I-Pats: An Intelligent Search System for Thai Patents

Marut Buranarach\textsuperscript{1} Choochart Haruechaiyasak\textsuperscript{2} Alisa Kongthon\textsuperscript{3}

\textsuperscript{1,2,3} Human Language Technology Laboratory, National Electronics and Computer Technology Center (NECTEC), Thailand Science Park, Klong Luang, 12120, Pathumthani, Thailand

Tel. +66-2-564-6900, Fax.: +66-2-564-6872
e-mail: {marut.buranarach, choochart.haruechaiyasak, alisa.kongthon}@nectec.or.th

Abstract

This paper describes development of I-Pats, an intelligent search system for Thai patents. One of the major goals is to improve efficiency, effectiveness and intelligence in searching Thai patent information. The system was built on top of a full-text search engine augmented for texts in Thai language. Phonetic-based index was used in improving effectiveness for searching names, i.e. patentee and inventor names. Patent analysis was provided in a multi-faceted view to allow flexible visualization and navigation of the patent data. Based on the initial testing on the system, the retrieval results of I-Pats in terms of efficiency and effectiveness are improved over a baseline database system.

1 Introduction

Patents, also known as patents for invention, are the most widespread means of protecting the rights of inventors. A patent may be granted for a new, useful, and non-obvious invention, which gives the inventors exclusive rights to exploit their works and invention for a limited period. With the growth of patented inventions, the need for effective patent search systems has become increasingly important. Patents are an important source of technological intelligence that organizations and companies can use to gain strategic advantage. Further, inventor normally conducts patent search as a preemptive measure to ensure that the same idea or a similar is not already in use by others. Available patent search systems include those provided by the European Patent Office (EPO)\textsuperscript{1} and the United States Patent and Trademark Office (USPTO)\textsuperscript{2}.

Patent database in Thailand is maintained by the Department of Intellectual Property\textsuperscript{3}, Ministry of Commerce. The patent information is provided in Thai language. Terms from foreign languages were transliterated into Thai including technical terms, person and organization names. There was a major limitation in developing a search system for Thai patents using database management system (DBMS). Typical full-text indexing option in DBMS normally cannot be applied to Thai text straightforwardly. This is primarily due to the non-segmented nature of the Thai writing system. The lack of full-text search support for Thai in DBMS has been a major drawback in gaining good retrieval efficiency in existing Thai patent search system.

Another problem in searching Thai patent database is searching based on person or organization names. In locating patents, users commonly use inventor or patentee names as search terms. However, names are often misspelled. In addition, transliterations of names in the patent database, e.g. from English, Japanese or Chinese into Thai, were done inconsistently. As a result, searching based on names often gives unsatisfactory results.

Patent analysis is one of the most important features in a patent database. Analyzing and visualizing patent information helps to provide the users with technological and business intelligence. Basic patent analysis, such as competitive analysis, allows the users to observe who are being active in a technology field. A more advanced analysis includes technology

\textsuperscript{1} http://ep.espacenet.com/
\textsuperscript{2} http://www.uspto.gov/patft/
\textsuperscript{3} http://www.ipthailand.org/
I-Pats (Intelligent Search System for Thai Patents) is an initiative to build a search system for Thai patents that focuses on efficiency, effectiveness and intelligence in retrieving Thai patent information. A prototype system was built and combined both language-dependent and language-independent techniques in improving retrieval performance. The system was built on top of Lucene\(^4\), a Java-based information retrieval (IR) library augmented by three sub-components: Thai-text analyzer, Thai Grapheme-to-Phoneme (G2P) and Multi-faceted statistical analyzer. The initial testing of I-Pats indicated improvement in terms of retrieval efficiency and effectiveness over a baseline DBMS system.

2 Design and Implementation

I-Pats supports three main user functions: full-text search for Thai texts, name search based on phonetic similarity, i.e. soundex search, and patent analysis and visualization. This section discusses the design and implementation of I-Pats. Section 2.1 provides a conceptual architecture. Sections 2.2-2.4 describe the design and development of each main function. Section 2.5 describes the user interface of I-Pats.

