Advanced Science Letters
Vol. 21, No. 10, 3308-3312, 2015

Clustering Bilingual Documents Using Various Clustering Linkages Coupled with Different Proximity Measurement Techniques

Rayner Alfred[1,2,3], Leow Ching Leong[1,2], Mohd Hanafi Ahmad Hijazi[1,3,] Joe Henry Obit[1] and Kim On Chin[1,2,3]

[1]Faculty of Computing and Informatics, Universiti Malaysia Sabah, 88400 Kota Kinabalu, Sabah, Malaysia
[2]Centre of Excellence in Semantic Agents, Universiti Malaysia Sabah, 88400 Kota Kinabalu, Sabah, Malaysia
[3]Artificial Intelligence Research Unit (AIRU), Universiti Malaysia Sabah, 88400 Kota Kinabalu, Sabah, Malaysia

With the rich data on the web, a documents clustering task for monolingual documents is insufficient in order to produce an efficient information retrieval system. A Multilingual Document Clustering (MDC) had been introduced and it is one of the most popular trends in the area of natural language processing (NLP). In this paper, the effects of applying different clustering linkages coupled with different proximity measurements on the clustering bilingual Malay-English documents in parallel are investigated. A Hierarchical Agglomerative Clustering (HAC) has been implemented and applied in clustering bilingual Malay-English documents. Several different linkages are used in the HAC method that includes Single, Complete, Centroid and Average linkages. Not only that, the cosine similarity and the extend Jaccard coefficient are also applied in order to investigate a proper proximity measurement that can be coupled with the different type of clustering linkages used for clustering bilingual news articles written in English and Malay. The HAC method coupled with the average linkage can be considered to produce reasonable clustering results even though the average DBI is a bit high. Now only that, the study also shows that the extend Jaccard coefficient proximity measurement can produce a better clustering results compared to the cosine similarity.

Keywords: Genetic Bilingual-Clustering, Proximity Measurement, Hierarchical Agglomerative Clustering, Document Clustering.

1.INTRODUCTION

There are multiple types of data that are available in the web with the advanced of the technology nowadays. One of the most common data that can be retrieved is the document articles. An efficient method is required to analyze large amount of documents so that the required information can be retrieved easily and effectively. The most common method that had been widely used is the document clustering, which groups the documents according to their similarities. There are lots of studies on the document clustering[1-6].

There are several aspects that need more attention when clustering documents which include the volume, the dimensionality and the semantics complexity of the documents[7]. It cannot be denied that the documents on the web are in multiple languages. This shows that clustering monolingual documents are currently insufficient for retrieving the useful information. This had leads to the study of multilingual document clustering.

There are several advantages that encourage the studies towards clustering multi-lingual documents compared to clustering monolingual documents. Clustering multi-lingual documents can be used as a tool

*Email Address: ralfred@ums.edu.my

 3308-3312/2015/21/3308/005 DOI: 10.1166/asl.2015.6476

to indirectly verify the categories or clusters that a document belongs to by mapping the clusters obtained from one language into the clusters obtained from different language. It will also lead towards the elimination of biased language-specific usages which will improve the grouping of the documents[8]. Clustering multilingual documents can be used in many applications that include the Cross-Lingual Information Retrieval applications, the training of parameters in Statistics Based Machine Translation, or the Alignment of parallel and non-parallel corpora, among others[9].

The main contribution of this paper is to study the effect of using different distance linkages coupled with different proximity measurements used for bilingual documents in parallel on the clustering results obtained. The remainder of the paper is organized as follow. Section 2 reviews some of the related work. Some of the cluster linkages and proximity measurements used in clustering documents are discussed in Section 3. The experimental set up is described and Section 4 and the evaluation and results are discussed in Section 5. Finally, Section 6 presents the conclusion and future works.

