Applications of Tolerance Rough Set Model Semantic Text Analysis

Hung Son Nguyen

Institute of Computer Science
The University of Warsaw
Banacha 2, 02-097, Warsaw Poland
Email: son@mimuw.edu.pl

Abstract. Tolerance Rough Set Model (TRSM) is an extension of Rough Set theory and can be used as a tool for approximation of hidden concepts in collections of documents. In recent years, numerous successful applications of TRSM in web intelligence including text classification, clustering, thesaurus generation, semantic indexing, and semantic search, etc., have been proposed. This paper revises the basic concepts of TRSM, some of its possible extensions and some typical applications of TRSM in text mining. We also discuss some further research on TRSM.

1 Extended Abstract

Rough set theory has been introduced by Pawlak [1] as a tool for concept approximation under uncertainty. The idea is to approximate the concept by two descriptive sets called *lower and upper approximations*. The fundamental philosophy of rough set approach to concept approximation problem is to minimize the difference between upper and lower approximations (the *boundary region*). This simple but brilliant idea leads to many efficient applications of rough sets in machine learning, data mining and also in granular computing. The connection between rough set and other computational intelligence techniques was presented by many researchers, e.g. [2] [3] [4] [5] [6] [7]. Numerous computational intelligence techniques based on rough sets including support vector machine [8], genetic algorithm [9] [10], modified self-organizing map [11] have been proposed. The rough set based data mining methods were applied to many real life applications, e.g., medicine [12], web user clustering [13] [11] [7] and marketing [10].

Tolerance Rough Set Model was developed in [14, 15] as a basis to model documents and terms in Information Retrieval, Text Mining, etc. With its ability to deal with vagueness and fuzziness, Tolerance Rough Set Model seems to be a promising tool to model relations between terms and documents. In many Information Retrieval problems, especially in document clustering, defining the relation (i.e. similarity or distance) between document-document, term-term or term-document is essential. In Vector Space Model, is has been noticed [15] that

a single document is usually represented by relatively few terms[1]. This results in zero-valued similarities which decreases quality of clustering. The application of TRSM in document clustering was proposed as a way to enrich document and cluster representation with the hope of increasing clustering performance.

In fact Tolerance Rough Set Model is a special case of a generalized approximation space, which has been investigated in [16]. as a generalization of standard rough set theory. Generalized approximation space utilizes every tolerance relation overs objects to determine the main concepts of rough set theory, i.e., lower and upper approximation.

The main idea of TRSM is to capture conceptually related index terms into classes. For this purpose, the tolerance relation R is determined as the co-occurrence of index terms in all documents from D. The choice of co-occurrence of index terms to define tolerance relation is motivated by its meaningful interpretation of the semantic relation in context of IR and its relatively simple and efficient computation.

1.1 Standard TRSM

Let $D = \{d_1, \ldots, d_N\}$ be a corpus of documents Assume that after the initial processing documents, there have been identified N unique terms (e.g. words, stems, N-grams) $T = \{t_1, \ldots, t_M\}$.

Tolerance Rough Set Model, or briefly TRSM, is an approximation space $\mathcal{R} = (T, I_\theta, \nu, P)$ determined over the set of terms T where:

- The parameterized **uncertainty function** $I_\theta : T \to \mathcal{P}(T)$ is defined by
$$I_\theta(t_i) = \{t_j \mid f_D(t_i, t_j) \geq \theta\} \cup \{t_i\}$$
where $f_D(t_i, t_j)$ denotes the number of documents in D that contain both terms t_i and t_j and θ is a parameter set by an expert. The set $I_\theta(t_i)$ is called the *tolerance class* of term t_i.
- **Vague inclusion function** $\nu(X, Y)$ measures the degree of inclusion of one set in another. The vague inclusion function is defined as $\nu(X, Y) = \frac{|X \cap Y|}{|X|}$. It is clear that this function is monotone with respect to the second argument.
- **Structural function:** All tolerance classes of terms are considered as structural subsets: $P(I_\theta(t_i)) = 1$ for all $t_i \in T$.

