Rough Set in Term Relation Analysis of Indonesian Language

by Gloria Virginia

Rough Set in Term Relation Analysis of Indonesian Language

Gloria Virginia[1] and Hung Son Nguyen[1],

[1] Faculty of Mathematics, Informatics, and Mechanics, University of Warsaw, Banacha 2, 02-097 Warsaw, Poland
virginia@icm.edu.pl, son@mimuw.edu.pl

Abstract. This paper presents the result of initial study about implementation of rough upper approximation and Tolerance Rough Set Model (TRSM) in generating association thesaurus automatically from a corpus. The main objective of our study is to investigate semantic relation between words in documents. Further, it should reveal the possibility of the methods to be implemented for the intended task. We use our own corpus which is a thousand emails of Indonesian Choral Lovers (ICL) Yahoo! Groups where topic(s) has been assigned into each email manually by choral experts. From annotation task we also have list of words that highly related with the given topics defined by the experts where the augmentation of those words is used as the ground truth of the study. Comparison between result of the first corpus and list of words of the second corpus is evaluated quantitatively and qualitatively.

Keywords: rough set, tolerance rough set model, thesaurus.

1 Introduction

Official estimation on the number of subscriber and internet user in Indonesia calculated by *Asosiasi Penyelenggara Jasa Internet Indonesia* (APJII - Indonesian Internet Service Provider Association) shows that the numbers never decrease since 1998[1]. Even the latest estimation is 2007, we are confident that recent number increase regarding the real situation in Indonesia (e.g. it is not difficult to get internet access freely at public area nowadays rather than 5 years ago). Considering that the population of Indonesia is more than 242 million people[2] and most of its inhabitants speak Indonesian language proficiently[3], these should support the fact that computer linguistic fields is important in Indonesia.

---

[1] http://www.apjii.or.id/dokumentasi/statistik.php. Updated on December 2007. Accessed on 14 August 2010.

[2] https://www.cia.gov/library/publications/the-world-factbook/geos/id.html. July 2010 estimation. Accessed on 14 August 2010.

[3] http://www.indonesia.go.id/. Updated on 2008. Accessed on 14 August 2010.

Influential effort of Indonesian research community since mid of 1990s is showed in [1], but there is none related with automatic ontology generation, even ontology has been proved to increase performance of retrieval system (e.g. [2], [3], [4]). For English, by taking advantage from newswire collection, this field of study is intensively investigated (e.g. [5], [6], [7]), however none using rough set theory so far.

Indonesian morphological rules is different from English' (e.g. affixes in Indonesian including prefixes, suffixes, infixes, and circumfixes [8]), therefore a study for Indonesian language should be designed specifically. Instead of newswire, a study using collection of emails as the corpus should bring more benefit with regard to the natural form of language, including the colloquial and slang words, in order to catch the semantic part of language.

This study uses the first continual 1,000 emails of Indonesian Choral Lovers (ICL) Yahoo! Groups where the topic(s) has been assigned manually for each by two choral experts. In previous study [9], knowledge acquisition has been doing along with the annotation task of emails and still proceed. Some findings showed that a topic assigned by the experts could be not written explicitly on the emails, hence it is more about semantic thing.

Therefore, our study tries to investigates semantic relation between words in documents to discover subtle meaning of a text conveyed by the written words. This focus should lead us to reveal the possibility of the implementation of upper approximation and Tolerance Rough Set Model (TRSM) in order to construct an association thesaurus automatically from Indonesian email-corpus for a retrieval system.

2 Methodology

Technically, there are two corpus in this study. The 1,000 emails of ICL is the first (that is *ICL-corpus*) while the second corpus is the augmented words that are defined by choral experts as highly related with the topic(s) given to each email (that is *words-corpus*).

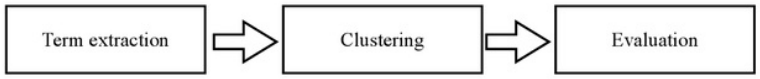

Fig. 1. The main phases of study consists of term extraction, clustering, and relation analysis.

Three main phases employed in our study as it is showed in Fig. 1. Term extraction consists of stemming, stopwords elimination, and index term selection. The stemming will use Confix-Stripping (CS) stemmer described in [10] and the stopwords elimination will use Vega's stopwords [11] which showed to give highest

precision and recall in [12]. We will filter out terms that contained in the query and terms/digits have character less then 2 .