2.RELATED WORKS

Two types of document that are in used for multilingual document clustering are the parallel and the comparable corpora. The parallel corpora are translated corpora where one corpus is the resource and others are the translation from the resource whereas the comparable corpora are just a collection of several corpora having the same or similar topic[10,11]. News articles are the example of the comparable corpora. The disadvantage of clustering multilingual document is that there are not many corpora that are exactly parallel in two languages. Bader and Chew have used different languages of the Bible and the Quran as their multilingual corpora[12]. The authors had chosen five different languages in their studies which are Arabic, English, French, Russian and Spanish. Leftin illustrated experiments on clustering Russian-English corpora using machine translation system[13]. Not only that, there is also study towards short-text corpora[28] where the effects of CLUPDIPSO on Spanish-English short texts corpora are studied.

Makita et al., have proposed an application that applies multilingual document clustering, which is known as the Patent Retrieval in Multilingual Environment (PRIME)[30]. In PRIME, besides the clustering algorithm, the system also used a translation function that translates the user query into the targeted language and translated the retrieved patents into user language. This helps researcher to retrieve patents that are filed in other languages. In another study, the Wikipedia is used as an external knowledge to enrich the information of the documents for clustering multilingual documents. The authors used English documents and Hindi documents as the basic documents. Additional information is been retrieved from the Wikipedia based on the basic documents. Based on the basic information of the documents and additional information from Wikipedia, the documents are clustered in each language. There are also some studies on multilingual document clustering such as English-Spanish[14-16], English-Japanese[17] and English-Bulgarian[18]. However there is not much study bilingual clustering towards English and Malay documents[19]. For this reason, this paper is proposed to analyse the effect of using different cluster linkages coupled with different proximity measurements in clustering bilingual English-Malay documents.

3. CLUSTERING AND PROXIMITY MEASUREMENT TECHNIQUES

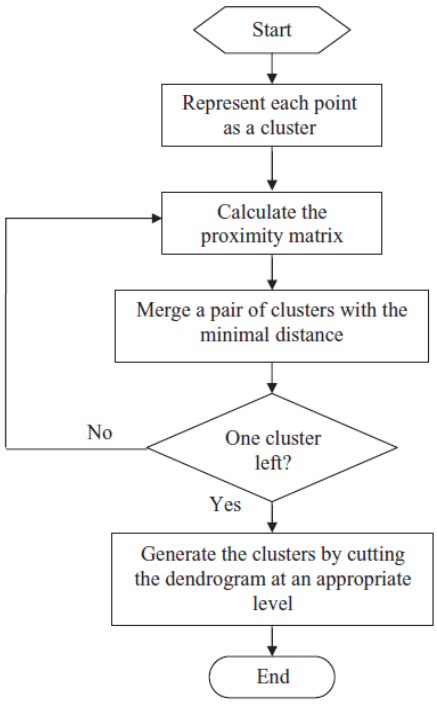

Fig.1. Flowchart of the HAC algorithm

A clustering method can be divided into two types which are hierarchical and partitional clustering. Hierarchical Agglomerative Clustering (HAC) is one of the well-known techniques used in document clustering. HAC starts with N point clusters where each cluster represents a document. The cluster will then merge until all the documents are grouped into the same cluster. The result of HAC is displayed as a dendogram, which is in a tree structure. Fig.1 shows the flowchart of the HAC algorithm[20]. The merging of the clusters depends on the distance function used in computing the distance between two clusters. There are several methods that can be used as distance function in HAC, namely Single, Complete, Average and Centroid Lingkages[21].

One aspect that needs to be considered in performing the clustering task is the type of proximity measurements used in the clustering algorithm. The proximity measurement is a generalization of both dissimilarity and similarity measurement technique. There are two factors that affect similarity and dissimilarity measurement values which are the properties of the two objects and the measurement itself[22].

