Chinese and Thai Bilingual Topic Detection Online

Ziqiang Rang[1,2], Lanjiang Zhou[1,2*], Jinpeng Zhang[3], Yantuan Xian[1,2] and Zhengtao Yu[1,2]

[1]School of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500, China
[2]The Key Laboratory of Intelligent Information Processing ,Kunming University of Science and Technology, Kunming, Yunnan 650500，China
[3]Information Management Center，Yunnan University Of Finance And Economics, Kunming, Yunnan 650500，China
Corresponding E-mail:915090822@qq.com

Abstract. Bilingual topic detection is a vital application of natural language processing in the Internet plus Era and trend of economic globalization. At present, the method of bilingual topic detection can't solve the problem of bilingual topic inconsistent distribution. Aiming at the shortcoming, this paper introduces a based on maximal clique method to find bilingual topic detection of Chinese and Thai feature words. First of all, extract the information of news with keywords of each Chinese and Thai documents through the TextRank algorithm. Next, disambiguate by means of the similarity combined with Chinese and Thai dictionary. Then, use credible association rules to cluster Chinese and Thai feature words, which generates maximal clique of bilingual topic. Finally, cluster similar maximal clique of topic to obtain the collection of final topic. According to the needs of users, the method can recommend a bilingual topic of different sizes. The test of Chinese and Thai news texts in January 2016 made good achievement. From the perspective of cross-language word clustering, the algorithm effectively solves the problem of inconsistency of bilingual topic distribution reasonably, and has the advantages of no need to estimate the number of topics and low time complexity, so it is suitable for the application of online discovery in bilingual topic.

Keywords. Chinese, Thai, maximal cliques, credible association rule, TextRank, bilingual topics detection.

1 Introduction

With the development of information technology, Internet has become an important channel for organizations or individuals to understand the trends of other countries. Now, Internet generates a massive amount of data every day, it is quite difficult that cover the useful news text, and there is a huge language barrier between Chinese and other languages. Therefore, it become a hot topic of public opinion monitoring, information competition, bilingual information retrieval that obtain the timely and effective foreign topic information in a massive amount of data. At present, in the field of cross-language topic detection, the mainly approaches are machine translation [1] or bilingual dictionary [2] and parallel corpus in the perspective of solving the language barrier. The method of based on machine translation or bilingual dictionary, it transforms cross-language texts into the texts of monolingual, then it analyzes

topics of single language. BNN (2000) proposed a Chinese-English topic detection and tracking system based on Machine Translation [1]. Ying-Ju Chen et al. (2002) used word vector cosine similarity based on English-Chinese dictionary to calculate the multi-language related topic detection, and they compared similarity of topics with Chinese to English and English to Chinese and not translation etc [3]. The method of based on parallel corpus, such as the LDA of bilingual, it is obtained the topic distribution of cross-language [4,5]. Mimno D et al. (2009) presented the multi-language topic model based on the extension LDA, and it had topic detection effect slightly worse than the monolingual LDA in the condition of large language differences.

The above methods had achieved some effect. But they didn't solve the bilingual topic inconsistent problem of feature word distribution in the process of topic detection, and the number of topics must be predicted in advance. The language barrier in the cross-language topic detection increases the difficulty of accurately predicting the number of topic. Therefore, the above methods are less effective, and they are not suitable for the application of the topic detection online. In order to solve the above problems, this paper found a method of Chinese-Thai bilingual topic detection online base on maximal clique clustering. First of all, this method adopts TextRank algorithm [6] to extract keywords of Chinese-Thai news text, then combination with named entities and the first paragraph information of topic generated feature word sequence of news. Next, the use of bilingual dictionaries and similarity disambiguation algorithm obtain the co-occurrence correlation relationship of the Chinese-Thai feature word in the news corpus to construct the co-occurrence adjacency matrix of the Chinese-Thai feature word, then to obtain the set of initial maximum clique topic through the maximum clique mining algorithm. Finally, to cluster similar maximal clique of topic obtain the collection of final topic. This paper provides a description of the ordered bilingual news topics with different sizes. In this paper, the experiment by Chinese-Thai news corpus of January 2016, compared the different generation ways of maximum clique topic and the comparison results with other methods of bilingual topic detection. The experimental results show that the time efficiency of the method is the best, and the F value is 69.03%, which is achieved by the method of this paper.

