Cross-language patent matching via an international patent classification-based concept bridge

Yen-Liang Chen
National Central University, Taiwan, Republic of China

Yu-Ting Chiu
National Central University, Taiwan, Republic of China

Abstract
Patent documents with sophisticated technical information are valuable for developing new technologies and products. They can be written in almost any language, leading to language barrier problems during retrieval. Traditionally, cross-language information retrieval and cross-language document matching have used text-translation-based or index-set-mapping methods. There are several challenges to the traditional methods, however, such as difficulties with natural language translation, complications owing to bilingual or multilingual translations (translating between two or more than two languages), and the unavailability of a parallel dual-language document set. This study offers a new and robust solution to cross-language patent document matching: the International Patent Classification (IPC) based concept bridge approach. The proposed method applies Latent Semantic Indexing to extract concepts from each set of patent documents and utilizes the IPC codes to construct a cross-language mediator that expresses patent documents in different languages. Experiments were carried out to demonstrate the performance of the proposed method. There were 3000 English patents and 3000 Chinese patents gathered as training documents from the United States Patent and Trademark Office and the Taiwan Intellectual Property Office, respectively. Another 30 English patents and another 30 Chinese patents were collected to be query patents. Finally, evaluations using an objective measure and subjective judgement were conducted to prove the feasibility and effectiveness of our method. The results show that our method out-performs the traditional text-translation methods.

Keywords
cross-language patent matching; latent semantic indexing; patent mining

1. Introduction
Patent documents issued by the Intellectual Property Office (IPO) are a treasure trove of novel technologies. A company should continuously analyse the knowledge within patent documents to create new product designs and to maintain competitive advantages [1]. Since patent documents are rich with information, inventors and researchers must search through a vast number of them to find documents related to their topic. Patent documents are not easy to retrieve without help from those familiar with the documents’ domain knowledge. It should also be noted that the irregular lengths of the documents and the variety of terminologies within them make analysis even harder [2]. Therefore, there is a need for proper retrieval and analysis techniques for patent documents, not only to discover critical technological research results, but also to increase the efficiency of patent prior art searches. According to Trappey et al., patent search prior, during and after the developing process of a new technology can prevent potential infringement of existing intellectual property rights and secure market advantage of newly developed intellectual property rights [3].

Textual documents are naturally written in different languages, depending on their country of origin. Every nation has its own Intellectual Property (or Patent) office that examines, preserves and protects these valuable assets that were
prepared in their native language. Inevitably, information is required by those who do not understand that language. According to Li and Shawe-Taylor [4], people often want to retrieve patent information in an unfamiliar language. Additionally, it is essential for organizations to obtain and manage knowledge in multiple languages [5]. This means that it is necessary to have a method for searching and retrieving multi-lingual information or documents. To meet commercial needs, inventors and assignees (people who apply for new patents) should be aware of relevant patents in other languages during the application process and maintenance stage. If a complete patent search could be carried out across multiple languages during the application process, it could prevent wasted efforts in applying for already-existing patents. Furthermore, discovering similar inventions described in different languages during the maintenance stage could help assignees become aware of potential competitors.

Traditional cross-language document retrieval and document matching methods are translation-based methods [6]. Since patents are a type of document, patent matching can be viewed as a type of document matching. Therefore, cross-language patent matching can be carried out using cross-language document matching methods. The translation-based methods can be further categorized into two groups, text-translation-based methods [6–9] and index-set-mapping methods [6, 10–12].

There are two major challenges in text-translation-based methods. First is the difficulty of natural language translation. For example, the different order of words in different languages is always a concern in any natural language processing [13, 14]. Even though we have software systems that can translate documents between different languages, the accuracy and quality are not acceptable for commercial applications. Problems that may occur during the translation process include lack of fidelity, information loss, translation ambiguity and others [6, 15, 16]. Second, when more languages are involved in document retrieval and matching, the task of language translation becomes even more complicated. For example, word sense disambiguation is a crucial problem in mapping similar concepts and measuring similar contexts, and is even harder to process across different languages [17, 18]. Index-set-mapping methods rely on statistical mapping relations found from document sets with parallel dual-language text, meaning that a prerequisite of this approach is the availability of a parallel dual-language document set. Unfortunately, this does not exist for patent documents. Additionally, the mapping relations are domain-dependent because they are learned from the parallel dual-language document set. When domain changes or topic drifts occur, the mapping relations change as well.

The main objective of this paper is to accomplish cross-language patent matching. In light of the drawbacks described above, it is clear that traditional approaches are not suitable, and so a concept-based bridge method is proposed. The proposed method is based on the distinctive attributes and structure of the IPC (International Patent Classification) code, which is unique to patent documents. Using the IPC code, a concept bridge can be built and applied to patent documents in different languages. Some advantages to this hierarchical structure are that it is officially published, adopted worldwide, constructed finely and changes only slightly. Most importantly, nearly every country includes it in their patent formats. In other words, every patent, no matter what language it was originally written in, contains at least one IPC code. Therefore, the IPC structure is suitable as a basis for a concept bridge between languages.

The method proposed in this study deals with the previously mentioned challenges. It can map patent documents, no matter the language, onto the same concept bridge. Patent documents can then be matched and compared across languages. This concept bridge can be viewed as a universal common framework for patents. In other words, any patent, no matter the language or domain, is represented by this universal common framework. This common representation allows users to determine the degree of similarity between patents. Compared with text-translation-based methods, the proposed method’s entire process can be conducted without dealing with challenges in natural language translation. Compared with index-set-mapping methods, the proposed method does not require a parallel dual-language document set in order to find mapping relations for index terms between different languages. These two differences are also two advantages of the proposed method.

