The Comparison of Semantic Suffix Tree Clustering and Suffix Tree Clustering Algorithm Influence on the Accuracy Rate of an Indonesian Question Answering System

IOP Conf. Series: Materials Science and Engineering **551** (2019) 012042 doi:10.1088/1757-899X/551/1/012042

The Comparison of Semantic Suffix Tree Clustering and Suffix Tree Clustering Algorithm Influence on the Accuracy Rate of an Indonesian Question Answering System

Dininta Isnurthina, Muhammad Ihsan Jambak, and Novi Yusliani

Faculty of Computer Science, Sriwijaya University, Palembang, Indonesia.

E-mail: dini.dininta@gmail.com

Abstract. This research analyses the comparison between Semantic Suffix Tree Clustering (SSTC) algorithm and Suffix Tree Clustering (STC) algorithm in clustering documents on Indonesian Question Answering System. The fundamental difference between the two algorithms is that SSTC considers the meaning of the words to generate clusters and readable labels rather than relying only on string matching like STC. As such, SSTC is able to return a more specific and accurate clusters, and has potential to increase the accuracy rate of a question answering system. However, contrary to the hypothesis, comparison results shows that the accuracy rate of Indonesian Question Answering System after the documents is clustered by SSTC is lower than by STC. The accuracy rate degradation occurred in almost every question category, except Definition category. In average, the accuracy rate obtained by Indonesian Question Answering System with SSTC is only 23.31%, while Indonesian Question Answering System with STC is able to obtain 83% accuracy rate. This significant difference indicates that Semantic Suffix Tree Clustering algorithm is not suitable in the context of document clustering on Indonesian Question Answering System.

1. Introduction

According to Zulen and Purwarianti [1], Question answering system is a system that automatically answers questions in natural language. A question answering system can utilize a database or documents collection (local or web) as a source for its answers, and typically, it consists of three main components; question analyser, document/passage retriever, and answer finder. There has been an extensive amount of research on question answering system, but only a few considered adding document clustering component, particularly on the Indonesian-based system. Purwarianti et al. [5] developed an Indonesian question answering system for factoid questions using inverse document frequency (idf) to retrieve the source documents. The experimental result shows that idf technique is more suitable than the term frequency-inverse document frequency (tf-idf) technique. On the other hand, Yusliani [6] did a research on Indonesian non-factoid questions and the document retriever component implemented tf-idf technique alongside cosine similarity. From 90 questions collected from 10 Indonesian people and 61 source documents, the system returned MRR value of 0.7689, 0.5925, and 0.5704 for each definition, reason, and method category respectively.

Using the same samples from the two previous researches, Rahmansyah [3] experimented on a new Indonesian question answering system by adding a document clustering component, implementing Suffix Tree Clustering (STC) algorithm from Zamir and Etzioni [2]. The STC algorithm is proven to be more feasible than standard clustering methods because it makes use of common phrases found in

group of documents to cluster and label them. By adding document clustering component using STC, Rahmansyah succeeded in developing an Indonesian question answering system with 83% accuracy rate. However, it can be assumed that the question answering system's accuracy rate could be further increased by considering the semantic similarity between each words from the source documents. Janruang and Guha [4] developed Semantic Suffix Tree Clustering (SSTC) algorithm that combines string matching and semantic similarity equation to cluster documents. The result of performance comparison between each algorithm in clustering 26.800 datasets from Dmoz.com is SSTC gained a precision rate of 0.81 while STC gained 0.68.

Based on the conclusion from Janruang and Guha [4], this research focuses on the addition of document clustering component using Semantic Suffix Tree Clustering algorithm in an Indonesian question answering system, assuming that it will increase the accuracy rate, and compares it with the result from Rahmansyah [3].

2. Related Works

2.1. Suffix Tree Clustering

According to studies from Zamir and Etzioni [2], STC algorithm has a higher average precision rate than other clustering methods that rely on term frequency distribution such as Single-Pass, K-Means, Buckshot, Fractionation, and Group-Average Agglomerative Hierarichal Clustering (GAHC). This is likely due to the phrases usage by the algorithm itself, as phrases are usually more informative than an unorganized string collection. In STC, phrases can be interpreted as a sequence of one or more words. STC algorithm has two main phases, the first one is base cluster identification for all documents using suffix tree. The suffix tree model considers document $d = w_1\ w_2 \ldots\ w_m$ as a words set w_i, not characters ($i = 1, 2, \ldots, m$). The second phase is combining base clusters that overlap in one cluster.

