Article

Constructing Patent Maps Using Text Mining to Sustainably Detect Potential Technological Opportunities

Hei Chia Wang *, Yung Chang Chi and Ping Lun Hsin

Department of Industrial and Information Management and Institute of Information Management, National Cheng Kung University, Tainan 701, Taiwan; charles.y.c.chi@gmail.com (Y.C.C.); cymbal0409@gmail.com (P.L.H.) * Correspondence: hcwang@mail.ncku.edu.tw; Tel.: +886-6-275-7575 (ext. 53724); Fax: +886-6-2362162

Received: 22 July 2018; Accepted: 11 October 2018; Published: 17 October 2018

Abstract: With the advent of the knowledge economy, firms often compete for intellectual property rights. Being the first to acquire high-potential patents can assist firms in achieving future competitive advantages. To identify patents capable of being developed, firms often search for a focus by using existing patent documents. Because of the rapid development of technology, the number of patent documents is immense. A prominent topic among current firms is how to use this large number of patent documents to discover new business opportunities while avoiding conflicts with existing patents. In the search for technological opportunities, a crucial task is to present results in the form of an easily understood visualization. Currently, natural language processing can help in achieving this goal. In natural language processing, word sense disambiguation (WSD) is the problem of determining which “sense” (meaning) of a word is activated in a given context. Given a word and its possible senses, as defined by a dictionary, we classify the occurrence of a word in context into one or more of its sense classes. The features of the context (such as neighboring words) provide evidence for these classifications. The current method for patent document analysis warrants improvement in areas, such as the analysis of many dimensions and the development of recommendation methods. This study proposes a visualization method that supports semantics, reduces the number of dimensions formed by terms, and can easily be understood by users. Since polysemous words occur frequently in patent documents, we also propose a WSD method to decrease the calculated degrees of distortion between terms. An analysis of outlier distributions is used to construct a patent map capable of distinguishing similar patents. During the development of new strategies, the constructed patent map can assist firms in understanding patent distributions in commercial areas, thereby preventing patent infringement caused by the development of similar technologies. Subsequently, technological opportunities can be recommended according to the patent map, aiding firms in assessing relevant patents in commercial areas early and sustainably achieving future competitive advantages.

Keywords: text mining; patent mapping; technological opportunities; competitive advantage

1. Introduction

In many cases, firms must expend substantial amounts of time and money if they are suspected of involvement in patent infringement [1]. To avoid this, firms often employ patent engineering teams to routinely retrieve and organize current patents relevant to the firm’s technologies [2]. This allows firms to understand which technologies are under patent protection, avoid conflicts between technologies from their research and development (R&D) and those in existing patents, and reduce the likelihood of patent infringement [2]. Analyzing the patent distribution of an industry is a vital task in preventing
infringement concerns. In addition, patent documents contain abundant credible technical information and key research results, which makes them a highly valuable and useful source of knowledge [3,4]. According to the World Intellectual Property Organization (WIPO), by searching for and reviewing patent literature, 90–95% of the world’s inventions can be understood, technology R&D time can be decreased by 60%, and research expenditures can be decreased by 40% [5]. Therefore, when firms intend to develop a new technology or product, they often first collect patents relevant to that technology. Through this collection process they accumulate new technical knowledge, which inspires innovation and assists firm decision-makers in developing a strategic direction and decreasing costs during the R&D process [6].

According to the 2016 WIPO report, from the initial implementation of the patent system to 2015, more than 75 million patent applications have been filed [7]. In practice, the precision and recall of patent retrieval systems have become increasingly incapable of meeting user expectations, resulting in information overload [3,8,9]. Although utilizing the International Patent Classification codes (IPC-codes) developed by the WIPO can limit the scope when searching for information, in practice these codes can only be used as a reference rather than an ultimate standard. In addition, patent documents can be found in numerous technical domains, but few people have professional knowledge in multiple domains [10]. To enable rapid user comprehension, it is convenient to represent the distribution of patents as a patent map.

Patent mapping is a common method that involves presenting patent information obtained from a retrieval system using various qualitative and quantitative analysis methods. Patent maps have several functions; for instance, the use of patent maps enables more efficient detection of patent infringement [1]. Furthermore, when competitors possess prospective patents and latecomers have no choice but to mimic necessary technologies within the patents, the latecomers can use patent maps to understand competitors’ patent distributions and attempt to design around such patents to avoid infringement. Because a standalone patent does not possess as much legal force as a group of related patents in a patent portfolio, patent maps can be used to develop a firm’s own patent distribution. In addition to increasing the number of competitors’ infringement cases and increasing settlement amounts, this can also render a competitors’ design-around strategies more difficult to execute [11]. Patent mapping can also be used to assess firms that wish to collaborate or merge [12] and to compare different technologies (structured data and unstructured text) to analyze aspects such as technical trends and interactions with competitors [12,13]. Patent mapping visualizations are one of the best ways to compare different technologies [14]. Therefore, at the national, industrial, and technical domain levels, as well as the product and firm levels, patent mapping can provide decision-makers and R&D professionals with comprehensive summaries of patent-related information. Using graphical representations of industry trends and technology distributions further provides them with comprehensive support during the development of business strategies and plans. This study proposes a visualization method that supports semantics, reduces the number of dimensions formed by terms, and can easily be understood by users.