2. Literature review

The literature review will first introduce the background and characteristics of the IPC. Then, since cross-language information retrieval (CLIR) and cross-language document matching (CLDM) are related to our work, we will review the work within these two fields. CLIR and CLDM are quite similar, but the former is more common and there are more relevant research studies in the literature. A review of the current state of CLIR and CLDM research is needed to understand the background of cross-language patent matching (CLPM). A brief introduction to compound nouns is also given in the last part of this section.
2.1. IPC

The IPC is a feature within patent documents and was established by the Strasbourg Agreement in 1971. Information about the IPC is currently published by the World Intellectual Property Organization. The ninth edition of the IPC report is the current version, released on 1 January 2009. This officially published IPC structure is well constructed and has strong public confidence. The classification has a five-level hierarchical structure, composed of section, class, sub-class, group and sub-group. The hierarchy contains a total of eight sections, more than 120 classes, more than 600 sub-classes, and approximately 70,000 groups [19]. Each group can be further divided into main groups and sub-groups, where 10% of the groups (nearly 7000) are main groups.

As illustrated in Figure 1, ‘H01L27/28’ is an example of an IPC code, where ‘H’ represents a section, ‘01’ is a class, ‘L’ is a sub-class, ‘27’ is a main group and ‘28’ is a sub-group. According to Kang et al. [20], the IPC is a standard taxonomy for sorting, organizing, classifying, determining and searching patent documents. The IPC code can be viewed as a topic label for the patent document’s contents. Therefore, we can determine the topic similarity of two patents based on the similarity of their IPC codes. The more similar two patent documents are, the longer a common prefix their IPC codes would have. For example, given three patents with IPC codes H01L27/18, H01L27/00 and H01L31/00, the first two patents are more similar than the last two.

2.2. CLIR

Göker and Davies [15] define CLIR as a technique that ‘addresses the problem of finding information in one language (e.g., Spanish) in response to queries in another (e.g., English)’. Lu et al. [21] used a semi-automatic term translation method to construct a Chinese–English MeSH (Medical Subject Headings) to use in a Cross-Language Medical Information Retrieval system for medical information. The best accuracy of automatically translating the English MeSH terms into Chinese was 63.9% in their experiment. Chen and Chen proposed a system to translate a Chinese named entity into English based on Google with a best accuracy of 87.6% [22]. Huang and Tsai [23] also developed a shopbot and built a semi-automatic multi-lingual ontology to assist customers in finding products in markets with different languages. Yang et al. applied the growing hierarchical self-organizing map to train and generate hierarchical feature map for retrieving multi-lingual information [24]. Al-Eroud et al. proposed an indexing method considering language and structure independently for the cross-language querying and searching tasks [25]. CLIR is also called ‘translingual information retrieval’ and closely related to ‘multilingual information retrieval’ [15]. As stated in Jiang and Tan [26], ‘For multilingual retrieval, transformations are usually needed for bridging the gap between different representation schemes based on different terminologies.’

The Introduction mentioned that traditional CLIR techniques and methods are designed based on translation, and can be further divided into text-translation-based methods and index-set-mapping methods. In addition, Oard and Diekema [27] mentioned four types of CLIR strategies: cognate matching, query translation, document translation and interlingual translation. In cognate matching, untranslatable terms are left unchanged, but might be matched to corresponding terms based on linguistic relationships. The steps in the other three translation-based strategies, query translation, document translation and interlingual translation [15, 27], are shown in Figure 2. The query translation approach translates the query keywords from one language (such as English) into another language (such as Chinese), and then uses the
translated terms (in Chinese) to discover the needed information (in Chinese). This approach is more efficient and flexible, but less accurate. The document translation approach translates the entire document’s contents (originally in Chinese) into the same language as the query keywords (such as English), and then the fully translated contents are used to discover the needed information. This approach is more accurate and translation ambiguities are resolvable, although time-consuming. The interlingual translation approach transforms and maps query keywords and document contents into a language-independent representation, and then uses this language-independent representation to find the required information.

There are several drawbacks and challenges to these current methods. The frequent occurrence of some poorly translated or untranslated key terms can impact retrieval performance or even cause correct terms to be discarded [28]. Therefore, correct translation of terms is a crucial issue. According to He et al. [16] and Göker and Davies [15], there are numerous ways that poor or erroneous translations can occur in translation-based cross-language information retrieval. For example, one word might be translated into more than a single word in another language, even when using a dictionary [26]. This is called the translation ambiguity problem [15]. In other words, problems come from the translating process, rather than the retrieval process. The translation ambiguity problem also occurs when translating terms without considering their context [15]. A simple example is translating the word ‘bank’. In the following three sentences, the word ‘bank’ has three different meanings: ‘She keeps her money in a bank’, ‘I took a walk on a river bank’ and ‘I saw a bank of flowers there’.