In TRSM model $\mathcal{R} = (T, I, \nu, P)$, the membership function μ is defined by

$$\mu(t_i, X) = \nu(I_\theta(t_i), X) = \frac{|I_\theta(t_i) \cap X|}{|I_\theta(t_i)|}$$

where $t_i \in T$ and $X \subseteq T$. The lower and upper approximations of any subset $X \subseteq T$ can be determined by the same maneuver as in approximation space [16]:

$$\mathbf{L}_\mathcal{R}(X) = \{t_i \in T \mid \nu(I_\theta(t_i), X) = 1\}$$
$$\mathbf{U}_\mathcal{R}(X) = \{t_i \in T \mid \nu(I_\theta(t_i), X) > 0\}$$

[1] In other words, the number of non-zero values in document's vector is much smaller than vector's dimension – the number of all index terms

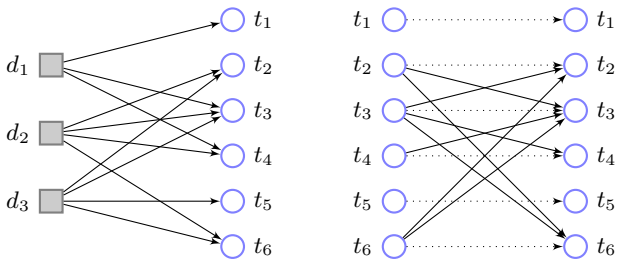


Fig. 1. Bag-of-words (left) determines the term co-location graph with $\theta = 2$ (right).

The standard TRSM was applied for document clustering and snippet clustering tasks (see [14], [15], [7], [17], [18]). In those applications, each document is represented by the upper approximation of its *set of words/terms*, i.e. the document $d_i \in D$ is represented by $\mathbf{U}_{\mathcal{R}}(d_i)$. For the example in Figure 1, the enriched representation of d_1 is $\mathbf{U}_{\mathcal{R}}(d_1) = \{t_1, t_3, t_4, t_2, t_6\}$.

1.2 Extended TRSM using Semantic Concepts

Let $D = \{d_1, \ldots, d_N\}$ be a set of documents and $T = \{t_1, \ldots, t_M\}$ the set of *index terms* for D. Let C be the set of concepts from a given domain knowledge (e.g. the concepts from DBpedia or from a specific ontology).

The extended TRSM is an approximation space $\mathcal{R}_C = (T \cup C, I_{\theta,\alpha}, \nu, P)$, where C is the mentioned above set of concepts. The uncertainty function $I_{\theta,\alpha} : T \cup C \to \mathbb{P}(T \cup C)$ has two parameters θ and α is defined as follows:

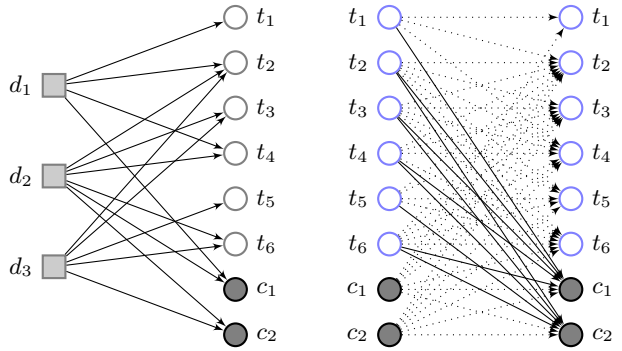


Fig. 2. Extended TRSM with $\theta = 1$, bag-of-words document representation (left) determines the structure of ESA model (right) when filtered to term $\to$ concept edges.

– for each term $c_i \in C$ the set $I_{\theta,\alpha}(c_i)$ contains α top terms from the bag of terms of c_i calculated from the textual descriptions of concepts.

– for each term $t_i \in T$ the set $I_{\theta,\alpha}(t_i) = I_\theta(t_i) \cup C_\alpha(t_i)$ consists of the tolerance class of t_i from the standard TRSM and the set of concepts, whose description contains the term t_i as the one of the top α terms.