2.1 Conceptual Architecture

The conceptual architecture of I-Pats is shown in Figure 1. I-Pats was designed and built on top of Lucene, a high-performance Java-based IR library. It provides an additional layer consisting of three sub-components: Thai-text analyzer, Thai Grapheme-to-Phoneme (G2P) and Multi-faceted statistical analyzer. The Thai-text analyzer mainly performed word-segmentation for Thai text and passed on the output tokens to Lucene for creating and searching full-text index. The Thai G2P converted input Thai word sequence into corresponding phonetic transcription. The converted text was then passed on to Lucene for creating and searching phoneme-based index. The multi-faceted statistical analyzer grouped and analyzed search results by facets. The analysis results were visualized using histogram charts.

\footnote{http://jakarta.apache.org/lucene/}
suffix array. The suffix array indexing approach can be viewed as a string-based approach.

We chose the inverted index approach in implementing full-text search for Thai text. There were two major reasons for the decision. First, based on evaluation results, the inverted index approach is more scalable in terms of indexing time and space (Haruechaiyasak et al., 2006). Second, the string-based matching technique used by the suffix array approach when applied to Thai text can sometimes result in matches of non-word tokens. For example, a query ‘รถ’ (‘Car’) can result in matching of the word ‘สามารถ’ (‘To be able to’). The two words have no semantic relations but share some common string of characters. This can lead to lower accuracy of the retrieval results. The problem can be prevented by the word-based approach used in the inverted index scheme.

Figure 2. Example of Creating Inverted Index for Thai Text

Creating an inverted index for Thai text must rely on language-dependent word segmentation technique. I-Pats contains the Thai-text analyzer module whose major task is to tokenize a Thai text into a set of words. The module was integrated with the Lucene API to enable preprocessing of Thai text before creating and searching from an index. Figure 2 illustrates an example of creating inverted index for Thai text.

2.3 Name Search based on Phonetic Similarity

In supporting search based on phonetic similarity, i.e. soundex search, for names, the Thai Grapheme-to-Phoneme (G2P) module is essential in converting Thai text in written form into corresponding phonetic transcription. The follows provide some background for Thai pronunciation system, which forms the basic knowledge in Thai G2P.

A basic Thai-pronunciation unit is a syllable that can be represented in the form of \( Ci \, V \, Cf \, T \), where \( Ci \) denotes an initial consonant, \( V \) denotes a vowel, \( Cf \) denotes a final consonant, and \( T \) denotes a tonal marker. \( Cf \) and \( T \) are optional. There are 44 consonants and 28 vowels in Thai. There are 21 phonemes for the consonants when used as initial consonants and nine phonemes when they are used as final consonants. Like Chinese, Thai is a tonal language. There are five tones in Thai, i.e. Mid (0), Low (1), Falling (2), High (3) and Rising (4). Four tonal markers and one non-mark are used to indicate the tone. A tone is determined by the combination of syllable structure, initial consonant and the tonal marker. Design and development of the Thai G2P module is discussed in (Tarsaku et al., 2001).

Figure 3. Example of Phoneme-based Indexing and Retrieval

A phoneme-based index was additionally created for the fields containing names, i.e. inventor and patentee names. The index stored phonetic representation of the field values and was created by mediating the Thai G2P module before Lucene indexer. For example, in Figure 3, three transliterations of the name “Smith” shared the same phonetic representation after being processed by the Thai G2P module. The converted values were passed to Lucene indexer and stored in the phoneme-based index. When the user queried the name “Smith” using a different spelling, the query was converted to the same phonetic representation before it was passed to Lucene searcher and searched in the
phoneme-based index. Thus, all the relevant names can be retrieved even though their spellings were different.

2.4 Patent Analysis and Visualization

Faceted classification is a representation that has been increasingly used in enabling semantic-based search in collection (Hearst, 2006). It provides flexible ways to navigate and access the contents of the underlying collection. In this scheme, a facet represents one dimension of the underlying data. The values in the facet may be hierarchically structured. By navigating the facet values, user can specify constraints on the items retrieved from the system in an interested dimension. Multiple facets further allow the user to define additional constraints on the retrieved results in some other dimensions.