Table 1. Algorithm of TRSM nonhierarchical clustering described in [13] for this study

Input	The set D documents and the number K of cluster
Result	K overlapping clusters of D associated with cluster membership of each documents
1.	Determine the initial representatives $R_1, R_2,..., R_K$ of cluster $C_1, C_2,..., C_K$.
2.	For each $d_j \in D$, calculate the similarity $S(U(R, d_j), R_k)$ between its upper approximation $U(R, d_j)$ and the cluster representative R_k, for $k = 1,..., K$. If this similarity is greater than a given threshold, assign d_j to C_k and take this similarity value as the cluster membership $m(d_j)$ of d_j in C_k.
3.	For each cluster C_k, re-determine its representative R_k.
4.	Repeat steps 2 and 3 until there is little or no change in cluster membership during a pass through D.
5.	Denote by d_u an unclassified document after step 2, 3, and 4, and by $NN(d_u)$ its nearest neighbor document (with non-zero similarity) in formed clusters. Assign d_u into the cluster that contains $NN(d_u)$, and determine the cluster membership of d_u in this cluster as the product $m(d_u) = m(NN(d_u)) \times S(U(R, d_u), U(R, NN(d_u)))$. Re-determine the representatives R_k, for $k = 1,..., K$.

The TRSM nonhierarchical clustering algorithm explained in [13] is implemented in clustering phase to form k cluster of a collection emails in ICL-corpus. Table 1 describes the algorithm for this study. Here, the initial representatives $R_1, R_2,..., R_K$ of cluster $C_1, C_2,..., C_K$ are not randomly selected documents in D, they are generated by calculating the representative of K categories in ICL-corpus. The categories are defined by choral experts based on topic assignment task.

$$w_{ij} = \begin{cases} (1 + \log(f_{d_j}(t_i))) \times \log \frac{M}{f_{\mathcal{D}}(t_i)} & \text{if } t_i \in d_j, \\ \min_{t_h \in d_j} w_{hj} \times \frac{\log(M/f_{\mathcal{D}}(t_i))}{1+\log(M/f_{\mathcal{D}}(t_i))} & \text{if } t_i \in \mathcal{U}(\mathcal{R}, d_j) \backslash d_j \\ 0 & \text{if } t_i \notin \mathcal{U}(\mathcal{R}, d_j) \end{cases} \quad (1)$$

Evaluation phase has three major tasks: term weighting, ranking, and relation analysis. Terms are weighted differently in specific corpus. In ICL-corpus, terms are weighted using a term-weighting method (equation 1) as in [13], while in words-corpus terms are weighted based on their frequency on a document. The assumption is the weight value of terms in ICL-corpus reveal the importance of those terms for a specific category. It means, the highest-weight-term seems to be the most important term for particular category. In words-corpus, terms are selected by choral experts manually as having highly relation with the given topic(s) of a document. It means,

the importance of these words have been *calculated* implicitly by the experts. Therefore the most frequent term seems to be the most important term for particular category. Let us say *List 1* for list of ranked-terms in ICL-corpus and *List 2* for list of ranked-terms in words-corpus.

Table 2. The main tasks of quantitative and qualitative analysis conducted in relation analysis.

Analysis	Main Tasks
Quantitative	1. Number of words in List 1 that come up in List 2
	2. Number of words in List 2 that cannot be found in List 1
Qualitative	1. Analysis the ranking of words in a category
	2. Analysis the distribution of a word in some categories

Quantitative and qualitative analysis employed in relation analysis. Table 2 describes the main tasks of those two.

Acknowledgments. This research has been supported by Specific Grant Agreement Number-2008-4950/001-001-MUN-EWC from European Union Erasmus Mundus "External Cooperation Window" EMMA and has been partially supported by grants N N516 368334 and N N516 077837 from Ministry of Science and Higher Education of the Republic of Poland.

