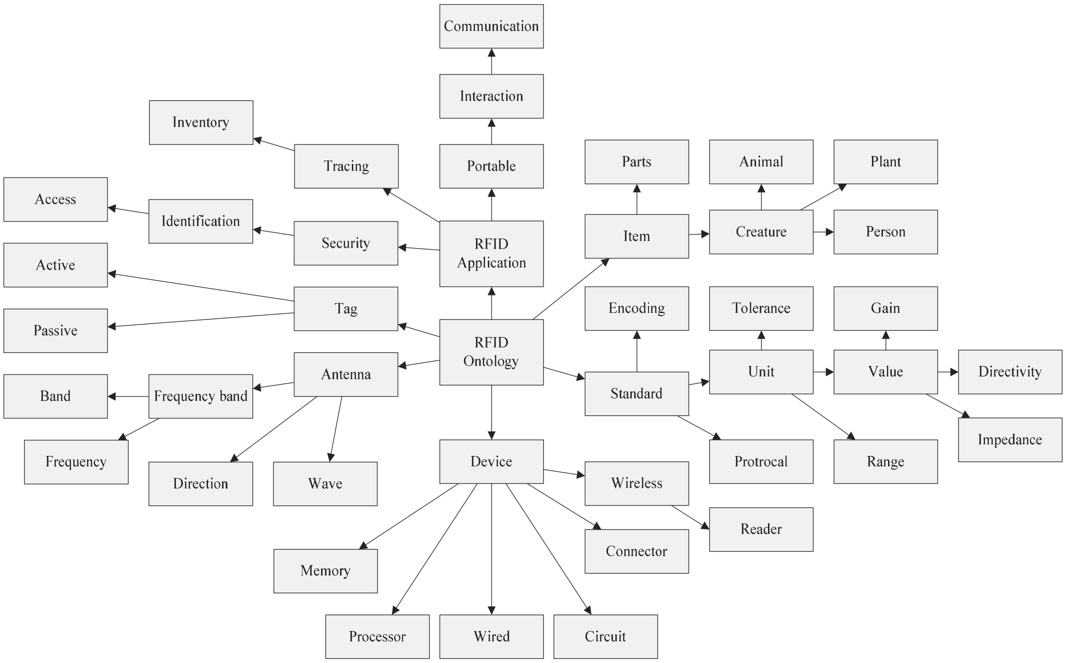

Figure 3 The RFID ontology tree structure (Trappey et al. 2007)

Table 4 The NTF-IDF of key phrases for patent document number JP 2008-204234

Sort	Phrase	TF	IDF	TF-IDF	NTF-IDF
1	rfid tag built	45	1.322	59.487	68.343
2	reader	26	1.322	34.37	39.487
3	sensor input terminal	23	1.322	30.404	34.931
4	driving signal	16	2.322	37.151	42.682
5	rfid chip	17	1.322	22.473	25.819
6	leader	14	1.322	18.507	21.262
7	phase difference sensor	11	2.322	25.541	29.344
8	reflector	9	2.322	20.897	24.009
9	switch control circuit	8	2.322	18.575	21.341
10	rfid tag communication	8	1.322	10.575	12.15
11	rfid communications department	6	1.322	7.932	9.112
12	power supply circuit	6	1.322	7.932	9.112
13	radio law	5	2.322	11.61	13.338
14	analog signal	5	2.322	11.61	13.338
15	communication range	6	2.322	13.932	16.006

Table 5 The partial NTF-IDF matrix for key ontology concepts extracted from US patents

Key Concepts\Patent No.	59386562	5939984	6100804	6677852	7516057	7518513
Antenna	2.2	2.9	14.9	1.6	5.9	94.1
Direction	0	0	0	4.2	4.2	0
Reader	1.1	1.0	0	6.3	0	37.2
Frequency	1.7	0	1.7	1.7	0	0
Identification	6.8	0	0	0	0	0
Tracing	0	0	3.7	1.8	1.8	1.8
Tag	50.1	41.2	100.8	144.5	83.1	222.9
Encoding	1.9	0	0	0	0	9.8
Range	0	0.9	0.9	0	5.9	0

