Investigating the Potential of Rough Sets Theory in Automatic Thesaurus Construction

Gloria Virginia and Hung Son Nguyen

Abstract. This paper presents the result of initial study about implementation of rough sets theory in generating a thesaurus automatically from a corpus. The main objective of this study is to investigate the relation between keywords (defined by human experts as highly related with particular topic) and the sets generated based on rough sets theory. Analysis was conducted into comparison results of all available sets. We concluded that implementing rough sets theory is a rational way to automatically construct a thesaurus, as it can enrich a concept and proved to be able to cover the keywords given by the human experts.

1 Introduction

Thesaurus is a type of lightweight ontology which provides some additional semantics in their terms relations, e.g. synonym relationships, and does not provide an explicit hierarchy [6]. In Information Retrieval system (IR), it is a significant tool used as a controlled vocabulary in indexing process and as a means for query expansion, such as in [7] and [3].

Constructing an ontology automatically has been studied for years. In [1], Crouch and Yang reported that the thesauri generated automatically based on document collection clustering substantially improved the retrieval effectiveness in their four test collections, although the implementation of term discrimination value theory (used to differentiate the classes produced between the useful-thesaurus-classes and the non-useful-thesaurus-classes) was unsuccessful. A study of Patry and Langlais in [10] presented an approach to automatically generate term extractor from a training

Gloria Virginia · Hung Son Nguyen
University of Warsaw, Faculty of Mathematics, Informatics and Mechanics,
Banacha 2, 02-097 Warsaw, Poland
e-mail: gloriavirginia@gmail.com, son@mimuw.edu.pl

F.L. Gaol (Ed.): Recent Progress in DEIT, Vol. 2, LNEE 157, pp. 107–112.
springerlink.com

corpus as well as proposed a way of combining some statistical metrics in order to extract the terms more efficient than when they were used in isolation.

The major issue of constructing a thesaurus automatically is identifying the semantically related terms. Term co-occurrence is one way that has been studied since 1960 [8]. Considering the semantic relatedness between words, rough set theory received our attention as it is a mathematical approach to vagueness [12]. Moreover, it has been successfully implemented in numerous areas of real-life applications [5].

The following section give a brief explanation of tolerance rough set model. Before delineating the phases of study, the experiment data is expounded. We discuss our findings at Sect. 5, then make the conclusion and propose some future works.

2 Tolerance Rough Set Model

Introduced by Pawlak [11] in 1982, rough set theory expresses vagueness of concept by means of a *boundary region* of a set. Suppose we have a concept, then the idea is to approximate the concept by two descriptive sets called *lower* and *upper approximations*. Intuitively, the lower approximation consists of all elements that *surely* belong to the set, the upper approximation consists of all elements that *possibly* belong to the set, whereas the boundary region consists of all elements that *cannot be classified uniquely* to the set or its complement, by employing available knowledge [12].

Tolerance rough set model (TRSM) is an extension of Rough Sets Theory introduced by Kawasaki, Nguyen, and Ho in [4] as a tool to model document in text mining. Basically, this method came from the *generalized approximation space* using *tolerance relation* described by Skowron and Stepaniuk in [13].

In order to enrich the document representation in terms of semantics relatedness, TRSM creates tolerance classes of terms and approximations of subsets of documents. The tolerance classes of terms in T was based on the co-occurrence of index terms in all documents from D, where $D = \{d_1, d_2, ..., d_N\}$ is a set of text documents and $T = \{t_1, t_2, ..., t_M\}$ is a set of index terms from D. Then, a weight vector is used to represent each document $d_i = \{w_{i,1}, w_{i,2}, ..., w_{i,M}\}$, where $w_{i,j}$ denotes the weight of term t_j in document d_i.

Defining the TRSM means defining the tolerance space $\mathbb{R} = \{U, I, \nu, P\}$ suitably for the information retrieval problem. Due to page limitation, we recommend [2] or [9] for detailed explanation of definition. After all, the lower approximation $L_{\mathbb{R}}(X)$, upper approximation $U_{\mathbb{R}}(X)$, and boundary region $BN_R(X)$ of any $X \subseteq T$ in tolerance space $\mathbb{R} = (T, I, \nu, P)$ are as follow

$$L_{\mathbb{R}}(X) = \{t_i \in T \mid \nu(I_\theta(t_i), X) = 1\} \tag{1}$$

$$U_{\mathbb{R}}(X) = \{t_i \in T \mid \nu(I_\theta(t_i), X) > 0\} \tag{2}$$

$$BN_R(X) = U_R(X) - L_R(X) \tag{3}$$

3 Experiment Data

We used ICL-corpus which consists of the first 1,000 emails of *Indonesian Choral Lovers* (ICL) Yahoo! Groups of Indonesian choral community. Each document of ICL-corpus was assigned topic(s) by choral experts (described in [15]) and this process yielded 127 topics.

During annotation process, in addition to decide topics, keywords were determined for each document in order to express the high related words with the given topic. We treated these keywords as the body text of each document in the second corpus, named *WORDS-corpus*. Thus, the WORDS-corpus also consists of 1,000 documents and its document id is in accordance with document id of ICL-corpus.