In current patent analysis, numerous patent documents use different words to describe the same events, resulting in semantic inconsistency [15] and polysemy due to the multiple meanings that may exist for one word. To resolve this, document analysis often necessitates the merging of synonyms into the same semantic dimension; a thesaurus can facilitate this process. These word sense disambiguation (WSD) methods decrease polysemy and allow the term similarities of patent documents to be calculated more precisely. This study uses WordNet [15], which is commonly used for synonym analysis, to calculate similarities between terms and merge similar terms. Additionally, multiple meanings may exist for one word by the average value formula (see Equation (2)) of semantic similarity. On the other hand, different words may be used to describe the same events and multiple meanings may exist for one word; WordNet can also calculate term similarity faster and more precisely in this case. Finally, to reduce the dimensions of documents for reading convenience, multidimensional scaling is used to simplify multidimensional research subjects into a low-dimensional space. The outlierness
of each document is also calculated: If the local density of a document is smaller than neighboring local densities, that document possesses higher outlierness, which indicates a lower number of similar patents and a gap in related technologies, which may indicate technological opportunity [16].

Term analysis can be implemented in different programming languages that are suitable for different fields. Multiple programming languages can be used for implementing different packages. In recent years, the R language has become increasingly popular, and developers are continuously developing new R language kits; consequently, the R language has become increasingly powerful for statistical processing, graphics, data mining, and big data. Therefore, R is the major programming language used in this study. In this paper, the R text mining package (tm) is used for term analysis and the R statistical analysis package is used for multidimensional scaling.

The three primary objectives of this study are (1) to enhance the effectiveness of patent retrieval using a semantic net, (2) to construct a patent map that distinguishes patent similarities to help firms avoid patent infringement caused by developing similar technologies, and (3) to use the outlier method to sustainably discover technological opportunities. The research contribution is a patent map that can assist firms in understanding patent distributions in commercial areas, thereby preventing patent infringement caused by the development of similar technologies. In addition, technological opportunities can be sustainably recommended using the patent map.

In Section 2, we discuss the limitations of existing methods through a review of the literature on word-sense disambiguation and patent mapping. We also present our specific research objectives and explain WordNet, a key technology for achieving those objectives. In Section 3, we present the basic concept and the detailed process of the proposed approach. Section 4 shows the term analysis of the proposed methodology, using real patents to derive and verify the results. In Section 5, limitations and areas for future research are discussed.

2. Related Work

2.1. Text Mining

Text mining extracts meaningful information from unstructured data and is used in many patent research fields because it can work with large amounts of text. Text mining can be divided into keyword-based analysis and word-based analysis [17]. As following Table 1, Song et al. [6] have applied text mining and T-term analysis to discover new technology opportunities. Walter et al. [18] have discussed text-mining analysis in relation to the beauty of the brimstone butterfly, the novelty of patents identified by near environment analysis. Alves et al. [19] have developed text-mining tools for information retrieval from patents. Kayser et al. [20] have extended the knowledge base of foresight and thereby the contribution of text mining. Shen et al. [21] have used text mining and cosine similarity to discover potential product opportunities. Kim et al. [22] have used text-mining patent keyword extraction for sustainable technology management. Roh et al. [23] have developed a methodology for structuring and layering technological information to apply text mining to patent documents.

<table>
<thead>
<tr>
<th>Authors</th>
<th>Year</th>
<th>Title</th>
<th>Methodology</th>
<th>Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Song et al.</td>
<td>2017</td>
<td>Discovering new technology opportunities based on patents: Text mining and F-term analysis.</td>
<td>Text mining and F-term analysis</td>
<td>The F-term can provide effective guidelines for generating new technological ideas.</td>
</tr>
<tr>
<td>Walter et al.</td>
<td>2017</td>
<td>The beauty of the brimstone butterfly: novelty of patents identified by near environment analysis based on text mining.</td>
<td>Text mining and near environment analysis</td>
<td>The approach taken can single out the content-wise novelty of patents in near environments.</td>
</tr>
</tbody>
</table>
Table 1. Cont.

<table>
<thead>
<tr>
<th>Authors</th>
<th>Year</th>
<th>Title</th>
<th>Methodology</th>
<th>Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alves et al.</td>
<td>2017</td>
<td>Development of Text Mining Tools for Information Retrieval from Patents.</td>
<td>Text mining</td>
<td>A patent pipeline was developed and integrated into @Note2, an open-source computational framework for BioTM (Biomedical text mining).</td>
</tr>
<tr>
<td>Kayster et al.</td>
<td>2017</td>
<td>Extending the knowledge base of foresight: The contribution of text mining.</td>
<td>Text mining</td>
<td>Text mining facilitates the detection and examination of emerging topics and technologies by extending the knowledge base of foresight. Hence, new foresight applications can be designed.</td>
</tr>
<tr>
<td>Shen et al.</td>
<td>2017</td>
<td>A cross-database comparison to discover potential product opportunities using text mining and cosine similarity.</td>
<td>Text mining and cosine similarity</td>
<td>Remote health monitoring technology is used as a case study. The results show four product opportunities, namely wireless sensor devices, telecommunication systems and technology, wearable devices and systems, and medical services and systems.</td>
</tr>
<tr>
<td>Kim et al.</td>
<td>2018</td>
<td>Patent Keyword Extraction for Sustainable Technology Management.</td>
<td>Text mining and keyword extraction</td>
<td>The proposed method improves the precision by approximately 17.4% over the existing method.</td>
</tr>
<tr>
<td>Roh et al.</td>
<td>2017</td>
<td>Developing a Methodology for Structuring and Layering Technological Information in Patent Documents through Natural Language Processing.</td>
<td>Text mining and natural language processing</td>
<td>The structured and layered keyword set does not omit useful keywords and the analyzer can easily understand the meaning of each keyword.</td>
</tr>
</tbody>
</table>

