Journal of Computer Science 7 (7): 1088-1093, 2011
ISSN 1549-3636

A Comparative Analysis of the Entropy and Transition Point Approach in Representing Index Terms of Literary Text

[1]Hayati Abd Rahman and [2]Shahrul Azman Noah
[1]Department of Computer Science, Faculty of Computer and Mathematical Sciences,
MARA University of Technology, 40450 Shah Alam, Malaysia
[2]School of Information Technology, Faculty of Information Science and Technology,
University Kebangsaan Malaysia, 43600 Bangi, Malaysia

Abstract: Problem statement: Concept hierarchy is a hierarchically organized collection of domain concepts. It is particularly useful in many applications such as information retrieval, document browsing and document classification. **Approach:** One of the important tasks in construction of concept hierarchy is identification of suitable terms with appropriate size of domain vocabulary. **Results:** One way of achieving such a size is by using term reduction. The aim of this study is to examine the effectiveness of reduction approach to reduce size of vocabulary using term selection methods for literary text. The experiment compares entropy method, transition point method and hybrid of transition point and entropy methods with the Vector Space Model (VSM). **Conclusion/Recommendations:** Results indicate the effectiveness of Transition Point method as compared to the others in reducing size of vocabulary but at same time preserve those important terms that exist in the literary documents.

Key words: Information retrieval, term reduction, concept hierarchy, Dominating Set Problem (DSP), Vector Space Model (VSM), Transition Point (TP)

INTRODUCTION

Concept hierarchy is a hierarchically organized collection of domain concepts. It has been used for organizing, summarizing and accessing to information and it is particularly useful in many applications such as information retrieval, document browsing and document classifications. Many research has embarked on the utilization of concept hierarchies; example are Lawrie *et al*. (2001); Lawrie and Croft (2003) and Sieg *et al*. (2004) which have benefited the topical elements that formed the hierarchy as compared to ranked list type. However, varieties of techniques have tested at different parts of the system in order to improve the concept hierarchy. Term selection is the most important part that needs to be considered during the development and has been support by Schultz (2003). He said that in the information retrieval domain documents are often abbreviated to their most salient terms in order to reduce storage requirements and processing time and also to make algorithms more efficient.

Generally, concept hierarchies provide ways to definite the terms that are available to describe the documents' concepts. The purpose is to automatically produce a better and richer collection of terms in a vocabulary. The main and important task in the construction of concept hierarchy is the identification of suitable terms with appropriate size of domain vocabulary. Sieg *et al*. (2004) has taken this matter with their own way. According to them, the effectiveness of web searching always distracted by the ambiguous problem and the inability of users to give appropriate query for searching. They had proposed Adaptive Retrieval based on Concept Hierarchies (ARCH), which able to enrich the set vocabulary using user profile and formulate for query based on relevance feedback. Lawrie *et al*. (2001) on the hand use probabilistic language model approach which simplified the summarization of the documents automatically and use it for exploring the generating of topic hierarchies using Dominating Set Problem (DSP) approach.

Salton and McGill (1983) noted that TF-IDF has been chosen as the basic method for term weighting (Baeza-Yates and Ribeiro-Neto, 1999). Terms with score above certain threshold will be given priority to

Corresponding Author: Shahrul Azman Noah, Department of Information Science,
Faculty of Information Science and Technology, University Kebangsaan Malaysia,
43650 Selangor, Malaysia Tel: +60389216178 Fax: +60389256732

be selected into the vocabulary set and this determine the accessible of document or information.

Massive literary texts are currently available and accessible in digital form. Eagleton (1996) mentioned that literary text can be regarded as a written work with peculiar use of language. An old poetry, classic English literature such as Shakespeare's novels, a collection of ancient writings such as the Bible can be considered as literary documents. In this research, the Quran has been chosen due to its availability and read by many Muslims in this country. The original Quran was written in Arabic and has been translated to other languages. We focused on the English translation of

Yusof Ali (1934) which was first published in 1938 and is the most popular English translation of the Quran. Based on the definition given by Moens (2000), term 'representation' or 'representative' are used for naming the condensed characteristic of the content. A good representative of terms in documents is one of the goals for building concept hierarchy. Rojas *et al*. (2007) highlighted that for unsupervised type of document, term selection is an important process for natural language processing tasks. So, this study will focus on term selection and term reduction, which are necessary task to extract representative concepts. Here, we report an experiment that has been carried on the Quran in order to assess the effectiveness of a number of selected term reduction methods mainly are the entropy method, the transition point method and the combination of both methods.

