Towards the Semantic Text Retrieval for Indonesian

Autoreferat

Gloria Virginia

University of Warsaw, Faculty of Mathematics, Informatics and Mechanics
Banacha 2, 02-097 Warsaw, Poland

May 2013

1 Information Retrieval

Information retrieval is concerned with representing, searching, and manipulating large collections of electronic text and other human-language data [1, p. 2]. Searching system, clustering systems, categorization systems, summarization systems, information extraction systems, topic detection systems, question answering systems, and multimedia information retrieval systems are other applications utilize IR service.

The main task of information retrieval is to retrieve relevant documents in response to a query[2, p. 85]. In a common search application, an *ad hoc* retrieval mode is applied in which a query is submitted (by a user) and then evaluated against a relatively static document collection. A set of query identifying the possible interest to the user may also be supplied in advanced and then evaluated against newly created or discovered documents. This operational mode where the queries remain relatively static is called *filtering*.

Documents (i.e. electronic texts and other human-language data) are normally modeled based on the positive occurrence of words while the query is modeled based on the positive words of interest clearly specified. Both models then are examined in similarity basis using a devoted ranking algorithm and the output of information retrieval system (IRS) will be an ordered list of documents considered pertinent to the query at hand.

In the keyword search technique commonly used, the similarity between documents and query is measured based on the occurrence of query words in the documents. Thus, if the query is given by a user, then the relevant documents are those who contain literally one or more words expressed by him/her. The fact is, text documents (and query) highly probable come up in the form of natural language. While human seems effortless to understand and construct sentences, which may consist of ambiguous or colloquial words, it becomes a big challenge for an IRS. The keyword search technique is lack of capability to capture the

meaning of words, wherefore the meaning of sentences, semantically on documents and query because it represents the information content as a syntactical structure which is lack of semantical relationship. For example, a document contains words *choir*, *performance*, and *ticket* may talk about a *choir concert*, in spite of the fact that the word *concert* is never mentioned on that particular document. When a user inputs the word *concert* to define his/her information need, the IRS which approximate the documents and query in a set of occurrence words may deliver lots of irrelevant results instead of corresponding documents.

We may expect better effectiveness to IRS by mimicking the human capability of language understanding. We should move from *keyword* to *semantic* search technique, hence the semantic IRS.

2 Philosophical Background

Language is one of the vehicles of mental states, hence linguistic meaning is a form of derived intentionality. Based on Searl's and Grice's accounts on meaning, there is distinction between intentional content and the form of its externalization. To ask for the meaning is to ask for an intentional content that goes with the form of externalization [3]. It is maintained that for a successful speech act, a speaker normally choose an expression which is conventionally fixed, i.e. by the community at large, to convey a certain meaning. Thus, before the selection process of appropriate expressions, it is fundamental for a speaker to know about the expression in order to produce an utterance, and consequently the audience is required to be familiar with those conventional expressions in order to understand the utterance.

Searl's and Grice's accounts pertaining the meaning suggest knowledge for language production and understanding. This knowledge should consists of concepts who are interrelated and commonly agreed by the community. The communication is satisfied when both sides are active participants and the audience experiences effects at some degree.

3 Challenges in Indonesian

Considerable effort with regard to information retrieval for Indonesian is showed by a research community in University of Indonesia (UI) since mid of 1990s [4]. Other significant studies conducted by Asian which proposed an effective techniques for Indonesian text retrieval [5] and published the first Indonesian testbed [6]. It is worth to mention that despite the long list of works ever mentioned, only limited number of the results is available publicly. Among those Indonesian studies, it is hardly to find a work pertaining to automatic ontology constructor specifically.

The latest data released by Statistics Board of Indonesia (BPS-Statistics Indonesia)[1] pertaining the population of Indonesia, showed that the number reached 237.6 million for the 2010 census. This number ranked Indonesia on the forth most populous country in the world after China, India, and United States[2].

According to Sneddon [7, p. 196], Indonesia has about 550 languages. However, *Bahasa Indonesia* or Indonesian language was chosen as the national language and nowadays most Indonesians are proficient in using the language; the number of speaker of Indonesian is approaching 100 percent [7, page 201].