Table 6 TCs, the cluster centers and the overlapping KPs belonging to the TCs

Technical Cluster	Center (KP No.)	Center (KP Name)	Other KPs in the Technology Cluster
TC1	KP8	connected	Housing, coupling coil, bar antenna, outer antenna, antenna resonance circuit
TC2	KP42	reader writer device	rfid tag, rfid reader writer, rfid tag device
TC3	KP1	drawing	rfid tag, type tactile sensor, antenna, rfid type tactile, reader writer, transponder, example, id supporting structure, plate, attachment, reader
TC4	KP22	transponder	Drawing, reader writer, reader
TC5	KP44	reader	Drawing, rfid tag built, transponder

After deriving the technical clusters with their key phrase groupings, patent documents use the TCs as features to process documents into clusters applying the same non-exhaustive overlapping approach. Using proposed approach, given threshold values influence the results of the clustered patent documents. If the similarity of the patent documents is greater than the set threshold value, the patent documents can be accepted into the given cluster. Otherwise, the documents will be excluded from the clusters. For the case study, the threshold value is set at 0.9 and twenty-two patents were included in five non-exhaustive document clusters (DCs). The results of clustering are shown in Table 7. Figure 5 shows the patent DCs generated after 100 iterations.

The first patent cluster contains antenna, circuit structure, transponder, and tag key phrases. This cluster aggregates innovations for antennas and antenna application related technology. The second cluster is related to RFID reader designs and systems that integrate readers, tags and antennas. The third cluster describes RFID reader and writer configuration technology. The fourth cluster represents RFID tag applications and related technical approaches. The fifth cluster groups RFID systems and methods with tags that utilize signal and information processing techniques. The listing the WIPO classifications and definitions by cluster as listed in Table 8. The table shows that

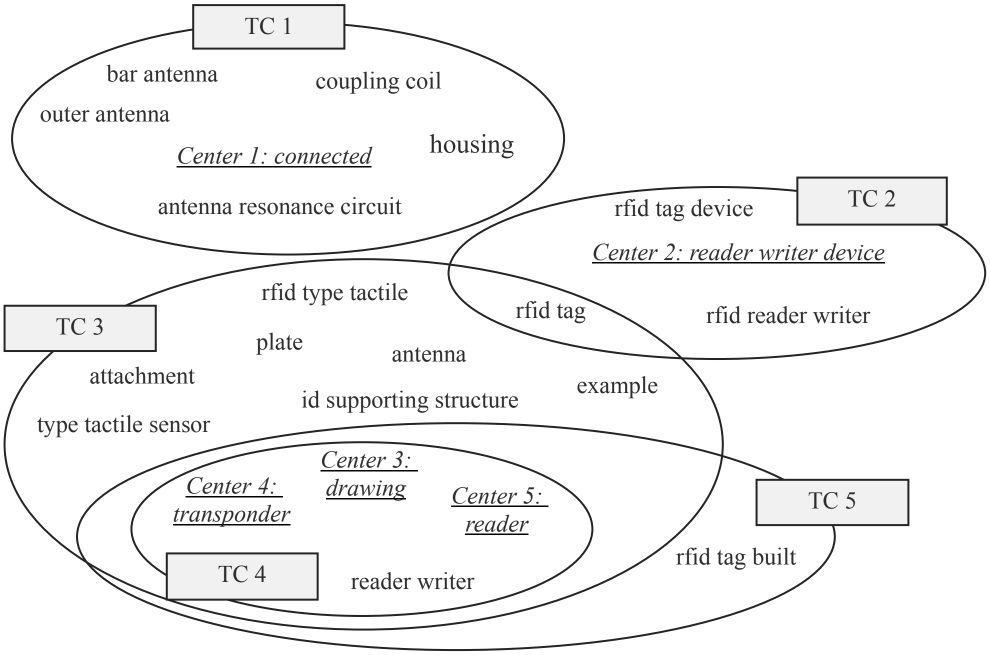

Figure 4 The technical clusters with overlapping key phrases

the fact that the overlapping technical nature of patents categorized into any given ICP code. WIPO classification overlaps in cluster 1 and cluster 5 for G08 class (signaling).

