AUTOMATIC IDENTIFICATION OF SPECIFIC WEB DOCUMENTS BY USING CENTROID TECHNIQUE

Udomsit Sukakanya, Kriengkrai Porkaew
School of Information Technology, King Mongkut's University of Technology Thonburi, Thungkru, Bangkok, Thailand

Keywords: Search, Web Mining, Centroid, Identification, Classification

Abstract: In order to reduce time to find specific information from high volume of information on the Web, this paper proposes the implementation of an automatic identification of specific Web documents by using centroid technique. The Initial training sets in this experiment are 4113 Thai e-Commerce Web documents. After training process, the system gets a Centroid e-Commerce vector. In order to evaluate the system, six test sets were taken under consideration. In each test set has 100 Web pages both known e-Commerce and non e-Commerce Web pages. The average system performance is about 90 %.

1 INTRODUCTION

Nowadays, the World Wide Web is the largest source of information available almost freely to the world. People throughout the world contribute information seen here. This makes it diverse, dynamic and ever changing. Most Internet users find the information on the web by using search engines. However, most search engines are based on the same query-list paradigm. The user types a list of words or keywords, and the system returns a ranked list of documents, which matches that query. This approach is acceptable since the user can refine his query to narrow the search space; however it is proven that most Internet users use short, generic queries of 1-3 words (Bharat and Henzinger, 1998), which usually produce millions of matching documents in the result list. Usually, users give up their search completely if they don't get a good result after a few tries of refining their queries. Designers of search engines have been struggling with this problem for years with better or worse results.

In Thailand, the latest status of Thai Web documents, based on databases crawling on June 2003 (Surasak and Kasom,2003) is summarized by NontriSpider. They focus on web servers whose names must be registered only under ".th" and they found that there are 3,589,961 HTML documents on 26,826 Web servers (all servers are categorized into 7 major domains: .ac.th, .co.th, .go.th, .in.th, .mi.th, .net.th and .or.th). In fact, the actual number of pages and servers should be bigger than this because there are still many web servers behind firewalls which are not intended for public access and some servers register the hostname under foreign domain names (.com, .org, .net, etc.). From the large size of Web documents in Thailand, it is difficult to identify specific Web documents manually. This paper tries to fill the gaps mentioned above by using the suitable algorithms to implement and evaluate an automatic identification system working for Thai specific Web documents.

The remainder of this paper is organized as follows. Section 2 provides literature review. Section 3 provides the proposed system. Section 4 provides the details in experiments and system performance evaluation. Finally, Section 5 provides the conclusion and future work.

2 LITERATURE REVIEW

Numerous methods for improving the quality of the ranked list presentation have been proposed from simple filtering techniques to complex algorithms employing Artificial Intelligence techniques. However, due to the rapidly growing and unstable characteristic of the Web, such directories very often point to outdated, even not existing documents. Data classification, one of the search techniques, is the process which finds the common properties among a set of objects and classifies them into different classes. The objective of classification is to analyze

Sukakanya U. and Porkaew K. (2005).
AUTOMATIC IDENTIFICATION OF SPECIFIC WEB DOCUMENTS BY USING CENTROID TECHNIQUE.
In *Proceedings of the First International Conference on Web Information Systems and Technologies*, pages 333-338
DOI: 10.5220/0001234903330338

the input data and to develop an accurate model for each class using the features presented in the data. The class descriptions are used to classify future test data for which the class labels are unknown. Web document classification is an attempt to merge the quality and user-friendliness of directories with the popular ranked list presentation. By classification the results, increasing their readability by showing thematic groups, instead of a mixture of documents on all possible subjects matching the query.

Web Mining is an important field that aims to make good use of the information available on the web and find the data that was either previously unknown or hidden. An important step in the mining process is information retrieval and extraction. The retrieval and extraction methods differ in what aspect of a document is used in extraction information (Lan and Bing, 2003). In general there are two schools of thought; natural language processing techniques and techniques that use the structure of the web. Natural language processing techniques involve using the data of the web using string manipulation. Structural methods build a structure from the structure of the document itself. The research in web mining also derives from the research in other fields like natural language processing, artificial intelligence and machine learning. The techniques that are dealt in these fields mostly deal with a subset of the web pages. Efforts to combine the content and structure of a web page to build a model that is suitable for mining a wide variety of web documents are few and certainly insufficient.