MATERIALS AND METHODS

Term reduction approach: The Vector Space Model is considered as a standard method for representing text separately which incorporate document indexing containing all term appears in each document. A quality of good documents is measured by the selected terms that are able to represent their documents. However, not all the indexed term are "necessarily used" to represent documents. Some can be considered as 'distortion' in their vocabulary.

Salton *et al*. (1975) has discussed in their study about discrimination value model in document indexing. The discrimination value has been defined as a measurement given to the term so that it able to increase the differences among document vector when assigning an index term to a given collection of documents. A high discrimination value indicates a good term where the similarity value between documents is decreased when the index term is assigned to the collection while a low discrimination value indicates vice versa. For instance, they had proposed that the "most discriminant" terms have a frequency within the range of n/100 and n/10 from a given collection of n documents.

In order to eliminate the unnecessary terms in the collection, term reduction approach seems to be useful to optimize the size of the term collection, by removing noisy terms which may increase the precision ratio but not to reduce the recall ratio. There are three term selection 8 methods proposed by Rojas *et al*. (2007) aimed to represent each document by the most important terms, which are: Entropy (H) method, Transition Point (TP) method and the hybrid of H and TP method. They believed these methods able to improve precision in document retrieval. The Entropy (H) method seems to be computationally expensive yet effective. Therefore, Transition Point (TP) method is used as the reference for measuring the term representativeness.

Effort on selecting the minimal set of index terms has been done by Rojas *et al*. (2007). Although VSM is a very well known classical term-document model, it exhibits noisy effect in the text representation (Sebastiani, 2002). It is the intention of this research to explore the capability of the TP and Entropy methods to discriminate the relevant and non-relevant terms 18 throughout the vocabulary set. These methods are described as follows:

Entropy: The entropy formula is extracted from Shanon's work on information theory which considers an information measure from different possibilities that a system has (Jimenez-Salazar *et al*., 2005).

Given a set of documents $D = \{D_1, D_2,\ldots, D_m\}$ and N_i, which represents the number of words in the document D_i. The relative frequency (f) of the word w_j in D_i is defined as:

$$f_{ij} = \frac{tf_{ij}}{N_i^t f_{ij}}$$

Where:
tf_{ij} = The frequency of term
w_j = In document D_i

The Probability (p) of the word w_j in document D_i is calculated as:

$$p_{ij} = \frac{f_{ij}}{\sum_{j=1}^{m} f_{ij}}$$

Then, the entropy of w_j will be obtained from the following equation:

$$H(w_j) = -i = 1\sum^{m} p_{ij} \log p_{ij}$$

The representation of a document D_i will be given by a vector in the VSM. The maximum entropy value of vocabulary terms, H_{max}, will be taken as a reference to identify the most important terms according to this criteria. For a document D_i, the most important terms are those 16 whose entropy value is greater than a ratio of H_{max}, which is formally defined as:

$$H_i \left\lfloor w_j \in D_i \mid H(wj) > H_{max}.u \right\rfloor$$

where, *u* is a threshold which defines the level of high entropy. For this experiment, *u* has been set as 0.5. This is based on the opinion at Rojas *et al.* (2007).

Transition point: The TP technique has shown to be effective in document retrieval (Rojas *et al.*, 2007), text categorization (Moyolt and Jimenez, 2005) and text clustering (Pinto *et al.*, 2006). The TP has been used in text indexing by Urbizagastegui (1999) based on Zipf Law of Word Occurrence (Zipf, 1949) and Booth (1967) (Moyolt and Jimenez-Salazar, 2004).

Basically, TP is a frequency value that divides a set of terms in the vocabulary into low frequency and high frequency terms (Moyolt and Jimenez-Salazar, 2004). Below is the typical formula used to obtain the value of TP.

$$TP = \frac{-1 + \sqrt{8.I_1 + 1}}{2}$$

where, I1 represents the number of words with frequency equal to 1 (Urbizagastegui, 1999).
The hypothesis which was made by Salton *et al.* (1975) mentioned that terms of medium frequency usually have high semantic content (Moyolt and Jimenez-Salazar, 2004). Therefore, the experiment using this method has been carried out to identify the terms with medium term frequency around the TP.

Basically, the equation based on Zipf Law (1949) is meant to demonstrate the mid-frequency terms that are closely related to the conceptual content of a document.