4.2 System Evaluation

The results of fuzzy partition clustering in Figure 5, is compared to the K-means centroid clustering approach in Figure 6. Figure 6 depicts the same patent documents from the test stage clustered using K-means, where k is assigned to 5. The K-means approach clusters RFID tag technical patent documents into the first cluster whereas RFID antennas are described in the second cluster. The third cluster refers to RFID reader applications and the fourth cluster contains the largest set of patents related to RFID systems, antennas, transponders and tags. The fifth cluster is a outlier containing only one patent related to tag holder design.

Given these two results, a hypothetical case for patent infringement litigation is proposed. In order to document evidence for the claim, lawyers must analyze and compare the claims of related patents and the prior art for the valid assessment of the claim. For instance, a local company manufactures and sells RFID tag reader systems. The reader system responds to a radio tag even if communication quality deteriorates due to interference the peripheral environment. Suppose that the assignee of the Japanese patent (JP 2007-065976) claims that the RFID tag reader systems and the applied technology of the local company have infringed upon their intellectual property. Thus, the local company must try to prove that the alleged infringement is invalid by analyzing and demonstrating prior art. Using the overlapping clusters as shown in Figure 7, the patent in question (JP 2007-065976, Application Date: 2005/08/31) belongs to two clusters: DC 2 and DC 4. Patents with similar prior art are depicted in DC 2 and DC 4 (i.e., US 5973599, Application Date: 1997/10/15, and JP

Table 7 The document clusters derived using the non-exhaustive overlaps

Clusters	Patent No.	Patent Title
DC 1	US 5206657 (Center)	Printed circuit radio frequency antenna
	US 4001830	Radio receiver set with rotatable antenna
	US 4625212	Double loop antenna for use in connection to a miniature radio receiver
	JP 2006-012129	Antenna array for RFID reader compatible with transponders operating at different carrier frequencies
	JP 2005-064865	Antenna for RFID compatible with three frequencies
DC 2	JP 2007-049554 (Center)	RFID reader management system, and program
	JP 2007-065976	Radio tag reader and RFID system
	US 5621199	RFID reader
	US 6429776	RFID reader with integrated display for use in a product tag system
	US 6903656	RFID reader with multiple antenna selection and automated antenna matching
	JP 2006-245951	Reader/writer device
	US 7515049	Extended read range RFID system
DC 3	JP 2006-025304 (Center)	Antenna for RFID reader/writer
	JP 2007-064765	RFID tag device, RFID reader writer device and distance measuring system
	US 5621199	RFID reader
DC 4	JP 2008-172500 (Center)	RFID tag and transmitting/receiving method in RFID tag
	US 5986562	RFID tag holder for non-RFID tag
	US 5973599	High temperature RFID tag
	JP 2008-186398	RFID tag detector, RFID tag detection system, and RFID tag detection method
	JP 2008-159007	RFID tag
	JP 2007-065976	Radio tag reader and RFID system
DC 5	US 7515047 (Center)	RFID conveyor system and method
	US 7515049	Extended read range RFID system
	JP 2006-319730	System and method for receiving RFID tag signal
	US 7518511	Dynamic product tracking system using RFID
	JP 2008-186398	RFID tag detector, RFID tag detection system, and RFID tag detection method