Pertaining to the growth of Internet users, the Internet World Stats[3] recorded that there are about 55 million internet users (with 22.4% penetration rate) and 43 million Facebook users (with 17.7% penetration rate) as of Dec. 31, 2011 in Indonesia. These facts are some indicators of the digital media usage proliferation in Indonesia which is considered to keep on growing.

4 Tolerance Rough Sets Model

Basically, an information retrieval system consists of three main tasks: *(1)* modeling the document; *(2)* modeling the query; and *(3)* measure the degree of correlation between document and query models. Thus, the endeavor of improving an IRS revolves around those three tasks. One of the effort is a method called tolerance rough set model (TRSM) which has performed positive results on some studies pertaining to information retrieval. In spite of the fact that TRSM does not require complex linguistic process, it has not been investigated at large extent.

Since it was formulated, tolerance rough sets model (TRSM) is accepted as a tool to model a document in a *richer* way than the base representation which is represented by a vector of TF*IDF-weight terms[4] (let us call it TFIDF-representation). The richness of the document representation produced by applying the TRSM (let us call it TRSM-representation) is indicated by the number of index terms put into the model. That is to say, there are more terms belong to TRSM-representation than its base representation.

The power of TRSM is grounded by a knowledge, i.e. thesaurus, which is comprised by index terms and the relationships between them. In TRSM, each set of terms considered as semantically related with a single term t_j is called the tolerance class of a term $I_\theta(t_j)$, hence the thesaurus contains tolerance classes of all index terms. The semantic relatedness is signified by the terms co-occurrence

[1] BPS-Statistics Indonesia. URL: `http://www.bps.go.id/`. Accessed on 25-10-2012.
[2] July 2012 estimation of The World Factbook. URL: `https://www.cia.gov`. Accessed on 25-10-2012.
[3] URL: `http://www.internetworldstats.com`. Accessed on 25-10-2012.
[4] Appendix A of the thesis provides an explanation of the TF*IDF weighting scheme.

in a corpus in which a tolerance value θ is set to define the threshold of co-occurrence frequency.

5 Research Objective and Approach

The research aims to investigate the tolerance rough sets model in order to propose a framework for a semantic text retrieval system. The proposed framework is intended for Indonesian language specifically hence we are working with Indonesian corpora and applying tools for Indonesian, e.g. Indonesian stemmer, in all of the studies.

The researches of TRSM ever conducted pertaining to information retrieval have focused on the system performance and involved a combination of mathematics and engineering in their studies [8–12]. In this thesis, we are trying to look at TRSM from a quite different viewpoint. We are going to do empirical studies involving observations and hypotheses of human behavior as well as experimental confirmation. According to the Artificial Intelligence (AI) view, our studies follow a human-centered approach, particularly the cognitive modeling[5], instead of the rationalist approach [14, p.1-2]. Analogous to two faces in a coin, both approaches would result in a comprehensive perspective of TRSM.

In implementing the cognitive approach, we start our analysis from the performance of an ad hoc retrieval system. It is not our intention to compare TRSM with other methods and determine the best solution. Rather, we will take the benefit of the experimental data to learn and understand more about the process and characteristic of TRSM. The results of this process function as the guidance for computational modeling of some TRSM's tasks and finally the framework of a semantic IRS with TRSM as its heart.

6 Thesis Structure

Our research falls under the information retrieval umbrella. Chapter 2 provides an explanation about the main tasks of information retrieval and the semantic indexing in order to establish a general understanding of semantic IRS.

Several questions are generated in order to assist us to scrutinize the TRSM. The issues behind the questions should be apparent when we proceed into the nature of TRSM that would be exposed on theoretical basis in Chapter 3 We have selected four subjects of question and will discuss them in the following order:

[5] The cognitive modeling is an approach employed in the Cognitive Science (CS). Cognitive science is an interdisciplinary study of mental representations and computations and of the physical systems that support those processes [13, p.xv].