2.2. Word-Sense Disambiguation

In computational linguistics, word-sense disambiguation (WSD) [24] is an open problem in natural language processing and ontology. WSD means identifying which sense of a word (i.e., meaning) is being used in a given sentence when the word has multiple meanings. The solution to this problem impacts other computer-related issues, such as discourse, improving the relevance of search engines, anaphora resolution, coherence, and inference. Fixed grammatical structures in a language can be used to determine which semantic meaning should be attributed to a term; for example, in English, prepositions must be followed by nouns, pronouns, or noun phrases. Neighboring words can also be referenced to determine semantic meaning. Additionally, a term usually only represents a single meaning in a document; this concept can be used to limit the meanings of terms. Through these WSD methods, term similarities can be more precisely calculated. So far, expert dictionaries have only a few fields for polysemous words that frequently occur in WordNet. This study proposes a term similarity calculation and term grouping method that relies on WSD. These methods make it easier to decrease the calculated degree of distortion between two terms so that they can be compared more precisely.

2.3. WordNet

WordNet constructs subsemantic networks for each of the following four parts of speech: nouns, verbs, adjectives, and adverbs [25]. Semantically similar terms in the semantic web (such as “kid” and “child”) are loosely grouped into synonym sets (synsets), whereas words with multiple meanings appear in several different synsets. The relationships among terms do not exist in isolation but also generate subsequent links among the four semantic networks for the synsets. In the noun semantic network, the links from synonymous word sets indicate hyponymy (super-subordinate relations) as well as meronymy (part-whole relations); all of these are used to form a hierarchical architecture.
In this study, based on the WordNet similarities between terms, the similarities between patents are quantified by calculating the degree of closeness based on the depth of the concepts in the hierarchy.

2.4. Patent Mapping

Qualitative analysis is used to analyze the contents of individual patent documents, such as technical content files, and the results usually contain the relevant individual patent document number. Although patent analysis by expert analysts can provide valuable information, it involves manual, detailed reading of individual patent documents, which is time-consuming. Typically, the results of such analysis are presented through patent mapping, which generates illustrations in the form of tree structures, tables, or matrices [26]. A matrix is the basic form of patent presentation. A patent map in the form of a tree structure can indicate the development of technology, the spread of technology, and the state of joint application [26].

Quantitative analysis involves the formation of a group of patents as a parent group of a given category of patents. It uses bibliographic information contained in the patent literature, including: distinctions among documents; document numbers; the patent classification; the applicant’s nationality, name, and address; the name of the inventor; and, the number of inventions. Additional information provided by the patent office and related to retrieval, prosecution, and cited documents is also used in quantitative analysis [18]. Detailed analysis additionally involves separate supplemental indices. As in qualitative analysis, quantitative analysis results are presented in various forms, including illustrations, graphs, tree structures, and matrices. Among these, graphics are the basic form of presentation and visualization. Patent analysis requires both qualitative and quantitative analysis [26]. In this study, patent mapping is the core of patent visualization and refers to a series of methods used to build patent maps based on differences and similarities among patents. Patent mapping visualizations are one of the best ways to compare different technologies [14].

2.5. Multidimensional Scaling

Multidimensional scaling (MDS) is a nonparametric, distance-based multivariate analysis technique that produces statistical maps from the main characteristics of the data. It has the advantage of making the results accessible to non-specialists in an intuitive manner [27]. In this study, after similarity calculations, multidimensional matrices that are difficult to read are generated from the similarities of patent documents. To facilitate understanding of the relationships among patents, multidimensional matrices should be reduced dimensionally. Patent maps are mostly two-dimensional or, at most, three-dimensional. Therefore, multidimensional matrices cannot be represented by patent maps; MDS must first be used to convert multidimensional matrices into two- or three-dimensional matrices.

2.6. Local Outlier Factor

Local outlier factor (LOF) is an anomaly detection algorithm that compares the local density of a point’s neighborhood with those of its neighbors. It indicates the degree of an object’s outlierness from a cluster [28]. LOF enables outlier data that could be more valuable than normal data to be extracted. Researchers have proposed numerous outlier mining algorithms that can effectively detect outlierness in a data set [29]. In this study, the final LOF output is intended to discover technological vacancies or technical opportunities.

3. Proposed Methods

The research structure of this study is divided into five modules, as shown in Figure 1. These are the “Document collection and preprocessing module”, the “Term similarity calculation and term grouping module”, the “Document similarity calculation module”, the “Multidimensional scaling-based dimensionality reduction module” and the “Document outlierness calculation module”. The function of each module is described below.
3.1. Document Collection and Preprocessing Module

This study used patent documents approved by the United States Patent and Trademark Office. Regular documents are unstructured data consisting of several terms. Patent documents are semi-structured data comprising the following fields: patent number (Pat. No.), patent title, patent abstract, patent assignee, references cited, IPC-code, patent claims, and patent description.

The module adopted in the present study references methods from relevant studies [4,12,30–32] to avoid incorporating an overly large number of words; only words in the patent title, abstract, and claims fields were retrieved for processing. All numbers, punctuation marks, and special symbols were removed, and after processing by the Stanford parser, numerous terms containing relevant parts of speech were acquired. This was followed by a removal of stop words from all terms.