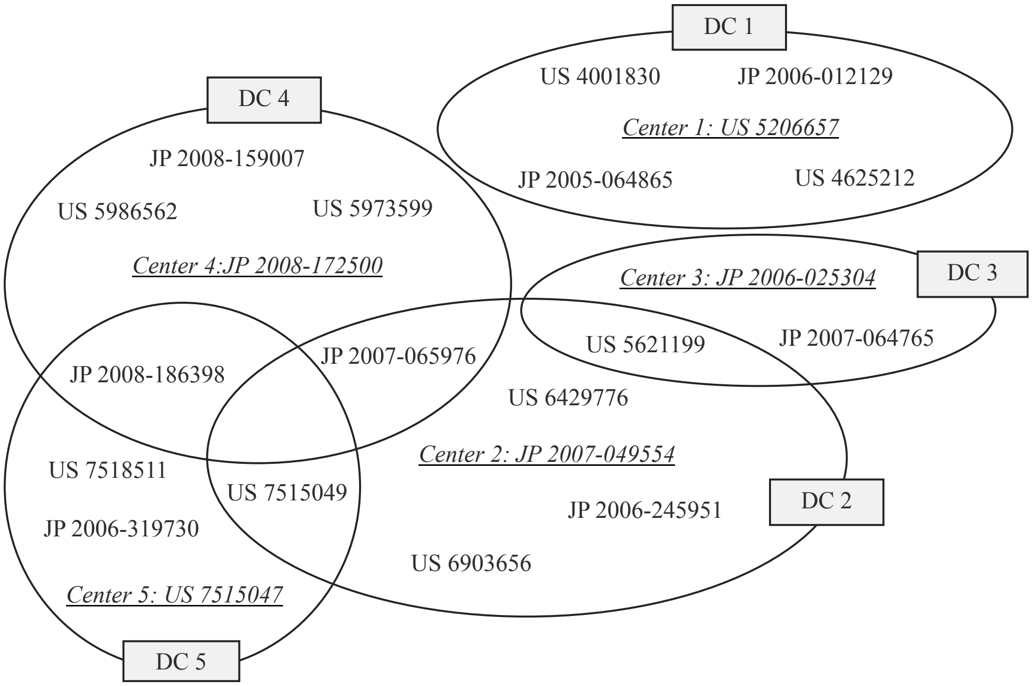

Figure 5 The non-exhaustive clustering result for patents

Table 8 The international patent classification analysis of the clusters

Patent Clusters	WIPO Classification and Definitions
DC 1	H01: Basic electric elements H01Q: Aerials
DC 2	G06: Computing; Calculating; Counting G07: Checking-devices G08: Signaling
DC 3	G01S: Radio direction-finding; Radio navigation; Determining distance or velocity by use of radio waves; Locating or presence-detecting by use of the reflection or reradiation of radio waves; Analogous arrangements using other waves
DC 4	G06: Computing; Calculating; Counting G06K: Recognition of data; Presentation of data; Record carriers; Handling record carriers
DC 5	G08: Signaling G08B: Signaling or calling systems; Order telegraphs; Alarm systems G08B 13: Burglar, theft, or intruder alarms

2007-049554, Application Date: 2005/08/11). Thus, the lawyers for the local company can analyze the claims of the Japanese patent and compares the technology of the overlapping patents. The discovery of overlapping prior art patents would not be possible using the K-means approach.