2 The cross-language topic detection online process

2.1 Extraction news keywords and generating news information by TextRank algorithm

To express the same theme, there is a certain semantic relationships between words in the same text words. In this paper, it extracts Chinese-Thai news keywords in each news by TextRank algorithm. The essence of the algorithm is a sorting algorithm based on graph. It seen the Chinese-Thai vocabulary as vertices in the graph and reflect the relationship between them by undirected edges between the vertices.

According to basic theory of Sort algorithm based on graph, you can establish a connection between the words having a semantic relationship, then to calculate number of scores words according to mutual "vote" between the words recursive, each word of the score depends on the other word vote and fraction size of it, finally, select high-scoring words for the important words. And which does not have any words to connect words is an isolated point [6,11]. Mihalcea R and others have already proven that whether the graph have directions will not have much impact on keyword extraction. In this paper, I adopt undirected graph model. To calculate each word score by using iteration in the figure, the higher the score is, the more important to the word is. Finally Each Chinese-Thai News Documents obtained most important keywords before K corresponds. Candidate word score is calculated as shown in Equation 1:

$$WS(V_i) = (1-d) + d * \sum_{V_j \in Link(V_i)} \frac{w_{ji}}{\sum_{V_k \in Link(V_j)} w_{jk}} WS(V_j) \qquad (1)$$

Where $WS(V_i)$ is the vertex V_j (That word score of J), d is the ratio of two parts of the score distribution, $Link(V_i)$ is a cable of the point of the set of points.

2.2 Based on credit association rule mining the topic maximum clique of Chinese-Thai cross-language

The theoretical basis for mining topic maximum clique is the co-occurrence relationship of the related topic feature words. This paper addressed the statistical problem of the Chinese-Thai feature word's co-occurrence relationship by Chinese-Thai bilingual dictionary and the disambiguation algorithm of Chinese candidate translation similarity.

2.2.1 Chinese-Thai feature words similarity disambiguation algorithm

In this paper, it considers translation of Thai to Chinese. First of all, it makes part-of-speech tagging and entities realize in Thai text through the lexical analysis tool of the laboratory development [19]. Through the process of 2.1 to generate a news feature words sequence, then words of the words sequence find the Chinese translation words with part-of-speech based on bilingual dictionary. Because of the difference of Chinese and Thai language, there are some ambiguous words in Chinese-Thai test feature words. Through TextRank algorithm to find 5 unambiguous Thai nouns of the highest score vote for Thai translation words, and the highest score indicates that the five nouns are the most closely related to the translation word. To translate the five nouns into Chinese words, then to judge the similarity between the candidate word and the five Chinese words, the word of the highest similarity is the Chinese translation word. Such as, the Thai word of นิวเคลียร์ has two meanings in the Thai-Chinese bilingual dictionaries, and the two meanings are atomic nucleus and cell nucleus [17]. But words if the five unambiguous Thai nouns translated into Chinese word for North Korea, country, Jeong-eun Kim, weapons and earthquake. It is obvious that the word distribution similarity between atomic nucleus and the five related words is higher than cell nucleus. So, it judges the word of Chinese translation as atomic nucleus.

In this paper, it calculated the word similarity through the Google word2vec [12] tool. The word2vec is a open-source tool for Google in 2013, it can be used to effectively characterize the real-value vector from word, then the similarity of vector space can be used to represent the semantic similarity of word. Using word2vec to train the news corpus of Sogou Laboratory. Each word is represented by 200-dimensional word vector distribution, and using cosine similarity to calculate the similarity between any two words.

2.2.2 The Chinese-Thai mapping dictionary of one-to-many

The differences of language culture between Chinese and Thai and the phenomena of logogram omission in news make some topic feature words between them be not the simple relationship of word alignment. There exists a phenomenon of verbal nominalization [13] and compound words formed by a combination of simple words [14] in Thai. For example, "express delivery" is a word in Chinese, while its meaning is "**จดหมาย**/piece, **ด่วน**/fast (this means something fast.)". Because some Chinese phrases correspond to a word of Thai, at the same time some Thai phrases may also correspond to a word of Chinese, this paper developed a special vocabulary of Chinese and Thai. The vocabulary contained some Chinese-Thai dictionaries which had segmentation of different sizes but the same meaning from Thai to Chinese. It was suitable for the translation of Chinese-Thai words in bilingual news. Some words are shown in Table 1.