2.3. CLDM

CLDM is a technique for finding documents in one language that are related to a target document in another language. Its fundamental purpose is to determine the similarity between two documents written in different languages. Kishida applied double-pass algorithm to cluster multi-lingual documents – English, French, German and Italian news articles – for document translation [29]. Either the text-translation-based approach or the index-set-mapping approach could be used to perform this task. As mentioned previously, text-translation-based methods translate all the contents of every document from one language (e.g. Chinese) into another (e.g. English), and then compute the degree of similarity between each pair of translated documents (e.g. English). In contrast, index-set-mapping methods first generate the index vectors for every document in the original language (e.g. Chinese). These index vectors are then mapped to another language (e.g. English), according to the statistical mapping relations between terms in both languages, which are obtained by analysing parallel dual-language documents. Finally, the similarities between each pair of documents can be computed by the correlation or distance between the corresponding mapped index vectors.

According to Kishida [6], there are three main approaches to CLDM translation: machine translation [30], translation by a bilingual machine-readable dictionary [7, 8] and parallel or comparable corpus-based methods [10, 31]. The first two are text-translation-based approaches and the last one is an index-set-mapping approach. Kishida [6] reviewed the research on the use of text-translation-based and index-set-mapping approaches to solve the CLDM problem, and stated that, in the future, it would be useful to discover ways to find resources for resource-poor languages. Gey et al. [32] also noted that new instruments and methodologies need to be created for CLIR and CLDM to overcome all language barriers.

Another possibility for solving the CLDM problem is using methods similar to the interlingual CLIR translation mentioned in Subsection 2.2. It seems to make more sense to transform all documents into common concepts and represent those documents with a concept bridge, rather than translate them into the same language. An advantage of using a concept bridge is that it avoids wording and syntactic relations during the translation process. There are some drawbacks, however, to using a concept bridge. Multi-lingual thesauri are needed for the construction of language-independent

Figure 2. Three strategies for translation-based CLIR.

---

Chen and Chiu

Journal of Information Science, 2013, pp. 1–17 © The Author(s), DOI: 10.1177/0165551513494641
representations, which are expensive to build; additionally, how to automatically map query keywords and document terms to a language-independent representation is still an open question [15, 33, 34].

2.4. CLPM

Patent documents are rich with information related to technology, novelty, inventions and utility. Previous researches often designed methods for single-language patent retrieval, such as the method proposed by the authors [35]. Cetintas and Si designed a method to generate and post-process a query automatically for prior art patent search [36]. In addition, Vrochidis et al. developed a PatMedia search engine to retrieve patent images on the basis of contents [37].

However, there is little literature on cross-language patent retrieval. Li and Shawe-Taylor [4] applied machine translation techniques to cross-language patent retrieval. They introduced a learning algorithm, based on Kernel Canonical Correlation Analysis, for training and retrieving bilingual patent documents. Fujii et al. [9] analysed the accuracy of machine translation for cross-language patent retrieval. Literature on cross-language patent matching, however, is rare.

Cross-language patent matching can be viewed as a special application of cross-language document matching, since patents are a special type of document. This research does not focus on translating documents into another language or dealing with language differences encountered in translation. To the best of our knowledge, this is the only work where concept bridges are used to solve the problem of cross-language patent matching. A concept bridge for CLPM is designed using an IPC hierarchical structure, an exclusive feature of patent documents.

2.5. A brief introduction to compound nouns

Compound nouns (also called n-gram words) are formed by joining two, or more than two, single and adjacent words [38]. Compound nouns are named as key-phrases (or keyphrases) in other research [39–41]. For instance, ‘information-retrieval’ is a compound noun formed from two single words: information and retrieval. An n-gram phrase can express a more comprehensive meaning, which helps avoid misunderstanding, while also retaining richer and more precise semantic content. Karanikolas and Skourlas proposed a method to extract the most frequent and discriminate terms from documents within a class to form key-phrases, and then use these key-phrases to classify a new document [39, 40]. According to Joorabchi and Mahdi, utilizing keyphrases is an effective way to reveal subjects in scientific and research documents, and to retrieve information from databases [41]. Protaziuk et al. introduced an approach to discover compound nouns, using grammatical patterns and a T-GSP (Text Generalized Sequential Patterns) algorithm to extract frequent text sequences that satisfy a given set of grammatical rules [42]. They stated that compound nouns contain more semantic content and are more meaningful than a single noun or term, which is why the n-gram process is used to assemble compound nouns in this study.

Based on the concept of Protaziuk et al. [42], we use n-gram phrases to replace single words or terms to prevent implication misunderstanding within a document. We also set n-grams with larger n values to have higher weightings, since the probability of an n-gram phrase appearance is lower than that of single words.

3. Research design

A flowchart of this research design is shown in Figure 3. It is divided into two parts; the training stage is shown on the left-hand side and the new patent testing stage is on the right-hand side. Phases 1–5 are performed during the training stage, while phases 2–6 are performed in the new patent testing stage, where we find patent documents similar to the target patent. In the following discussion, an $A * B$ vector means that an object $A$ is expressed by a vector of $B$s (i.e. a vector containing all $B$ attributes). For example, a document * keyword vector refers to a document that is represented by a vector of keywords. Similarly, a document * concept vector means a document that is represented by a vector of concepts, while a document * category vector means a document that is represented by a vector of IPC categories.

The training stage steps are further illustrated in Figure 4. The process of constructing a document * category vector to represent a patent document using IPC categories can be divided into five phases. The similarity between two patent documents is computed in phase 6. The details of each phase are explained in the following sub-sections.