A document D_i and its vocabulary V_i is represented as $\{(w_i, tf_i, (w_j)) \mid w_j \in D_i\}$, where $tf_i, (w_j) = tf_i$. TP_i will be the transition point of D_i and a set of terms which represent D_i is calculated as follows:

$$R_i \left\{ \begin{array}{c} w_j \mid ((w_j, tf_{ij}) \in V_i \\ TP_i .(1-u) \le tf_{ij} \le TP_i(1+u)) \end{array} \right\}$$

where, u = 0.4 which claimed to be a good value for the threshold (Rojas *et al.*, 2006).

Important terms that are closer to the TP will be considered. R_i set remarks: a term with frequency very 'close' to TP will get a high weight and those 'far' to TP will get a weight close to zero.

Union of entropy and transition point: All terms represented by both approaches, TP and H, will be combined. The selected terms must satisfy either of two conditions. Basically, the representation of a document D_i, is given by:

$$H_i^{'} = H_i \cup R_i$$

The Quran as a Literary Text: Natural language is a primary source in literary text. The use of suitable data analysis tools is necessary to discover the natural language words and sentences within its context in a literary corpus. The Quran is an example that contains literary text. Most of the text is unstructured, formless and difficult to be interpreted directly. Theoretically, Hanauer (1998) have summarized the differences of the character of the genre of poetry texts (or literary texts) and encyclopedic texts. The characteristic of literary text more or less focus on linguistic features; have multiple levels of meaning, contains multiple understandings of text, depends on the internal structure of meaning construction, the direction of reading process is non-linear and it depends on human significance.

In this experiment, the context from each document can be determined directly by the content of the documents which contained any important terms in its vocabulary. Noun phrase is been exploited as the key phrase that is useful in many applications such as text categorization, concept hierarchy and clustering.

Even though there are no specific number of noun phrase contain in other corpus, but in this experiment there are only 2,220 noun words out of 29,465 words in the corpus. Therefore, only 7.3% terms has been extracted to form a set of vocabulary.

Experiment: An experiment has been performed using four different methods; they are VSM, TP, H and the union of both H and TP (H+TP). The selected terms have been listed and evaluated using the ratio of representativeness of the index terms by the selection methods. We compare the terms extracted from these four methods against a manually extracted index of the Quran available from Islamicity.com and Submission.org.

The Islamicity.com and the Submission.org have been used as the publicly available Quran index in the internet, which was established in 1995 and 1997 respectively. Both websites contained information about Islam and Quran in text, audio and video besides using multi-language including English. From our

observation, both websites provide a Quran index glossary which based on Quran texts in English. Therefore we decided to choose them as a benchmark.

The ratio of terms given by an index terms in the vocabulary obtained by each method is calculated using the following equation:

$$\frac{\#\text{ terms exist in vocabulary and the index terms}}{\#\text{ terms in the vocabulary}}$$

For instance the ratio of representativeness of index terms using TP method is calculated according to the number of terms that exist both in the vocabulary of TP method and the manual index (Islamicity.com or Submission.org) divided by the number of terms in the vocabulary of the TP method.

As previously mentioned the Quran index chosen are from http://www.islamicity.com/mosque/and http://www.submission.org/quran/koran-index.html, which represent as a set of important keywords contained in the Quran. The total numbers of index in both sites are 792 and 3200 respectively.

Data collection: Testing has been done based on natural language processing on a 6,666 documents from unsupervised English Translated Quran text corpus. It contained a literary kind of text which has been indexed using an inverted file. On average, each document consists of 430.29 words.

RESULTS

Evaluation: The purpose of the comparison is to see how well the performance of those methods as compared to VSM and also to see which method shows the best for term reduction for the Quran.

Table 1 and 2 show the comparison among the four methods and the result is very encouraging. The TP method shows the best performance for both data index which is due to the use of mid-frequencies terms as compared to the classical VSM representation.

Accordingly, Table 3 shows that TP method performed the highest percentage of reduction from the benchmark which contains only 41.71% of vocabulary from the whole vocabulary set while H+TP method is the lowest percentage of vocabulary reduction.

In principle, the Quran index acts as a benchmark which contained only the important keywords represented in the Quran. Thus, the more terms appeared in the index indicates the method has the capacity of identifying suitable or important keywords. From the analysis, the TP method performs better compared to the others not only in terms of term reduction but also the representativeness of selected term exists in the Quran index.