In the extended TRSM, any document $d_i \in D$ can be represented by

$$\begin{aligned}\mathbf{U}_{\mathcal{R}_C}(d_i) &= \mathbf{U}_{\mathcal{R}}(d_i) \cup \{c_j \in C \mid \nu(I_{\theta,\alpha}(c_j), d_i) > 0\} \\ &= \bigcup_{t_j \in d_i} I_{\theta,\alpha}(t_i)\end{aligned}$$

1.3 Weighting Schema

Any text d_i in the corpus D can be represented by a vector $[w_{i1}, \ldots, w_{iM}]$, where each coordinate $w_{i,j}$ expresses the significance of j-th term in this document. The most common measure, called *tf-idf* index (term frequency-inverse document frequency) [19], is defined by:

$$w_{i,j} = tf_{i,j} \cdot idf_j = \frac{n_{i,j}}{\sum_{k=1}^{M} n_{i,k}} \cdot \log\left(\frac{N}{|\{i : n_{i,j} \neq 0\}|}\right) \tag{1}$$

where $n_{i,j}$ is the number of occurrences of the term t_j in the document d_i.

Both standard TRSM and extended TRSM are the conceptual models for the Information Retrieval. Depending on the current application, different extended weighting schema can be proposed to achieve as highest performance as possible. Let us recall some existing weighting scheme for TRSM:

1. The extended weighting scheme is inherited from the standard TF-IDF by:

$$w_{ij}^* = \begin{cases} (1 + \log f_{d_i}(t_j)) \log \frac{N}{f_D(t_j)} & \text{if } t_j \in d_i \\ 0 & \text{if } t_j \notin \mathbf{U}_{\mathcal{R}}(d_i) \\ \min_{t_k \in d_i} w_{ik} \frac{\log \frac{N}{f_D(t_j)}}{1+\log \frac{N}{f_D(t_j)}} & \text{otherwise} \end{cases}$$

   This extension ensures that each term occurring in the upper approximation of d_i but not in d_i itself has a weight smaller than the weight of any terms in d_i. Normalization by vector's length is then applied to all document vectors: $w_{ij}^{new} = w_{ij}^* / \sqrt{\sum_{t_k \in d_i} (w_{ij}^*)^2}$ (see [14], [15]). The example of standard TRSM is presented in Table 1.
2. Explicit Semantic Analysis (ESA) proposed in [20] is a method for automatic tagging of textual data with predefined concepts. It utilizes natural language definitions of concepts from an external knowledge base, such as an encyclopedia or an ontology, which are matched against documents to find the best associations. Such definitions are regarded as a regular collection of texts, with each description treated as a separate document. The original purpose of ESA was to provide means for computing semantic relatedness between texts. However, an intermediate result – weighted assignments of concepts

Table 1. Example snippet and its two vector representations in standard TRSM.

Title: EconPapers: Rough sets bankruptcy prediction models versus auditor
Description: Rough sets bankruptcy prediction models versus auditor signalling rates. Journal of Forecasting, 2003, vol. 22, issue 8, pages 569-586. Thomas E. McKee. ...

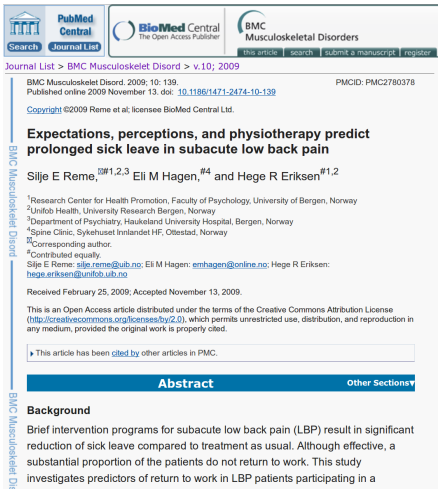
Original vector		**Enriched vector**	
Term	Weight	Term	Weight
auditor	0.567	auditor	0.564
bankruptcy	0.4218	bankruptcy	0.4196
signalling	0.2835	signalling	0.282
EconPapers	0.2835	EconPapers	0.282
rates	0.2835	rates	0.282
versus	0.223	versus	0.2218
issue	0.223	issue	0.2218
Journal	0.223	Journal	0.2218
MODEL	0.223	MODEL	0.2218
prediction	0.1772	prediction	0.1762
Vol	0.1709	Vol	0.1699
		applications	0.0809
		Computing	0.0643