3.1. Phases 1a and 1b: collect patent documents

The goal of this study is to provide a cross-language document matching model for patent documents. Two document sets in two different languages (English and Chinese) were separately collected and prepared from the USPTO (United States Patent and Trademark Office) and the TIPO (Taiwan Intellectual Property Office).
Let \( D = \{d_1, d_2, \ldots, d_n\} \) be the body of patent documents, where \( n \) is the total number of patent documents in set \( D \). Every single document in \( D \) is denoted as \( d_j \). Since the set of patent documents contains both English and Chinese documents, it can be further divided into \( D_E \) and \( D_C \). Hence, \( D_E = \{d_{E1}, d_{E2}, \ldots, d_{Ep}\} \) is the set of English patent documents and \( D_C = \{d_{C1}, d_{C2}, \ldots, d_{Cq}\} \) is the set of Chinese patent documents, where \( D_E \cap D_C = \emptyset \), and \( D_E \cup D_C = D \). The total numbers of English and Chinese patent documents are \( p \) and \( q \), respectively, and \( |D| = |D_E| + |D_C| = p + q = n \).

### 3.2 Phases 2a and 2b: perform data preprocessing

After collecting the raw data (i.e. patent documents), several preprocessing operations are needed to clean the terms. The main objective of this phase is to obtain meaningful keywords. First, part-of-speech tagging [43] is performed on the USPTO patents and CKIP tagging [31, 44] is performed on the TIPO patents. Tagging identifies the syntactical functions and morphological features of every word in each sentence. Only adjectives and nouns are selected and retained to form \( n \)-gram phrases; they are combined based on the rules illustrated in Table 1 to construct the final set of \( n \)-gram phrases. Some \( n \)-gram phrases which might be too general or insignificant to be keywords are eliminated. The details of processing \( n \)-gram phrases are described in Sub-section 4.1.
At the end of this phase, the set of keywords is obtained and defined as $K = \{k_1, k_2, \ldots, k_r\}$, where $r$ is the total number of keywords in set $K$. Since patent documents in both English and Chinese are stored in set $D$ in the form of text files, the keywords for patent document $d_i$ can be represented as $K(d_i) = \{k_i \mid k_i \in d_i\}$. Keyword set $K$ is further divided into $K_E$ and $K_C$ to represent the sets of keywords in English and Chinese, respectively. There are English words within documents in $D_E$ and Chinese words within documents in $D_C$. Hence, $K_E = \{k_{E1}, k_{E2}, \ldots, k_{Es}\}$ is the set of English keywords and $K_C = \{k_{C1}, k_{C2}, \ldots, k_{Cs}\}$ is the set of Chinese keywords, where $K_E \cap K_C = \emptyset$, and $K_E \cup K_C = K$. The total number of English keywords $k_{Ei}$ in set $K_E$ is $s$ and the total number of Chinese keywords $k_{Ci}$ in set $K_C$ is $t$.

### 3.3. Phases 3a and 3b: build document * keyword vectors

In the third phase, document * keyword vectors are built by generating a document-by-keyword matrix. The document-by-keyword matrix is denoted as $A_{doc \times keyword} = [w_{doc \times keyword}]_{n \times r}$, where there are a total of $n$ documents and $r$ keywords in the matrix. The weight ($w_{doc \times keyword}$) of keyword $y$ in document $x$ is calculated with the Term Frequency–Inverse Document Frequency (TF-IDF) formula [45, 46]. For example, with English patent documents, document $d_{E1}$ can be represented by the keyword vector in Figure 5.

### 3.4. Phases 4a and 4b: transform to document * concept vectors

According to Manning et al. [47], using all the terms within a document set to represent every document could lower the accuracy rate. In contrast, representing a document with only a few key concepts will increase the accuracy, representing its real semantic contents. Therefore, in this phase, fewer but more representative concepts are used to represent documents in order to increase the matching accuracy.

Obviously, a document that is expressed by a set of keywords may contain multiple concepts. Furthermore, these keywords are usually closely related. Thus, document representation can be improved if these closely related keywords can be simplified by substituting them with independent concepts. This suggests that a document can be represented by a list of concepts with different weighting scores.

In this work, the well-known and convincing methods of latent semantic analysis (LSA) and latent semantic indexing (LSI) [48, 49] are adopted to group similar keywords together and a dummy index is used to express these keywords. Past work states, ‘Latent Semantic Analysis is a theory and method for extracting and representing the contextual-usage meaning of words by statistical computations applied to a large corpus of text’ [49, 50]. LSA is based on singular value decomposition, and can perform dimensionality reduction for text corpora. The LSI method has also been applied to CLIR. In Wei et al. [51], an LSI-based MLDC (multi-lingual document clustering) technique was designed to cluster multi-lingual documents and generate knowledge maps. In this research, the LSI method converts and transforms the patent documents from document * keyword vectors to document * concept vectors. Then, each document is represented by a set of concepts, and each concept has its respective weighting score.

In this phase, the important concepts from all documents are found and are used to represent every document. To this end, the LSI algorithm is applied to the keywords in the target document set. In other words, the goal of this phase is to use LSI to reduce the keyword dimensions and construct the document * concept vectors for all patent documents.
3.5. Phase 5: construct a cross-language mediator (IPC-based concept bridge)

Although all the documents in $D_E$ and $D_C$ have been transformed into document * concept vectors, the concepts extracted from $D_E$ and $D_C$ differ and cannot be communicated directly. Since the goal of this research is to design and construct a robust cross-language mediator, the IPC category codes, which are language-independent terminology and indicators, are used to represent documents. This cross-language mediator is named the IPC-based concept bridge. Utilizing a common labelling system like the IPC codes makes this cross-language mediator robust. Not only can it be used to link English patents with Chinese ones, but it can also be used in other languages.