Table 1: Percentage of Representativeness of Index Term as compared to Islamic.com Number of Index Term in Islamic.com = 792

Method	Vocabulary size	Number matched	(%)
VSM	2220	248	11.17
H	1601	203	12.68
TP	926	133	14.36
H + TP	1772	216	12.19

Table 2: Percentage of Representativeness of Index Term as compared to Submission.org Number of Index Term in Submission.org = 3200

Method	Vocabulary Size	Number Matched	(%)
VSM	2220	752	33.87
H	1601	630	39.35
TP	926	418	45.14
H + TP	1772	662	37.36

Table 3: The vocabulary size which obtained from Quran corpus

Method	Vocabulary size	Percentage of reduction
VSM	2220	0.00
H	1601	27.88
TP	926	58.29
H+TP	1772	20.18

DISCUSSION

The main concern of this study is to evaluate the representativeness of terms using the term selection methods in the context of literary text. Thus, the term selection and tern reduction seems to give significant results. Moens (2000) has identified some characteristics of good term representations and the experiment has performed two of them which are term reduction and term representativeness (aboutness) of document text.

As compared to the other research done by Rojas *et al*. (2007), entropy method has performed better than TP in term of percentage of reduction over the result of recall and precision. However, they commented that, TP method still has a good performance because of the low computational cost as compared to entropy. In addition, the effectiveness of TP method also has been tested for term selection process by Pinto *et al*. (2006) based on clustering approach and their finding is encouraging even though the analysis on the stability of the method need to done. Besides that, the text categorization approach also has been experimented to TP method (Moyotl-Hernandez and Jimenez-Salazar, 2004) and the result is encouraging.

For this experiment over the literary text, TP method also performed better in term selection as well as term reduction. The features of TP that can be easily computed give advantage to it. As stated by Salton *et al*. (1975), there are more than 70% of terms in collection described as poor discriminator. Thus, within the context vector space, TP method is found able to enhance the property of in formativeness. Within the

context of H and TP methods, the unification of both methods has been tested but the result is not very encouraging. Ideally, the enrichment of terms selected from H and TP property should be able to optimize the term representativeness. However, the unification of both methods resulted in an increased number of terms but the terms that matched with the benchmark index do not increased at a similar ratio. Therefore, the H+TP method is unable to achieve a better result as compared to the TP and H methods.

Most probably the H and TP method are in such a manner, by instance, that selecting different set of terms which may cause one of them are noise to the other. This would be the reason of lower performance of H+TP method than TP method. Therefore, the H+TP method is unable to achieve a better result as compared to the TP and H methods.

TP method seems to be a promising technique for term selection in term of providing some property of informative content of terms for documents. As compared to entropy method, TP is more practical and easy to be used (in terms of computation).

CONCLUSION

In this study, a set of Quran corpus originally from Al-Bayan CD-ROM which contains very short texts had been filtered using basic procedure of text pre-processing, such as part-of-speech tagging and stemming. The representativeness of texts is based on unsupervised nature.

The experiment was then conducted using method used in Rojas *et al*. (2007). The result of testing indicates that TP method has better performance in terms of its capability to extract important terms. Besides, the computational work is much easier than the entropy. However, there are still many important terms in a document that have a frequency far from TP.

The highest percentage of representativeness illustrated in this experiment is only 45.14% which indicate the needs to further explore new approach in order to increase the percentage of representativeness. As a result, further experiments will be conducted that will consider external sources such as the WordNet (Miller, 1995), ConceptNet (Liu and Singh, 2004) and other thesauri.

REFERENCES

Baeza-Yates, R. and B. Ribeiro-Neto, 1999. Modern Information Retrieval. 1st Edn., Addison Wesley Longman Limited, New York, ISBN: 0-201-39829-X, pp: 544.

Booth, A., 1967. A “Law” of Occurrence of Words of Low Frequency. Inform. Control, 8 10: 383-396. http://dblp.uni-trier.de/db/journals/iandc/iandc10.

Eagleton, T., 1996. Literary Theory. 2nd Edn., Wiley-Blackwell, Oxford, ISBN: 0-8166-13 1251-X, pp: 147.