Perera [9] implements the STC algorithm in his question answering system to cluster its information. The system is evaluated with MRR and has 0.73 average accuracy rate for 10 sets of questions where each set consists of 35 questions.

2.2. Semantic Suffix Tree Clustering

Janruang and Guha [4] compared SSTC algorithm and Semantic Lingo, a technique developed by Sameh and Kadray [7] which uses semantic similarity to extract phrases in the documents into cluster labels as well. The experiment result showed that SSTC algorithm is able to cluster the documents into a more specific and readable labels compared to Semantic Lingo. On the other hand, Shabbir et al. [8] compared SSTC to conventional Lingo algorithm by clustering web search results using 'scholarships' as its topic. After testing with 5 different queries, the average amount of cluster labels returned by SSTC is 4, due to its nature to return more specific clusters, while Lingo returned 2 clusters. In terms of precision rate, SSTC excelled with average precision of 79%, leaving Lingo with average precision of 75%.

3. Research Methodology

3.1. Data

The data that are used in this research are collected from following researches.

- 3116 pairs of Factoid questions and answers sample, 221 source documents, and research result from Purwarianti et al. [5], conducted in Department of Information and Computer Science, Toyohashi University of Technology.
- 350 pairs of Non-factoid questions and answers sample, 61 source documents and research result from Yusliani [6], conducted in School of Electrical and Informatics Engineering, Institut Teknologi Bandung.
- Research result from Rahmansyah [3], conducted in Faculty of Computer Science in Universitas Sriwijaya.

3.2. Pre-processing

Pre-processing is the first step to process data before the main process is conduced. Question inputs are pre-processed to help the Question Analyser process in obtaining the expected answer types (EAT) and keywords. Answer types are used in Answer Finder process, while keywords are needed by Document Retriever process. In this stage, there are four stages of pre-process, consist of case folding, tokenizing, stop words removal, and stemming.

3.3. Question Analyzer

The aim of Question Analyser is to analyse the questions entered by user to the system and obtain the keywords, and expected answer types (EAT).

3.3.1. Expected Answer Types (EAT).

To determine the EAT, the system will accept only eight question words in Indonesian, which are "*siapa*" (who), "*siapakah*" (who), "*apa*" (what), "*apakah*" (what), "*dimana*" (where), "*kemana*" (where), "*bagaimana*" (how), dan "*bagaimanakah*" (how). Clue words are also used to determine the EAT. This research implements the method of classifying EAT from previous research by Zulen and Purwarianti [1].

3.4. Document Retriever

The Document Retriever process will retrieve all documents that contain the keywords which has been obtained previously. The retrieving method only checks the keywords occurrences in each document. If a document has one of the keywords inside, then the document will be considered to have the required answer.

3.5. Document Clustering using SSTC

SSTC algorithm clusters semantically similar documents into the same cluster. SSTC consists of three phases; pre-processing and semantic suffix tree construction, tree pruning, and cluster identification.

3.5.1. Semantic Suffix Tree Construction.

Semantic suffix tree is a new data structure that extends the suffix tree and is constructed based on the meaning of word strings. To determine whether a pair of words is similar semantically, the synonym sets of each word are compared. Basically, if there is any intersection between the synonym sets, the pair of words is deemed similar. The synonym sets can be obtained from an online dictionary or database. The semantic similarity equation is shown as equation (1).

$$SemSim(w_a, w_b) = 1 \ if \ |synset(w_a) \cap synset(w_b)| \geq 1$$
$$SemSim(w_a, w_b) = 0 \quad otherwise \tag{1}$$

Where w_a is first word to be compared, w_b is second word to be compared, $synset(w_a)$ is synonym set of w_a, and $synset(w_b)$ is synonym set of w_b.

3.5.2. Tree Pruning.

The first semantic suffix tree constructed after the previous phase is too huge to be executed. Therefore, tree pruning process is conducted to reduce the number of nodes without omitting the concept or meaning of each node and sub-tree. The tree pruning process uses pre-order traversal that traverse from left to right to update and delete a branch using the depth-first traversal method. Additionally, after the pruning process is finished, the tree is compactified to further reduce the length of tree by merging a sub tree that has only one child or no documents and suffix link.

3.5.3. Cluster Identification.
In this algorithm, each sub-tree is assumed to be a concept cluster and each node is a member that has a set of documents. The post-order traversal technique is used to calculate cluster and label. To reduce the number of useless clusters, the similarity of each cluster is checked by comparing the documents set between two cluster and finding if there is any intersection, as shown by equation (2).

$$\begin{aligned} ClusSim(C_a, C_b) = 1 \; if \; |C_a \cap C_b| = |C_a|, \; C_a \; is \; deleted \\ or \; |C_a \cap C_b| = |C_b|, \; C_b \; is \; deleted \\ ClusSim(C_a, C_b) = 0 \quad otherwise \end{aligned} \tag{2}$$

Where C_a is first cluster to be compared and C_b is second cluster to be compared.