Assuming that each topic given by the experts are a concept, we consider the agglomeration of keywords for each topic as the terms variants of the concept that semantically related. Therefore, the WORDS-corpus became the ground truth of this study.

4 Main Phases of the Study

There were three phases conducted in this study, those were *extraction*, *rough sets*, and *evaluation*. Fig. 1 depicts the whole process including the resulted sets of each phase. The rectangle represents the phase while the circle represent the result.

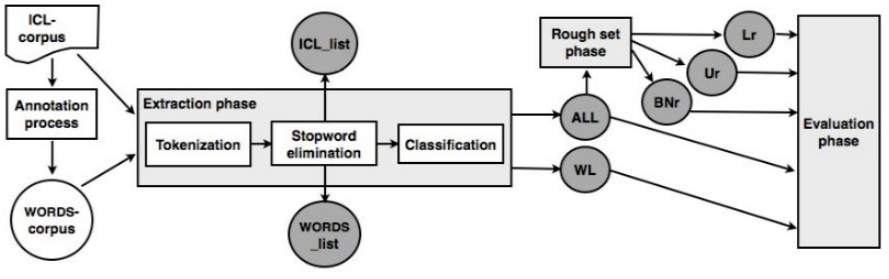

Fig. 1 Main phases of the study

From *extraction phase* it was identified that there were 12,363 unique words in ICL-corpus (called *ICL_list*) and 4,281 unique words in WORDS-corpus (called *WORDS_list*). We then classified our corpora based on the 127 topics and considered all terms appears in each class as the terms of its representative vectors. A frequency matrix of topic-term was created based on these classes. The resulted classes of ICL-corpus (called *ALL*) were then further processed in rough set phase whereas resulted classes of WORDS-corpus (called *WL*) was used later in evaluation phase.

In *rough set phase* the upper set *Ur*, lower set *Lr*, and boundary set *BNr* of each class from ICL-corpus were generated using (1), (2), and (3) respectively; assume that *RS* refers to those three sets. In order to ensure this job could be done quickly,

the occurrence binary matrix (OC), co-occurrence matrix (COC), and tolerance matrix (TOL) were created based on algorithm described in [9]. For evaluation purpose, RS were generated with co-occurrence threshold θ between 1 to 30.

Finally, comparison between all available sets were conducted in *evaluation phase* to get the amount of terms appear in both compared sets. From these comparisons we got values for each topic, then the average computed for each θ value.

5 Analysis

Comparison between ICL_list and WORDS_list shows that ICL_list consists of almost all of WORDS_list terms (99.6%). Analysis of 17 words of WORDS_list which not appear in ICL_list shows that it is caused by typographical error (7 words, all appear in ICL_corpus), informal words (5 words), derived words (4 words), and foreign language (1 word) that emerge in both corpus. For at least 8 of them should be resolved by the stemming process which has not been employed in this study.

The difference between ICL_list and WORDS_list is 8,082 terms, that is 65.37% of ICL_list. Further data related with this are comparison results between ALL and WL, which calculate the same term for each topic. We noticed that the average percentage of same word between ALL and WL in each topic is only 14.56%.

Table 1 Average of same word between sets

Set	Ur (%)	Lr (%)	BNr (%)
ALL	100.00	5.00	95.00
WL	97.74	4.89	92.85

Table 1 shows the results of sets comparison. The value in each cell is the average of number of same word between two sets. The value of ALL-Ur in Table 1 (100%) could be used as an indicator that the RS sets were generated in right manner. Ur is the upper approximation of ALL set hence should consist of all elements belongs to the ALL set. With regard to the number of same terms between ALL and Ur, there are only 42.35% of ALL terms appear in Ur. From this, we can say that the implementation of rough sets theory definitely enrich a concept.

The other values of the first row of Table 1 show that only few terms of ICL_corpus (about 5%) actually could be classified as belong to specific topic while most of them (about 95%) cannot be classified uniquely into a specific topic. These values are somehow similar with comparison results of WL and RS, presented at the second row of Table 1, although the Ur does not cover all terms of WL (only 97.74% of WL). This is possibly the case happen in classification, that limited number of terms could be considered precisely belong to a particular class while numerous of them are in uncertain condition.

Regarding that WL consists of keywords define as highly related with particular concept, the high value of WL-Ur (97.74%) and the small value of WL-Lr (4.89%),

attest that topic assignment is beyond the written terms on a text, i.e. the number of terms being considered during decision of a topic is more than the number of terms written on a text, hence automatic topic assignment task cannot be simply relied on the particular written text. Related with the previous finding that average percentage of same word between ALL and WL in each topic is only 14.56%, then reduction of index term seems to be compulsory.

The experiment to alter the co-occurrence threshold θ value in range 1 to 30 shows a great improvement in decreasing the number of Ur, as it is clearly depicted in Fig. 2. From this figure, we can see that the dramatic change starts to be stable at θ value around 19. Analyzing the other graphs (i.e. the graph of all comparison made) yielded similar result, thus it is suggested to set the θ value ≥ 19. This finding is supported by the high average of same word between WL and Ur that is still larger than 90%, up to θ value 40.