Construction of the IPC-based concept bridge is shown in Figure 8. Each IPC category code, denoted as $g_t$, is one of the elements in the bridge. The set of IPC category codes is denoted as $G = \{g_1, g_2, \ldots, g_m\}$, with a total of $m$ IPC category codes. There are two steps to the fifth phase, which allows us to transform every document, whether in English or in Chinese, into a vector of IPC category codes using the IPC-based concept bridge.
The first step (steps 5a.1 and 5b.1) is to compute the concept-category weight. The concept-by-category matrix with a total of \( n \) concepts and \( m \) categories is denoted as \( A_{\text{con} \times \text{cat}} = [w_{\text{con} \times \text{cat}}]_{n \times m} \), where \( w_{\text{con} \times \text{cat}} \) is the weight of category \( y \) in concept \( x \). The concept-by-category matrix is obtained through a conversion process that transposes the category-by-concept matrix \( A_{\text{cat} \times \text{con}} \). The transposition is performed by interchanging the rows and columns in the matrix. For instance, the transposition of matrix \( A = [a_{x,y}]_{p \times r} \) can be written as \( A^T = [b_{y,x}]_{r \times p} \) and \( b_{x,y} = a_{y,x} \). After transposition, the concept * category weights \( w_{\text{con} \times \text{cat}} \) are obtained by \( w_{\text{con} \times \text{cat}}^T \).

Two methods to obtain the category * concept weights are proposed. The first method is described below.

\[
W_{\text{con} \times \text{cat}}^T(p, q) = \frac{\sum_{x \in D(p)} W_{\text{doc} \times \text{con}}(x, q)}{|D(p)|}
\]

The weight of concept \( c_i \) in category \( g_i \) in the cat * con matrix is the average weight of concept \( c_i \) from all documents \( d_j \) belonging to category \( g_i \) in the doc * con matrix. In other words, a concept’s weight with respect to a category is obtained by averaging the concept weights of the documents in this category.

The concepts of two separate patents may be very similar or dissimilar. With the help of the IPC hierarchy, determining to what extent a concept belongs to a category can be calculated more sophisticatedly. For example, if a concept belongs to category H01L31, it will have a decreasing extent of belonging to its parent level category (H01L), and its grandparent level category (H01). Thus, it would be more reasonable to consider the partial contributions coming from categories closer to the current category when calculating the weight of concept \( c_i \) in category \( g_i \). For this reason, we designed a second method for computing the weight of concept \( c_i \) in category \( g_i \) in the cat * con matrix. In this second method, the degree to which a document \( d_j \) belongs to category \( g_i \), which is denoted as \( W_{g_i}(d_j) \), is taken into consideration.

The following formula is used to compute the weights of concept \( c_i \) in category \( g_i \) in a category-by-concept matrix \( A_{\text{cat} \times \text{con}} \):

\[
W_{\text{cat} \times \text{con}}(g_i, c_i) = \sum_{d_j \in D(g_i)} W_{\text{doc} \times \text{con}}(d_j, c_i) \times W_{g_i}(d_j)
\]

In the formula \( W_{\text{cat} \times \text{con}}(g_i, c_i) \), the value of \( W_{g_i}(d_j) \) can be computed using the following \( W_{g_i} \) formula, and the value of variable \( r \) is an integer ranging from 0 to 2. Here, the value of variable \( r \) is 0, 1 or 2 if document \( d_j \) and category \( g_i \) have the same IPC code through the group level, the sub-class level and the class level, respectively. Otherwise, we set the value of \( W_{g_i}(d_j) \) to zero when they are further away.

\[
W_{g_i} = 1.0 \times (0.5)^r
\]

Consequently, the concept * category weight \( W_{\text{con} \times \text{cat}}(c_i, g_i) \) for each category in a concept can be obtained from the concept * category matrix through transposition.

In the second step (steps 5a.2 and 5b.2), the document-category weight is computed. The document-by-category matrix is denoted as \( A_{\text{doc} \times \text{cat}} = [w_{\text{doc} \times \text{cat}}]_{r \times m} \). Additionally, the weight of each element in the document * category matrix is computed by
Since the concepts obtained from the English keywords, the weights \( W_{d_j|g_i}^{\text{doc}*\text{cat}} \) for patents in English and Chinese should be computed independently. The weight of a category (i.e. the IPC code) \( g_i \) in an English patent document \( d_j \) is calculated by \( W_{d_j|g_i}^{\text{doc}*\text{cat}} = \sum_{i=1}^{a} W_{d_j|c_i}^{\text{doc}*\text{con}} \times W_{c_i|g_i}^{\text{con}*\text{cat}} \), and the weight of a category (i.e. the IPC code) \( g_i \) in a Chinese patent document \( d_j \) is calculated by \( W_{d_j|g_i}^{\text{doc}*\text{cat}} = \sum_{i=1}^{b} W_{d_j|c_i}^{\text{doc}*\text{con}} \times W_{c_i|g_i}^{\text{con}*\text{cat}} \).