3.2. Term Similarity Calculation and Term Grouping Module

The number of terms generated by the previous module is too large for easy interpretation by users, and the problem of synonyms persists. Studies have indicated that semantic analysis can be used to solve this problem [33]. Regarding the semantic associations among terms, this study followed the example of Miller [25] to calculate the similarities between pairs of terms. Several methods exist for using WordNet to perform this calculation, such as PATH (A simple node-counting scheme) [34], WUP (Wu & Palmer measure) [35], and LESK (Lesk algorithm) [36]. Among these, WUP is based on the measurement of the depth of each concept in use; that is, it measures the length of the path from the root node or the nearest common ancestor to two concepts, or the depth of the lowest common subsume (LCS) of the two concepts. Because it can reflect the specific level of the concept, WUP was selected. The WUP method calculates the similarity between two terms according to the following formula:

\[
\text{Similarity}(\text{Term}_1, \text{Term}_2) = \frac{2 \times \text{Depth}(\text{LCS}(\text{Term}_1, \text{Term}_2))}{\text{Depth}(\text{Term}_1) + \text{Depth}(\text{Term}_2)}
\] (1)
The Depth() function returns the depth of the collection of synonyms of the obtained terms in the WordNet hierarchical framework, and the LCS() function returns the LCS of the respective synonym collections of the two obtained terms. For instance, to calculate the similarity between “CPU” and “RAM,” because the depth of “CPU” is 8 and the depth of “RAM” is 10, the LCS of “CPU” and “RAM” is “hardware,” whose level value is 7, as shown in Figure 2. Therefore:

\[
\text{Similarity(“CPU”, “RAM”)} = \frac{2 \times 7}{8 + 10} = 0.778
\]

Figure 2. Example of the WordNet hierarchical framework.

Additionally, multiple meanings may exist for one word. For instance, in WordNet, “processor” has three different noun meanings, which are “a business engaged in processing agricultural products,” “someone who processes things,” and “central processing unit”; similarly, five different noun meanings exist for “memory” (Figure 3). For this type of situation, this study took the mean value of maxa (3, because “processor” has 3 noun meanings) multiplied by that of maxb (5, because “memory” has 5 noun meanings) to represent the 15 total semantic similarities, as presented in Table 2. The average value of the semantic similarity between “processor” and “memory” is as follows:

\[
\text{Average Similarity(“processor”, “memory”) = } \frac{\sum_{a=1}^{\max_a} \sum_{b=1}^{\max_b} \text{Similarity(“processor #n #a”, “memory #n #b”)}}{\max_a \times \max_b}
\]

Table 2. Semantic Similarity.

<table>
<thead>
<tr>
<th>Memory#n#1</th>
<th>Memory#n#2</th>
<th>Memory#n#3</th>
<th>Memory#n#4</th>
<th>Memory#n#5</th>
</tr>
</thead>
<tbody>
<tr>
<td>processor#n#1</td>
<td>0.133</td>
<td>0.133</td>
<td>0.133</td>
<td>0.125</td>
</tr>
<tr>
<td>processor#n#2</td>
<td>0.143</td>
<td>0.143</td>
<td>0.143</td>
<td>0.133</td>
</tr>
<tr>
<td>processor#n#3</td>
<td>0.133</td>
<td>0.133</td>
<td>0.133</td>
<td>0.875</td>
</tr>
</tbody>
</table>
In the original vector space model, distinct terms such as “CPU” and “processor” formed independent dimensions. However, “CPU” and “processor” should possess a certain degree of semantic similarity. For instance, if a user intends to search for patents related to “CPU”, patents related to “processor” must not be omitted because overlooking technology patents that use synonyms could result in infringement. Therefore, in this study, terms with a certain degree of semantic similarity (i.e., in the same cluster) were viewed as identical terms. In Figure 4, for instance, because “CPU” and “processor” were grouped into the same cluster, they were considered identical terms and shared the same dimension in the vector space model. This method was used to enhance the precision of calculating patent document similarities.

The mean value of the semantic similarity of “processor” and “memory,” which is 0.179, is obtained by dividing the sum of the 15 values in Table 2 by 15.

After obtaining the term similarity matrix, Equation (3) can be used to obtain the distance matrix between terms. This module can subsequently utilize the distance matrix to group terms, as shown in Figure 4. We have:

\[
\text{Distance}(\text{Term}_1, \text{Term}_2) = 1 - \text{Similarity}(\text{Term}_1, \text{Term}_2)
\]  

(3)
3.3. Document Similarity Calculation Module

In this module, terms in the same cluster (generated in the previous module) were considered related synonyms, and the term frequency-inverse document frequency (TF-IDF) method of information retrieval was used to calculate their weights. The concept of term frequency (TF) states that the more frequently a term appears in a document, the higher its weight should be. In contrast, in inverse document frequency (IDF), terms occurring in a greater number of documents are relatively less relevant and should be weighted less. In the TF-IDF formula, \( \text{freq}_{ij} \) represents the number of occurrences of the word \( j \) in the file \( i \), \( M \) represents the number of files, \( m_j \) represents the number of files containing the word \( j \), and \( W_{ij} \) is the weight of word \( j \) in file \( i \). The TF-IDF formula is as follows:

\[
\text{tf}_{ij} = \frac{\text{freq}_{ij}}{\text{MAX}(\text{freq}_{ij})}
\]

\[
\text{idf}_j = \log \frac{M}{m_j}
\]

\[
W_{ij} = \text{tf}_{ij} \times \text{idf}_j
\]

(4)

The simple structure and ease of use of this method have enabled various applications of patent text analysis [23,37].