Similarity Measurement

A numerical value is normally used to measure the degree of similarity for two objects. It is a non-negative value and its range is from 0 to 1. A higher value indicates the two objects are very alike. A similarity of value 1 indicates that the two documents are the same. A similarity value of zero shows that two documents do not share any similar characteristics. Similarity measurement satisfies two conditions 1) s(x,y) = 1 only if x=y $(0 \leq s \leq 1)$ and 2) s(x,y) = s(y,x) for all x and y (Symmetry). A document consists of thousands of words in its content and each word is known as a term. However, when dealing with multiple documents there is a possibility that some words may exist in a document but not in the other documents. The cosine similarity and the extend jaccard coefficient are methods used to measure the similarity between documents having the same words. Equation shown below defines the formula for cosine similarity measurement,

$$\cos(x, y) = \frac{x \bullet y}{||x|| ||y||}$$

where *x* and *y* is the vectors of the respective documents and ||x|| = sqrt(x*x). The Extended Jaccard Coefficient is the extend version of Jaccard Coefficient which just deals with binary attributes. The extended version of Jaccard Coefficient is capable to deal with continuous or discrete non-negative feature. Extended Jaccard Coefficient is also known as Tanimoto Coefficient. It compares the sum weight of shared terms to the sum weight of terms that are present in either of the two documents but are not the shared terms[22] as shown below,

$$EJ = \frac{x \bullet y}{||x||^2 + ||y||^2 - x \bullet y}$$

where *x* and *y* is the vectors of the respective documents and ||x|| = sqrt(x*x).

4.EXPERIMENTAL SETUP

The objective of this experiment is to study the effects of using different distance linkages coupled with different proximity measurements used for bilingual documents in parallel on the clustering results obtained. A Hierarchical Agglomerative Clustering (HAC) is implemented in order to cluster English and Malay news articles in parallel[19]. For instance, each language will go through all the pre-processes in parallel that include stop words elimination, stemming, NER elimination, TF-IDF matrix construction and the clustering process. In order to make sure that the contents of news articles for both languages are similar, 500 news articles for both languages are retrieved manually from the Bernama archive and theStar website.

In this study, the Davies-Bouldin Index (DBI) will be used in order to measure the cluster quality produced in both languages. In addition to that, a standard deviation of the number of documents for all clusters is also computed in order to qualify the amount of variation of the numbers of members for all clusters. A good clustering result can be obtained by having a low standard deviation of the number of documents for all documents and also having a low DBI measurement. A low standard deviation of the number of documents for all clusters indicates that the documents are evenly distributed in all clusters. Similarly, a low DBI measurement indicates that the clusters produced are compacted and well separated.

The flow of work consists of data pre-processing, weighting scheme, clustering and cluster validation. The data pre-processing stage consists of stopword elimination, symbol elimination, stemming and elimination of Named-Entity Recognition (NER) and finally the construction of the TF-IDF matrix. The stopword elimination process removes non-important words that reduces the number of terms and increases the efficiency of the clustering process[23]. The symbol elimination process eliminates symbols that exist in the news articles.

Stemming is a process of converting words to its root words[24]. By stemming words that exist in documents, the number of unique terms or features can be reduced. The Porter stemming algorithm is normally used for stemming English words and a rule-based Malay stemming algorithm is used for stemming Malay words[25]. A named-entity recognition algorithm may also be used to reduce the number of terms. For English articles, a Stanford NER is used for annotating the English named-entities[26] whereas for Malay articles, a rule-based Malay NER is used for annotating the Malay named-entities[27].

In this study, the weights of terms are computed by using the term frequency-inverse document frequency (TF-IDF). The computation of TF-IDF will reveal the list of terms that are considered important for clustering purposes. The weighting scheme of TF-IDF is shown as follow,

$$tfidf\ (d, t) = tf(d, t) * \log \frac{|D|}{df(t)}$$

where *df*(*t*) is the frequency of documents that term *t* appears. A Hierarchical Agglomerative Clustering (HAC) is used in this experiment and the effects of using different distance linkages on the quality of clustering results will be investigated. There are four types of linkages used in HAC are been tested in this study. These linkages are Single, Complete, Average and Centroid linkages.