Table 1. The one-to-many bilingual dictionary of Chinese-Thai.

Thai word	Chinese word
วัน day / เสาร six	Saturday
คน man / จีน China	Chinese
ตลาด bazaar / ขาย sell / ดอกไม flower	flower fair

คน man / โบราณ ancient times	the ancients
โทรศัพท์	call / phone
โทรกลับ	call back / phone

2.2.3 The semantic similarity translation of Chinese-Thai words

Because it is different of the description way and level of detail for the same event in news text of Chinese and Thai. It leads to the differences of news words in the two languages. Since the WordNet could better reflect the semantic relationship between words and the semantic information resource of the English WordNet was abundant, this paper used the synset_id correspondence of the WordNet synonym sets of different language versions to build the correspondence between the Chinese-Thai WordNet and the English WordNet, thus inquiring the synonymous and part-whole semantic relationship of the Chinese-Thai words on the English WordNet.

2.2.4 The construction of two-item credible set of Chinese-Thai feature words

The trusted association rule and the maximum clique algorithm [9] construct the two-item credible set and the trusted association rule according to the adjacency matrix. This paper did the processing of deduplication on the sequence set of Chinese-Thai feature words generated in section 2.1, and then the paper found the Chinese words corresponded to the Thai words according to the section 2.2.1-2.2.3, thus forming Chinese-Thai equivalent word pairs. If the unique word in Chinese or Thai didn't have a suitable translation item, the paper characterized it only in the form of Chinese or Thai, thus forming a characteristic equivalent vocabulary. The vocabulary was numbered sequentially. The paper regarded the adjacency matrix as A, the items of A as a_{ij}. Among them, i was the row index and j was the column index. They both started from 0. i and j both represented the id of a word in the Chinese-Thai equivalent vocabulary. The id was the characteristic equivalent word pairs of i, denoted $wordi$. If the part of Chinese or Thai in the equivalent word pairs appeared in the corpus, it was considered that the equivalent word appeared. The paper constructed the a_{ij} as follows:

(1)When $i = j$, if the $wordi$ appeared in the text of a document, a_{ii} + 1. If the $wordi$ appeared in the title of a document or the first sentence of the first paragraph, $a_{ij}+2$. Each document only counted once for a_{ii}, and finally the maximum bonus was taken. For example, the $wordi$ appeared twice in a document, once in the title and the other in the text, this document added 2 to a_{ii}, not repetitive;

(2)When $i \neq j$, if the $wordi$ and the $wordj$ never appeared in the same news, the a_{ij} was 0;

(3)When $i \neq j$, if the $wordi$ and the $wordj$ simultaneously appeared in the title of a news text, $a_{ij}+2$. If the $wordi$ and the $wordj$ appeared in the text or only one word appeared in the title or the first sentence of the first paragraph, while the other word appeared in the text, $a_{ij}+1$. Similarly, each document only added 1 to a_{ij} and the maximum bonus was also taken.

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n-1} & a_{1n} \\ a_{21} & \cdots & & & \\ \cdots & & a_{ij} & & \\ a_{n-11} & & & \cdots & \\ a_{n1} & a_{n2} & \cdots & a_{nn-1} & a_{nn} \end{bmatrix}$$

The paper constructed the adjacency matrix A through the above process. a_{ii} reflected the text frequency of the $wordi$ appearing in the Chinese-Thai news corpus. a_{ij} reflected the text frequency of the $wordi$ and the $wordj$ appearing simultaneously. The paper further constructed the two-item credible set of matrix C through the adjacency

matrix A. The element c_{ij} of C reflected the co-occurrence relationship of the $wordi$ and the $wordj$. The initial definition of c_{ij} was calculated as follows:

$$C_{ij} = \frac{a_{ij}}{a_{ii} + a_{jj} - a_{ij+1}} \qquad (2)$$

Then the paper determined the c_{ij} was greater than the lowest confidence level or not. If the value of c_{ij} was greater than θ, there existed a trusted relationship between the $wordi$ and the $wordj$ and the c_{ij} was updated as 1, otherwise the c_{ij} was 0. Finally, the paper generated the two-item credible set of matrix C. The paper could find all the two-item feature word trusted relationship sets through the traversal on the matrix C.