Additionally, Ref. [38] purports that employing a cosine measure to calculate the similarity between two documents in a vector space model generally results in better performance. Therefore, this study utilized a cosine measure [21] with the following formula:

\[
\text{similarity}(\text{Doc}_1, \text{Doc}_2) = \frac{\sum w_{1i} w_{2i}}{|\text{Doc}_1| \times |\text{Doc}_2|}
\]

(5)

Here, \( \text{Doc}_1 \) is expressed as \( \{w_{11}, w_{12}, w_{13} \ldots w_{1n}\} \) and \( \text{Doc}_2 \) as \( \{w_{21}, w_{22}, w_{23} \ldots w_{2n}\} \); \( n \) represents the number of terms; and, \( w_{ij} \) represents the weight of term \( j \) in document \( i \) generated using TF-IDF. Thus, the similarity matrix of a document can be obtained.

3.4. Multidimensional Scaling-Based Dimensionality Reduction Module

The multidimensional similarity matrix resulting from the similarity calculation for a regular document is difficult to read. To render the relationships among patents more understandable, multidimensional matrices should be converted into low-dimensional patent maps. To reduce the number of dimensions for this purpose, we used MDS [39]. However, reducing the number of dimensions in the source data while retaining the original relationships among the data doubtlessly generates information loss problems. Therefore, the quality of the results obtained from using multidimensional scaling was measured using stress values calculated as follows:

\[
\text{Stress} = \sqrt{\frac{\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} (d_{ij} - d_{ij}')^2}{\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} d_{ij}^2}}
\]

(6)

\[
d_{ij} = \sqrt{\frac{1}{n} \sum_{k=1}^{p} (x_{ik} - x_{jk})^2}
\]

(7)

Here, \( n \) represents the number of data, \( p \) the number of dimensions, \( x_{ik} \) \(-\) \( x_{jk} \) the gap between the data points \( x_i \) and \( x_j \) in dimension \( k \), \( d_{ij} \) the distance between the two data points prior to dimensionality reduction, and \( d_{ij}' \) the distance following dimensionality reduction. Following the use of multidimensional scaling, stress values range between 0 and 1. According to [40], if a stress value is less than 0.2, the result following dimensionality reduction is acceptable. The closer a stress value is to
0, the more precisely a result has retained the original relationships among data. This study employed classic MDS from Quick-R in the R language to reduce dimensionality.

In Figure 5, each point represents a patent document plotted in R’s plot function. This figure retains the similarity relationships between documents; that is, a shorter distance between two points indicates a higher similarity between those two documents. Documents 122 and 15 are closer to each other than are documents 57 and 15. This patent map can assist firms in understanding the distribution of technology while they establish development strategies to avoid developing technologies that would result in patent infringement.

![Figure 5. Patent map showing MDS results. Each point represents a patent document number.](image)

### 3.5. Document Outlierness Calculation Module

In the patent map, to detect technological opportunities within a high-density cluster of patents, the cluster must first be searched for patents with lower similarity, which indicates fewer R&D individuals involved in the technologies covered by similarly few patents. Although the data have different areas of density, the LOF method can still operate favorably [41]. Therefore, this module employed the LOF method proposed by Breunig et al. [42] to calculate the outlierness of each document. The concept of the LOF is that if the local density of a document is less than the local densities of its neighbors, that document possesses higher outlierness. The LOF can be calculated by the following equations.

**k-distance**: The distance between the kth nearest point and the point \( doc' \) that is closest to the data point \( doc \) is called the k-distance of the point \( doc \) and denoted k-distance \( (doc) \).

Reachability distance: The reachability distance is related to the k-distance. When the parameter \( k \) is given, the reachable distance from the data point \( doc \) to the data point \( doc' \) is called Reachability Distance \( (doc ← doc') \). It equals the data point \( doc' \) of the k-adjacent distance of the data point \( doc \) and the maximum distance between the data points \( doc \) and \( doc' \). We have

\[
\text{ReachabilityDistance}_k(doc ← doc') = \max\{ d(doc, doc'), \text{Distance}_k(doc) \} \quad (8)
\]

Local reachability density: The definition of local reachability density is based on the reachable distance. For the data point \( doc' \), the data point whose distance from \( doc \) is less than or equal to
$k$-distance$(doc)$ is called its $k$-nearest-neighbor and denoted $N_k(doc)$. The local reachability density of the data point $doc$ is the reciprocal of its average reachability distance from adjacent data points:

$$\text{Local Reachability Density, } lrd_k(doc) = \frac{\|N_k(doc)\|}{\sum_{doc' \in N_k(doc)} \text{ReachabilityDistance}_k(doc \rightarrow doc')}$$ (9)

Local outlier factor: According to the definition of local reachability density, if a data point is far from other points, then its local reachability density is obviously small. However, the LOF algorithm measures the anomaly of a data point: not its absolute local density, but the relative density of its neighboring data points. The advantage of this is that it allows for an uneven distribution of data and different densities. The local anomaly factor is defined by the local relative density. The local relative density of the data point $doc$ (local anomaly factor) is the ratio of the average local reachability density of the neighbors of $doc$ to the local reachability density of $doc$. We have:

$$\text{Local Outlier Factor, } \text{LOF}_k(doc) = \frac{\sum_{doc' \in N_k(doc)} lrd_k(doc')}{lrd_k(doc)} \frac{\|N_k(doc)\|}{\sum_{doc' \in N_k(doc)} \text{ReachabilityDistance}_k(doc \rightarrow doc')}$$ (10)

where $doc$ is the current document for which outlierness is being calculated, $d(doc,doc')$ is the Euclidean distance between $doc$ and $doc'$, $\text{Distance}_k(doc)$ is the Euclidean distance between $doc$ and another neighbor $k$, and $N_k(doc)$ is the collection of all documents whose Euclidean distance from $doc$ is less than $\text{Distance}_k(doc)$. Following the calculation of the outlierness of each patent document, an outlierness ranking can be obtained. A higher outlierness ranking indicates a lower number of similar patents. The outlier patents, in an overall sense, were more novel than non-outlier patents in terms of related technologies and potential business opportunities [24].