In this study, all these clustering linkages will be coupled with two proximity measurements that include the cosine and extend Jaccard coefficient similarities. These proximity measurements have been outlined and described in Section 3. The HAC method will be used as the main clustering method in this experiment. In order to evaluate the quality of the clustering results obtained, Davies-Bouldin Index (DBI) had been implemented. It measured the compactness of the documents in a cluster and separation between clusters. A good quality clustering results will produce low DBI values. The formula of DBI is as shown below,

 3308-3312/2015/21/3308/005 DOI: 10.1166/asl.2015.6476

$$DBI = \frac{1}{c}\sum_{i=1}^{c} Max_{i \neq j}\left(\frac{d(x_i)+d(x_j)}{d(c_i,c_j)}\right)$$

where $d(c_i, c_j)$ is the distance between centroid of cluster i and cluster j, $d(x_k)$ is the distance of documents in cluster x_k and c is the number of clusters. Other than that, the formula for standard deviation is shown as below,

$$SD = \sum_i \sqrt{\frac{\sum(x-\bar{x})^2}{n}}$$

where n is the number of clusters, x is the number of documents in the clusters and $\bar{x}$ is the mean.

5. RESULTS

The DBI value for English language clustering and Malay language clustering will be used to determine the quality of clustering results. The average DBI values are obtained from cluster size of 2 up to cluster size of 498. Table 1 shows the average DBI values for all cluster sizes. Based on results obtained, it is obvious that the HAC method coupled with the Extend Jaccard coefficient produced much better clustering results compared to the clustering results obtained when the HAC method coupled with cosine distance method. Overall, the HAC method coupled with the complete linkage distance produced the worst clustering results compared to other clustering linkages distance methods. This is because the average DBI measurement obtained is large compared to the other clustering linkages. The performance of the HAC method is better when the Single and Centroid linkages are used.

The DBI value measures the cluster quality by measuring the compactness of the clusters and the average separation distance among clusters. A lower average DBI value obtained when using single and centroid linkages shows that the documents are closely clustered. The HAC method coupled with the complete linkage shows the highest DBI value because clusters are merged based on furthest distance between two documents that belong to different clusters. The merging of furthest documents between clusters produced a large size cluster. Nevertheless, based on the patterns of documents distribution in all clusters obtained, the HAC method coupled with the Single or Centroid linkage produced clusters in which most of the documents are clustered into the first cluster. As a result, the standard deviation of the number of documents for all clusters is high based on the results obtained which are shown in Table 1.

Based on singleton pattern that can be observed for most of the clusters produced by using the HAC method coupled with the Single or Complete linkage. This is due to the fact that the list of terms used to cluster these documents is too long that has caused the similarity of two documents is too small to be differentiated.

By comparing several thousand terms with just a few similar terms between the news, this lowers the similarity measurement between the news. As a result, most of the news articles are the same and near to each other. On the other hand, most of the clusters produced by the HAC method coupled with the average linkage or the complete linkage contain several news articles in each cluster. Based on the standard deviation shown in Table 1, it can be obtained that the HAC method coupled with the average or the complete linkage produced a better clustering result in which the document are not clustered into singleton cluster.

Table.1. The average DBI and standard deviation values for various cluster size

Linkage	Cluster Size	Davies-Bouldin Index		Standard Deviation	
		Cosine	Jaccard	Cosine	Jaccard
Single	100	**1.256**	1.279	**72.390**	73.355
	200	1.347	**1.314**	**42.297**	42.957
	300	1.308	**1.254**	**29.556**	30.007
	400	1.231	**1.154**	**22.653**	22.993
	498	1.107	**1.039**	**18.345**	18.614
Average Result		**1.250**	**1.208**	**37.048**	**37.585**
Centroid	100	1.408	**1.380**	62.167	**58.831**
	200	1.408	**1.345**	34.071	**32.094**
	300	1.346	**1.244**	23.748	**22.367**
	400	1.243	**1.144**	18.218	**17.151**
	498	**1.035**	1.118	15.723	**14.889**
Average Result		**1.282**	**1.240**	**31.829**	**30.486**
Complete	100	**2.962**	3.027	**15.363**	15.406
	200	2.305	**2.297**	9.125	**9.102**
	300	1.973	**1.908**	6.620	**6.596**
	400	1.752	**1.643**	5.206	**5.174**
	498	1.526	**1.425**	4.547	**4.529**
Average Result		**2.104**	**2.060**	**8.172**	**8.162**
Average	100	2.422	**2.384**	**30.481**	31.175
	200	1.965	**1.900**	**17.693**	18.059
	300	1.726	**1.622**	**12.713**	12.962
	400	1.562	**1.428**	**9.907**	10.091
	498	1.372	**1.254**	**8.085**	8.228
Average Result		**1.810**	**1.718**	**15.776**	**16.103**

In short, the HAC method coupled with the average linkage and the extend Jaccard coefficient can be considered as a acceptable technique to cluster bilingual documents that take into consideration the low DBI value and the low standard deviation of the number of documents in all clusters.