2.1 Centroid Technique

In centroid-based classification algorithm, the Web documents are represented using the vector-space model (Salton, 1989) (Raghavan and Wong, 1986). In this model, each Web document is considered to be the term-frequency vector as following equation.

$$\vec{W}_{tf} = (tf_1, tf_2, ...tf_n)$$

where $\vec{W}_{tf}$ *is Web document vector*

tf_i *is the frequency of the ith term* (1)

A widely used refinement to this model is to weigh each term based on its inverse document frequency (IDF) in the Web document collection (Salton, Wong, and Yang, 1975). The motivation behind this weighting is that terms appearing frequently in many Web documents have limited discrimination power, and for this reason they need to be de-emphasized. This is commonly done by multiplying the frequency of each term i by $\log(N/df_i)$, This leads to the tf-idf representation of the Web document as equation 2 .

$$\vec{W}_{tfidf} = (tf_1 \log(N/df_1), tf_2 \log(N/df_2), ...tf_n \log(N/df_n)) \quad (2)$$

where df_i *is the number of documents that contain the ith term*

N is the total number of documents in training set

In order to account for documents of different lengths, the length of each Web document vector is normalized so that it is of unit length. Given a set N of N Web documents and their corresponding vector representations, the centroid vector (Han and Karypis, 2000) is described as equation 3.

$$\vec{W}_C = \frac{1}{N} \sum_{W_{tfidf} \in N} \vec{W}_{tfidf} \quad (3)$$

Equation 3 is nothing more than the vector obtained by averaging the weights of the various terms presented in N Web documents. N is referred as the supporting set for the centroid. In the vector-space model, the similarity between two Web documents $\vec{W}_i$ and $\vec{W}_j$ is commonly measured using the cosine function, given by equation 4

$$\cos(\vec{W}_i, \vec{W}_j) = \frac{\vec{W}_i . \vec{W}_j}{\left\|\vec{W}_i\right\|^2 * \left\|\vec{W}_j\right\|^2} \quad (4)$$

where "·" denotes the dot - product of the two vectors

The advantage of the summarization performed by the centroid vectors is that the computational complexity of the learning phase of this centroid-based classifier is linear on the number of Web documents and the number of terms in the training set. Moreover, the amount of time required to classify a new Web document *x* is at most *O(km)*, where *k* is the number of centroids and *m* is the number of terms present in *x* .

2.2 Web Document Indexing

In order to reduce the complexity of the Web documents and make them easier to handle, they have to be transformed to the vectors. The vector space model procedures can be divided in to three steps. The first step is content extraction where content bearing terms are extracted from each Web page. The second step is term weighting to enhance retrieval of Web document relevant to the user. The last step ranks the Web document with respect to the query according to similarity measure.

Web document indexing is based on term frequency, which has both high and low frequency. In practice, term frequency has been difficult to implement in automatic indexing. However, the use of a stop list which holds common words to remove high frequency words (stop words), which makes the indexing method language dependent. In general, 40-50% of the total numbers of words in a Web document are removed with the help of a stop list.

Non linguistic methods for indexing have also been implemented. Probabilistic indexing is based on the assumption that there is some statistical difference in the distribution of content bearing words and stop words. Probabilistic indexing ranks the terms according to the term frequency in the whole collection. The stop words are modeled by a Poisson distribution over all Web documents, as content bearing terms cannot be modeled. The use of Poisson model is expanded to Bernoulli model. Recently, an automatic indexing method which uses serial clustering of words in text has been introduced. The value of such clustering is an indicator if the word is content bearing.

3 THE PROPOSED SYSTEM

In order to identify any type of Web documents, a large amount of Web pages were collected manually. This is a difficult task since it required a precise search over the Internet. This system will focus on Thai e-Commerce Web documents. The sample have to be representative, consistent and rather large in order to form the Thai e-Commerce Vector sufficiently. The content of each Web page is classified to the meta-tags, to some special tags and to disseminated plain text. Figure 1 presents the details in the Thai e-commerce identification system. The main processing steps in this system involve:

Training process

1. Define the Thai e-Commerce Web Document specifications. (Table 1)
2. Set the Thai e-Commerce Web document training set. (Section 3.1)
3. Execute English and Thai word segmentation. (Section 3.2)
4. Create the Thai e-Commerce Vector then do normalization (Section 2.1).
5. Find the Centroid of all vectors. (Section 2.1)
6. Remove stop words or common words then repeat step 4 to 6

Testing process

7. Test and evaluate the system. (Section 4)

3.1 Web Document Collection

For the purpose of information identification in the Thai e-Commerce Web documents and with a view to facilitate the implementation of the identification system, a more formal definition of Thai e-Commerce is followed. It helps to decide cleary which “Thai electronic business activities” fall into a well-defined, quantifiable framework of Thai e-Commerce. Thus, an implicit distinction of Thai e-Commerce from the structures of the traditional commerce is required, since new ways are invented in order to measure the revolutionary elements and methods. In an attempt to investigate and measure the ways Thai e-Commerce has fundamentally changed the way of transactions, a sharp criterion in distinguishing the Thai e-Commerce pages is taken under consideration.

