A New Synthesizing Cluster Labels Algorithm for Thai Web Search Results

Nawaporn Leardtharatat[1] and Worapoj Kreesuradej[2]
Faculty of Information Technology
King Mongkut's Institute of Technology Ladkrabang University
Bangkok 10520, Thailand
E-mail: [1]nawaporn@dss.go.th, [2]worapoj@it.kmitl.ac.th

Abstract: Currently, there is a large amount of information on internet. Users usually find information using search engine. But, number of results returned from search engine is often irrelevant information to users wanted. As a solution to the problem, clustering web search results will help users finding relevant information to user's goal quickly. Suffix tree clustering (STC) technique is the most known for clustering web search results. However, when the technique is applied to cluster web search results for Thai language web pages, STC technique usually gives incomplete phrases for cluster labels. Therefore, this paper is proposed a new approach for synthesis cluster labels obtained from STC technique. The proposed technique can give more readable and complete phrases for cluster labels than that from STC technique.

1. Introduction

Search engines such as Google [9] and Yahoo [10] often return a long list of search results. Users are often forced to sift through to find a long ordered list of snippets returned by search engines. As a result, finding information using those web search engines is not always successful. Web search result clustering is an approach to deal with such problem. Several web search result clustering methods for English web pages such as Carrot [11] and SHOC [3] have been proposed. However, those web search result clustering methods can not work efficiently when are applied to Thai web pages. Those clustering methods usually give incomplete Thai phrases as cluster labels. Thus, this paper proposed a new approach for synthesizing cluster labels for Thai web search results.

Our approach is composed of three major steps. The first step is preprocessing segment all the phrase from snippet using maximal matching algorithm [5] and removing stopwords. Then, the second step is clustering snippet using suffix tree clustering algorithm and ranking cluster labels. The third step is synthesizing cluster labels.

2. Suffix Tree Clustering

Many algorithms clustering web search results have been developed including K-means [6] or Hierarchical agglomerative clustering [6], cluster Web search results based on hyperlinks [6] and Suffix Tree Clustering (STC) [1], present by Zamir and O. Etzioni's [2] STC automatically group Web search results and works in two main phases: base cluster discovery phase that contain at least one phrase in common and using Minimal Base Cluster Score and base cluster merging phase using common document, it uses phrases rather than words and allow clusters to overlap.

3. Web search result clustering for Thai web pages

Unlike English sentences, Thai words in a sentence are written with no separation between each word. A segmentation method is necessary to obtain Thai words from a sentence. Thus, web search result clustering methods for English language can not be applied to Thai web pages directly.

Our algorithm is composed of four phases: (1) Preprocessing, (2) Discovering base Clusters, (3) Discovering common phrases using a join phrase algorithm and (4) Ranking phrase label. The algorithm is shown in Figure 1.

```
Search result fetching and split snippet
Phase 1   Preprocessing
    - The non-word tokens are strip
      (HTML Tag, punctuation, number etc.)
    - Word segment of Thai snippet into tokens on space with
      maximal matching algorithm
    - Remove Stop words , if the English word are stem tokens
      to "root" word
Phase 2 Discovering Base Clusters
  2.1 Creation of a suffix tree with n-gram technique of all
      sentence
      for each sentence {
        split sentence into n-gram (phrase)
        for each phrase{
          insert phrase into each node of suffix tree
          update internal node with the index to current
          document while rearranging the tree have common
          phrase, number of document }
      }
  2.2 Build base cluster if each node in tree have common phrase,
      number of document > 1
Phase 3 Discovering true common phrases using a join phrase
algorithm
      for each base cluster {
  3.1 Delete cluster B if phrase cluster B is subset of phrase
      cluster A or delete document of cluster B if phrase of
      cluster B length =1 and Overlap with cluster A }
      for each base cluster {
  3.2 Joint base cluster A and B if phrase cluster A and B length
      > n-gram and number of document in cluster is subset and
      word's position in document in pair cluster is join }
```

Figure 1. As shown the pseudo-code of web search result clustering for thai web pages.