to documents (induced by the term-concept weight matrix) may be interpret as a weighting scheme of the concepts that are assigned to documents in the extended TRSM.
Let $W_i = [w_{i,j}]_{j=1}^{N}$ be a bag-of-words representation of an input text d_i, where $w_{i,j}$ is a numerical weight of term t_j expressing its association to the text d_i. Let $s_{j,k}$ be the strength of association of the term t_j with a knowledge base concept c_k, $k \in \{1, \ldots, K\}$ an inverted index entry for t_j. The new vector representation, called a *bag-of-concepts* representation of d_i, is denoted by $[u_{i,1}, \ldots u_{i,K}]$, where: $u_{i,k} = \sum_{j=1}^{N} w_{i,j} s_{j,k}$. For practical reasons it is better to represent documents by the most relevant concepts only. In such a case, the association weights can be used to create a ranking of concept relatedness. With this ranking it is possible to select only top concepts from the list or to apply some more sophisticated methods that involve utilization of internal relations in the knowledge base. An example of top 20 concepts for an article from PubMed is presented in Figure 3

The described above weighting scheme naturally utilized in Document Retrieval as a semantic index [21, 22]. A user may query a document retrieval engine for documents matching a given concept. If the concepts are already assigned to documents, this problem is conceptually trivial. However such a situation is relatively rare, since employment of experts who could manually labelled documents from a huge repository is expensive. On the other hand, utilization of an automatic tagging method, such as ESA, allows to infer labeling of previously untagged documents. More sophisticated weighting schema have been proposed in, e.g. [23], [24].

1.4 The applications of TRSM in Semantic Web

Let us now briefly describe some applications of TRSM in semantic text analysis



The list of top 20 concepts:
"Low Back Pain", "Pain Clinics", "Pain Perception", "Treatment Outcome", "Sick Leave", "Outcome Assessment (Health Care)", "Controlled Clinical Trials as Topic", "Controlled Clinical Trial", "Lost to Follow-Up", "Rehabilitation, Vocational", "Pain Measurement", "Pain, Intractable", "Cohort Studies", "Randomized Controlled Trials as Topic", "Neck Pain", "Sickness Impact Profile", "Chronic Disease", "Comparative Effectiveness Research", "Pain, Postoperative"

Fig. 3. An example of a document and the list of top 20 concepts assigned by the semantic tagging algorithm in SONCA.

TRSM-base search: Let us recall that in TRSM, the upper approximations of documents can be used as an enriching bag-of-word document representations, and it can be applied in information retrieval systems. In [25], we supplement TRSM by a weight learning method in an unsupervised setting and apply the model to the problem of extending search results. We also introduce a method for a supervised multi-label classification problem and briefly compare it to an algorithm described in [23], which is based on Explicit Semantic Analysis [20]. The same model structure (defined by tolerance relations) can be also used for different searching tasks, e.g. inference of authors by defining a different structurality function.

Semantic indexing: document databases use external knowledge bases to facilitate the searching process. For example, bio-medical documents in PubMed are semi-manually tagged with concepts from MeSH. Queries sent to the database are then automatically extended by the corresponding MeSH headings. Indeed, the ontological part of our data model supports storage of information from different external knowledge bases, such as MeSH or DBpedia. Therefore, we may implement some universal methods for detecting associations between documents and concepts. The obtained tags can be then utilized in various processes, such as grouping of search results or topical classification (e.g.: automatic classification of documents into MeSH's topics).

