A Survey of Automatic Indexing Techniques for Thai Text Documents

Todsanai Chumwatana*

Abstract

With the rapidly increasing number of Thai text documents available in digital media and websites, it is important to find an efficient text indexing technique to facilitate search and retrieval. An efficient index would speed up the response time and improve the accessibility of the documents. Up to now, not much research in Thai text indexing has been conducted as compared to more commonly used languages like English or other European languages. In Thai text indexing, the extraction of indexing terms becomes a main issue because they cannot be specified automatically from text documents, due to the nature of Thai texts being non-segmented. As a result, there are many challenges for indexing Thai text documents. The ma-jority of Thai text indexing techniques can be divided into two main categories: a language-dependent technique and a lan-guage-independent technique as will be described in this paper.

Keywords: Thai Text Indexing, Language-Dependent Technique, Language-Independent Technique.

1. Introduction

There is an ongoing challenge to develop more efficient text indexing techniques for Thai text documents, in order to enhance the performance of Thai information retrieval. To meet the challenges, a number of indexing techniques have been proposed for the Thai language. These techniques were designed to be used in information retrieval systems, to find required information from the huge amount of text documents. Figure 1 provides an overview of a general indexing and retrieval system, before the detail of the Thai text indexing will be described.

From Figure 1 it can be seen that indexing is one of the more important processes for managing information of text documents in order to facilitate the information retrieval process. Text documents are by necessity indexed using some indexing techniques for efficient retrieval [1], [2]. However, it has long been known that the advancement of Thai text indexing is challenging due to its nature of being a non-segmented language.

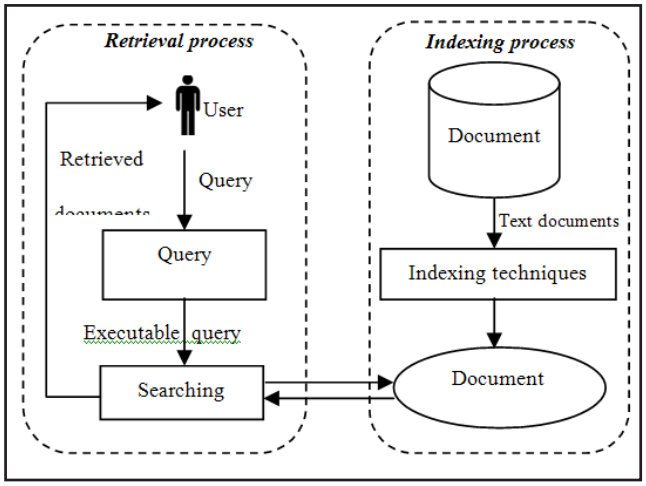

Figure 1 *A general indexing and retrieval system.*

The Thai language belongs to a class of non-segmented languages, the sentences of which consist of a string of symbols without explicit word delimiters. Words in the Thai language are not naturally separated by any word-delimiting symbols. Due to this characteristic, indexing Thai text documents is a challenging task and has become one of the research focuses in the area of Thai information retrieval. Many techniques have been proposed to index Thai text documents, including n-gram inverted index [3], word inverted index and suffix array approaches [4].

In the following sections, overviews of Thai text indexing techniques used in the area of Thai information retrieval systems are presented. These will offer help to understand the challenges present in Thai text indexing.

2. Overview of Thai Text Indexing

Indexing Thai text documents is an essential process in Thai information retrieval [4], [5], [6]. Indexing is a process that creates the necessary data structure, known as indices, for mapping keywords (also called indexing terms) to those text docu-ments containing the keywords. In an information retrieval system, it is necessary to index a text document first to enable efficient lookup and retrieval of the text document later. A number of data structures are used in indexing Thai text documents. They include an inverted index, also called inverted file, and suffix array. Inverted index is currently

* *Faculty of Information Technology, Rangsit University.*

regarded as one of the better index data structures for most applications [1].

However, a challenging task for Thai text indexing is extracting the indexing terms, because Thai text documents are non-segmented. Most of the semantic indexing terms are usually carried by nouns [7], although a sentence in natural language text is composed of nouns, pronouns, articles, verbs, adjectives, adverbs, and connectives. In Thai text documents, the extraction of indexing terms becomes a main issue because they cannot be specified automatically from text documents, due to the nature of Thai texts being non-segmented. Although indexing terms can be manually specified by experts, this process is very time consuming and dictionaries can be costly to maintain [7], [8]. As a result, there are many challenges for indexing Thai text documents. Thai text indexing techniques can be divided into two main categories: a language-dependent technique and a language-independent technique. When indexing Thai text documents using a language-dependent technique, a word segmentation technique [9], [10], [11], [12] is generally essential during the pre-processing stage for extracting the indexing terms before an inverted index can be constructed. This technique is known as the word inverted index. The techniques for Thai word segmentation can be broadly classified into four approaches: dictionary based [13], [14], rule based [15], [16], [17], hybrid [18], [19], and machine learning based approaches [5], [20]. These four approaches require complex language analysis and lengthy computational time to segment Thai text documents into indexing terms before constructing an index to allow retrieval. The success of a word segmentation approach relies very much on the language analysis techniques or on the use of an appropriate dictionary or corpus.

On the other hand, n-gram inverted index [21], [22] and suffix array [23], [24] were proposed as the alternative indexing techniques which do not require linguistic knowledge of a language. These techniques are language-independent and most widely used to tackle Asian languages, many of which are un-delimited languages. An n-gram inverted index is the main in-dexing method that has been widely used in many Asian language text retrieval systems [25], [26], [27], [28], [29], [30] such as Chinese, Japanese and Korean (CJK). This is because these Asian languages share similar difficulties in segmenting texts and specifying indexing terms. Therefore, this technique is acknowledged by many Asian researchers as a workable solution to information retrieval problems for many Asian languages [30], [21], [25]. Another approach is by using a data structure called suffix array for indexing Thai text documents [31], [4]. In the suffix array, a given text document is viewed as a se-quence of characters that can be constructed as an array containing character occurrences without indexing term extraction requirements.