2.2.5 Unearthing the maximum clique of topic feature words:

The paper regarded the two-item credible set matrix generated in the progress above as C_2. Since a topic needed more feature words to express, the paper needed to complete a credible set of items from k to $k+1$, that is, the matrix $C_k \to C_{k+1}$. This paper generated the k-item credible set through the following algorithm:

(1)The initial state was C_2, which was the two-item credible set obtained through calculation;

(2)This step required setting up a generating from C_k to C_{k+1}. Firstly, the paper arbitrarily selected two k-item sets. If the front $k-1$ items were completely the same and the last item of the two k-item sets appeared in the two-item credible set, the front $k-1$ items and the last item of the two k-item sets were regarded as the candidate set of C_{k+1}. Then the paper determined that all other items were credible or not. If they were credible, the paper generated C_{k+1} and marked all of its k-item subsets. When all the work was completed, if there were k-item sets which could not be trusted, they could not be merged.

2.3 The maximum clique topic generated the quasi maximum clique topic through clustering

This paper had a lot of topics through analyzing the topics generated in the above process. However, because the language culture between Chinese and Thai, the translation deviation of media and the public opinion of the two countries are different about the angle and the degree of emotional stance for the same topic, it leads to the different representations between Chinese and Thai for the same topic, that is, the topic distributions are inconsistent. For example, some Chinese-Thai news topics were shown in Table 2.

Table 2. The maximum clique news topic.

1.	North Korea / **เกาหลีเหนือ**, hydrogen bomb / **ระเบิดไฮโดรเจน**, test / **การทดสอบ**, success / **ประสบความสำเร็จ**, arrangement / **รูปแบบ**, security / **การรักษาความปลอดภัย**, America / **ในสหรัฐอเมริกา**, menace / **การคุกคาม**, Russia / **รัสเซีย**, Republic of Korea / **เกาหลี**, European Union / **สหภาพยุโรป**, expression / **กล่าวว่า**, condemn / **ประณาม**,sanction / **การลงโทษ**
2.	North Korea / **เกาหลีเหนือ**, hydrogen bomb / **ระเบิดไฮโดรเจน**, test / **การทดสอบ**, China / **จีน**, Ministry of Foreign Affairs / **กระทรวงการต่างประเทศ**, statement / **ประกาศ**, oppose / **การต่อต้าน**, urge / **เรียกร้องให้**, denuclearization / **โครงการอาวุธนิวเคลียร์**, peninsula / **คาบสมุทร**, obligation / **หน้าที่**, strive / **ความพยายาม**
3.	Jeong-eun Kim / **คิมจองอึน**, North Korea / **เกาหลีเหนือ**, international / **ระหว่างประเทศ**, test / **ทดสอบ**, hydrogen bomb / **ระเบิดไฮโดรเจน**, celebrate / **เฉลิมฉลอง**, commend / **ยกย่อง**

This paper did clustering on the topics meeting certain similarity to create a quasi maximum clique. The topic similarity was calculated as follows:

$$\text{sim}ilarity(topici, topicj) = \frac{wi \cap wj}{\min(wi, wj)} \qquad (3)$$

Among them, $wi \cap wj$ was the number of synonymous and similar words between the two topics. If there was a pair of synonyms, $wi \cap wj + 1$. If there was a pair of similar words, such as China and the United States, opposition and condemnation, $wi \cap wj + 0.5$. The paper inquired the similarity of Chinese words mainly through the Word2vec. If the similarity of the two words exceeded 0.5, the two words were similar. $\min(wi, wj)$ was the number of words in the word-fewer document. The steps of merging maximum clique were as follows:

(1)Firstly, the paper loaded all maximum clique sets. Each group was placed into a word queue and the initial weight of each word was counted as 1. Then the data structure was placed into a list;