The key concept of semantic indexing process is to assign to each document a new representation called *the bag-of-concepts*. As a step toward this direction, we implemented the extended TRSM algorithm, where natural language definitions of concepts from an encyclopedia or an ontology are matched against texts to find the best associations. Thus, we can easily construct an inverted semantic index that maps words occurring in such descriptions into related concepts. For each

Table 2. Exemplary tags assigned to documents by PubMed experts and SONCA. The "∗" in the "MeSH tags by PubMed" column indicates the primary headings.

Document title	MeSH tags by PubMed	MeSH tags by TRSM
Cockroaches (Ectobius vittientris) in an intensive care unit, Switzerland.	Cockroaches*, Insect Control*, Intensive Care Units*, Cross Infection, Insect Vectors	Cockroaches, Intensive Care Units, Klebsiella Infections, Pest Control, Cross Infection
Serotonin transporter genotype, morning cortisol and subsequent depression in adolescents.	Depressive Disorder*, Genetic Predisposition to Disease*, Serotonin Plasma Membrane Transport Proteins*, Genotype, Multilevel Analysis	Depressive Disorder, Genome-Wide Association Study, Multilevel Analysis, Cohort Studies, Adolescent Psychiatry
Capacity of Thailand to contain an emerging influenza pandemic.	Disaster Planning*, Health Policy*, Disease Outbreaks, Health Resources, Influenza Human	Health Care Rationing, Health Resources, Epidemics, Evidence-Based Medicine, Influenza B virus

new document, concepts that correspond to its words basing on such inverted index are retrieved and aggregated to form an extended bag-of-concepts.

Online document grouping. Online grouping methods utilize content of usually up to several hundreds snippets (contexts for the searched term occurrences) returned by the Web search engines. The output is a list of labeled groups assigned with some objects (typically Web pages). The goal of grouping is then to provide a navigational rather than a summary interface [26]. On the other hand, a document retrieval system can usually access higher quality information about documents, which sets up expectations at a different level. In such a case, the groups based merely on snippets' content may not be informative enough to provide a meaningful overview of documents returned by the query. This suggests that enriching snippets may lead to a higher quality clustering.

1.5 The accuracy and performance

The performance and quality tests undertaken so far on over 200K full-content articles resulting in 300M tuples confirm SONCA's scalability, which should be investigated not only by means of data volume but also ease of adding new types of objects that may be of interest for specific groups of users.

We applied the semantic indexing methods in combination with MeSH and DBpedia to index PubMed documents. We verified effectiveness of our approach in two ways. First, we clustered small subsets of documents represented by bag-of-words and bag-of-concepts using a simple k-means algorithm and found out that the semantic representation frequently yields better results [24]. We also compared the key MeSH concepts assigned to selected documents with the corresponding tags assigned by the PubMed experts. Preliminary results of this analysis reveal that the ESA method produces quite reasonable tags (see Table 2).

Table 3. A cluster labeled "Body Weight" discovered after a baseline document representation was extended with citation information. Column "Grouping (abstract)" shows original (baseline) groups assigned to each document (two of them were previously unassigned to any group). The third column lists MeSH terms associated with each document (these terms were unavailable for the fourth document). We emphasized concepts that seem (subjectively) to be similar to the group label.

Title	Grouping (abstracts)	MeSH keywords
Effects of antenatal dexamethasone treatment on glucocorticoid receptor and calcyon gene expression in the prefrontal cortex of neonatal and adult common marmoset monkeys.	Molecular; Dexamethasone	Age Factors; Animals; Animals, Newborn; **Body Size**; **Body Weight**; Callithrix; Dexamethasone; Female; Glucocorticoids; Male; Membrane Proteins; Prefrontal Cortex; Pregnancy; Prenatal Exposure Delayed Effects; Receptors, Glucocorticoid; Receptors, Mineralocorticoid; RNA, Messenger
The body politic: the relationship between stigma and obesity-associated disease		Adiposity; Age Factors; **Body Mass Index**; Electric Impedance; Female; Humans; Male; **Obesity**; Prejudice; Risk Factors; Sex Factors; Stress, Psychological
Prenatal Stress or High-Fat Diet Increases Susceptibility to Diet-Induced Obesity in Rat Offspring.	High-fat Diet	Animals; Child; Diabetes Mellitus, Type 2; **Dietary Fats**; **Energy Intake**; Female; Genetic Predisposition to Disease; Humans; Infant; Male; **Obesity**; Pregnancy; Prenatal Exposure Delayed Effects; Rats; Rats, Sprague-Dawley
The TNF-α System: Functional Aspects in Depression, Narcolepsy and Psychopharmacology.		