Before Thai text indexing techniques can be described in more detail, the linguistic characteristics of the Thai language are first provided in the next section to assist with understanding of the problem.

3. Linguistic of Characteristics of The Thai Language

In the Thai writing system, sentences are formed as a long sequence of characters without word boundaries or sentence separators to delimit words, phrases or sentences. The spaces in the Thai language are usually used to interrupt an idea or to help the reader pay attention to the text, but they do not signify a split between words, phrases or sentences [8]. Additionally, the Thai language has no capital letters to identify proper nouns or starting points of sentences like the English language. Figure 2 shows an example of the Thai language.

ข้อคิดในการใช้ชีวิต เป็นข้อคิดที่เก็บไว้เมื่อนานมาแล้ว แต่เมื่อเอามาอ่านอีกทีก็พบว่ายังเป็นเรื่องที่ดีมาก ๆที่ควรแบ่งปันให้รับรู้โดยทั่วกัน ซึ่งเราสามารถนำข้อคิดเหล่านี้มาเป็นพื้นฐานในการดำเนินชีวิต

Figure 2 *Example of the Thai language.*

In the Thai character set, there are a total of 76 characters, consisting of 44 consonants, 14 vowels, 4 tone marks, 10 Thai digits, and 4 special characters [11].

Jaruskulchai [32] showed that the Thai language has similar grammatical categories to the English language in term of the parts of speech. Thai words can be classified into 14 categories: nouns, pronouns, verbs, auxiliary verbs, determiners, adjec-tives, adverbs, classifiers, prepositions, conjunctions, interrogatives, prefixes, suffixes, and negative verbs.

The Thai writing system exhibits SVO (Subject-Verb-Object) word-order [33], and it reads from left to right and from top to bottom. There are four levels of the appearance of Thai characters: the tone, upper vowel, middle, and lower vowel as shown in Figure 3. These levels can be used to indicate the character types. For instance, the middle level usually contains the consonants but sometimes contains vowels, Thai digits and special characters. The tone marks are always located in the tone level and most vowels are usually placed on the upper vowel and lower vowel levels.

For other characters, they also have their specific positions. Figure 3 shows the position in order from the top level to lower levels of the phrase "ข้อมูลเนื้อสัตว์" (meat information).

Figure 3 *Four levels of appearance of Thai characters.*

According to W. Aroonmanakun [11], most Thai words are defined as a linguistic unit that is usually made up of simple words or monosyllables. However, Thai words can also be a combination of two or more morphemes, these are known as compound words. Therefore, Thai words can be viewed as a combination of syllables, and they can be distinguished into two types of words under Thai grammar rules, which are Simple words and Compound words.

Although the criteria of Thai grammar rules seem to be clear, when looking at the real data, it is not always easy to determine the number of words in a given text. For example, หม้อหุงข้าว (rice cooker) can be analyzed as one compound word, or three simple words หม้อ (pot), หุง (cook), and ข้าว (rice). As a result, the problem of defining the word exists because of compound words that can be formed from the combination of two or more simple words. Because of these problems, Thai language processing is a challenge due to its nature of being a non-segmented language. This has called for the need to research and develop effective Thai word segmentation and Thai text indexing techniques.

4. Thai Text Indexing Techniques

In the area of Thai information retrieval, many techniques have been proposed for solving the problems of extraction of Thai indexing terms and producing Thai text indexing algorithms [34]. The existing indexing techniques used in Thai information retrieval can be divided into two main categories: a language-dependent technique and language-independent technique, and they are subdivided into three approaches: the word inverted index, the n-gram inverted index and suffix array approach as shown in Figure 4.

4.1 Using Language-Dependent Technique for Thai Text Indexing

In using a language-dependent technique, the prominent approach is to use the more widely adopted solution called the word inverted index [1], [4], [24], [35]. The word inverted index can be viewed as a word based approach which has been shown to work effectively for segmented languages