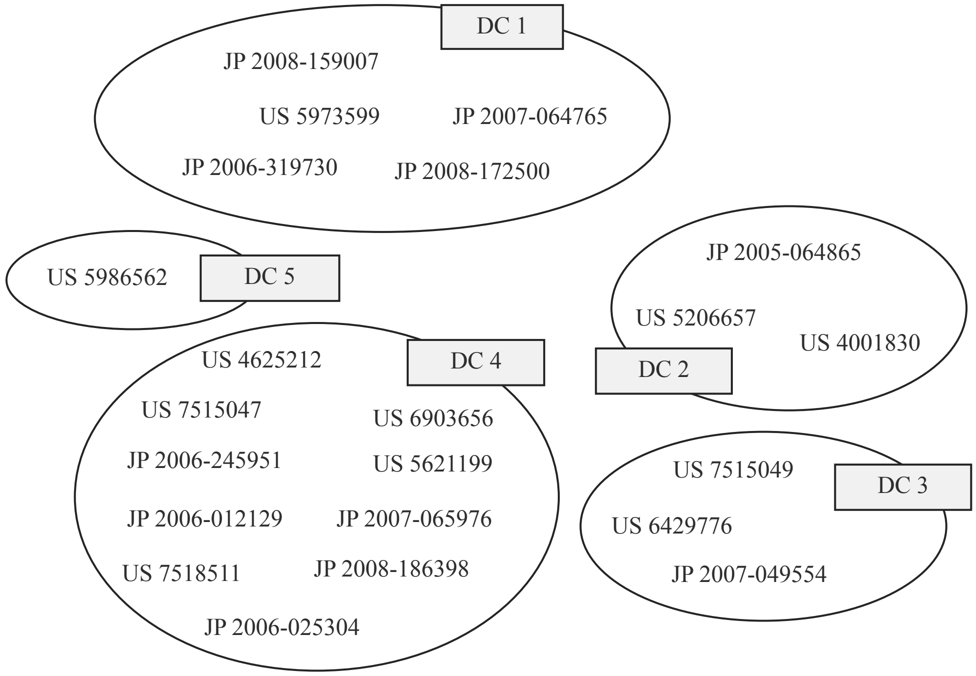

Figure 6 The patent documents clustered into 5 exclusive groups using K-means

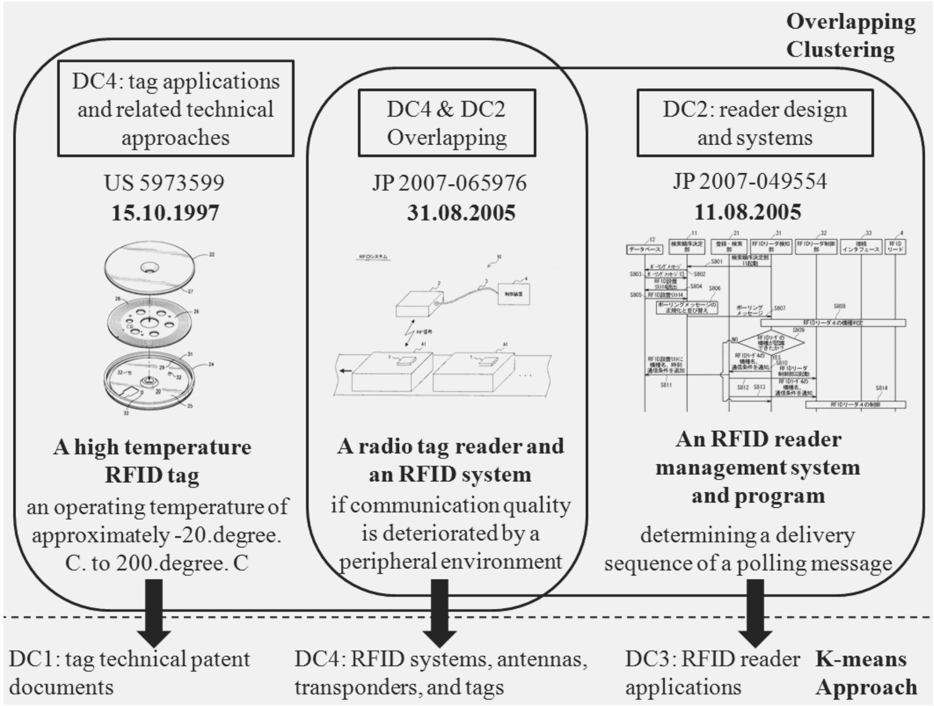
Overlapping Clustering
DC4: tag applications and related technical approaches
US 5973599
15.10.1997
A high temperature RFID tag
an operating temperature of approximately -20.degree. C. to 200.degree. C
DC4 & DC2 Overlapping
JP 2007-065976
31.08.2005
A radio tag reader and an RFID system
if communication quality is deteriorated by a peripheral environment
DC2: reader design and systems
JP 2007-049554
11.08.2005
An RFID reader management system and program
determining a delivery sequence of a polling message
DC1: tag technical patent documents
DC4: RFID systems, antennas, transponders, and tags
DC3: RFID reader applications
K-means Approach

Figure 7 The infringement assessment of patents within overlapping clusters

In addition to patent infringement assessment, overlapping clustering can assist R&D engineers to identify related technologies that can be used to an advance technology for a new patent. Hypothetically, a R&D engineer would like to analyze the target patent (US 5986562) for improving the design of an RFID tag holder (DC 4 in Figure 8). The international patent classification "G06K 19/077" in Table 8 includes recognition and presentation of data, record carriers and handling record carriers. The patent invention is a tag holder with a non-RFID tag. RFID circuitry embedded in the housing is connected to an antenna to provide the non-RFID tag with RFID capabilities. The R&D engineer uses the overlapping clustering approach to find applications for related techniques (Figure 8). For instance, the techniques of protecting, transmitting, and receiving data for RFID tags DC 4 are found in Table 7. The protection method of patent JP 2008-159007 decreases bending strength on an RFID tag circuit chip to prevent an antenna break. Further, the transmitting and receiving method of patent JP 2008-172500 shows how an RFID tag can increase communication distances without increasing the electric field strength of radio waves used for power supply and signaling. Using the information, the R&D engineer explores new designs for RFID tag holder with increased protection capabilities, high security, and low power communications. On the other hand, the K-means approach, only patent US 5986562 is identified as belonging to DC 5 (in Figure 8) and limits the related patents identified for exploratory design.