3.6. Answer Finder
The clustered documents become the source to find the answer for entered question. This research uses the surface expression method based on research by Yusliani [6] to find the answer.

4. Results
The experimental result is achieved by comparing answers returned by the system to the actual answers from the questions and answers collection to determine whether or not the returned answer is correct. Afterward, the percentage result is calculated using equation (3).

$$Percentage = \frac{Amount\ of\ correct\ answers}{Amounr\ of\ returned\ answers\ by\ the\ system} x\ 100\% \tag{3}$$

Subsequently the average of all percentage result from each question type is calculated to become the result of this research. The average equation is described as follows.

$$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{4}$$

Where x_i is the value of x in i index, $\bar{X}$ is average, and n is total amount of samples.
The entire percentage result of all question types is shown in the Table 1 below.

Table 1. Experimental result of Indonesian Question Answering System usingSSTC algorithm

Question Type	Returned Answers by the System	Correct Answers	Percentage (%)
Factoid			
People	546	82	15,01
Organization	527	120	22,77
Location	524	123	23,47
Name	508	86	16,92
Date/Time	506	138	27,2
Quantity	505	119	23,56
Non-Factoid			
Definition	110	83	75,45
Reason	119	29	24,3
Method	121	28	23,14
Total	**3466**	**808**	**23,31**

Table 2. Experimental result of Indonesian Question Answering System using STC algorithm [a]

Question Type	Returned Answers by the System	Correct Answers	Percentage (%)
Factoid			
People	500	447	89
Organization	500	421	84
Location	500	405	81
Name	500	435	87
Date/Time	500	410	82
Quantity	500	396	79
Non-Factoid			
Definition	110	107	97
Reason	119	102	85
Method	121	79	65
Total	**3350**	**2795**	**83**

[a] Source: Rahmansyah [3]

5. Conclusions

The comparison between Indonesian Question Answering System with SSTC and with STC showed a huge difference in terms of the accuracy rate, particularly in answering the factoid questions. However, for the non-factoid questions of definition, the accuracy rate gap is not overly significant. The reason is the question sentences in definition category satisfy the surface expression rules, hence they are easier to find in the documents.

The poor accuracy rate of Indonesian Question Answering System with SSTC in answering questions from other categories is caused by several factors as follows.

1. The documents order affects the order of the returned answers by system. After the documents are clustered, their order follows the order of their cluster respectively. As a result, the correct answer could also be placed in the bottom of answers order if its source document is in the bottom of documents order.
2. The incomplete stop words dictionary could prompt the returning of irrelevant keywords and cause the retrieval of irrelevant documents. Problem arises when an incorrect answer sentence contains more keywords than the actual answer sentence.
3. The surface expression method from Rahmansyah [3] is not the proper method to be implemented in this research's Question Analyzer component. In some cases, the question analyzer result obtained a false expected answer type (EAT), therefore the system returned a false answer.

References

[1] Zulen A *et al* 2011 *25th Pacific Asia Conf on Lang, Info and Comp.* 622
[2] Zamir O *et al* 1998 *Proc. of the 21st Annual Inter ACM SIGIR Conf on Res and Develop in Info Retrieval.* Melbourne, Australia. 46
[3] Rahmansyah A 2015 *Implementasi Suffix Tree Clustering Pada Sistem Tanya Jawab Bahasa Indonesia Untuk Pertanyaan Factoid Dan Non-Factoid* (Indonesia: Universitas Sriwijaya)
[4] Janruang J *et al* 2011 *First IRAST Int. Conf. on Data Eng. and Int Tech(DEIT).* Bali, Indonesia.
[5] Purwarianti A *et al* 2007 *Int Conf on Art Intel and App. 25th Multi-Conf.* Innsbruck, Austria.
[6] Yusliani N 2010 *Sistem Tanya Jawab Bahasa Indonesia untuk Non Factoid Question* (Indonesia: Institut Teknologi Bandung)
[7] Sameh A *et al* 2010 *Int J. of Res and Rev in Comp Sc (IJRRCS).* **1** 2.
[8] Shabbir U *et al* 2015 *13th Int Conf on Frontiers of Inf Tech.* Islamabad, Pakistan.
[9] Perera, R 2012 *IEEE 4th Int Conf on Tech for Edu*. Hyderabad.