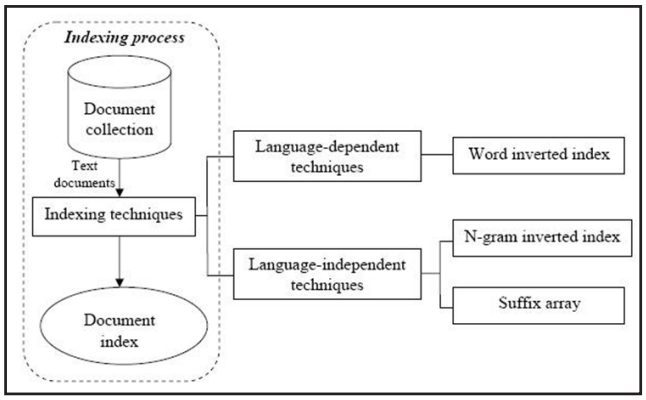

Figure 4 *Existing Thai text indexing techniques.*

such as English or other European languages. The word inverted index is a popular and important index data structure used in the area of information retrieval. It is used to speed up the process of building the indexing structure and it is efficient at the retrieval of text documents containing a query. This indexing structure usually collects indexing terms from text documents that describe the contents of the documents. Therefore, when using the word inverted index for Thai text indexing, Thai text documents need to be parsed and tokenized into individual words. The word segmentation technique is generally an essential part in performing the indexing term extraction, before the word inverted index can be constructed. The techniques for Thai word segmentation can be sorted to Dictionary based [13], [14], Rule based [15], [16], [17], Hybrid [18], [19] and Machine learning based approaches [5], [20]. Most of these approaches are language-dependent, they rely on language analysis or on the use of dictionary or corpus. Dictionary based methods match each word of the dictionary against the text document and their performance depends on the size and the quality of the dictionary. The morphology of Thai enables to use rule based techniques, but the accuracy then depends again on hand-crafted rules. Hybrid technique is combination of the two approaches: dictionary and rule based techniques. Machine learning techniques use tagged training corpora to build a statistical model able to identify boundaries between words in the text document. Although, this approach does not require the use of dictionary or language analysis, it still needs corpus and its performance depends critically on the characteristics of the document domain and the size of the training corpus, and also the preparation of this approach is time consuming. Additionally, stopword removal is another text processing task, which is also normally performed before constructing the word inverted index, in order to reduce noisy words. In Figure 5, Thai text indexing architecture using the word inverted index

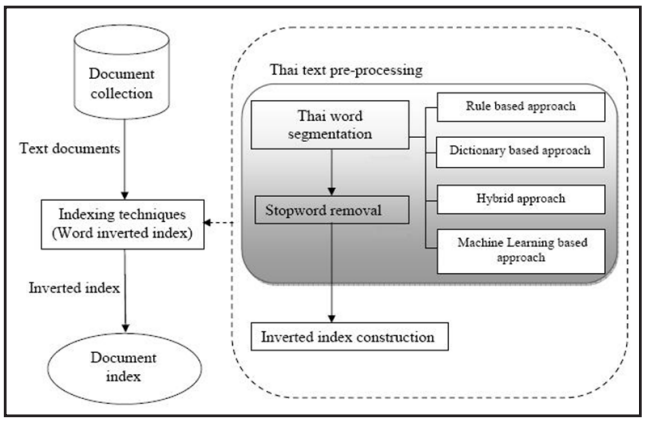

Figure 5 *General process of indexing Thai text documents using word inverted index.*

technique is depicted.

To illustrate the word inverted index technique in more detail, a typical process of the word inverted index approach, which is created on the document d1 containing the strings "การประกอบการ" that means "engage in business" in English, is shown in Figure 6.

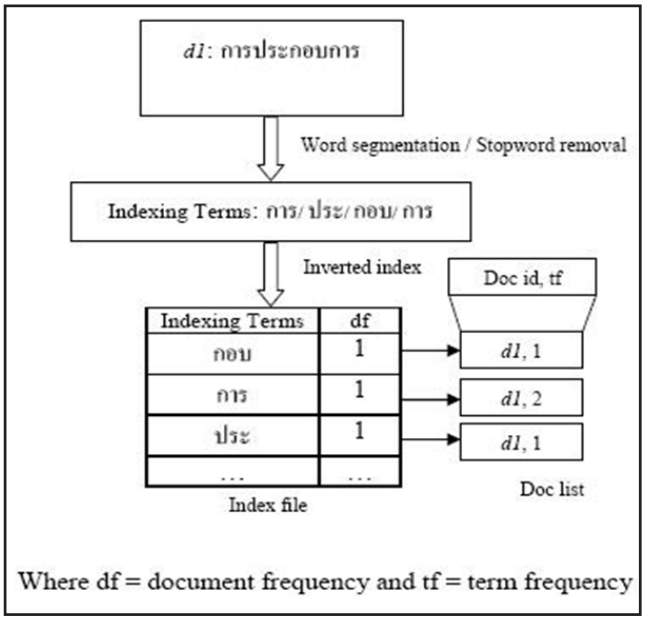

Figure 6 *Example of the word inverted index.*

For Thai information retrieval, after Thai text documents are segmented into a sequence of indexing terms using word seg-mentation and stopwords removal, all tokenized indexing terms are then stored in alphabetical order in the inverted index for fast retrieval. The inverted index is composed of two elements: the vocabulary and posting file. The vocabulary contains the set of all distinct indexing terms that occur in the documents. The posting file contains a list of pointers or indexing term posi-tions where they appear in the text documents. The posting file also consists of the identifier of the document that contains the indexing terms and the list of the offsets where the indexing terms occur in the text document. For each indexing term t, there is a posting list that contains postings <d, f, [o1, ..., of]>, where d is a document identifier (document ID), f is the frequency of the indexing term t in the document d and [o1, ..., of] is a list of offsets o that can refer to indexing term or char-acter positions. The posting file can be used to provide faster and more accurate information retrieval. The following shows the organization of the indexing terms into the inverted index. An example of the vocabulary and posting file of indexing terms, which are created on the document d1 containing the string s "การประกอบการ", can be shown in Figure 7

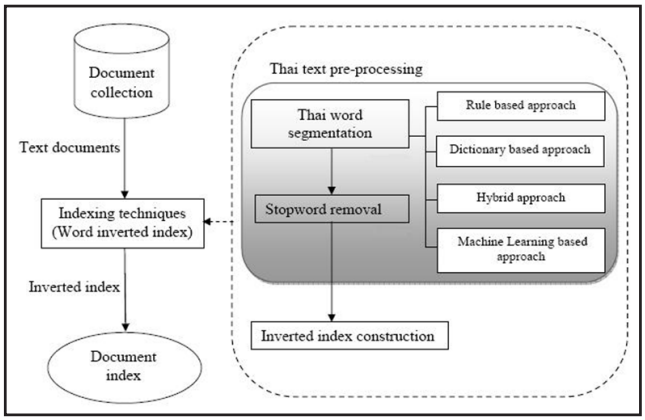

Figure 7 *Example of vocabulary and posting file of document containing the string s "การประกอบการ".*

4.2 Using Language-Independent Technique for Thai Text Indexing

In this section, the details of two approaches for indexing Thai text documents using language-independent techniques: an n-gram inverted index and suffix array techniques are described.