References

1. Adriani, M. and Manurung, R.: A Survey of Bahasa Indonesia NLP Research Conducted at the University of Indonesia. In: Proceedings of the 2nd International MALINDO Workshop Cyberjaya, Malaysia, June 12-13, 2008). (2008)
2. Burke, R. D., Hammond, K. J., and Cooper, E.: Knowledge-Based Information Retrieval from Semi-Structured Text. In AAAI Workshop on Internet-based Information System. AAI, 15-19. (1995)
3. Ismail, M. A., Yaacob, M., Kareem, S. A., and Halim, A. H. A.: Semantic Search Engine in Institutional Repository: An Ontological Approach. In Building an Information Society for All: Proceedings of the International Conference on Libraries, Information and Society (Petaling Jaya, Malaysia, June 26-27, 2007). ICoLIS 2007. Kuala Lumpur, 55-63. (2007)
4. Tang, B. and Hodges, J.: Knowledge Representation, Learning, and Reasoning in WebDoc - A Web Document Classification System. In: Proceedings of the AAAI-2000 Workshop on tificial Intelligence for Web Search (Austin, Texas, July 30, 2000). (2000)
5. Jing, Y. and Croft, W. Bruce.: An Association Thesaurus for Information Retrieval. RIAO 94 Conference Proceedings, Rockefeller University, New York City, 11-13 October 1994, pp. 146-160. (1994)

6. Kaji, H., Morimoto, Y., Aizono, T., and Yamasaki, N.: Corpus-dependent Association Thesauri for Information Retrieval. In: Proceedings of the 18th Conference on Computational Linguistics (Saarbrücken, Germany, 31 July - 04 August 2000). (2000)
7. Lee, H., Lin, S., and Huang, C.: Interactive Query Expansion Based on Fuzzy Association Thesaurus for Web Information Retrieval. In: Proceedings of the 10th IEEE International Conference on Fuzzy Systems (Australia, December 2-5, 2001), 724-727. (2001)
8. Alwi, H., Sardjowidjojo, S., Lapoliwa, H., and Moeliono, A.: Tata Bahasa Baku Bahasa Indonesia Edisi Ketiga. Pusat Bahasa dan Balai Pustaka, Jakarta. (2003)
9. Virginia, G., Nguyen, H. S.: Automatic Ontology Constructor for Indonesian Language. In: Workshop proceedings of Web Intelligence-Intelligent Agent Technology. IEEE Computer Society Press (August 31 - September 6, 2010). (2010) (Accepted)
10. Adriani, M., Asian, J., Nazief, B., Tahaghogi, S. M. M., and Williams, H. E.: Stemming Indonesian: A Confix-Stripping Approach. ACM Transaction on Asian Language Information Processing. 6, 4 (December 2007). (2007)
11. Vega, V. B.: Information Retrieval for the Indonesian Language. Master's thesis. National University of Singapore. (2001)
12. Asian, J.: Effective Techniques for Indonesian Text Retrieval. Doctor of Philosophy Thesis. School of Computer Science and Information Technology. RMIT University. (2007)
13. Ho, T. B. and Nguyen, B. N.: Nonhierarchical document clustering based on a tolerance rough set model. International Journal of Intelligent Systems 17, 2. 199-212. (2002)

Rough Set in Term Relation Analysis of Indonesian Language

ORIGINALITY REPORT

13%	10%	10%	8%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	myais.fsktm.um.edu.my Internet Source	1%
2	Qinmin Hu. "A Term Association Approach for Genomics Information Retrieval", 2011 IEEE International Conference on Bioinformatics and Biomedicine, 11/2011 Publication	1%
3	Submitted to UT, Dallas Student Paper	1%
4	www.readbag.com Internet Source	1%
5	Platkowski, T.. "Evolution of populations playing mixed multiplayer games", Mathematical and Computer Modelling, 200405 Publication	1%
6	www.mimuw.edu.pl Internet Source	1%
7	Lukasz Neuman. "Information Retrieval Using Bayesian Networks", Lecture Notes in	1%

Computer Science, 2004
Publication

8	Karsono, O. M. F.. "Juncture Patterns of the Surabaya-citizen's Speech", k@ta, 2012. Publication	1%
9	200.46.218.164 Internet Source	1%
10	Submitted to Institute of Graduate Studies, UiTM Student Paper	1%
11	P. G. Aaron. "Qualitative and quantitative differences among dyslexic, normal, and nondyslexic poor readers", Reading and Writing, 1989 Publication	1%
12	Submitted to University of Hong Kong Student Paper	1%
13	"Strategies for Regenerating the Library and Information Profession", Walter de Gruyter GmbH, 2009 Publication	1%
14	Blaszczynski, J.. "Monotonic Variable Consistency Rough Set Approaches", International Journal of Approximate Reasoning, 200907 Publication	<1%

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	Off		