1. **Is TRSM a viable alternative for a semantic IRS?**
   The simplicity of characteristic and positive result of studies makes TRSM an intriguing method. However, before moving any further, we need to ensure that TRSM is reasonable to be the ground floor of the intended system. This issue will be the content of Chapter 4.
2. **How to generate the system knowledge automatically?**
   The richer representation of document yielded by TRSM is achieved fundamentally by means of a knowledge, which is a thesaurus. The thesaurus is manually created, in the sense that a parameter, namely *tolerance value* θ, is required to be determined by hand. In Chapter 5 we would propose an algorithm to select a value for θ automatically.
3. **How to improve the quality of the thesaurus?**
   The thesaurus of TRSM is generated based on a collection of text documents functions as a data source. In other words, the quality of document representation should depend on the quality of data source at some degree. Speaking of which, the TRSM basically works based on the raw frequency of terms co-occurrence, and it arises an assumption that other co-occurrence data might bring a benefit for the effort to optimize the thesaurus. These presumptions would be reviewed and discussed in Chapter 6.
4. **How to improve the efficiency of the intended system?**
   The TRSM-representation is claimed to be richer in the sense that it consists of more terms than the base representation. Despite the fact that the terms of TRSM-representation are semantically related, more terms on document vector results in more cost of computation. In other words, system efficiency becomes the trade-off. We came into an idea of a compact document representation that would be explained in Chapter 7.

This thesis proposes three methods based on TRSM for the mentioned problems. All methods, which are discussed in Chapter 5 to 7, were developed by the use of our own corpus, namely ICL-corpus, and evaluated by employing an available Indonesian corpus, called Kompas-corpus[6]; Chapter 8 describes the evaluation process. The evaluation on the methods achieved satisfactory results, except for the compact document representation method; this last method seems to work only in limited domain.

The final chapter provides our conclusion of the research as well as discussion of some challenges that lead to advance studies in the future.

[6] Explanation about all corpora used in this thesis is available in Appendix C.

7 Conclusion

7.1 The TRSM-Based IRS

The research of extended TRSM, along with other researches of TRSM ever conducted, acted in accordance with the rational approach of AI perspective. This thesis presented studies who complied with the contrary path, i.e. a cognitive approach, for an objective of a modular framework of semantic text retrieval system based on TRSM specifically for Indonesian.

Figure 1 exhibits the schema of the intended framework which consists of three principal phases, namely preprocessing phase, TRSM phase, and retrieval phase. The mapping phase is included for an alternative and subject to change.

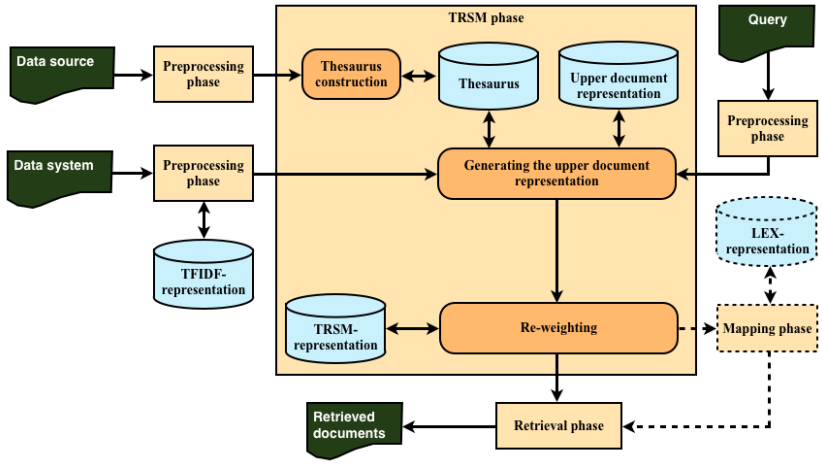

Fig. 1. The schema of text retrieval system based on TRSM. The dashed line shapes are optional.

The proposed framework is in Java and takes a benefit of using Lucene 3.1 while indexing. Indonesian stemmer (i.e. CS stemmer), lexicon (i.e. created by University of Indonesia), and stopword (i.e. Vega's stopword) which are embedded make the framework works specifically for Indonesian language; altering them specific to one language would make the framework dependent to that particular language.

7.2 Novel Strategies for The TRSM-Based IRS

With regard to the framework of retrieval system, we delved into four issues based on the nature of TRSM. The very first issue questioned about the capacity

of TRSM for the intended system, while the other three touched the system effectiveness.

In order to answer the first question, we did a feasibility study whose aim was to explain the meaning of *richness* of the TRSM-representation, rather than listing the strengths and weaknesses of TRSM. By working in close cooperation with human experts, we were able to reveal that the representation of document produced by TRSM does not merely contain more terms than the base representation, it rather contains more semantically related terms. Concerning our approach, we deem this as a stronger affirmation for the meaning of *richness* of the TRSM-representation as well as a satisfactory indicator in an endeavor to have a semantic retrieval system. Moreover, our analysis confirmed that rough sets theory intuitively works as the natural process of human thought.