4.2.1 An n-gram inverted index technique

Besides the word inverted index, an n-gram inverted index is one of many indexing techniques that uses n-gram terms as an indexing term [21], [22], [7]. This method can be viewed as an n-gram based approach and was first introduced and tested as indexed terms by Adams in 1991 [36]. The n-gram inverted index is a language-independent approach, which does not require the use of language analysis, or a dictionary or corpus. In information retrieval systems, the n-gram inverted index is often used for Asian languages where extraction of words is not simple [37]. The n-gram inverted index has been acknowledged as a viable solution for indexing non-segmented languages such as Chinese, Japanese and Korean (CJK) [21], [22]. It is also used for other

non-segmented text in the area of bioinformatics [38], [39]. This method has been experimented with in many related IR fields [40], [41]. There are a number of techniques implemented in Chinese, Japanese and Korean information retrieval systems. It is also used as a possible solution for Thai in the same way as Chinese, Japanese and Korean.

Although the Thai language is as linear as other Asian languages such as Chinese or Japanese and the same n-gram inverted index used in Chinese and Japanese can be used for Thai, the parameters for the n-gram method should be adjusted to be appropriate for the Thai language. Determining the dimensions of the gram term for using the n-gram inverted index should be considered so that they are appropriate for each application. Before the detail of the n-gram inverted index is described, the general process of the n-gram inverted indexing technique is first provided in Figure 8.

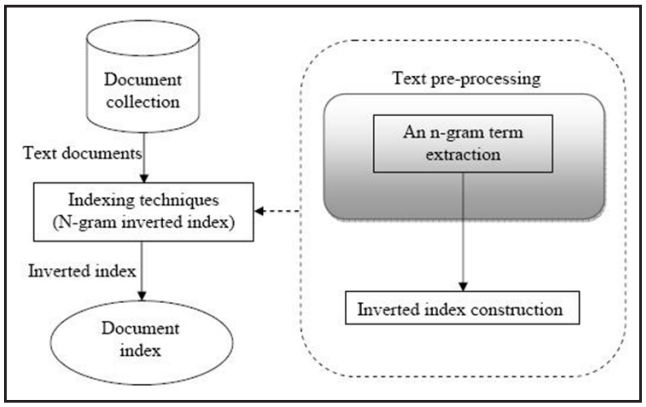

Figure 8 *General process of indexing Thai text documents using n-gram inverted index.*

An n-gram term extraction

In order for the n-gram inverted index to be successful, n-gram term extraction is essential and it has to be done before in-dexing can be performed. Selection of the dimension of the gram term is important for these languages. For instance, it has been shown that the bi-gram term is effective for indexing Chinese text documents [42], [43], [44], [45]. Furthermore, most Chinese bi-gram terms do not lose the semantics of words. In Japanese, the dimension of the gram term had also been found to be equal to two [46]. In bioinformatics, CAFE [38] is a well known method which uses the n-gram base approach. It uses 9-gram terms for the genome sequence and 3-gram terms for the protein sequence as indexing terms.

Meanwhile the n-gram based approach with n equal to three and four characters seems to have the best parameters to achieve retrieval effectively for the Thai language [3], [7]. This is because the top 20 of the high frequencies' 3-gram and 4-gram terms are complete words in the Thai language, where Thai words have varying lengths. As a result, three and four are both used as the best parameters in the n-gram inverted index for the Thai language. In the Thai language, Jaruskulchai [3] showed that the more probable *n* for the Thai language should be greater than two. By selecting the *n* greater than two, one could increase the possibility of achieving the effective retrieval, since the minimum number of characters for Thai word ap-pearance is two, with at least one of them being a consonant. Furthermore, each Thai character cannot represent a word or a meaning like Chinese or Japanese. In Thai, the smallest unit which can represent a word or a meaning is a syllable. Due to the above reasons, there is no single parameter for n-gram that is best for all un-delimited texts and applications. The following paragraphs will describe the process of n-gram term extraction.

Assume that document *d* consists of a string of characters *a1, a2, ..., aN*. An n-gram term is a substring of n overlap or non-overlap successive characters extracted from the string. Extracting a set of n-gram terms from the documents *d* can be done by using the 1-sliding technique [29]. That is, sliding a window of length *n* from a_1 to a_N and storing the characters located in the window. Therefore, the ith n-gram term extracted from document *d* is the substring a_i, a_{i+1}, ..., a_{i+n}. Figure 9 shows 1-gram, 2-gram, 3-gram, 4-gram, …, N-gram overlap sequence of the document d containing the string s "การประกอบการ".

1-gram terms	ก, า, ร, ป, ร, ะ, ก, อ, บ, ก, า, ร
2-gram terms	กา, าร, รป, ปร, ระ, ะก, กอ, อบ, บก, กา, าร
3-gram terms	การ, ารป, รปร, ประ, ระก, ะกอ, กอบ, อบก, บกา, การ
:	
N-gram terms	การประกอบการ

Figure 9 *Sets of 1-gram, 2-gram, 3-gram, ..., N-gram overlap sequence of document d containing the string s "การประกอบการ".*

To construct the n-gram inverted index, the same technique used in the word inverted index construction is employed. Af-ter Thai text documents are segmented into a series of indexing terms or n-gram terms using the n-gram term extraction, all tokenized indexing terms are then stored in alphabetical order in the inverted index.