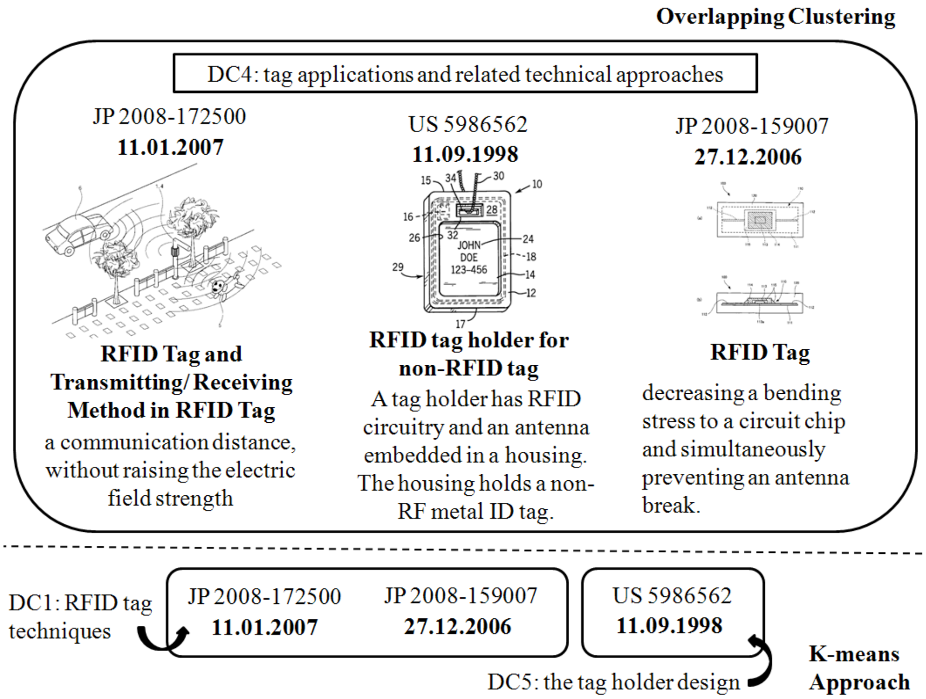
Overlapping Clustering
DC4: tag applications and related technical approaches
JP 2008-172500
11.01.2007
US 5986562
11.09.1998
JP 2008-159007
27.12.2006
15
34
30
10
16
28
26
32
24
JOHN
DOE
123-456
18
14
29
12
17
RFID Tag and Transmitting/ Receiving Method in RFID Tag
a communication distance, without raising the electric field strength
RFID tag holder for non-RFID tag
A tag holder has RFID circuitry and an antenna embedded in a housing. The housing holds a non-RF metal ID tag.
RFID Tag
decreasing a bending stress to a circuit chip and simultaneously preventing an antenna break.
DC1: RFID tag techniques
JP 2008-172500
11.01.2007
JP 2008-159007
27.12.2006
US 5986562
11.09.1998
DC5: the tag holder design
K-means Approach

Figure 8 Referencing patents in the same cluster

Table 9 The comparison of overlapping clusters and K-means clusters

Characteristics	K-means	Overlapping Clusters
Model	Partitioning	Fuzzy partitioning
Input Data	Numeric	Numeric
Cluster Center	Pseudo centroid	Object
Outlier	Influenced	Excluded
Clustered Object	Exhaustive	Non-exhaustive
Clustered Result For IPC Analysis	Difficulty to cluster clearly and meanings of some terms are lost	Each cluster represents a focused meaning and can contain patents with overlapping characteristics

The characteristic comparison between non-exhaustive overlapping and K-means is summarized in Table 9. The traditional K-means approach provides an exhaustive partition where each object belongs to only one cluster. The K-means cluster center is a pseudo centroid that represents the objects located in the same cluster based on their similarity and not following a key patent as the cluster center. Overlapping clusters for multiple characteristics and excludes the influence of outliers. Therefore, the overlapping cluster approach is better suited for patent analysis when a patent often consisting of multiple techniques.