We conducted experiments which utilized document representations based on inbound and outbound citations (i.e.: the lists of documents that are referenced by and that reference each given paper), semantic indexes described earlier in this section, as well as snippets extended by document abstracts. MeSH terms assigned by the PubMed domain experts to documents provided natural means of validation for each of clustering methods, as ideally the system would group documents in a similar way that the experts would do it [26, 24]. Table 3 shows an example of cluster that was discovered after extending document representations by information about citations. We expect that extraction of more meaningful snippets can further improve our results in the nearest future.

The relational data model employed within DocDB enables smooth extension of the set of supported types of objects with no need to create new tables or attributes. It is also prepared to deal on the same basis with objects acquired at different stages of parsing (eg concepts derived from domain ontologies vs. concepts detected as keywords in loaded texts) and with different degrees of information completeness (eg fully available articles vs. articles identified as bibliography items elsewhere). However, as already mentioned, the crucial aspect is freedom of choice between different data forms and processing strategies while optimizing Analytic Algorithms, reducing execution time of specific tasks from (hundreds of) hours to (tens of) minutes.

1.6 Further Perspectives and Conclusions

SONCA (Search based on ONtologies and Compound Analytics) platform is developed at the Faculty of Mathematics, Informatics and Mechanics of the University of Warsaw. SONCA is expected to provide interfaces for intelligent algorithms identifying relations between various types of objects. It extends typical functionality of scientific search engines by more accurate identification of relevant documents and more advanced synthesis of information. To achieve this, concurrent processing of documents needs to be coupled with ability to produce collections of new objects using queries specific for analytic database technologies.

Ultimately, SONCA should be capable of answering the user query by listing and presenting the resources (documents, Web pages, etc.) that correspond to it *semantically*. In other words, the system should have some *understanding* of the intention of the query and of the contents of documents stored in the repository as well as the ability to retrieve relevant information with high efficacy. The system should be able to use various knowledge sources related to the investigated areas of science. It should also allow for independent sources of information about the analyzed objects, such as, e.g., information about scientists who may be identified as the stored articles' authors.

Our primary motivation to develop SONCA is to extend functionality of the currently available search engines towards document based decision support and problem solving, via enhanced search and information synthesis capabilities, as well as richer user interfaces. For this purpose, we have been seeking for inspiration in many projects and approaches, related to such fields as, e.g., semantic

web , social networks or hybrid information networks.Surely, there are plenty of aspects to be further investigated, in particular, in what form the results should be transmitted between modules and eventually reported to users. With this respect, we can refer to some research on, e.g., enriching original contents and linguistic summaries of query results.

Another challenge is how to manage a hierarchy of computational tasks in order to assembly the answers to compound queries. Basing on initial observations in Section 1.4, we can see that the framework for specifying intermediate components of search and reasoning processes is crucial for both performance and extendability of the system [27, 28]. The chain of computational specifications may follow a way human beings interact with standard search engines in order to summarize knowledge they are truly interested in. Thus, it is crucial to know how to represent and learn behavioral patterns followed by domain experts while solving problems [29]. Some hints in this area may come out from our previous research related to ontology-based approximations of compound concepts and identifying behavioral patterns in biomedical applications [30].