Since the TRSM was introduced, no one has ever discussed or examined TRSM's parameter (i.e. tolerance value θ) pertaining to its determination, whereas we consider it as fundamental for TRSM implementation. Obadi et. al. [15] seemed to realize this particular issue by stating that TRSM is very sensitive to parameter setting in their conclusion, however they did not explain or suggest anything about how to initiate it. In Chapter 5 we proposed a novel algorithm to define a tolerance value automatically by learning from a set of documents; and later we named it *TolValGen*. The algorithm was a result from careful observation and analysis performed through our corpora (i.e. ICL-corpus and WORDS-corpus) in which we learned some principles for a tolerance value resolution. The TolValGen was evaluated using another Indonesian corpus (i.e. Kompas-corpus) and yielded positive result. It was capable to produce an appropriate tolerance value for each variants of Kompas-corpus.

We recognized that the thesaurus dominates TRSM in its work, hence optimizing the quality of thesaurus became another important issue we discussed. We admit that our idea to enhance the quality of thesaurus by adding more documents specifically for data source of thesaurus did not come up with a promising result as of Nguyen et al. [16, **?**] which performed much more clever idea by extending the TRSM such that it accommodates more than one factors for a composite weight value of document vector. However, from the analysis carried out through several corpora (the variants of ICL-corpus and the Wiki_1800), we learned that tolerance value, data source of thesaurus, and semantic measure determine the quality of thesaurus. Specifically, a data source which is in a corresponding domain with the system data and is larger in number might bring more benefit. We also found that the total number of terms and index terms contribute more to the quality of thesaurus, despite the size of corpus. Finally, we suggested to keep the raw frequency of co-occurrence to define the semantic relatedness between terms for it gave better results in experiment rather

than other measure, i.e. Cosine. All of these findings were validated by means of evaluation using Kompas-corpus.

The last issue discussed in this thesis associated with both the effectiveness and efficiency of system. It was motivated by a fact that the richer representation of TRSM is indicated by the larger number of index terms put into the model. Concerning the size of vector dimension, we came into an idea of a compact model of document based on the mapping process between index terms and lexicon after the document enriched by TRSM. The experimental data over ICL-corpus expressed a promising result, however the evaluation through Kompas-corpus remarked differently. Even though numerous irrelevant terms successfully removed from LEX-representation by the lexicon, we learned that our model cannot be applied for general use. The LEX-representation might be easily corrupted and thus become much less reliable when a query comprised of many terms which are not part of the lexicon and those terms are considered significant. Whereas, this particular situation is highly probable to occur in natural language.

8 Future Directions

The proposed framework is lack of comparison result. The studies presented in this thesis focused only on the use of TRSM which were compared to the result of TF*IDF. Comparison studies of methods, such those explained in Chapter 2 for semantic indexing, would put TRSM on certain position and bring some suggestion for further development.

The high complexity of our framework is the consequence of TRSM implementation. The application of Lucene module supports the indexing task in preprocessing phase of the framework, however we failed in the attempt to alter the index directly after TRSM phase which forced us to store the revised-index in different space. We found that it reduced the efficiency of IRS significantly, even though index file was applied. Studies focus on indexing in TRSM implementation is thus essential.

The proposed framework was developed for laboratory environment which is effective for restricted format and type of documents, i.e. follow the TREC-format and written in a *.txt* file. For a real application, our proposed framework should be extended to have the ability to deal with various format and type of documents. Much further, we should consider the recent phenomena of big data[7].

The TolValGen has showed to work on our corpora and their variations. However, it suffers from the expensive time and space to operate. In order to have cheaper complexity for tolerance value generator, further study on this

[7] Big data is a term to describe the enormity of data, both structured and unstructured, in volume, velocity, and variety [17].

theme with different methods is needed. We might expect some advantage by the use of machine learning method that accommodates the dynamic change of data source.

The lexicon-based document representation is an attempt on system efficiency. Despite the result of evaluation in Chapter **??** which signifies that it is lack of scalability, the fact that we did not implement any other linguistic methods arose our confident that those computations (such as tagging, feature selection, n-gram) might give us benefit in the effort of refining the thesaurus that serves as the basis of tolerance rough sets model, and thus the knowledge of our IRS.