5. Conclusion

This research develops a clustering system to improve the classification of patents with multiple claims and multiple technical innovations. A non-exhaustive overlap algorithm for patent document analysis is derived using key phrase extraction, an ontology, and term frequency weights. The ontology is used to identify domain concepts as key phrases and then the normalized term frequency and inverse document values are used for patent document clustering. The proposed algorithm differs from the traditional clustering since the clusters are allowed to overlap while reducing the effects of outlying patents.

Non-exhaustive overlap clustering maximizes the distance among cluster centers and maximizes the average number of objects contained in a cluster. The system implementation was tested on a set of patents collected from the USPTO and JPO databases. As a result, the proposed method clusters five overlapping key phrase groups as technical centers for cluster formation. These five technical centers represent the key attributes of the patent document clusters. After patent document clustering, the results help identify patent documents with similar technologies. For example, the RFID reader and writer configuration technology overlaps with the information receiving techniques of RFID tags. On the other hand, RFID tag applications, RFID systems integration, and the innovation of antennas have significantly different underlying technologies. Consequently, non-exhaustive overlapping clustering better enables R&D teams to better understand competing patent portfolios. While developing new market strategies, legal advisors can use the approach to identify potential cases of patent infringement or devise strategies for avoiding litigation.

Acknowledgments

This research is partially supported by the National Science Council research grant. Mr. Y.-S. Huang, a M.S. student in IEEM, National Tsing Hua University (Hsinchu, Taiwan), has jointly worked with the authors during the initial case implementation. His contribution is acknowledged with appreciation.

References

[1] Anderberg, M. (1973). Cluster Analysis for Applications. Academic Press, New York

[2] Berkhin, P. (2002). Survey of clustering data mining techniques. Technical Report, Accrue Software, Inc.

[3] Berry, M.J.A. & Linoff, G. (1997). Data Mining Techniques: For Marketing, Sale, and Customer Support. John Wiley & Sons Inc.

[4] Chen, B., Tai, P.C., Harrison, R. & Pan, Y. (2005). Novel hybrid hierarchical k-means clustering method (H-K-means) for microarray analysis. In: Proceedings of Computational Systems Bioinformatics Conference, Sandford CA, USA, August 8-11, 2005

[5] Chen, E. & Wu, G. (2005). An ontology learning method enhanced by frame semantics. In: Proceedings of the Seventh IEEE International Symposium on Multimedia (ISM), 374-382

[6] Chen, T.S., Tsai, T.H., Chen, Y.T., Lin, C.C., Chen, R.C., Li, S.Y. & Chen, H.Y. (2005). A combined k-means and hierarchical clustering method for improving the clustering efficiency of microarray. In: Proceedings of 2005 International Symposium on Intelligent Signal Processing and Communication Systems, 405-408, December 13-16, 2005

[7] Grandstrand, O. (1999). The Economics and Management of Intellectual Property: Toward Intellectual Capitalism. Edward Elgar Publishing

[8] Grilliches, Z. (1990). Patent statistics as economic indicators: a survey. Journal of Economic Literature, 28 (4): 1661-1707

[9] Gupta, V.K. (1999). Technological trends in the area of fullerenes using bibliometric analysis of patents. Scientometrics, 44 (1): 17-31

[10] Han, J. & Kamber, M. (2000). Data Mining: Concepts and Techniques. Morgan Kaufman

[11] Hong S. (2009). The magic of patent information. World Intellectual Property Organization (WIPO). Available via DIALOG. http://www.wipo.int/sme/en/documents/patent_information.htm. Cited December 1, 2009

[12] Hu, H.L. (2006). Optimization in data mining an overlapping cluster algorithm to provide non-exhaustive clustering. Ph.D. Thesis, Department of Information Management, National Central University, Chung-Li, Taiwan, China