The advantage of the n-gram inverted index is language-independence [1], [47], [48]. Hence, it was one of the

promising alternatives for indexing Thai text documents [7]. This n-gram inverted index can be used to search for the query where dictionary or language analysis may not be used. There are many applications of the n-gram based approach such as string searching, approximate string matching, and similar sequence matching in bioinformatics [49]. However, this technique suffers from a larger index size and poor retrieval time [47], [7], [29] when compared to the word inverted index. Like the word inverted index, the n-gram inverted index requires query processing and text pre-processing to extract n-gram terms before retrieval and indexing can be performed.

4.2.2 Suffix array approach

Another indexing approach is by using a data structure called suffix array. Within the suffix array scheme, Thai texts are viewed as a sequence of characters which can be structured by using an array. Suffix array approach can be viewed as a full text indexing technique. This technique does not require indexing term extraction like word inverted index and n-gram inverted index approaches. Suffix array approach is one of the more efficient methods for computing terms and document frequency for all substrings (i.e. keywords) from the text. This technique was proposed in 2001 [23] by Yamamoto and Church. The algorithm is based on suffix arrays [50] for computing *tf* (term frequency) and *df* (document frequency). Suffix array can also be used to solve substring problems. Term frequency (*tf*) is the standard notion of frequency in corpus-based natural language processing. It counts the number of times that a type (term-word-n-gram) appears in a text.

Suffix array construction

The suffix array data structure makes it convenient to compute the frequency and locations of a substring in a text. The lexicographical ordering technique is used to group all suffixes together in the suffix array, and can be found efficiently with a search algorithm. This technique constructs a suffix array that contains all suffixes that are sorted alphabetically. A suffix, also known as a semi-infinite string, is a string that starts at position *i* in the text and continues to the end of the text. Therefore, the constructed suffix array shows all possible substrings [23]. The constructed suffix arrays of a string can be used as an index to locate all occurrences of a substring within the string. Finding all occurrences of the substring is equivalent to finding every suffix that begins with the substring. This enables the algorithm to compute the term frequency using overlapping computation. As a result, suffix array can be used to search substrings efficiently. Figure 10 shows an illustration of suffix array from string s = "การประกอบการ"

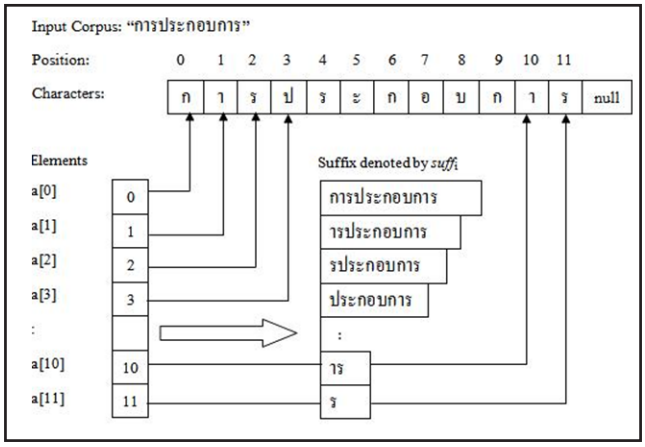

Figure 10 *Illustration of suffix array from string s = "การประกอบการ".*

From Figure 10 the suffixes are enumerated by using suffix array, but elements in the suffix array have not been initialized and sorted. For each element in the suffix array, a[*i*] is an integer denoting a suffix, starting at position *i* in the text and extending to the end of the text. The elements in the suffix array are then sorted in alphabetical order for the next process as shown in Figure 11.

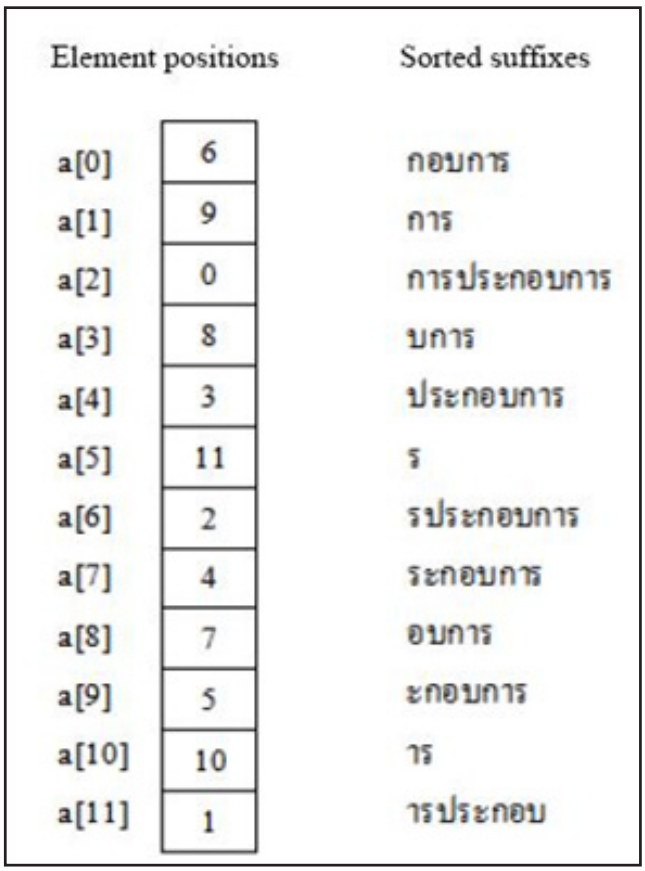

Figure 11 *Illustration of suffix array from Figure 12, which has been sorted in alphabetical order*