3. 1 Preprocessinng

First, all snippets returned by a search engine are detected for duplicated snippets. Then, the second step is to segment Thai words using maximal matching algorithm. This

algorithm first generates all possible segmentations for a sentence and then selects the one that contains the fewest words. As an example, "ไปหามเหสี" (go to see the queen) is segmented as ไป (go) หา (to see) มเหสี (the queen). As the last step, all stop words are removed.

3. 2 Discovering Base Clusters

Base Clusters are found by using suffix tree with n-gram technique. As a example, a set of 4 snippets shown in Table 1 are used to find base clusters. A suffix tree with n-gram <=3 gram shown in Figure 2 is constructed. Then, the internal nodes that do not contain snippets and have one linkage are compacted. The label of a compacted internal node is defined to be the concatenation of the edge-labels on the path from the root to that node. The suffix tree after compacting internal nods is shown in Figure 3. According to Figure 3, base clusters shown in Table 2, are discovered.

Table 1. List four snippets.

Document	Snippet
D0	จัดการเรียนการสอนผ่านเครือข่ายอินเทอร์เน็ต
D1	การพัฒนาระบบการจัดการเรียนการสอนผ่านทางเครือข่ายอินเทอร์เน็ต
D2	บทเรียนคอมพิวเตอร์ช่วยสอนแบบออนไลน์
D3	บทเรียนคอมพิวเตอร์ผ่านอินเทอร์เน็ต

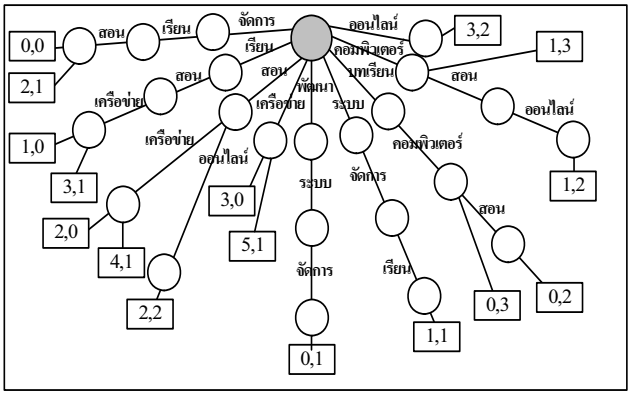

Figure 2. A suffix tree with n-gram <= 3.

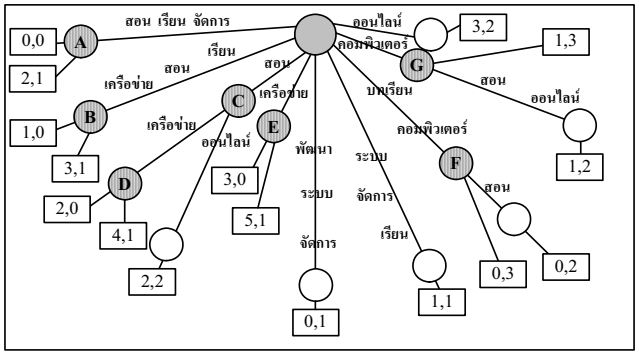

Figure 3. The suffix tree after compacting internal nodes.

From Table 2. Base clusters have some overlap of snippets. An overlap of irrelevance snippets in multiple clusters can hurt cluster quality.

Table 2. List of base clustering.

Cluster	Node	Label base cluster	Document (position,document)
1	A	จัดการ (manage) เรียน (learn) สอน(teach)	(0,0),(2,1)
2	B	เรียน (learn) สอน(teach) เครือข่าย(network)	(1,0),(3,1)
3	C	สอน (teach)	(2,0), (4,1),(2,2)
4	D	สอน (teach) เครือข่าย (network)	(2,0),(4,1)
5	E	เครือข่าย (network)	(3,0),(5,1)
6	F	บทเรียน (tutorial) คอมพิวเตอร์ (computer)	(0,2),(0,3)
7	G	คอมพิวเตอร์(computer)	(1,2),(1,3)

In addition, STC with n-gram may give too many base clusters and STC with n-gram can not discover a true common phrase when the length of n-gram is shorter than the length of true common phrases. Thus, some base clusters need to be merged together and their labels need to be joined together.