(4)The paper took the first group from the list, and calculated its similarity with other groups;

(3)If they were not similar, the paper placed the first group into the result set and returned back to the step 2 to continue;

(2)If the first group was similar with the m -th group after calculating through the formula 3, the paper did clustering statistics on the words of these two groups. The specific practices were as follows: The weight of the same word was added 1. Similar words were combined into a phrase, and then the weight was added 0.5. The weigh of the words neither different nor similar was changeless. The paper filtered the word sequence after statistics and abandoned the word whose weight was less than 1.5. The filtered phrase was added into the new group. Then the new group was placed into the m -th position of the list and the original group was replaced. Then the paper deleted the first group and returned back to the step 2 to continue;

(5)These steps ended until the list was empty.

The paper merged the bilingual topics whose distribution of feature words was inconsistent but satisfied the semantic relationship through the clustering process of maximum clique. The description purity of the quasi maximum clique topic was more higher after merger. It reduced the impact of noise word on the topic description and truthfully represented the bilingual content of the topic between Chinese and Thai.

2.4 Personalized recommendation of topics

Now, Internet will produce massive data everyday. But people often have no way to start, in the face of a various data. In this paper, it designed a method of Chinese-Thai bilingual topic detection online, and the method recommended the description of expectant maximum clique topic to the user after the merger of topics. If the user is interested in a topic, then the method pushes a maximum clique queue of the topic to the user. The method calculates the similarity between bilingual topics of the maximum clique queue and the news corpus texts, and the news text time of the highest similarity is defined as the topic build time. The method provides user with a topics set of the topic, sort it by topic build time. Personalized recommendation result of the North Korea H-bomb test as shown in Table 3.

Table 3. The personalized recommendation example for bilingual news.

quasi maximal group topic	langua ge	before clustering topic
North Korea, hydrogen bomb, **เกาหลีเหนือ**[3], **ระเบิดไฮโดรเจน**[3],	Chinese:	1.(01/07/2016 08:34) North Korea, hydrogen bomb, test, success, arrangement, security, America, menace, Russia, Republic of Korea, European Union, expression, condemn, sanction
test,	Thai:	**เกาหลีเหนือ**, **ระเบิดไฮโดรเจน**, **การทดสอบ**,

การทดสอบ[3], { America, China, Russia,		ประสบความสำเร็จ, รูปแบบ, การรักษาความปลอดภัย, ในสหรัฐอเมริกา, การคุกคาม, รัสเซีย, เกาหลี, สหภาพยุโรป, กล่าวว่า, ประณาม, การลงโทษ
{ ในสหรัฐอเมริกา, จีน, รัสเซีย, Republic of Korea, European Union },	Chinese:	2.(01/08/2016 17:11) North Korea, hydrogen bomb, test, China, Ministry of Foreign Affairs, statement, oppose, urge, denuclearization, peninsula, obligation, strive
เกาหลี, สหภาพยุโรป}[3] , { expression, statement }, { กล่าวว่า, ประกาศ}[1.5] ,	Thai:	เกาหลีเหนือ, ระเบิดไฮโดรเจน, การทดสอบ, จีน, กระทรวงการต่างประเทศ, ประกาศ, การต่อต้าน, เรียกร้องให้, โครงการอาวุธนิวเคลียร์, คาบสมุทร, หน้าที่, ความพยายาม
{ condemn, oppose }, { ประณาม, การต่อต้าน}[1.5] ,	Chinese:	3. (01/12/2016 10:00) Jeong-eun Kim, North Korea, international, test, hydrogen bomb, celebrate, commend
{ celebrate, commend } { เฉลิมฉลอง, ยกย่อง}[1.5]	Thai:	คิมจองอึน, เกาหลีเหนือ, ระหว่างประเทศ, ทดสอบ, ระเบิดไฮโดรเจน, เฉลิมฉลอง, ยกย่อง

3 Experiments and results

In order to investigate the effect of a bilingual topic for online discovery, this paper has designed two sets of tasks.1. Through different keyword extraction methods to compare the effect of a maximum clique of bilingual topic for discovery; 2. Compare this method with other mainstream bilingual topic discovery methods. The experimental corpus has selected 161 bilingual topic news corpus from the internet between China and Thailand in Jan 2016, which contains 86 Chinese News and 75 Thai News. The experimental corpus involve 16 aspects of international, ASEAN, economic and trade, cultural and entertainment four topics.