Therefore, following the concepts of the Business Media Framework (BMF) (Klose, Lechner and Ulrike, 1999) and Thai e-Commerce properties, a Thai e-Commerce model can be analyzed into a series of concurrent processes, while these processes are implemented by elementary transactions. According to the accepted transaction cost theory, four phases distinguish an e-Commerce transaction. The knowledge phase, the intention phase, the contract phase and the settlement phase. The distinction of these four phases serves as an analytical tool in order to identify Thai e-Commerce pages. Table 1 presents Thai e-Commerce Web Document Training Sets that correspond to these phases respectively. The initial Thai e-Commerce Web document collection in this system is 4113 Web pages. However, it must be noted that some Web pages were categorized into more than one type, during the first review. To avoid any potential misidentification, these Web pages were once more examined in order to be finally categorized into the most appropriate type.

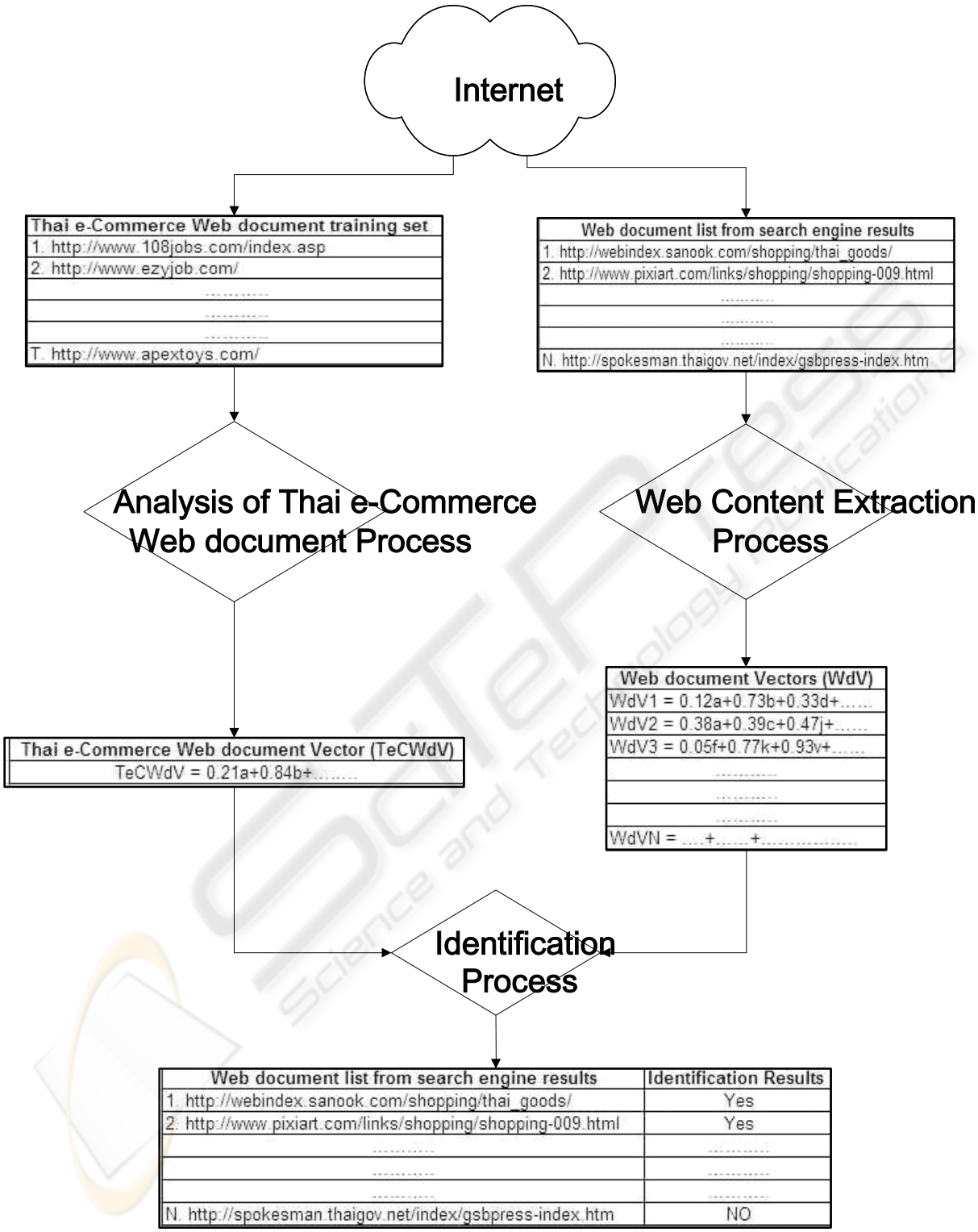

Figure 1: Thai e-commerce identification system

Table 1: Web Document Training Sets

Thai e-Commerce Web Document Training Set
1. Product detail, Categories, Homepage
2. Order detail, Order form
3. Payment detail
4. Delivery detail