Table 1 *Advantages and disadvantages of Thai text indexing techniques*

Thai text indexing techniques	Advantages	Disadvantages
The word inverted index technique	• Requires less storage space for index-ing when compared to the n-gram in-verted index and the suffix array ap-proach	• Requires word segmentation as text pre-processing before indexing • Requires long computation time for indexing when compared to the n-gram inverted index and the suffix array techniques • Requires additional space for storing a dictionary or corpus or manually hand crafted rules. • Can only support one language or application depending on the dictionary or corpus or language knowledge used • Requires word segmentation to perform query processing before searching
The n-gram inverted index technique	• Language-independent technique • Supports any language or application	• Requires n-gram term extraction to perform text pre-processing before indexing • Requires n-gram term extraction to perform query processing before searching • Requires more storage space for indexing when compared to the word inverted index technique
The suffix array technique	• Language-independent technique • Does not require text pre-processing and query processing before indexing and searching • Supports any language or application	• Requires more storage space for indexing when compared to the word inverted index and the n-gram inverted index

5. Comparison of Advantage and Disadvantage of Thai Text Indexing Techniques

While a number of indexing techniques have been proposed for the Thai text documents in order to enhance the perfor-mance of Thai information retrieval, this paper also revealed several underlying limitations of these techniques. In this section, the discussion on advantages and disadvantages of Thai text indexing techniques is provided. In order to compare the Thai text indexing techniques, Table 1 points out the advantages and disadvantages of three indexing tech-niques: the word inverted index, the n-gram inverted index and the suffix array approach as following.

5.1 The word inverted index technique

The word inverted index can be regarded as the more widely used indexing technique in Thai information retrieval. The main advantage of this technique is that it requires less space for indexing and storing the indexing terms when compared to the n-gram inverted index and suffix array approaches. However, indexing Thai text documents using the word inverted index has shown several limitations. The main limitation of the word inverted index is that the process of constructing the word inverted index is very time consuming. This limitation is caused by the need for word segmentation to perform text pre-processing in extracting the indexing terms before the inverted index can be constructed. Most word segmentation approaches require complex language analysis and thus require long computation time, and sometimes require dictionaries or corpora that are costly to maintain. Beside this, the word inverted index also requires additional storage space for storing a dictionary or corpus or manually hand crafted rules to perform word segmentation. Another limitation of the word inverted index is that it is language-dependent. The word inverted index needs knowledge of the language to extract the indexing terms before indexing can be performed. From a search point of view, the limitation of the word inverted index is that this technique requires query processing before the searching process can be performed. To search the word inverted index, it is necessary to apply Thai word segmentation to the query before it is sent to the search process to look up the relevant documents. Note that the word segmentation technique applied to the query has to be the same word segmentation technique used to extract the indexing terms from the text documents.

5.2 The n-gram inverted index technique

The main advantage of the n-gram inverted index is that this technique supports any language or application due to its

being a language-independent technique. Due to this advantage, the n-gram inverted index has been one of the most often used indexing techniques for many Asian documents, and it has also been used in analyzing genome sequences in bioinformatics. However, some limitations of the n-gram inverted index still exist. Its first disadvantage is that the n-gram inverted index requires the indexing term extraction using the n-gram term extraction method before the inverted index can be constructed. Determining the dimensions of the gram term is essential, so that they are appropriate for each application.

Additionally, the n-gram inverted index has limitations in terms of storage space. The n-gram inverted index requires larger space for storing indexing terms when compared to the word inverted index, because the number of indexing terms extracted by the n-gram inverted index is usually more than the number of indexing terms extracted by the word inverted index. Furthermore, from a search point of view, a limitation of the n-gram inverted index is that it requires query processing in extracting the n-gram terms from the query before searching can be performed.

5.3 The suffix array technique

The suffix array approach is one of the language-independent techniques, which do not require the use of a dictionary or corpus or grammatical knowledge of a language. Due to being language-independent, the main advantage of the suffix array approach is that it is applicable for any language or application. The suffix array also does not require text pre-processing and query processing before indexing and searching can be performed. However, the main limitation of this approach is that it requires more storage space for indexing when compared to the word inverted index and the n-gram inverted index. One of the drawbacks is that storage space, in terms of the index size of the suffix array technique, could be critical. Since the size of electronically stored information in the Thai language has grown exponentially, the method of suffix array may not be practical for use in some applications.

6. Conclusion

In conclusion, research relating to Thai text indexing has been provided in this paper. In Thai text indexing, the methodologies can be categorized into two main techniques: language-dependent methods and language-independent methods. For language-dependent methods, the word inverted index technique was discussed. For language-independent methods, the n-gram inverted index and suffix array approaches were described. The disadvantages of the existing indexing techniques were examined. In the word invert index, word segmentation is required to extract the indexing terms before the inverted index can be constructed. However, most word segmentation approaches require complex language analysis and long computation time, and sometimes dictionaries or corpora that are costly maintain. Success of the word inverted index relies on the accuracy of word segmentation. The word inverted index also needs knowledge of an individual language in terms of extracting indexing terms, due to being language-dependent. While the n-gram inverted index is language-independent, it still requires indexing term extraction using the n-gram term extraction method before the inverted index can be constructed. Although the n-gram inverted index can be applied to many Asian languages and other sequence patterns due to its being language-independent, determining the appropriate dimensions of the gram term is problematic. This method also requires more space for storing indexing terms when compared to the word inverted index. Regarding the suffix array approach, this refers to a language-independent technique that can be applied to any language and other sequence patterns. However, one of its drawbacks is that this method obviously requires a large amount of storage space for indexing because it generates and keeps all suffixes from text documents during the indexing process. Although the suffix array approach does not require text pre-processing in terms of extracting the indexing terms before the suffix array can be constructed, one of the drawbacks in terms of index size seems to be very critical. This makes the suffix array approach impractical at times to be used in the Thai environment, as the amount of the Thai language that is electronically stored has grown exponentially.

