International Journal of Approximate Reasoning 82 (2017) 210–226

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

www.elsevier.com/locate/ijar

A three-way approach for learning rules in automatic knowledge-based topic models ☆

Muhammad Taimoor Khan [a,b,*], Nouman Azam [a], Shehzad Khalid [b], JingTao Yao [c]

[a] National University of Computer and Emerging Sciences, Pakistan
[b] Bahria University, Islamabad, Pakistan
[c] Department of Computer Science, University of Regina, Regina, SK, S4S 0A2, Canada

ARTICLE INFO

Article history:
Received 20 July 2016
Received in revised form 15 December 2016
Accepted 19 December 2016
Available online 29 December 2016

Keywords:
Topic models
Automatic knowledge-based models
Game-theoretic rough sets
Three-way decisions

ABSTRACT

Topic modeling aims to uncover hidden thematic structures in a collection of documents by representing them as a set of topics. Automatic knowledge-based topic models are recently introduced to meet the demands of processing large-scale text collections. They are based on automatic extraction of rules from multiple domain corpuses. Generally, the extracted rules are large in number and some thresholds are used to select only a small number of useful rules. There are two shortcomings in this for selecting important rules. Firstly, they are based on fixed thresholds for extracting rules from all domain corpuses. Secondly, the thresholds are predefined or explicitly set by expert opinions and are not based on automated mechanisms. In this article, we address these shortcomings by considering a three-way approach based on rules having strong positive associations, rules having strong negative associations and rules having weak associations. A pair of thresholds defines and controls the three-way partitioning of the rules. It is argued that the domain specific and automated selection of thresholds in the three-way framework may be approached from the viewpoint of a tradeoff between the quantity of rules and the quality of rules. We apply the game-theoretic rough set (GTRS) model to implement this tradeoff. Algorithms using the GTRS are introduced for automatically determining the thresholds. Experimental results on Chen2014 dataset suggest an average improvement of 52.82 points in topic coherence by increasing the quantity of rules to 17.93%.

1. Introduction

Topic modeling deals with extracting and uncovering hidden thematic structures in collection of documents (also known as a domain) [12]. The data to be processed may consist of multiple domains, each having documents about a subject. The essential idea is to form groups of words that exhibit strong relationship across the documents in a domain [58]. Each group represents a realization of a conceptual topic. The words within the group or topic depict different ways of referring to the topic [9,12,39,46,59]. For instance, a group containing the words such as *dna, gene, genetic, life, organism* and another group containing the words *computer, data, processing, storage* may represent topics related to *biology* and *computing*, respectively.

☆ This paper is part of the virtual special issue on tri-partition, edited by Davide Ciucci and Yiyu Yao.

* Corresponding author at: National University of Computer and Emerging Sciences, Pakistan.
E-mail address: taimoor.khan@nu.edu.pk (M.T. Khan).

http://dx.doi.org/10.1016/j.ijar.2016.12.011
0888-613X/

The words in a topic are typically ordered based on their relevance and relationship to the topic. Topic models are frequently used in applications such as automated domain exploration, text summarization, aspect extraction, sentiment analysis and other natural language processing tasks [9,28,46,47].

The conventional model for topic extraction is known as the Latent Dirichlet Allocation (LDA) model [9–12]. It is based on a sampling technique for combining words under the topics by considering the document-topic and word-topic relationships. Document-topic relationship considers the association and distribution of topics across the documents. Similarly, the word-topic relationship shows the distribution of words into topics. By exploiting these two relationships iteratively, the words are assigned to different topics until they do not move between the topics [9,10,40,46,73]. There are several extensions of the LDA model including supervised models, hybrid models, semi-supervised models, transfer learning models, knowledge-based models and automatic knowledge-based topic models [50]. Due to automated nature, they are better suited to meet with the demands of processing large scale text collections [16]. The importance of automatic knowledge-based topic models is that it requires least user involvement compared to other techniques while producing comparable results [27]. For this reason, we consider automatic knowledge-based topic models in this research.