4. Term Analysis and Results

4.1. Data Set

The data set for this study was a collection of patent documents from the United States Patent and Trademark Office (USPTO) that had the strings “USB connector” or “Universal Serial Bus connector” in the title fields and had patent issue dates between 2005 and 2014. A total of 152 documents meeting these criteria were retrieved. Twenty-eight documents were design patents without text in the patent abstract or patent application fields, so they were excluded. A final 124 invention patents were used as the data set for this study.

4.2. Assessment Indicators

The assessment indicators for the term analysis were computed using the R package ROCR and examined whether the inclusion of a semantic net could enhance the effectiveness of patent retrieval. The assessment indicators were precision, recall, and F-measure. Precision refers to how many documents retrieved by the system were relevant, and recall is defined as how many of the existing relevant documents were retrieved [22]. The F-measure is the harmonic mean of precision and recall. The formulas are presented below and in Table 3. TP means true positive and corresponds to the number of positive examples correctly predicted by the classification model. FP is false positive and corresponds to the number of negative examples wrongly predicted to be positive by the classification model. We have

$$\text{Precision} = \frac{TP}{TP + FP} = \text{the fraction of the relevant documents that have been retrieved.}$$ (11)

$$\text{Recall} = \frac{TP}{TP + FN} = \text{the fraction of the retrieved documents that are relevant.}$$ (12)
F-measure = \[
\frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}
\] (13)

Table 3. Confusion Matrix.

<table>
<thead>
<tr>
<th>Test Outcome</th>
<th>Gold Standard</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Positive (TP)</td>
</tr>
<tr>
<td>Positive</td>
<td>True Positive</td>
</tr>
<tr>
<td>Negative</td>
<td>False Positive</td>
</tr>
</tbody>
</table>

4.3. Term Analysis 1: Examining the Effect of Semantic Nets on Patent Retrieval

In this term analysis, similarities between the patent \(\text{Doc}_x\), which was marked as most familiar by patent engineers, and 10 other patents \((\text{Doc}_a-\text{Doc}_j)\) were indicated from 0 to 1. These indicators are presented as the first column in Table 4. According to the patent engineers, the documents marked with similarities of 0.6 and higher were documents that required retrieval, which included \(\text{Doc}_a\), \(\text{Doc}_b\), \(\text{Doc}_c\), and \(\text{Doc}_d\). Obtained similarities that were not included in the WordNet calculations are shown as the second column in Table 4. Under the same condition (a similarity threshold of 0.6), the documents retrieved by this system were \(\text{Doc}_a\), \(\text{Doc}_b\), and \(\text{Doc}_f\). Therefore, without inclusion in the semantic net, the precision was 0.667, the recall was 0.5, and the F-measure was 0.571. The similarities obtained with inclusion in the WordNet system search are shown as the third column in Table 4. Given the same similarity threshold (0.6), the documents retrieved by this system were \(\text{Doc}_a\), \(\text{Doc}_b\), \(\text{Doc}_c\), \(\text{Doc}_f\), and \(\text{Doc}_g\). Therefore, with inclusion in WordNet, the precision was 0.6, the recall was 0.75, and the F-measure was 0.667. Finally, because the F-measure value with inclusion in WordNet was higher than the F-measure value without inclusion in the semantic net, inclusion in WordNet demonstrably increases the effectiveness of patent searches.

Table 4. Semantic Net Similarity Comparison.

<table>
<thead>
<tr>
<th>Similarity Marked by Patent Engineers</th>
<th>Similarity Obtained without Inclusion in Semantic Net</th>
<th>Similarity Obtained with Inclusion in Semantic Net</th>
</tr>
</thead>
<tbody>
<tr>
<td>Similarity ((\text{Doc}_a, \text{Doc}_b))</td>
<td>0.9</td>
<td>0.6</td>
</tr>
<tr>
<td>Similarity ((\text{Doc}_a, \text{Doc}_c))</td>
<td>0.8</td>
<td>0.7</td>
</tr>
<tr>
<td>Similarity ((\text{Doc}_a, \text{Doc}_d))</td>
<td>0.7</td>
<td>0.5</td>
</tr>
<tr>
<td>Similarity ((\text{Doc}_b, \text{Doc}_c))</td>
<td>0.6</td>
<td>0.4</td>
</tr>
<tr>
<td>Similarity ((\text{Doc}_b, \text{Doc}_d))</td>
<td>0.5</td>
<td>0.4</td>
</tr>
<tr>
<td>Similarity ((\text{Doc}_c, \text{Doc}_d))</td>
<td>0.4</td>
<td>0.6</td>
</tr>
<tr>
<td>Similarity ((\text{Doc}_c, \text{Doc}_e))</td>
<td>0.3</td>
<td>0.5</td>
</tr>
<tr>
<td>Similarity ((\text{Doc}_c, \text{Doc}_f))</td>
<td>0.2</td>
<td>0.4</td>
</tr>
<tr>
<td>Similarity ((\text{Doc}_c, \text{Doc}_g))</td>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>Similarity ((\text{Doc}_c, \text{Doc}_h))</td>
<td>0.0</td>
<td>0.2</td>
</tr>
</tbody>
</table>

4.4. Term Analysis 2: Examining Patent Documents with Higher Outlierness

After the document similarity matrix undergoes dimensionality reduction using MDS, we can construct a patent map that can distinguish patent document similarities. Figure 6 shows the patent map constructed in this study. The numerical portion of the text string in the figure indicates the U.S. patent number of the patent document; patents with higher similarity are also closer in the figure. For future assessments of new patents, the patent must only be added to the data set, followed by applying the same steps. Thus, this patent map can be referenced to perform a preliminary judgement on patents with high similarity to this patent.
Figure 6. Semantic net similarity comparison (Table 4: Similarity Comparison).