3. 3 Type-style and fonts Discover true common phrases using a join phrase algorithm

According to Jongkol [4]., join phrase equation is shown in (1).

$$A \oplus B = \left\{ \begin{array}{ll} a_0 & \oplus \\ a_1 = b_0 & \\ a_2 = b_1 & \\ \vdots & \\ a_n = b_{n-1} & \\ \oplus & b_n \end{array} \right\} if \ \mathrm{A}_{(d)} \in \mathrm{B}_{(d)} > 1 \qquad (1)$$

Where A and B is base clusters, $A_{(d)}$ is snippets in A cluster, $B_{(d)}$ is snippet in B cluster, $\{a_0, a_1, \ldots, a_n\}$ is a set of terms that appear in label A cluster and $\{b_0, b_1, \ldots, b_n\}$ is the sets of terms that appeared in label of B cluster. Results of combining pairs of similarity base clusters shown in Table 3

Table 3. List of base clusters is joint phrase.

Cluster	Label base cluster	Document
1	จัดการ (manage) เรียน (learn) สอน (teach) เครือข่าย(network)	0, 1
2	บทเรียน (tutorial) คอมพิวเตอร์ (computer)	2, 3

After merged base clusters, all base clusters are scored by the following equation:

$$S(c) = |d| * f(|p|) * \sum_{i=1}^{|d|} tfidf(p, d_i) \qquad (2)$$

$$\mathrm{f}(|\mathrm{p}|) = \begin{cases} 0, \text{if} |\mathrm{p}| = 1 \\ |\mathrm{p}|, \text{if} 2 \le |\mathrm{P}| \le 6 \\ \propto, \text{if} |\mathrm{p}| > 6 \end{cases}$$

Equation (2) is calculated for cluster labels, balance the length of the phrase where $|d|$ is the number of snippet in p cluster, $|p|$ is the number of word in p phrase of cluster and $tfidf(p,d)$ is an inverse phrase frequency.

Score of cluster labels use ranking cluster. Ranking cluster is reordering cluster according to their interesting scores and present for user.

4. Synthesizing Thai cluster label

Cluster labels obtained from previous step are not complete phrase. The cluster labels that obtained from previous step are just separate Thai keywords. The cluster labels are often unreadable. Thus, this paper proposes a new approach to synthesize cluster labels. Our approach is based on the assumption that a good label for a Thai web search result cluster is a longest common phase of snippets in the cluster. Therefore, a new algorithm for synthesizing Thai cluster labels is proposed in this paper. Here, the algorithm has two steps. In the first step, longest left-right common phrase are discovered and the second step, the longest left-right common phrases are combined into cluster labels.

4. 1 Discovering the longest left-right common phrases

To discover the longest common phrases, the left longest common phrases and right longest common phrases are found separately. Here, we propose a phrase tree to obtain the left longest common phrases or right longest common phrases. As an example for constructing left and right phrase trees, "พัฒนา (develop) ระบบ (system) การ (act) จัดการ (manage) เรียน (learn) การ สอน (tech) ผ่าน (pass) เครือข่าย (network) อินเทอร์เน็ต (internet)" (a snippet) of "จัดการ (manage) เรียน(learn) สอน (tech) เครือข่าย (network)" (a cluster label) can be formed left and right phrase tree as shown in Figure 4. From the Figure, the left phrases tree is constructed by selecting the first word of the cluster label to the first word of the snippet. Similarly, the left phrases tree is constructed by selecting the first word of a cluster label to the last word of the snippet. Then, a longest common phrase can be obtained from a node in phrase trees that contain the most snippets. As an example, the results of the longest left-right common phrases are shown in Table 4.