### 3.6. Phase 6: compute similarity

After completing the Training Stage (phases 1–5), every patent document, whether in Chinese or English, is transformed into a document * category vector. The New Patent Testing Stage begins when a user starts a search for patent documents similar to his or her own specific patent document. The specific patent document given by the user is a new document called the query patent document. This query patent document is transformed into a document * concept vector using the permanently recorded left-singular vectors \( K_c \) and diagonal matrix \( S_c \). The category * concept vector is then constructed by calculating the weights for each category in every concept using the two \( W_{c_i|g_i}^{\text{cat}*\text{con}} \) formulas. Finally, the document * category vector for the query patent document is constructed and ready for patent document similarity comparison (phase 6). The similarity score for each pair of patent documents (the query patent document \( d_j \) and every patent document \( d_i \) in the patent document collection) is computed using the traditional cosine similarity formula:

\[
\text{sim}(d_i, d_j) = \frac{\overrightarrow{d_i} \cdot \overrightarrow{d_j}}{|\overrightarrow{d_i}| \times |\overrightarrow{d_j}|} = \frac{\sum_{i=1}^{m} W_{d_i|c_i}^{\text{doc}*\text{cat}} \times W_{d_j|c_i}^{\text{doc}*\text{cat}}}{\sqrt{\sum_{i=1}^{m} (W_{d_i|c_i}^{\text{doc}*\text{cat}})^2} \times \sqrt{\sum_{i=1}^{m} (W_{d_j|c_i}^{\text{doc}*\text{cat}})^2}}
\]

### 4 Experiment and evaluation

We wanted to demonstrate the proposed method’s performance compared with traditional text-translation-based approaches. Two experiments were conducted. One found similar patent documents in an English patent corpus using Chinese query patent documents (named CqEc), and the other found similar patent documents in a Chinese corpus using English query patent documents (named EqCc). The details of the experiments are described in the following subsections.

#### 4.1. Data collection and preprocessing

The raw data, the English and Chinese patent documents, were collected from the USPTO and TIPO databases, respectively. The format of each collected patent document is presented in a txt file. There were 3000 English patents and 3000 Chinese patents gathered as training documents, and stored in \( D_E \) and \( D_C \). These two sets of patents were viewed as matching patent document sets in the query document testing stage. In order to create some diversity within the document set, half of the 3000 patents were randomly selected from section ‘A’, and the other half were randomly selected from section ‘H’. Additionally, another 30 patent documents were randomly collected from the USPTO and another 30 from the TIPO to act as query patent documents. Half of the 30 query documents were from section ‘H’ and half were from section ‘A’. In our experiment, there were a total of 30,847 n-gram phrases in the English patent document set, and 24,239 n-gram phrases in the Chinese patent document set. After stopword elimination and filtering out phrases with lower TF-IDF values, 3283 English n-gram phrases and 2397 Chinese n-gram phrases were retained for the following operations.

#### 4.2. Vector transformation

While the data was being prepared, the following steps transformed all the patent documents, regardless of whether it was from the USPTO or TIPO, into document * category vectors. In this experiment, the number of concepts in the LSI algorithm was set at 70. Consequently, 3283 English keywords from the set of English patents were transformed into 70 dummy concepts, and 2397 Chinese keywords from the set of Chinese patents were transformed into another 70 dummy
concepts. Afterwards, an IPC-based concept bridge (the document category vectors) was generated as a cross-language mediator between the English and Chinese patent documents.

4.3. Proposed method and comparison methods

In the experiment, there were two proposed and three comparison methods for seeking and matching similar patent documents across different languages: IPC, the proposed IPC method without hierarchy weight consideration; IPC-H-Weight, the proposed IPC method with hierarchy weight consideration; TT-Google, text-translation method using Google Translator [52]; TT-Yahoo, text-translation method using Yahoo Translator [53]; and TT-Microsoft, text-translation method using Microsoft Translator [54]. Owing to the unavailability of a parallel dual-language document set for patents, the index-set-mapping approach was excluded from the comparison methods.

As laid out in the research design, the IPC method was used in both experiments CqEc and EqCc to construct document category vectors for all patents in the matching patent document sets and the query patent document sets. In CqEc, the similarity score for each Chinese query patent and every patent in DE was computed; also, in EqCc, the similarity score for each English query patent and every patent in DC was computed. Using these similarity scores, the similarity of a patent to a query patent was ranked in descending order. The top k similar patent documents in the ranking list would be the matching results for the query patent document. The IPC-H-Weight method was conducted similarly, but with the computation of hierarchical weight. The calculation of the category concept weight in the IPC-H-Weight method took into consideration the hierarchical IPC relationships and used formulas (2) and (3); the category concept weight in the IPC method was calculated without considering the hierarchical IPC relationships using formula (1).

The other three text-translation-based methods are representative of traditional methods. The solutions obtained with these methods are compared with those obtained with the proposed methods. With the three traditional methods, all the patents in the matching patent document set and the query patent document set were transformed into document key-word vectors. In CqEc and EqCc, the query patent documents were translated with Google Translator, Yahoo Translator and Microsoft Translator. In other words, the Chinese query patent documents in CqEc were translated into English prior to matching and seeking similarities, and vice versa. In CqEc, the similarity score for each query patent written originally in Chinese and every patent in DE was computed. In EqCc, the similarity score for each query patent written originally in English and every patent in DC was calculated. Based on the similarity scores, patents similar to each query patent were ranked in descending order. The top k similar patent documents in the ranking list most closely matched the query patent document.