3.2 Word Segmentation

Extracting the content of Web document in this system, the problem found is Thai word segmentation. Thai writing system has some minor similarities to English; however, it is more similar to Chinese and Japanese scriptures due to its possession of the same system of sequential continuances. It is done as a sequence of continuous symbols, without word boundaries or punctuation marks to indicate words, phrases, or sentences. Word building in Thai writing system is based on phonetic rules. The smallest linguistic unit in Thai is called a monosyllable. Most Thai words are monosyllable. However, Some Thai words can also be formed by combining two or more words. SWATH algorithm (Meknavin, Charoenpornsawat, and Kijsirikul, 1997) is used to perform Thai word segmentation.

3.3 The Identification Threshold

This system uses a relevance threshold in the similarity value between Web page vectors and Centroid vector as explained in Section 2.1. This threshold is considered to be the minimum value, which will define Thai e-Commerce Web documents. In other words, Web pages with score above the threshold are considered to be Thai e-Commerce web page, while those with lower scoring values are not.

In the proposed work, 95% of the selected Thai e-Commerce Web pages presented a scoring value above 0.10. The rest of the Web documents presented lower scores. Thus, the threshold value for Thai e-Commerce Web documents identification in the proposed system was set to 0.10.

4 SYSTEM PERFORMANCE

After the training process (Section 3.1 to 3.3), the system get a Centroid vector of all Thai e-Commerce Web document vectors. In order to evaluate the system performance in Thai e-Commerce Web document identification, 6 test sets (600 Web pages- 300 known e-Commerce Web pages and 300 non e-Commerce Web pages) were taken under consideration.

The corresponding e-Commerce identification performance for each test is presented in Table 2. The average system performance for Thai e-Commerce Web documents is about 90%.

5 CONCLUSION

This paper proposes the implementation of a system and experiment, which identify Thai e-Commerce Web documents. The test Web pages are considered as vectors and are compared with Thai e-commerce vector under the centroid algorithm for knowledge representation. The similarity metric used for comparing these vectors is the cosine coefficient. The average system performance is about 90%.

Future work involves another specific Web document types such as government, education, and others. Moreover, Thai e-Commerce Web document identification can be recommended in the experiment of another algorithm such as Self Organized Map (SOM).

Table 2: Thai e-Commerce Web document identification performance

Thai e-Commerce Web document identification performance						
Test	e-commerce Web documents		%	Non e-commerce Web documents		%
	Total	Success		Total	Success	
1	0	-	-	100	95	95.00
2	20	18	90.00	80	78	97.50
3	40	37	92.50	60	59	98.33
4	60	53	88.33	40	40	100.00
5	80	72	90.00	20	19	95.00
6	100	92	92.00	0	-	-
Average			**90.57**	**Average**		**97.17**

REFERENCES

Bharat, K., Henzinger, M., 1998. Improved Algorithms for Topic Distillation in a Hyperlinked Environment. In *Proceedings of the 21st ACM SIGIR Conference on Research and Development in Information Retrieval*.

Han, E.H., Karypis, G., 2000. Centroid-Based Document Classification: Analysis & Experimental Results, *Technical Report TR-00-017*, Department of Computer Science, University of Minnesota, Minneapolis.

Klose, M., Lechner and Ulrike., 1999. Design of Business Media- An Integrated Model of Electronic Commerce, *Proceeding of the fifth America Conference on Information Systems (AMCIS'99)*, Milwaukee, WI.

Lan, Y., Bing, L.,2003. Web Page Cleaning for Web Mining through Feature Weighting. In *Proceedings of Eighteenth International Joint Conference on Artificial Intelligence (IJCAI-03)*, Acapulco, Mexico.

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

Raghavan, V.V., Wong, S.K.M.,1986. A Critical Analysis of the Vector Space Model for Information Retrieval, *Journal of the American Society for Information Science (JASIS)*.

Salton, G., 1989. Automatic Text Processing: *The Transformation, Analysis, and Retrieval of Information by Computer*. Addison-Wesley, 1989.

Salton, G., Wong, A., and Yang, C.S., 1975. A Vector Space Model for Automatic Indexing, *Communication of the ACM*.

Surasak, S., Kasom, K., 2003 Structure Properties of the Thai WWW: The 2003 Survey, *The Conference on Internet Technology (CIT2003), Asian Institute of Technology*., Thailand.