7. References

[1] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. New York: ACM Press, 1999.

[2] A. Califano and I. Rigoutsos. "FLASH: A Fast Look-Up Algorithm for String Homology." *In Proceedings of the 1st International Conference on Intelligent Systems for Molecular Biology*, Bethesda, Maryland, 1993.

[3] C. Jaruskulchai. "Thai Text Segmentation: Problems and Potential Solutions." *In the Sixth Annual Workshop on Science and Technology Exchange between Thai Professionals in North America and Thailand*, Edmonton,

Alberta, Canada, 1996.

[4] C. Haruechaiyasak, C. Damrongrat, C. Sangkeettrakarn, S. Kongyoung, and N. Angkawattanawit. "Sansarn Look!: A Platform for Developing Thai-Language Information Retrieval Systems." *In 21st International Technical Conference on Circuits/Systems*, Computers and Communications, Chiang Mai, Thailand, 2006.

[5] C. Haruechaiyasak, S. Kongyoung, and C. Damrongrat. "LearnLexTo: A Machine-Learning Based Word Segmentation for Indexing Thai Texts." *In ACM 17th Conference on Information and Knowledge Management*, 2008.

[6] W. Kanlayanawat and S. Prasitjutrakul. "Automatic Indexing for Thai Text with Unknown Words Using Trie Structure." *In Proceeding of NLP Pacific Rim Symposium*, 1997, vol. 14, no. 2, pp. 153-172, 2001.

[7] J. Chuleerat. "An Automatic Indexing for Thai Text Retrieval." PhD thesis, George Washington University, USA, 1998.

[8] C. Jaruskulchai and C. Kruengkrai. "A practical text summarizer by paragraph extraction for Thai." *In Proceedings of the Sixth International Workshop on Information Retrieval with Asian Languages*, Sappro, Japan, pp. 9-16, 2003.

[9] C. Haruechaiyasak, S. Kongyoung, and M. N. Dailey. "A Comparative Study on Thai Word Segmentation Approaches." *In Proceedings of Electrical Engineering/ Electronics, Computer, Telecommunications, and Information Technology*, 2008.

[10] S. Meknavin, P. Charoenpornsawat, and B. Kijsirikul. "Feature-based Thai Word Segmentation." *In Proceedings of the Natural Language Processing Pacific Rim Symposium (NLPRS'97)*, Phuket, Thailand 1997.

[11] W. Aroonmanakun. "Collocation and Thai word segmentation." *In Proceedings of the 5th SNLP & 5th Oriental COCOSDA Workshop*, pp. 68-75, 2002.

[12] S. Luksaneeyanawin. "A Thai Text to Speech System." *In Proceedings of the Conference on Electronics and Computer Research and Development*, 1992.

[13] V. Sornlertlamvanich. "Word Segmentation for Thai in Machine Translation System." Bangkok.

[14] Y. Poovorawan. "Dictionary-based Thai Syllable Segmentation (in Thai)." *In 9th Electrical Engineering Conference*, Bangkok, 1986.

[15] Y. Thairatananond. "Towards the Design of a Thai Text Syllable Analyzer." Asian Institute of Technology, 1981.

[16] S. Charnyapornpong. "A Thai Syllable Separation Algorithm." Asian Institute of Technology, 1983.

[17] T. Theeramunkong, V. Sornlertlamvanich, T. Tanhermhong, and W. Chinnan. "Character-Cluster Based Thai Information Retrieval." *In Proceedings of the Fifth International Workshop on Information Retrieval with Asian Languages,* Hong Kong, pp. 75-80, 2000.

[18] K. Asanee, T. Chalathip, and S. Sapon. "A Statistical Approach to Thai Word Filtering." *In Proceedings SNLP'95, the 2nd Symposium on Natural Language Processing*, Bangkok, Thailand, August 2-4, pp. 398-406, 1995.

[19] V. Ruttikorn, S. Waraporn, J. Somsak, and T. Sakchai. "An Analysis on Correct Sentence Selection by Word's General Usage Frequency." *In Natural Language Processing: Multi-lingual Machine Translation and Related Topics*, pp. 291-300, 1994.

[20] C. Kruengkrai and H. Isahara. "A Conditional Random Field Framework for Thai Morphological Analysis." *In Proceeding of the Fifth International Conference on Language Resources and Evaluation (LREC-2006),* 2006.

[21] P. Majumder, M. Mitra, and B. B. Chaudhuri. "N-Gram: A Language Independent Approach to IR and NLP." *In International Conference on Universal Knowledge*, 2002.

[22] W. Cavnar and J. Trenkle. "N-Gram Based Text Categorization." *In Proceedings of the 3rd Annual Symposium on Document Analysis and Information Retrieval*, Las Vegas, 1994, pp. 161-175.

[23] M. Yamamoto and K. W. Church. "Using Suffix Arrays to Compute Term Frequency and Document Frequency for All Substrings in a Corpus." *Computational Linguistics*, Vol. 27, pp. 1-30, 2001.

[24] S. J. Puglisi, W. F. Smyth, and A. Turpin. "Inverted Files Versus Suffix Arrays for Locating Patterns in Primary Memory." *In SPIRE 2006*, pp. 122-133, 2006.