Firstly, through different keyword extraction methods to compare the effect of a maximum clique of bilingual topic for discovery. The experiment are planed to compare the effects of generating the topic through three different ways: TextRank keyword extraction method, mainstream TF-IDF keyword extraction algorithm and not extracting keyword. TextRank keyword extraction method sets 5 co-occurrence window, d of the damping coefficient is 0.85, and confidence level of credible association rule is 0.22. The evaluation indicators include the precision rate of topic detection, recall rate and F-measure.

Table 4. Results of different extraction methods.

index / method	Precision rate	Recall rate	F-measure
not extracting keyword	67.09%	73.96 %	70.35%
TF-IDF	54.62%	62.37 %	58.23%
TextRank	65.54%	72.85 %	69.03%

Based on the Table 4, it is showed that the effect of TF-IDF keyword extraction algorithm is the worst but not extracting key words is the best. The reason of this result is that when we use TF-IDF in a text keyword extraction process, we need to consider the word frequency in the anti-external knowledge of other text. However, this will inevitably result in characteristic value deviations in the process of lexical translation between Chinese and Thai. This paper thinks that TF-IDF is not suitable to describe the task of a single text, it is suitable for the task of finding out the characteristic of text difference. However maximum clique method need to find out the commonness between the text from Chinese-Thai news, TextRank is more suitable for the discovery of a bilingual group than TF-IDF. Full text vocabulary without the extraction of keywords after removal of the stop words and generating the topic of the maximum clique contains more comprehensive information, so it's the best. But because it also contains some noise and translation errors, the final F value is only slightly better than TextRank. Considering that not extracting key words will increase the dimension of adjacency matrix and the number of translation disambiguation, which greatly increased time complexity. In this paper, we use TextRank keyword extraction method to generate topic.

In order to verify the effectiveness and performance of this experiment, Experiment 2 is based on machine Translation k-means [16] text clustering and bilingual LDA [4,5] to compare the method of this paper with the mainstream of bilingual topic. The text clustering and bilingual LDA methods need to point out the number of K values in advance, K value is set to 16. This paper select 621 on the Chinese encyclopedia Chinese-Thai parallel texts from China Radio International as the training data set of LDA model. $\alpha = 50/K$, $\beta = 0.01$. LDA method can only filter the topic distribution probability weight of more than 0.03 of document.

Table 5. Results of bilingual topic detection methods.

Index / Method	Precision rate	Recall rate	F-measure	Consuming time
Text Clustering (k-means)	51.89%	45.31%	48.37%	819.53s
Bilingual LDA	62.13%	69.20%	65.46%	118.72s
TextRank	65.54%	72.85%	69.03%	96.48s

We can find out the Table 5:

(1)The method of this paper is the highest value of F, then bilingual LDA , the last is text clustering. When we use the text clustering, we need to compare the similarity between the topic vector and the document vector and the document vector. Due to the feature vector clustering on text level need to cross, it inevitably increases the noise characteristics and translation errors.

(2)The effect of bilingual LDA bilingual topic discovery is not very ideal. The analysis considers that the training data set of LDA does not cover all the topic distribution and its ability to discover new topics is poor. Each of these topics has a fixed number of words, while the topic in the actual description of the number is not fixed. The alignment effect of the bilingual topic is entirely dependent on the quantity and quality of the training set of the LDA parallel corpus. Despite the massive collection of Chinese-Thai News parallel text training topic model can improve the F value of bilingual LDA, the Internet Chinese-Thai parallel corpus, especially news parallel corpus resources are scarce and we need to invest more in human and linguistic knowledge. The instability of quality parallel materials influence the effect of bilingual LDA.

(3)In time complexity, the method is the lowest. Owing to the extraction of the keywords of the news document, we can significantly reduce the number of word processing objects and do not need to carry out the convergence process of k-means and LDA text iteration. This method is most suitable for a bilingual online topic discovery application.