In accordance with Searl's and Grice's accounts on meaning, Ingwersen [18, p. 33] defined that the concept of information, from a perspective of information science, has to satisfy dual requirements: *(1)* being the result of a transformation of generator's knowledge structures (by intentionality, model of recipients' states of knowledge, and in the form of signs); and *(2)* being something which when perceived, affects and transforms the recipients's state of knowledge. Thus, the endeavor of a semantic IRS is the effort to retrieve *information* and not merely terms with similar meaning. This thesis is a step toward the objective.

References

1. Büttcher, S., Clarke, C.L.A., Cormack, G.V.: Information Retrieval: Implementing and Evaluating Search Engine. MIT Press, Cambridge, Massachusetts (2010)
2. Weiss, S.M., Indurkhya, N., Zhang, T., Damerau, F.J.: Text Mining - Predictive Methods for Analyzing Unstructured Information. Springer, New York (2005)
3. Searle, J.R.: Intentionality: An Essay in the Philosophy of Mind. Cambridge University Press, Cambridge (1983)
4. Adriani, M., Manurung, R.: A survey of bahasa indonesia nlp research conducted at the university of indonesia. In: Proceedings of the 2nd International MALINDO Workshop. (2008)
5. Asian, J.: Effective Techniques for Indonesian Text Retrieval. PhD thesis, School of Computer Science and Information Technology, RMIT University (March 2007) Doctor of Philosophy Thesis.
6. Asian, J., Williams, H.E., Tahaghoghi, S.M.M.: A testbed for indonesian text retrieval. In Bruza, P., Moffat, A., Turpin, A., eds.: ADCS, University of Melbourne, Department of Computer Science (2004) 55–58
7. Sneddon, J.: The Indonesian Language: It's History and Role in Modern Society. UNSW Press (2003)
8. Kawasaki, S., Nguyen, N.B., Ho, T.B.: Hierarchical document clustering based on tolerance rough set model. In: Proceedings of the 4th European Conference on Principles of Data Mining and Knowledge Discovery. PKDD '00, London, UK, Springer-Verlag (2000) 458–463

9. Ho, T.B., Nguyen, N.B.: Nonhierarchical document clustering based on a tolerance rough set model. International Journal of Intelligent Systems **17**(2) (February 2002) 199–212
10. Nguyen, H.S., Ho, T.B.: 47. In: Rough Document Clustering and the Internet. John Wiley & Sons Ltd. (2008) 987–1003
11. Wu, Y., Ding, Y., Wang, X., Xu, J.: On-line hot topic recommendation using tolerance rough set based topic clustering. Journal of Computers **5** (April 2010) 549–556
12. Gaoxiang, Y., Heping, H., Zhengding, L., Ruixuan, L.: A novel web query automatic expansion based on rough set. Wuhan University Journal of Natural Sciences **11**(5) (2006) 1167–1171
13. Bly, B.M., Rumelhart, D.E., eds.: Cognitive Science: Handbook of Perception and Cognition. Second edn. Academic Press, California (1999)
14. Russell, S., Norvig, P.: Artificial Intelligence: A Modern Approach. Third edn. Pearson Education, Inc., New Jersey (2010)
15. Obadi, G., Dráždilová, P., Hlaváček, L., Martinovič, J., Snášel, V.: A tolerance rough set based overlapping clustering for the dblp data. In: Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and International Conference on Intelligent Agent Technology - Workshops. Volume 3 of WI-IAT '10., IEEE (2010) 57–60
16. Nguyen, S., Świeboda, W., Jaśkiewicz, G.: Extended document representation for search result clustering. In Bembenik, R., Skonieczny, Ł., Rybiński, H., Niezgódka, M., eds.: Intelligent Tools for Building a Scientific Information Platform. Volume 390 of Studies in Computational Intelligence. Springer Berlin Heidelberg (2012) 77–95
17. Troester, M.: Big data meets big data analytics. `http://www.sas.com/resources/whitepaper/wp_46345.pdf` (2012) Copyright ©SAS Institute Inc.; Accessed 22-February-2013.
18. Ingwersen, P.: Information Retrieval Interaction. First edn. Taylor Graham, London (1992)