Hanauer, D., 1998. The genre-specific hypothesis of reading: Reading poetry and encyclopedic. Poetics, 26: 63-80. DOI: 10.1016/S0304-422X(98)00011-4

Jimenez-Salazar, H., M. Castro, F. Rojas, E. Minon and D. Pinto *et al*., 2005. Unsupervised Term Selection using Entropy. In: Advance in Artificial Intelligence and Computer Science, Gelbukh, A., C. Yanez and O. Camacho (Eds.). Research on Computing Science, pp: 163-172. ISSN: 1665-9899.

Lawrie, D. and W.B. Croft, 2003. Generating hierarchical summaries for web searches. Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Jul. 28- Aug. 1, ACM New York, Toronto, Canada, pp: 457-458. ISBN: 1-58113-646-3

Lawrie, D., W.B. Croft and A. Rosenberg, 2001. Finding Topic words for hierarchical summarization. Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Sept. 9-12, New Orleans, Louisiana, ACM New York, pp: 349-357. ISBN: 1-58113-331-6

Liu, H. and P. Singh, 2004. Concept Net: A practical common sense reasoning toolkit. BT Technol. J., 22: 211-226. DOI: 36 10.1023/B:BTTJ.0000047600.45421.6d) 37

Miller, G.A., 1995. Electronic Sources: WordNet. http://wordnet.princeton.edu/

Moens, M.F., 2000. Automatic Indexing and Abstracting of Document Texts. 2nd Edn., Kluwer Academic Publishers, Massachusetts, ISBN: 0-7923-7793-1, pp: 265.

Moyotl, E. and H. Jimenez, 2005. Enhancement of DPT feature selection method for text categorization. Lecture Notes Comp. Sci., 3406: 719-722. DOI: 10.1007/b105772

Moyotl-Hernandez, E. and H. Jimenez-Salazar, 2004. An analysis on frequency of terms for text categorization. XX Congreso de la Sociedad Española para el Procesamiento del Lenguaje 5 Natural, July 22-24, Universitat de Barcelona, pp: 141-146. ISSN: 1135-6 5948

Pinto, D., H. Jimenez and P. Rosso, 2006. Clustering abstracts of scientific texts using the transition point technique. Lecture Notes Comp. Sci., 3878: 536-546. DOI: 10.1007/11671299

Rojas, F., H. Jimenez and D. Pinto, 2007. A competitive term selection method for information retrieval. Lecture Notes Comp. Sci., 4394: 468-475. DOI: 10.1007/978-3-540-70939-8

Rojas, F., H. Jimenez, D. Pinto and A. Lopez, 2006. Dimensionality Reduction for Information Retrieval. In: Advances in Artificial Intelligence, Gelbukh, A., S. Torres and I. Lopez (Eds.). Research on Computing Science, pp: 107-112. ISSN: 1665-9899

Salton, G. and M. McGill, 1983. Introduction to Modern Information Retrieval. 2nd Edn., McGraw-Hill Book Company, New York, ISBN: 0070544840, pp: 448.

Salton, G., A. Wong and C.S. Yang, 1975. A vector space model for automatic indexing. Commun. ACM, 18: 613-620. ISSN: 0001-0782. DOI: 30 http://doi.acm.org/10.1145/361219.361220

Schultz, J.M., 2003. Term selection for information retrieval applications. PhD Dissertations. University of Pennsylvania. http://proquest.umi.com/pqdlink?Ver=1&Exp=10-27-38 2014&FMT=7&DID=765024601&RQT=309&attempt=1

Sebastiani, F., 2002. Machine learning in automated text categorization. ACM Comput. Surveys, 34: 1-47. DOI: 10.1.1.17.6513

Sieg, A., B. Mobasher, S. Lytinen and R. Burke, 2004. Using concept hierarchies to enhance user queries in web-based information retrieval. Proceedings of the IASTED International Conference on Artificial Intelligence and Applications, Feb. 16-18, Innsbuck, Austria. ISSN: 1027-2666.

Urbizagastegui, A.R., 1999. Las Posibilidades de la Ley de Zipf en la Indizacion Autom´atica (In Spanish). http://b3.bibliotecologia.cl/ruben2.htm

Yusof Ali, A., 1934. The Meaning of Holy Qur'an. 10th Edn., Amana Corporation, ISBN-10: 0915957760, pp: 1762.

Zipf, G.K., 1949. Human Behavior and the Principle of Least Effort-An Introduction to Human Ecology. 1st Edn., Hafner Pub. Co, Massachusetts, pp: 573. http://www.amazon.com/Human-behavior-principle-least-effort/dp/B0007FMHHW