6. CONCLUSION

As a conclusion, the HAC method coupled with the single linkage has produced lower average DBI values compared to the other methods. However, the clustering results produced are not ideal. Similar patterns can be observed for the HAC method coupled with the centroid linkage. Hence, the HAC method coupled with the average linkage can be considered to produce a reasonable clustering result even though the average DBI value is high. This can be observed based on the results obtained that a low standard deviation of the number of documents for all clusters is obtained for the HAC method coupled with the average linkage. This experiment has also proven that the extend Jaccard coefficient proximity measurement can produce a better

clustering result compared to the cosine similarity. In future works, a genetic algorithm can be proposed to optimize the terms to increase the clusters mapping results and indirectly improve the efficiency of Malay-English bilingual clustering

ACKNOWLEDGMENTS

This work has been supported partially by the Long Term Research Grant Scheme (LRGS) project funded by the Ministry of Higher Education (MoHE), Malaysia under Grants No. LRGS/TD/2011/UiTM/ICT/04.

REFERENCES

[1] Saad, F. H., Mohamed, O. I. E., Al-Qutaish, R. E.: Comparison of Hierarchical Agglomerative Algorithms for Clustering Medical Documents. In International Journal of Software Engineering & Applications (IJSEA). Vol. 3. pp. 1-15. (2012)

[2] Deshmukh, D. B., Pandey Y.: A Review on Hierarchical Document Clustering. Journal of Data Mining and Knowledge Discovery. 3(2):65-68. Bioinfo Publications. (2012)

[3] Mugunthadevi, K., Punitha, S.C., Punithavalli, M.: Survey on Feature Selection in Document Clustering. International Journal on Computer Science and Engineering. 3(3):1240–1241. (2011)

[4] Gautam, D. P., Shrestha, D. Members IAENG: Document Clustering Through Non-Negative Matrix Factorization: A Case Study of Hadoop for Computational Time Reduction of Large Scare Documents. In Proc of the International MultiConference of Engineers and Computer Scientists (IMECS). Vol1. (2010)

[5] Premalatha, K. Natarajan, A. M.: A Literature Review on Document Clustering. Information Technology Journal 9(5). pp. 993-1002. Asian Network for Scientific Information. (2010)

[6] Park, S., An, D. U., Cha D. R., Kim, C. W.: Document Clustering with Semantic Features and Fuzzy Association. In Information Systems, Technology and Management Communications in Computer and Information Science. Vol. 54. pp. 167-175. Springer Berlin Heidelberg. (2010)

[7] Punitha, S. C., Mugunthadevi, K., Punithavalli, M.: Impact of Ontology based Approach on Document Clustering. International Journal of Computer Applications. 22(2):22-26. (2011)

[8] Wang, Y. Y., Lafferty, J., Waibel, A.: Word Clustering with Parallel Spoken Languag Corpora. In Proc. of 4th International Conference on Spoken Language Processing (ICSLP). (1996)

[9] Montalvo, S., Martinez, R., Casillas, A., Fresno, V.: Multilingual News Document Clustering: Two Algorithms Based on Cognate Named Entities. Text, Speech and Dialogue. LNAI. Vol. 4188. pp. 165-172. Springer Berlin Heidelberg. (2006)

[10] Lee, C. H., Yang, H. C.: Text Mining of Bilingual Parallel Corpora with a Measure of Semantic Similarity. Systems, Man and Cynernetics. Vol. 1. pp. 470-475. IEEE. (2001)