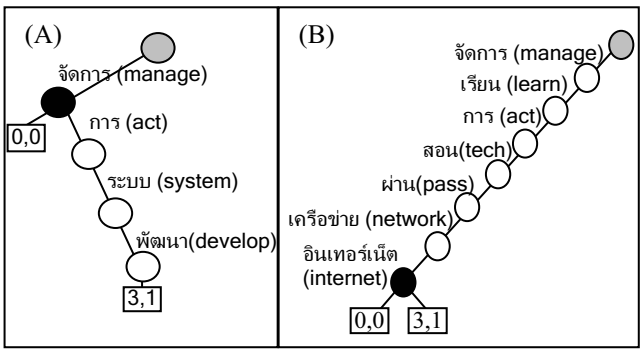

Figure 4. (A) left common phrase tree, (B) right common phrase tree.

Table 4. The longest left-right common phrases.

Longest left common phrase	Longest right common phrase
จัดการ (manage)	จัดการ (manage) เรียน (learn) การ สอน (tech) ผ่าน (pass) เครือข่าย (network) อินเทอร์เน็ต (internet)

4. 2 Discovering longest common phrases

A longest common phrase can be found by joining the longest left and right common phrases together. As a example, the longest left-right common phrases from Table 4. can be joined as shown in Figure 5.

Figure 5. Joining longest left-right common phrase.

5. Experimental result

Due to lacking of standard dataset for testing web search results clustering, we have to build a test dataset. For this purpose, we have collected search result dataset by querying from Sanook.com which is a Thai portal website. We select three types of queries: ambiguous queries such as "พวงมาลัย" which means a wheel or lei of flower in Thai language, entity names such as "การเขียนโปรแกรม" which means programming in Thai language and general terms such as "แผนที่" which means a map in Thai language. The test dataset from the query collection consists of 2,490 snippets.

We use 3 methods for evaluate performance our system is modified from Jianchao Li and Tianfang Yao at [7].

They are cluster label quality, cluster overlap and cluster precision. Cluster label quality are calculation of some manually synthesizing cluster label result sets for different queries. Cluster label quality measure the reasonable and readability of the cluster labels set by our method, such that

$$Q = M/N \quad (3)$$

Where M is the number of readable cluster labels and N is total number of cluster labels in of cluster labels. Readable cluster labels as cluster label is meaning and present information in cluster.

The cluster overlap describe snippet to more than one group, such that

$$V = (A/S)-1 \quad (4)$$

Where A is the number of snippet assign in group and S is the total number of snippet.

The cluster precision of cluster j with class i is define as :

$$P = \sum_i \left(|i| * \frac{N_{ij}}{N_j} \right) \Big/ \sum_i |i| \qquad (5)$$

Where N_j is the number of member of class j and N_{ij} is the number of member of class i in cluster j.

To evaluate the performance of our technique, we use the top 20 ranking cluster labels from clustering of each queries from Table 6. The results are shown in Table 7. and Figure 7.

Table 7. Result of cluster labels.

Query	Cluster label Quality	Cluster Overlap	Cluster Precision
พวงมาลัย (Wheel Or Lei of Flowers)	0.85	0.47	0.89
ปาล์ม (Palm Tree Or Palm PDA)	0.90	0.11	0.86
ฟอร์ด (Ford)	1.00	0.31	0.79
การเขียนโปรแกรม (Programming)	1.00	0.18	0.70
ขนมไทย (Thai dessert)	0.95	0.27	0.79
พอเพียง (Sufficiency)	0.65	0.49	0.72
ข้าวหอมมะลิ (Dormitory jasmine rice)	0.90	0.24	0.64
ประเทศลาว (Lao)	0.95	0.43	0.90
แผนที่ (Map)	1.00	0.09	0.85
สุนัข (Dog)	0.95	0.11	0.89
Average	**0.92**	**0.27**	**0.80**

We can see that synthesizing cluster labels are readable by the cluster label quality 92% , cluster precision is about 80% .