We also need to work on completion of the list of query types that should be supported. Besides examples mentioned in the previous sections, one may be interested in questions such as: "Who specializes in the treatment of a given condition (countries, states, hospitals)?"; "What are the current and past methods of diagnosis and treatment (e.g.: links to patient histories and medical images)?"; "Which pharmaceutical patents are relevant to treatment of the condition?".

Furthermore, the user-system dialog may go beyond answering to queries (see e.g. [31]). The system may be actually more active by means of proposing solutions, suggesting additional pieces of information that should be completed, or even identifying the existing pieces that might need to be reexamined. For example, let us imagine a SONCA-based diagnostic support system based on a repository of medical documents and clinical data sets, where a medical doctor should be able to enter information about a patient's history and, within a context of specific queries, expect some guidelines with regards to further medical treatment and, if necessary, further data acquisition and verification.

References

1. Z. Pawlak, *Rough sets: Theoretical aspects of reasoning about data.* Kluwer Dordrecht, 1991.
2. ——, "Granularity of knowledge, indiscernibility, and rough sets," in *Proceedings: IEEE Transactions on Automatic Control 20*, 1999, pp. 100–103.
3. L. T. Polkowski and A. Skowron, "Towards adaptive calculus of granules," in *Proceedings of the FUZZ-IEEE International Conference, 1998 IEEE World Congress on Computational Intelligence (WCCI'98)*, 1998, pp. 111–116.
4. H. S. Nguyen, A. Skowron, and J. Stepaniuk, "Granular computing: a rough set approach," *Computational Intelligence: An International Journal*, vol. 17, no. (no. 3), pp. 514–544(31, August 2001.
5. J. F. Peters, A. Skowron, Z. Suraj, W. Rzasa, and M. Borkowski, "Clustering: A rough set approach to constructing information granules," in *Soft Computing and*

Distributed Processing. Proceedings of 6th International Conference, SCDP, 2002, pp. 57–61.
6. H. S. Nguyen, "Approximate boolean reasoning: Foundations and applications in data mining," in *Transactions on Rough Sets V*. Springer, 2006, pp. 334–506.
7. H. S. Nguyen and T. B. Ho, "Rough document clustering and the internet," in *Handbook of Granular Computing*, W. Pedrycz, A. Skowron, and V. Kreinovich, Eds. Wiley & Sons, 2008, pp. 987–1004.
8. S. Asharaf, S. K. Shevade, and M. N. Murty, "Rough support vector clustering." *Pattern Recognition*, vol. 38, no. 10, pp. 1779–1783, 2005.
9. P. Lingras, "Unsupervised rough set classification using gas," *Journal of Intelligent Information Systems*, vol. 16, no. 3, pp. 215–228, 2001.
10. K. Voges, N. Pope, and M. Brown, "Cluster analysis of marketing data: A comparison of k-means, rough set, and rough genetic approaches," *Heuristics and Optimization for Knowledge Discovery, Idea Group Publishing*, vol. 208216, 2002.
11. P. Lingras, M. Hogo, and M. Snorek, "Interval set clustering of web users using modified kohonen self-organizing maps based on the properties of rough sets," *Web Intelligence and Agent Systems*, vol. 2, no. 3, pp. 217–225, 2004.
12. S. Hirano and S. Tsumoto, "Rough clustering and its application to medicine," *Journal of Information Science*, vol. 124, pp. 125–137, 2000.
13. P. Lingras and C. West, "Interval set clustering of web users with rough k-means," *Journal of Intelligent Information Systems*, vol. 23, no. 1, pp. 5–16, 2004.
14. S. Kawasaki, N. B. Nguyen, and T. B. Ho, "Hierarchical document clustering based on tolerance rough set model," in *Proceedings of PKDD 2000, Lyon, France*, ser. Lecture Notes in Computer Science, D. A. Zighed, H. J. Komorowski, and J. M. Zytkow, Eds., vol. 1910. Springer, 2000.
15. T. B. Ho and N. B. Nguyen, "Nonhierarchical document clustering based on a tolerance rough set model," *International Journal of Intelligent Systems*, vol. 17, no. 2, pp. 199–212, 2002.
16. A. Skowron and J. Stepaniuk, "Tolerance approximation spaces," *Fundamenta Informaticae*, vol. 27, no. 2-3, pp. 245–253, 1996.
17. S. H. Nguyen, G. Jaśkiewicz, W. Świeboda, and H. S. Nguyen, "Enhancing search result clustering with semantic indexing," in *Proceedings of the Third Symposium on Information and Communication Technology*, ser. SoICT '12. New York, NY, USA: ACM, 2012, pp. 71–80.
18. G. Virginia and H. S. Nguyen, "Investigating the effectiveness of thesaurus generated using tolerance rough set model," in *ISMIS*, ser. Lecture Notes in Computer Science, M. Kryszkiewicz, H. Rybinski, A. Skowron, and Z. W. Ras, Eds., vol. 6804. Springer, 2011, pp. 705–714.
19. R. Feldman and J. Sanger, Eds., *The Text Mining Handbook*. Cambridge University Press, 2007.
20. E. Gabrilovich and S. Markovitch, "Computing semantic relatedness using wikipedia-based explicit semantic analysis," in *Proceedings of the 20th international joint conference on Artifical intelligence*, ser. IJCAI'07. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2007, pp. 1606–1611.
21. A. Hliaoutakis, G. Varelas, E. Voutsakis, E. G. M. Petrakis, and E. Milios, "Information retrieval by semantic similarity," *Int. Journal on Semantic Web and Information Systems (IJSWIS). Special Issue of Multimedia Semantics*, vol. 3, no. 3, pp. 55–73, 2006.
22. A. M. Rinaldi, "An ontology-driven approach for semantic information retrieval on the web," *ACM Trans. Internet Technol.*, vol. 9, pp. 10:1–10:24, July 2009.