[11] Tao, T., Zhai, C. X.: Mining Comparable Bilingual Text Corpora for Cross-Language Information Integration. Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining. pp. 691-696. ACM. (2005)

[12] Bader, B. W., Chew, P. A.: Algebraic Techniques for Multilingual Document Clustering. In Text Mining: Applications and Theory. pp. 21-26. John Wiley & Sons Ltd. (2010)

[13] Leftin, L.J.: News blaster russian-english clustering performance analysis. Technical report, Columbia computer science Technical Reports (2003)

[14] Montalvo, S., Martinez, R., Casillas, A., Fresno, V.: Multilingual Document Clustering: an Heuristic Approach Based on Cognate Named Entities. In Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL. pp. 1145-1152. Association for Computational Linguistics. (2006)

[15] Montalvo, S., Martinez, R., Fresno, V.: NESM: a Named Entity based Proximity Measure for Multilingual News Clustering. Procesamiento de Languaje Natural. pp. 81-88. Sociedad Espanola para el Procesamiento de Lenguaje Natural. (2012)

[16] Montalvo, S., Martinez, R., Casillas, A., Fresno, V.: Bilingual News Clustering Using Named Entities and Fuzzy Similarity. Proceedings of the 10th international conference on Text, speech and dialogue. pp. 107-114. Springer Verlag Heidelberg. (2007)

[17] Yamamoto, H., Sumita, E.: Bilingual Cluster Based Models for Statistical Machine Translation. In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language and Computational Natural Language Learning. pp. 514- 523. Association for Computational Linguistics. (2007)

[18] Alfred, R.: A Parallel Hierarchical Agglomerative Clustering Technique for Bilingual Corpora Based on Reduced Terms with Automatic Weight Optimization. In Proceedings of the 5th International Conference on Advanced Data Mining and Applications. pp. 19-30. Springer Berlin Heidelberg. (2009)

[19] Alfred, R., Chan, C. J., Ng, Z. W., Tahir, A., Obit, J. H.: Optimizing Clusters Alignment for Bilingual Malay-English Corpora. Journal of Computer Science 2012 8(12). pp. 1970-1978. Science Publications. (2012)

[20] Xu, R., Wunsch, D. C.: Clustering. pp. 31-62. John Wiley & Sons Inc. (2009)

[21] Tan, P. N., Steinbach, M., Kumar, V.: Introduction to Data Mining. Pearson International Edition. (2006)

[22] Huang, A.: Similarity Measures for Text Document Clustering. In the Proceedings of the New Zealand Computer Science Research Student Conference. pp. 49-56. (2008)

[23] El-Khair, I. A.: Effects of Stop Words Elimination for Arabic Information Retrieval: A Comparative Study. International Journal of Computing & Information Sciences. Vol. 4. No. 3. pp. 119-133. (2006)

[24] Alfred, R. Leow, C. L., Chin, K. O., Anthony, P.: A Literature Review and Discussion of Malay Rule-based Affix Elimination Algorithms. The 8th International Conference on Knowledge Management in Organizations. pp. 285-297. Springer Netherlands. (2014)

[25] Leow, C. L., Basri, S., Alfred, R.: Enhancing Malay Stemming Algorithm with Background Knowledge. PRICAI 2012: Trends in Artificial Intelligence Lecture Notes in Computer Science. Vol. 7458. pp. 753-758. (2012)

[26] Finkel, J. R., Grenager, T., Manning, C.: Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics. pp. 363-370. Association for Computational Linguistics. (2005)

[27] Alfred, R. Leow, C. L., Chin, K. O., Anthony, P.: Malay Named Entity Recognition Based on Rule-Based Approach. International Journal of Machine Learning and Computing. Vol. 3. No. 4. pp. 300-306. (2014)

[28] Ingaramo, D., Errecalde, M., Cagnina, L., Rosso, P.: A Particle Swarm Optimizer to Cluster Parallel Spanish-English Short-text Corpora. In the Proceedings of the Workshop on Iberian Cross-Language Natural Language Processing Task. pp. 43-48. (2011)

 3308-3312/2015/21/3308/005 DOI: 10.1166/asl.2015.6476