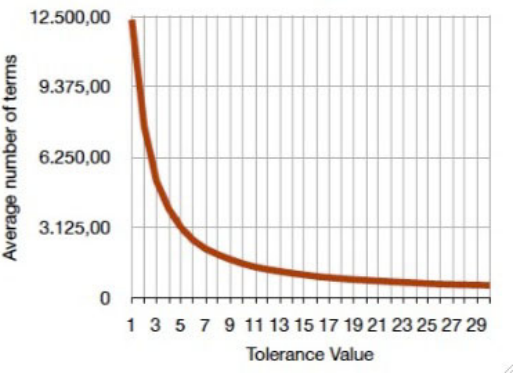

Fig. 2 Total terms of Ur. Total number of terms in upper sets decrease dramatically at the beginning and then become stable at tolerance value around 19.

6 Conclusion

By employing the TRSM, some sets consist of terms were generated from ICL-corpus and WORDS-corpus. Comparison between resulted sets were conducted in order to get the total number of terms that occur on each topic in both sets being compared. We analyzed the average value of each comparison with θ value in range 1 to 30.

We concluded that implementing rough sets theory is a rational way in order to automatically construct a thesaurus, as it can enrich a concept and proved to be able to cover the keywords given by the human experts. However, further study that employs Indonesian stemming and feature selection in rough set theory are requisite.

Acknowledgements. Specific Grant Agreement Number-2008-4950/001-001-MUN-EWC from European Union Erasmus Mundus External Cooperation Window EMMA, research grant of Duta Wacana Christian University, Indonesia, and grants from Ministry of Science and Higher Education of the Republic of Poland (N N516 368334 and N N516 077837).

References

1. Crouch, C., Yang, B.: Experiments in automatic statistical thesaurus construction. In: Proc. The 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 77–88. ACM Publisher, New York (1992)
2. Ho, T.B., Nguyen, N.B.: Nonhierarchical document clustering based on a tolerance rough set model. International Journal of Intelligent System 17, 199–212 (2002)
3. Imran, H., Sharan, A.: Thesaurus and query expansion. International Journal of Computer Science & Information Technology (IJCSIT) 1, 89–97 (2009)
4. Kawasaki, S., Nguyen, N.B., Ho, T.-B.: Hierarchical Document Clustering Based on Tolerance Rough Set Model. In: Zighed, D.A., Komorowski, J., Żytkow, J.M. (eds.) PKDD 2000. LNCS (LNAI), vol. 1910, pp. 458–463. Springer, Heidelberg (2000)
5. Komorowski, J., Pawlak, Z., Polkowski, L., Skowron, A.: Rough sets: a tutorial. In: Rough Fuzzy Hybridization: A New Trend in Decision-Making, pp. 3–98. Springer, Singapore (1998)
6. Lassila, O., McGuinness, D.: The role of frame-based representation on the semantic web. Technical Report KSL-01-02, Knowledge System Laboratory, Standford University
7. Lee, H., Lin, S., Huang, C.: Interactive query expansion based on fuzzy association thesaurus for web information retrieval. In: Proc. of the 10th IEEE International Conference on Fuzzy Systems, vol. 3, pp. 724–727 (2001)
8. Maron, M.E., Kuhns, J.K.: On relevance, probabilistic indexing and information retrieval. Journal of the ACM 7, 216–244 (1960), doi:10.1145/321033.321035
9. Nguyen, H.S., Ho, T.B.: Rough document clustering and the Internet. In: Pedrycz, W., Skowron, A., Kreinovich, V. (eds.) Handbook of Granular Computing, ch. 47, pp. 987–1003. John Wiley & Sons Ltd. (2008), doi:10.1002/9780470724163
10. Patry, A., Langlais, P.: Corpus-based terminology extraction. In: 7th International Conference on Terminology and Knowledge Engineering (TKE 2005), pp. 313–321 (2005)
11. Pawlak, Z.: Rough sets. International Journal of Computer and Information Science 11, 341–356 (1982)
12. Pawlak, Z.: Some Issues on Rough Sets. In: Peters, J.F., Skowron, A., Grzymała-Busse, J.W., Kostek, B.z., Świniarski, R.W., Szczuka, M.S. (eds.) Transactions on Rough Sets I. LNCS, vol. 3100, pp. 1–58. Springer, Heidelberg (2004)
13. Skowron, A., Stepaniuk, J.: Tolerance approximation spaces. Fundam. Inf. 27, 245–253 (1996)
14. Vega, V.B.: Information retrieval for the indonesian language. Master thesis. National University of Singapore (2001) (unpublished)
15. Virginia, G., Nguyen, H.S.: Automatic ontology constructor for Indonesian language. In: Proc. 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT 2010), pp. 440–443. IEEE Press (2010), doi:10.1109/WI-IAT.2010.122