23. A. Janusz, W. Swieboda, A. Krasuski, and H. S. Nguyen, "Interactive document indexing method based on explicit semantic analysis," in *RSCTC*, ser. Lecture Notes in Computer Science, J. Yao, Y. Yang, R. Slowinski, S. Greco, H. Li, S. Mitra, and L. Polkowski, Eds., vol. 7413. Springer, 2012, pp. 156–165.
24. M. Szczuka, A. Janusz, and K. Herba, "Clustering of Rough Set Related Documents with use of Knowledge from DBpedia," in *Proc. of the 6th Int. Conf. on Rough Sets and Knowledge Technology (RSKT)*, ser. LNAI, vol. 6954. Springer, 2011, pp. 394–403.
25. W. Swieboda, M. Meina, and H. S. Nguyen, "Weight learning for document tolerance rough set model," in *Rough Sets and Knowledge Technology 2013, LNAI 8171*, 2013, pp. pp. 385396,.
26. H. S. Nguyen and T. B. Ho, "Rough Document Clustering and the Internet," in *Handbook of Granular Computing*, W. Pedrycz, A. Skowron, and V. Kreinovich, Eds. New York, NY, USA: John Wiley & Sons, Inc., 2008, pp. 987–1003.
27. J. Barwise and J. Seligman, *Information Flow: The Logic of Distributed Systems*. Cambridge University Press, 1997.
28. L. G. Valiant, "Robust Logics," *Artif. Intell.*, vol. 117, no. 2, pp. 231–253, 2000.
29. V. Vapnik, "Learning Has Just Started (An interview with Vladimir Vapnik by Ran Gilad-Bachrach)," 2008. [Online]. Available: http://seed.ucsd.edu/joomla/index.php/articles/12-interviews/9-qlearning-has-just-startedq-an-interview-with-prof-vladimir-vapnik
30. J. G. Bazan, "Hierarchical Classifiers for Complex Spatio-temporal Concepts," *Transactions on Rough Sets*, vol. 9, pp. 474–750, 2008.
31. J. M. Tenenbaum and J. Shrager, "Cancer: A Computational Disease that AI Can Cure," *AI Magazine*, vol. 32, no. 2, pp. 14–26, 2011.