[13] Japan Patent Office (JPO). (2009). Available via DIALOG. http://www.jpo.go.jp/. Cited December 1, 2009

[14] Jones, K.S. (1972). A statistical interpretation of term specificity and its application in retrieval. Journal of Documentation, 28 (1): 11-20

[15] Kim, Y.G., Suh, J.H. & Park, S.C. (2008). Visualization of patent analysis for emerging technology. Expert Systems with

Applications, 34: 1804-1812

[16] Lai, K.K. & Wu, S.J. (2005). Using the patent co-citation approach to establish a new patent classification system. Information Processing and Management, 41: 313-330

[17] Luhn, H.P. (1957). A statistical approach to mechanized encoding and searching of literary information. IBM Journal of Research and Development, 1 (4): 309-317

[18] MacQueen, J.B. (1967). Some methods for classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1: 281-297

[19] Mogee, M.E. (2000). Foreign patenting behavior of small and large firms. International Journal of Technology Management, 19: 149-164

[20] Paci, R., Sassu, A. & Usai, S. (1997). International patenting and national technological specialization. Technovation, 17 (1): 25-38

[21] Salton, G. & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. Information Processing and Management, 24 (5): 513-523

[22] Sedding, J. & Kazakov, D. (2004). WordNet-based text document clustering. In: Proceedings of the Third Workshop on Robust Methods in Analysis of Natural Language Data, 104-113, Geneva

[23] Trappey, A.J.C., Huang, C.-J. & Wu, C.-Y. (2008). Building a formal ontology engineering methodology for knowledge definition and representation in design knowledge management. In: Proceedings of Management International Conference (MIC 2008), Barcelona, Spain, November 26-29, 2008

[24] Trappey, A.J.C., Trappey, C.V., Hsu, F.-C. & Hsiao, D.W. (2009). A fuzzy ontological knowledge document clustering methodology. IEEE Transactions on Systems, Man, Cybernetics: Part B, 39 (3): 806-814

[25] Trappey, A.J.C., Trappey, C.V. & Wu, C.Y. (2009). Automatic patent document summarization for collaborative knowledge systems and services. Journal of Systems Science and Systems Engineering, 18 (1): 71-94

[26] Trappey, C.V., Taghaboni-Dutta, F., Wu, H.Y. & Trappey, A.J.C. (2009). China RFID patent analysis. In: Proceedings of the ASME International Manufacturing Science and Engineering Conference, West Lafayette, Indiana, U.S.A., October 4-7, 2009

[27] United States Patent and Trademark Office (USPTO). (2009). Available via DIALOG. http://www.uspto.gov/. Cited December 1, 2009

[28] World Intellectual Property Organization (WIPO). (2009a). What is a patent? Available via DIALOG. http://www.wipo.int/patentscope/en/patents_faq.html#patent. Cited December 1, 2009

[29] WIPO. (2009b). International classifications. Available via DIALOG. http://www.wipo.int/classifications/fulltext/new_ipc/ipcen.html. Cited December 1, 2009

[30] WIPO. (2009c). IP and business: managing patent costs. Available via DIALOG. http://www.wipo.int/wipo_magazine/en/2006/05/article_0010.html. Cited December 1, 2009

[31] WIPO. (2009d). What does a patent do? Available via DIALOG.

http://www.wipo.int/patentscope/en/patents_faq.html#patent_role. Cited December 1, 2009

[32] WIPO. (2009e). Available via DIALOG. http://www.wipo.int/portal/index.html.en. Cited December 1, 2009

Charles Trappey is a professor of marketing in the Department of Management Science at the National Chiao Tung University.

Amy J.C. Trappey is chair professor in the Department of Industrial Engineering and Management and Dean, College of Management at the National Taipei University of Technology. She is also a faculty member of the Department of Industrial Engineering and Engineering Management, the National Tsing Hua University. Dr. Trappey is an ASME Fellow.

Chun-Yi Wu is a doctoral student in the Department of Industrial Engineering and Engineering Management at National Tsing Hua University and a system analyst and engineer at Avectec, Inc. His research interests include the development of computerized intelligent systems and the knowledge management of patents and intellectual properties.