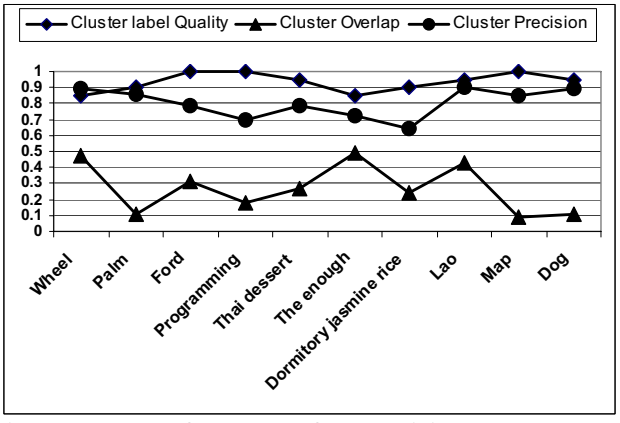

Figure 7.The performance of synthesizing cluster labels.

Table 8. shows some examples of the synthesizing cluster labels from "Programming", "Palm" and "Thai dessert" query.

Table 8. Example Synthesizing cluster labels.

Cluster Label	Synthesizing Cluster Label
Programming	
ja va sk ri (จา วา สค ริ)	java script (จาวาสคริป)
Palm	
P A Palm (พี เอ ปาล์ม)	about information PDA and Palm (ข้อมูลเกี่ยวกับพีดีเอปาล์ม)
mu Island in Surat Thani Province (เกาะ มุ จังหวัด สุราษฎร์ธานี)	Samui Island in Surat Thani Province (เกาะสมุยจังหวัดสุราษฎร์ธานี)
Thai dessert	
teach Thai food (สอน อาหาร ไทย)	teach making Thai food (สอนการทำอาหารไทย)
be ke ree (เบ เก รี่)	bakery (เบเกอรี่)

6. Conclusion

This paper is proposed a new technique for synthesizing cluster labels obtained from Suffix tree clustering (STC) technique. The proposed technique can give more readable and complete phrases for cluster labels than that from Suffix tree clustering (STC) technique. Performance of our method at cluster label quality higher about 92 % and cluster precision 80 % In the future, some comprehensive experiments will be conducted to evaluate the performance of the proposed technique.

References

[1] O. Zamir and O. Etzioni, "Document Clustering : A Feasibility Demonstration," Proceedings of the 19th International ACM SIGIR Conference on Research and Development of Information Retrieval, pp. 46-54, 1998.

[2] O. Zamir, "Clustering Web Document : A Phrase-Based Method for Grouping Search Engine Results," Doctoral Dissertation, University of Washington, 1999.

[3] D. Zhang and Y. Dong, "Semantic, Hierarchical, Online Clustering of Web Search Results," Proceeding of the 6th of Asia Pacific Web Conference (APWEB), Hangzhou, China, April 2004.

[4] J. Jongkol, "A New STC for Web Search Result Clustering," Master thesis, University of King Mongkut's Institute of Technology Ladkrabang, 2006.

[5] C. Paisarn, "Feature-Based Thai Word Segmentation," Chulalongkorn University, 1998.

[6] D. Crabtree, X. Gao and P. Andreae, "Improving Web Clustering by Cluster Selection," Proceeding of the 5th of Web Intelligence Conference (WI), Compiegne University of Technology, France, September 2005.

[7] Li Jianchao and Yao Tianfang, "An Efficient Token-based Approach for Web-Snippet Clustering," Proceeding of the Second International Conference on Semantic, Knowledge, and Grid (SKG'06).

[8] http://www.sanook.com/

[9] http://www.google.com/

[10] http://www.yahoo.com/

[11] http://www.Carrot-search.com/