In Figure 7, each point represents a patent document, and the surrounding area shows USB-related technologies. By comparing the density of each point and its neighboring points, we can determine whether the point is abnormal. If the density of the point is lower, it is more likely to be abnormal. The density is calculated based on the distance between the points. If the distance is higher, the density is lower; if the distance is lower, the density is higher. If the LOF score of the data point is approximately 1, the local densities of the data point with its neighbors are similar; if the data point’s LOF score is less than 1, the data point is in a relatively dense area, unlike an abnormal point; and, if the data point’s LOF score is much larger than 1, the data point is more alienated from other points, which indicates that it is an abnormal point. Outlier patents, in an overall sense, are more novel than non-outlier patents in terms of related technologies and potential business opportunities [20]. Figure 7 implements its own package to calculate the results; its results are different from those plotted by R’s plotting function package.

Table 5. Outlierness.

<table>
<thead>
<tr>
<th>Item</th>
<th>U.S. Patent Number</th>
<th>Outlierness</th>
<th>Item</th>
<th>U.S. Patent Number</th>
<th>Outlierness</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>8672692</td>
<td>1.915</td>
<td>k</td>
<td>8062067</td>
<td>1.31</td>
</tr>
<tr>
<td>b</td>
<td>7234963</td>
<td>1.552</td>
<td>l</td>
<td>7645154</td>
<td>1.30</td>
</tr>
<tr>
<td>c</td>
<td>6902432</td>
<td>1.432</td>
<td>m</td>
<td>8043101</td>
<td>1.29</td>
</tr>
<tr>
<td>d</td>
<td>7699633</td>
<td>1.411</td>
<td>n</td>
<td>8308507</td>
<td>1.28</td>
</tr>
<tr>
<td>e</td>
<td>8900012</td>
<td>1.386</td>
<td>o</td>
<td>7921233</td>
<td>1.28</td>
</tr>
<tr>
<td>f</td>
<td>8814583</td>
<td>1.384</td>
<td>p</td>
<td>7591657</td>
<td>1.28</td>
</tr>
<tr>
<td>g</td>
<td>7811111</td>
<td>1.379</td>
<td>r</td>
<td>8900016</td>
<td>1.26</td>
</tr>
<tr>
<td>h</td>
<td>8172585</td>
<td>1.369</td>
<td>s</td>
<td>8047877</td>
<td>1.26</td>
</tr>
<tr>
<td>i</td>
<td>7172460</td>
<td>1.352</td>
<td>t</td>
<td>6808400</td>
<td>1.26</td>
</tr>
<tr>
<td>j</td>
<td>8900017</td>
<td>1.320</td>
<td>u</td>
<td>8235746</td>
<td>1.25</td>
</tr>
</tbody>
</table>
According to the definition of the local anomaly factor, if the LOF score of the data point $doc$ is approximately 1, the local density of the data point $doc$ is similar to that of its neighbors; if the LOF score of $doc$ is less than 1, $doc$ is in a relatively dense area, unlike an abnormal point. If the LOF score of the data point $doc$ is much larger than 1, $doc$ is farther away from other points and is likely to be an abnormal point. For each data point, we calculate its distance from all other points and sort from the nearest data point to the farthest data point. We then find its $k$-nearest-neighbor and calculate the LOF score.

Table 5 displays the top 20 patent documents in terms of outlierness ranking obtained after using the LOF method to calculate the outlierness of each patent document. Patent documents with high outlierness rankings indicate fewer similar patents, suggesting a gap in related technologies and concomitant business opportunities. Not all outlier patents deliver new approaches to technological development; some provide fresh or unusual signals for further technological development. In the competitive technological environment, an early grasp of potential technological opportunities is important for developing technologies that can increase the competitiveness of a business [13].

In our numerical analysis of the outlierness rankings, only the top two outlier values were higher than 1.5. Furthermore, a detailed reading of these 20 patents indicated that all 18 patents with outlier values lower than 1.5 were design patents related to the body structures of USB connectors. The two patents with outlier values higher than 1.5 were more closely oriented toward applicability and convenience. The USB connectors currently on the market are extremely similar, and most consumers would not notice design variations related to the body structures; however, increased applicability and convenience can become selling points. This study was based on patent documents from 2005 to 2014. A subsequent query using Google Advanced Patent Search retrieved 31 USB patents that were approved between January and December 2015 worldwide. Among these, only 4 were related to body structure; the remaining 27 were related to increasing applicability and convenience.

Patent documents, which contain abundant highly credible technical information and crucial research results, are highly useful, valuable sources of knowledge. Therefore, this study employed the LOF method to calculate the outlierness of each patent document (Table 5). Only the top two outlier values were higher than 1.5, so we used the level of outlierness to discover potential technological or business opportunities related to the following two patents.
U.S. Patent No. 8672692 is primarily a USB connector structure. There is a component that can be inserted into a USB port, which has a terminal on one side, and a connector that links the terminal unit and the internal USB circuit. A reinforcing element is provided on the other side of the plug for protection, and an insulator is placed on its surface.