4.4. Experimental results

In order to explore the performance of our proposed method, two types of measurements, objective measure and subjective measure, were used to assess the experiment. The objective measure is the average IPC common prefix length (AICPL) between the query document and the retrieved documents. Since the IPC hierarchy is a classification scheme that organizes and classifies all patents, two patents that are more similar will have a longer common prefix in the hierarchy, or a larger AICPL. Therefore, if a method’s query results return larger AICPLs, the retrieved documents are more similar to the query document. In this way, the AICPL can objectively verify if one method is better than another. User judgement via user surveys is the subjective measure used to assess performance. It reflects users’ subjective feelings (level of satisfaction) regarding the query results. On the other hand, the AICPL is objective and can avoid human biases.

4.4.1. AICPL measure. In the IPC hierarchy, two patent documents with the same IPC code are conceptually similar. Furthermore, the contents of two patent documents are even more similar if their IPC codes have a longer common prefix, or a larger AICPL. For example, ‘H02B01’ and ‘H01L27’ have the same prefix to section level, while ‘H01L31’ and ‘H01L31’ have the same prefix to group level. Therefore, the latter pair is more similar than the former pair. The AICPLq,dj score is designed to measure the commonality between a query patent document qj and a matched patent document dj.

\[
\text{AICPL}_{q_j,d_j} = \text{Avg} \times (\text{Com})
\]

where Com equals 1, 0.5, 0.25 and 0.125 if the query document and the matched document have the same IPC code through the bottom level, the sub-class level, the class level and the section level, respectively. Otherwise, we set the
weight at zero. For example, the AICPL score for ‘H02B01’ and ‘H01L27’ is 0.125 and the score for ‘H01L31’ and ‘H01L31’ is 1.

The results of CqEc are illustrated in Figure 9 below. The average performance for the top $k$ similar patent documents, from $k$ equals 10 to 100, is shown in the graph. The IPC method and the IPC-H-Weight method performed the best. The TT-Google method performed worse than the first two methods, but slightly better than the other two text-translation methods.

The experimental results of EqCc are illustrated in Figure 10. Each point on a line indicates the average performance of the corresponding top $k$ similar patent documents. For example, on the IPC-H-Weight method line, the first point (0.1949) means that it performed better than the other four methods for the top 10 similar patent documents. The IPC method and the IPC-H-Weight method achieved the best performances. The TT-Google method performed worse than the IPC-H-Weight method and similar to the IPC method, but did perform slightly better than the other two text-translation-based methods.

4.4.2. User judgement. In the subjective evaluation, we measured user satisfaction regarding a query’s result list. In order to avoid overwhelming participants by asking them to read numerous patent documents not written in their native...
language, we only conducted EqCc to ask them which method found similar Chinese patents more accurately with a given English query document. The user evaluation consisted of eight sub-experiments. Each sub-experiment had a target patent document and a list of the most similar patent documents found using the five different methods (IPC, IPC-H-Weight, TT-Google, TT-Yahoo and TT-Microsoft). The English target patent document was one of the 30 patent documents randomly collected from the USPTO, and the similar patent documents were retrieved from the 3000 Chinese patent documents in $D_C$.

For each sub-experiment, 20 participants (PhD students and academics) were invited to participate in the user evaluation. During this process, participants were asked to critique the performance of each method for each sub-experiment. They were asked to judge the retrieved similar patent documents from the different methods, and to score the methods according to the similarities between the found patents and the target patent. The scores ranged from 1 to 5 (1 being the worst and 5 being the best). Figure 11 reveals the outcomes of the user evaluation. According to the results of user judgement, the performances of the proposed methods were better than those of the traditional methods. This result demonstrates the feasibility and usefulness of the proposed method.

The results of all sub-experiments in user judgement can be classified into three categories, where the first category includes sub-experiments 1, 3, 4 and 7, the second category includes sub-experiments 5 and 8, and the last category includes sub-experiments 2 and 6. In the first category, method IPC-H-Weight performed better than the traditional methods and the IPC method. By examining the candidate documents in detail, we found that the candidate documents that are similar to the query document occur not only in the IPC category of the query document but also in other nearby categories. That is why weighting can result in a better performance in the IPC-H-Weight method. In the second category, all methods performed poorly. In fact, since no method can achieve a satisfactory result, there are no true winners in sub-experiments 5 and 8, because even the winner is a loser. By examining the candidate documents in detail, we found that no candidate documents are similar to the query document in sub-experiments 5 and 8. Owing to the low similarity, the two sub-experiments become like a random test. Lastly, in the third category IPC method performed better than the traditional methods and the IPC-H-Weight method. By examining the candidate documents in detail, we found that in sub-experiments 2 and 6 the candidate documents that are similar to the query patent document are mostly

Table 2. t-Test result for user judgement.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>2.69</td>
<td>2.19</td>
<td>2.69</td>
</tr>
<tr>
<td>Variance</td>
<td>0.66</td>
<td>0.56</td>
<td>0.66</td>
</tr>
<tr>
<td>Observations</td>
<td>20</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>Pearson correlation coefficient</td>
<td>0.54</td>
<td>0.41</td>
<td>0.66</td>
</tr>
<tr>
<td>Hypothesized mean difference</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Degrees of freedom</td>
<td>19</td>
<td>19</td>
<td>19</td>
</tr>
<tr>
<td>t Statistic</td>
<td>2.93</td>
<td>2.07</td>
<td>6.76</td>
</tr>
<tr>
<td>$p (T \leq t)$ two-tailed</td>
<td>0.008525</td>
<td>0.052769</td>
<td>0.000002</td>
</tr>
<tr>
<td>t Critical two-tailed</td>
<td>2.093024</td>
<td>2.093024</td>
<td>2.093024</td>
</tr>
</tbody>
</table>
localized at the IPC category of the query document. Owing to the locality property, the IPC method could produce a better query result because it did not give weights to other nearby categories in the hierarchy.