The multi-faceted statistical analyzer permits grouping of patent search results into four facets, i.e. International Patent Classification (IPC) code, Patentee, Inventor and Year. The values in each facet are sorted by frequency and are visualized using histogram charts. The user can further refine query by browsing the facet values. Figure 4 illustrates analyzing and visualizing search results using the faceted classification scheme.

2.5 User Interface

One of the design goals of I-Pats was to make it simple and flexible for the users to search and navigate the patent information. The user interfaces for search and navigation consist of three main screens. The first screen is an advanced search form, which permits keyword-based, Boolean and full-text search with the soundex search option available for the patentee and inventor fields. The second screen is search result display, where the users can view the patent information as well as links to patent document copies. The final screen is search result summary which analyzes and visualizes patent records in the search results. It also provides simple faceted browsing interface, which allows the users to additionally refine queries by navigation. Figure 5 shows an
example of using I-Pats in searching and analyzing the patent information.

3 Evaluation Results

Retrieval performance of I-Pats was assessed in terms of response time, result accuracy for full-text search and result improvement for name search based on phonetics. The performance was evaluated in comparison with a baseline system, i.e., PostgreSQL 8.2 DBMS. The test was made to emphasize the improvement gained by the approaches used in I-Pats rather than benchmarking on the systems. The test was conducted in a Pentium M 1.8GHz system with 512 MB RAM and Windows XP platform.

Response time was assessed by querying over the ‘abstract’ field. The field contains long Thai text strings, and thus can not benefit from special data structure, i.e. B+-Tree or full-text index, offered in the database system. Test sets were reproduced from the patent database in five different sizes, i.e. 23K, 47K, 70K, 94K, 117K records. The average response time of both systems for the test sets is shown in Figure 6. The result shows overall improvement in average response time for I-Pats over the baseline system when querying over the test field (0.1 vs. 12.9 seconds).

Result accuracy for Thai full-text search was assessed using sample query terms that are normally ineffective in querying Thai database. The five sample query terms were ‘ไหม’ (‘silk’), ‘จอ’ (‘monitor’), ‘ราว’ (‘handle’), ‘ปลา’ (‘fish’), and ‘ข้าว’ (‘rice’). The terms were used in querying over the ‘title’ field. The result accuracy of both systems for the test queries is shown in Figure 7. The result shows overall improvement in result accuracy for I-Pats over the baseline system for the test queries (97% vs. 58%).

Result improvement for name search based on phonetic similarity was assessed in terms of Novelty Ratio (Korfhage, 1997). Novelty ratio is the proportion of the relevant retrieved items that were previously unknown to the user. The five sample query terms used in the test were ‘แพคเกจจิง’ (‘Packaging’), ‘วิจัย’ (‘Research’), ‘ซิส’ (‘System’), ‘ดิเวลลอปเมนท์’ (‘Development’), and ‘อิเล็กทรอนิก’ (‘Electronic’). These terms were transliterated terms and were found to have various spelling forms in the patent database. The average novelty ratio was calculated by measuring the novelty of the search results using the phonetic option in reference to each spelling form found in the database. The average novelty ratios with standard deviations for the test queries are shown in Figure 8. The result shows overall improvement for name searching using the phonetic option for the test searching (Avg. Novelty Ratio = 82%).
4 Conclusion

We demonstrated some intelligent approaches in searching Thai patent database. The system used the language-dependent techniques, i.e. Thai word-segmentation and grapheme-to-phoneme conversion, in improving retrieval efficiency and effectiveness. The system also combined a language-independent technique, i.e. faceted analysis and browsing, in providing flexible and intelligent analysis of the patent information. Some future plans include automatic data inconsistency detection and cleaning, which can further improve the effectiveness of search and analysis.

Acknowledgement

This work was supported by the National Science and Technology Development Agency (NSTDA), Thailand under the Thailand's Research Information Portal and Search Engine (ThaiReSearch) project. The authors would like to thank Phubate Udomsaph, for his significant contribution on the design of the system, Nattanun Thatphithakkul, Chai Wutiwiwatchai, Rungkarn Siricharoenchai, Sirichai Lerdvorawut and Chatchawal Sangkeetrakarn for their technical support.

References