[25] J. H. Lee and J. S. Ahn. "Using n-Grams for Korean Text Retrieval." *In Proceedings of the19th Annual International Conference on Information Retrieval, ACM SIGIR*, Zurich, Switzerland, 1996, pp. 216-224.

[26] K. L. Kwok. "Comparing Representations in Chinese Information Retrieval." *In Proceedings of the 20th Annual International ACM SIGIR Conference on*

Research and Development in Information Retrieval, Philadelphia, USA, pp. 34-41, 1997.

[27] H. Fujii and W. B. Croft. "A Comparison of Indexing Techniques for Japanese Text Retrieval." *In Proceedings of ACM SIGIR 16th Annual International Conference on Research and Development in Information Retrieval*, pp. 237-246, 1993.

[28] P. McNamee. "Knowledge-Light Asian Language." *In Proceedings of the Third NTCIR Workshop on Research in Information Retrieval, Automatic Text Summarization and Question Answering Text Retrieval, at the NTCIR-3. Workshop*, 2002.

[29] M. S. Kim, K. Y. Whang, J. G. Lee, and M. J. Lee. "n-Gram/2L: A Space and Time Efficient Two-Level n-Gram Inverted Index Structure." *In VLDB, Trondheim*, Norway, pp. 325-336, 2005.

[30] M. M. Hasan and Y. Matsumoto. "Chinese-Japanese Cross Language Information Retrieval: A Han Character Based Approach." *In Proceedings of the SIGLEX Workshop on Word Senses and Multi-linguality, ACL-2000*, Hong Kong, pp. 19-26, 2000.

[31] P. Srichaivattana. "Dictionary-Less Search Engine for the Collaborative Database." *In Proceeding of the 3rd International Symposium on Communications and Information Technologies*, September, 2003.

[32] C. Jaruskulchai. "Thai Text Segmentation: Problems and Potential Solutions." *In The Sixth Annual Workshop on Science and Technology Exchange between Thai Professionals in North America and Thailand*, Edmonton, Alberta, Canada, 1996.

[33] R. Pankhuenkhat. "Thai Linguistics": Chulalongkorn, 1998.

[34] J. Chuleerat, *Dictionary-Based Thai CLIR: An Experimental Survey of Thai CLIR*, Vol. 2406/2002: Springer Berlin/Heidelberg, 2002.

[35] A. Kawtrakul, C. Thumkanon, and P. McFetridge. "Automatic Multilevel Indexing for Thai Text Information Retrieval." *In IEEE Asia Pacific Conference on Circuits and Systems*, 1998.

[36] E. Adams. "A Study of Trigrams and Their Feasibility as Index Terms in a Full Text Information Retrieval System." PhD thesis, George Washington University, USA, 1991.

[37] Y. Ogawa and T. Matsudua. "Optimizing query evaluation in n-gram indexing." *In Proceedings of International Conference on Information Retrieval*, ACM SIGIR, Melbourne, Australia, pp. 367-368, 1998.

[38] H. E. Williams and J. Zobel, "Indexing and Retrieval for Genomic Databases." *In IEEE Transaction on Knowledge and Data Engineering*, pp. 63-78, 2002.

[39] H. E. Williams. "Genomic Information Retrieval." *In Proceedings of the 14th Australasian Database Conferences*, 2003.

[40] J. D. Cohen. "Highlights: Language - and Domain-Independent Automatic Indexing Terms for Abstracting." *Journal of The American Society for Information Science*, Vol. 46, No. 3, pp. 162-174, 1995.

[41] W. B. Cavnar. "Using an N-Gram-Based Document Representation with a Vector Processing Retrieval Model." *In Proceedings of the Third Text REtrieval (TREC-3)*, pp. 269-277, 1994.

[42] L. F. Chien."Fast and Quasi-Natural Language Search for Gigabytes of Chinese Texts." *In Proceedings of 18th ACM SIGIR Conference on Research and Development in Information Retrieval*, New York, USA, pp. 112-120, 1995.

[43] T. Liang, S. Y. Lee, and W. P. Yang. "Optimal Weight Assignment for a Chinese Signature File." *Journal of Information Processing and Management*, Vol. 32, No. 2, pp. 227-237, 1996.

[44] Y. T. Lin, Chinese English Dictionary of Modern Usage. Hong Kong: Chinese University of Hong Kong Press, 1972.

[45] H. Jiao, Q. Liu, and H.-b. Jia. "Chinese Keyword Extraction Based on N-Gram and Word Co-Occurrence." *In Proceedings of 2007 International Conference on Computational Intelligence and Security Workshops (CISW 2007)* China, 2007.

[46] Y. Ogawa, A. Bessho, M. Iwasaki, M. Nishimura, and M. Hirose. "A New Indexing and Text Ranking Method for Japanese Text Databases Using Simple-Word Compounds as Keywords." *In Proceedings of the Third International Symposium on Database Systems for Advanced Applications*, pp. 197-204, 1993.

[47] J. Mayfield and P. McNamee. "Single N-Gram Stemming." *In Proceedings of the 26th Annual International Conference on Information Retrieval*, ACM SIGIR, Toronto, Canada, pp. 415-416, 2003.

[48] E. Miller, D. Shen, J. Liu, and C. Nicholas. "Performance and Scalability of a Large-Scale N-Gram Based Information Retrieval System." *Journal of Digital Information*, Vol. 1, No. 5, pp. 1-25, 2000.

[49] J. D. Cohen. "Recursive Hashing Functions for n-Grams." *ACM Transactions on Information Systems*, Vol. 15, No. 3, pp. 291-320, July, 1997.

[50] U. Manber and G. Myers. "Suffix Arrays: A New Method for On-Line String Searches." *In the First Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 319-327, 1990.