U.S. Patent No. 7234963 is a design that corrects the orientation of the wire on a USB plug connector. It has a top cover, a connector, a wire, a wire slot, a cable rotating seat, a bottom cover, etc. It can prevent the fall of the USB transmission line due to rotation during use.

5. Conclusions

The semantic net similarity comparison table indicates that in patent documents, semantic inconsistencies are caused by different modifiers being used to describe the same operation. When the system is searching, it cannot distinguish synonyms, which causes patents that should have been retrieved to be overlooked. At present, no worldwide uniform or similar adjective guidelines exist for patent documents. In addition, patent specifications must describe the contents of the patent in detail; once a patent has been published, all the technical details of the patent are made public. Therefore, to protect patent-holder interests, some patent applications are reserved in their descriptions, play word games, or even contain traps. These all create obstacles during system searches and substantially decrease search precision. However, given that the purpose of patents is to protect the rights and interests of inventors, these obstacles may be another method of protecting patents. To alleviate them, in addition to unifying term modifiers in relevant standards and regulations, technical means can be applied to enhance search precision. This study used WordNet synonyms to calculate the similarities between pairs of terms and merge synonyms. Further investigation is warranted to further increase the precision of system searches.

Since polysemous words frequently occur in WordNet, this study proposed a different method than word-sense disambiguation (WSD) to decrease the calculated degree of distortion between two terms. For instance, “watch” can be interpreted as the noun meaning “wristwatch” or the verb meaning “to pay attention to.” The word “star” can be interpreted as “a star in space” or “a celebrity.” WSD determines the correct semantic meaning of a term in a document from numerous possibilities. A fixed grammatical structure in a language can be used to determine which semantic meaning should be attributed to a term; for example, in English, prepositions can be followed only with nouns, pronouns, or noun phrases. Neighboring words can be referenced to determine semantic meaning as well. Consider the term “pine cone.” “Pine” has two meanings in the dictionary: “a type of evergreen tree with needle-shaped leaves” and “to waste away through sorrow or illness.” “Cone” has three meanings: “a solid body which narrows at a point,” “something of this shape whether solid or hollow,” and “the fruit of a certain evergreen tree.” Therefore, the intersecting semantic meanings, namely, “evergreen” and “tree,” should be selected. Additionally, a term usually represents only one meaning in a document, which can also be used to limit the meanings of terms. Finally, if expert dictionaries can be established in the future for different areas of expertise, these dictionaries can be used to limit the meanings of terms or determine technical terms more precisely. Through these WSD methods, term similarity can be more precisely calculated.

The rapid development of technology and the accumulation of patents have led to an immense number of patent documents, and sorting through them directly results in information overload. Therefore, this study proposed a more efficient method to distinguish patent document similarities [38]. It involves extracting the titles, abstracts, and application fields of USPTO patent documents, preprocessing the text in these fields, and using the WUP method to calculate similarities between terms to obtain term similarity matrices. These matrices are used to group terms, after which the TF–IDF method is used to calculate term weights and a cosine measure is used to calculate similarities between two patent documents. After the document similarity matrices have been obtained, a patent map can be constructed by MDS.
Therefore, calculating outlier values to sustainably detect technological opportunities is a viable approach when new patents appear. However, further investigation is warranted to improve the precision of verifying assessment indicators of patent outlierness. Manual verification remains the most precise method. However, if the number of documents is large, a substantial amount of time is required for review; if the number of documents is too small, the results do not possess adequate integrity, which results in an inconsistency between the intended level of persuasiveness and the available sample size. Thus, in the future, the number of patent documents in a patent’s citation field and measures of the increase in patent documents related to the patent’s technology field can serve as references in the development of relevant assessment indicators for verifying document outlierness. The indicators thus developed can be made more persuasive.

This study integrated all operations into a single development environment. Because different programming languages are suitable for different fields, this study used multiple programming languages for different modules. In the future, if the functions of data mining and WordNet kits become increasingly comprehensive, the modules employed in this study can be integrated into a single development environment to reduce the complexity of development and enhance the overall performance of the implementation. This research limitation required us to avoid working with too many words; only the patent title field, the patent summary field and the text of the patent application scope field were used for processing.

U.S. Patent Nos. 8672692 and 7234963 are USB body structure patents. In Term Analysis 2, we mentioned that there were only four body structure patents in the world in 2015. The LOF patent map outliers and surrounding areas, which represent technologies related to USB body structure, indicate potential technological and business opportunities.

The primary goal of this study was to propose a method to reduce term dimensionality, specifically by grouping terms using term similarity matrices and merging semantically similar terms. In the related fields of information retrieval, data mining, and text mining [19], an immense number of eigenvalues or sparse matrices formed from terms tend to substantially reduce the effectiveness of the overall implementation. The method proposed in this study can reduce term dimensionality and facilitate user understanding through visualization. This method can also assist firms in formulating development strategies for avoiding patent infringement while sustainably discovering technological opportunities to achieve future competitive advantages.

Author Contributions: H.C.W. designed this research and collected the data set for the term analysis. Y.C.C. and P.L.H. analyzed the data to show the validity of this study. Y.C.C. wrote the paper and performed the research steps. In addition, the authors have all collaborated in revising the paper.

Funding: This research was funded by the Taiwan Ministry of Science and Technology grant number 107-2410-H-006-040-MY3.

Conflicts of Interest: The authors declare no conflicts of interest.

References


18. Walter, L.; Radauer, A.; Moehrle, M.G. The beauty of brimstone butterfly: Novelty of patents identified by near environment analysis based on text mining. Scientometrics 2017, 111, 103–115. [CrossRef]


© 2018 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).