To measure if the IPC-H-Weight method is significantly better than traditional methods, we compare the method with the three traditional methods by t-test. Table 2 above indicates the results of the t-test between the IPC-H-Weight method and the other three traditional text-translation-based methods. The t-test results show that the proposed method performs significantly better than the TT-Google and TT-Microsoft methods. As for the TT-Yahoo method, our method is the winner lying on the border line of statistical significance.

5. Discussion

The aim of the experiments was to evaluate the performance of the proposed methods. Since a larger AICPL score indicates higher similarity between a query document and a retrieved document, the retrieval results are more accurate when the AICPL score is larger. According to Figures 9 and 10, the proposed method IPC-H-Weight retrieved the most accurate results compared with the other four methods. Between the proposed methods of IPC-H-Weight and IPC, the performance of the former is better than that of the latter. This result indicates that the weighting scheme can improve the performance of retrieving the results.

In addition, this research also analysed whether there is any difference between the performance of CqEc (retrieving similar English patent documents via Chinese patent documents) and EqCc (retrieving similar Chinese patent documents via English patent documents). The performances of the IPC-H-Weight approach in CqEc and EqCc are quite similar. However, the performance of traditional translation approach in EqCc is better than that in CqEc. This interesting discovery indicates that the correctness and completeness of machine translation ability is unequal between the Chinese-to-English translation and the English-to-Chinese translation. According to this interesting discovery, a research question arising immediately is why cross-language retrieval between two particular languages is more effective than that between another two? In the future, we could attempt to study the IPC-based cross-language retrieval problem between two languages other than Chinese and English. For example, the grammatical order between Chinese and English is more similar than that between English and Korean or Japanese. Similarly, the grammatical order between Korean and Japanese is also similar. What if this study was about two languages having quite different grammatical order? Could the IPC-based method be still effective for such a case? This is an interesting question and could be further investigated.

The contributions of this research are not only that the proposed method improves the retrieval performance of previous translation-based methods but also that our proposed method provides an intermediary bridge for communicating documents in different languages. In other words, no translation or dictionary is needed when retrieving documents using the query document in a different language. No matter how many languages there are in the patent documents to be retrieved, our proposed method can work properly without conducting translation. However, if we still use a traditional translation-based approach to do cross language query, we will need \( C_n^2 \) translators for \( n \) different languages.

6. Conclusion

This study proposed a new method that generated an IPC-based concept bridge as a means to search for and match patent documents. Although patent documents are written in different languages, they can all be transformed into document category vectors using the IPC-based concept bridge. The proposed method can help overcome the two main challenges encountered with text-translation-based methods and index-set-mapping methods.

There are three limitations while doing this research and the experiments. They are the subjective-answer issue, the time-consuming issue and the user-fatigue issue. First of all, since it is subjective to rank the similarities of the retrieved patent documents for each query patent document, it is impossible to calculate the scores of recall and precision after each retrieval task in the experiment. This subjective-answer issue caused us to evaluate the performance of the proposed methods via AICPL measure and user judgement. Secondly, because of the time-consuming issue, we collected 3000 patent documents in English and another 3000 patent documents in Chinese as training documents to conduct the experiments. The size of the data collection can be enlarged in the future to test the performance and robustness of the proposed method. The third limitation is about user fatigue. Since the sub-experiment in the user judgement is complex and needs a long time, users feel too fatigued to concentrate their attentions. Hence, we only asked the participants to do eight sub-experiments in the user judgement to avoid the issue of user fatigue. The experimental and evaluation results demonstrated that the proposed methods performed better than any of the three text-translation-based methods (which used Google Translator, Yahoo Translator and Microsoft Translator). The results also demonstrated performance reliability.
In the future, this IPC-based concept bridge can be further extended. Feasibility could be tested by building representative vectors for more diverse content fields in patent documents (e.g., description and claims). The concept bridge approach could also be used to represent patents in languages other than Chinese and English and gathered from other patent databases, such as the JPO (Japan Patent Office) or EPO (European Patent Office). The JPO issues patent documents mainly written in Japanese. Users can browse or collect patent documents from the JPO web site. The EPO, on the other hand, issues patent documents mainly written in English, French or German. Users can browse or collect patent documents from EPO web site. This new method could be adapted for practical use in a cross-language document matching system to improve retrieval efficiency. Currently, search interfaces or services for similar patent documents are mostly monolingual. This deficiency suggests the need to develop a search interface in the future that is based on the proposed method to provide the service of retrieving patent documents in one language using a patent document in another language. This kind of system is helpful for a company that needs to understand whether other companies infringe upon its own intellectual property rights (i.e. have applied for a patent outside their country in other languages), or whether other companies are developing similar technologies and have applied for a patent abroad in other languages. Retrieving patent documents across languages will become more crucial when the patent war between companies become more intense.

References


[34] Loukachevitch NV and Dobrov BV. Cross-language information retrieval based on multilingual thesauri specially created for automatic text processing. In *Proceedings of workshop on cross-language information retrieval: A research roadmap, SIGIR*, 2002.