4 Conclusion

In this paper, we cluster the cross language words to generate a bilingual topic through maximal cluster. The method of this paper is a reasonable solution to the problem of the distribution of the topic in the bilingual environment and it is effective to solve the traditional topic detection methods to estimate the number of topics at the same time. The best results are obtained when the time complexity is lowest. This method provides a reasonable way of thinking for the application of bilingual topic on line. The next step is to consider the time factor into the topic discovery model and research the more reasonable way to solve a topic distribution of feature words inconsistent problem. We look forward to a better effect on the discovery of bilingual topics.

Acknowledgment

This paper is supported by China National Natural Science Foundation （61562049), （61662040）.

References

1. Leek, T., Jin, H., Sista, S., & Schwartz, R. (2000). The BBN crosslingual topic detection and tracking system. Tdt Evaluation System Summary Papers, 894--01.
2. Wactlar, H. D. (1999). New directions in video information extraction and summarization. Proceedings of Delos Workshop, 64(8), 229–232.
3. Y.J. Chen and H.H. Chen(2002). NLP and IR approaches to monolingual and multilingual link detection. Proceedings of the 19th international conference on Computational linguistics-Volume 1, pages 1–7.
4. Mimno, D., Wallach, H. M., Naradowsky, J., Smith, D. A., & Mccallum, A. (2009). Polylingual Topic Models. Conference on Empirical Methods in Natural Language Processing, EMNLP 2009, 6-7 August 2009, Singapore, A Meeting of Sigdat, A Special Interest Group of the ACL(Vol.2, pp.880--889).
5. Ni, X., Sun, J. T., Hu, J., & Chen, Z. (2009). Mining multilingual topics from wikipedia. Www, 1155-1156.
6. Mihalcea, R., & Tarau, P. (2004). TextRank: Bringing Order into Texts.Conference on Empirical Methods in Natural Language Processing, EMNLP 2004, A Meeting of Sigdat, A Special Interest Group of the Acl, Held in Conjunction with ACL 2004, 25-26 July 2004, Barcelona, Spain(pp.404-411).
7. Agrawal, R., Imieliński, Tomasz, & Swami, A. (1993). Mining association rules between sets of items in large databases. Acm Sigmod Record,22(2), 207-216.
8. Agrawal, R., & Srikant, R. (1994). Fast Algorithms for Mining Association Rules in Large Databases. International Conference on Very Large Data Bases (Vol.7, pp.487-499). Morgan Kaufmann Publishers Inc.
9. Bo, X., & Beijing. (2008). Credible association rule and its mining algorithm based on maximum clique. Journal of Software, 19(10), 2597-2610.
10. Chunying, L. I., Tang, Y., Tang, Z., Huang, Y., Yuan, C., & Zhao, J. (2015). Community detection model in large scale academic social networks. Journal of Computer Applications.
11. Jie Yang(2009). The research of keyword extraction technology in multi-document. Shenyang Institute of Aeronautical Engineering.
12. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionfouality. Advances in Neural Information Processing Systems,26, 3111-3119.
13. Yiyuan Luo(2011). Based on Thai. World Publishing Corporation.
14. Xiaorui Pei(2001). New Thai grammar. Peking University Press.
15. Thoongsup, S., Robkop, K., Mokarat, C., Tan, S., Charoenporn, T., & Sornlertlamvanich, V., et al. (2009). Thai wordnet construction. 139-144.

16. Yuan, F., Zhou, Z., & Song, X. (2007). K-means clustering algorithm with meliorated initial center. Computer Engineering, 33(3), 65-66.
17. Chen, A. L., Tang, C. J., Tao, H. C., Yuan, C. A., & Xie, F. J. (2004). An improved algorithm based on maximum clique and fp-tree for mining association rules. Journal of Software, 15(8), 1198-1207.
18. ----"Thai Lexical Analysis Tool".
    </>.
19. ----"Thai-Chinese Dictionary".
    <http://www.thai-language.com/>.
20. Han, Z. M., Zhang H., Zhang M. (2015). Fast topic detection and evaluation towards massive short texts. Computer Application Research, 32(3), 717-722.