The automatic knowledge-based topic models are based on selection of important rules for extracting topics. Typically, the selection of rules are based on some evaluation criteria and thresholds [18,22,32]. The evaluation criteria reflect the quality of rules and the thresholds are used to select most important ones. The evaluation criteria are commonly based on association measures [52]. Some typical evaluation measures include point-wise mutual information, mutual information, and others adopted from frequent pattern mining and frequent itemset mining [1,13,21,23]. The determination of thresholds plays a critical role in selection of rules. The existing studies however suffers from two limitations with regards to thresholds selection. The first limitation is the use of fixed thresholds for processing and extracting rules from all domain corpuses. This means that individual characteristic of domain corpuses are ignored in determining the thresholds, such as, vocabulary size, document sample size and organization of words in documents. The second limitation is the lack of automated mechanisms for determining the thresholds. The thresholds are either predefined or set based on experts opinions. This means that we have to rely on experts' intuitions rather than some scientific justification for explaining the choice of using certain thresholds. Please be noted that the word automatic in automatic knowledge-based model refers to the fact that the rules are not given by experts. The expert based thresholds is associated with other difficulties such as scalability, i.e., when we are interested in domain specific thresholds, it may not be always possible to find experts for providing domain specific thresholds for all the domains. Moreover, the cost of expert consultation may not be always reasonable.

In this article, we address these limitations by considering a three-way approach. The three-way approach employs a pair of thresholds to determine three types of rules, namely, rules having strong positive associations, rules having strong negative associations and rules having weak associations. The rules having strong positive and negative associations are selected and the rules with weak associations are ignored. The configuration of thresholds controls the selection of different types of rules. We argue that the effective selection of thresholds may be realized as a tradeoff between two important aspects of rules, i.e., the quantity of rules and the quality of rules. Adjusting the thresholds to increase or improve one aspect may lead to a decrease in another. We apply the game-theoretic rough set (GTRS) model to determine an effective tradeoff solution between the two aspects. The three-way approach based on the GTRS overcomes the two limitations by determining the thresholds for each domain automatically and provides a tradeoff perspective justifying the selection of thresholds. New algorithms based on GTRS are presented for selecting rules in automatic knowledge based models. Experiments are performed on Chen2014 dataset [20]. The proposed model provides an average of 52.82 points improvement in topic coherence (TC) when the rules are increased by 17.93%.

2. Background

This section elaborates on the literature background of topic modeling and rough sets. Topic modeling identifies unknown structures from known data and has various extensions to support different types of analysis. Rough sets offer a three-way interpretation of the problem space. They are discussed in the following subsections.

2.1. Topic models

The Latent Dirichlet Allocation (LDA) model is the most common and fundamental model for topic modeling [9–12]. It is a probabilistic model that treats data as arising from a generative process [9]. The process consists of identifying hidden topic structures based on observable information such as words in a document [64,73]. The model is initialized by randomly choosing a topic for each word in each document [50]. A sampling technique is then used to iteratively update the word-topic association for a chosen word based on per-document topic distribution and per-topic word distribution. Topics associated with most occurrences of the sampled word and having higher representation in the document of the sampled word are more likely chosen as new topic [14,40]. The words are chosen randomly from the corresponding distribution over the vocabulary [10]. The procedure stops when the newly chosen topics for sampled words are the same as their existing topics [41]. Another interesting aspect is that given the topic structure, topic models can be used to generate new documents [27].

The basic LDA model has been extended in a variety of ways and lead to models such as supervised models, hybrid models, transfer learning models, semi-supervised models, knowledge-based models and automatic knowledge-based mod-

els. Supervised topic models restrict the model to select from a set of topics labeled for each document [46,54]. Hybrid models incorporate both the supervised and unsupervised models. In particular, supervised model is used to train on syntactic features and the unsupervised model make use of it for topic modeling [30,43,57,72]. Semi-supervised models are manually tuned with seed words for each topic by a domain expert [44,49,63]. In transfer learning models, the results obtained from one domain are stored and later on applied to another similar domain [31,65]. In knowledge-based models, rules are being manually supplied by domain experts for extracting topics [2,3,19]. Automatic knowledge-based topic models have a self learning mechanism to select rules based on the data itself [16–18,22,32]. We consider automatic knowledge based topic models due to the volume and variety of data in large-scale datasets.

Automatic knowledge-based topic models were introduced as an extension to knowledge-based topic models for scaling them to large scale textual data [18]. Knowledge-based topic models require the domain experts to provide knowledge rules [26,29,49]. The knowledge rules are of two types, i.e. positive rules and negative rules. They are in the form of word pairs. The words in positive rules are positively associated, which indicates their higher coexistence in documents across all domains and vice versa. When the probability of a sampled word w is increased for a topic t. Using the rules containing word w, the probability of the accompanying words in positive rules is increased for topic t as well. While the accompanying words in negative rules have their probabilities decreased for topic t. The positive and negative rules are also sometimes referred to as must links and cannot links in the literature [3]. The rules add bias to the inference model and lead to topics containing words having higher semantic relevance and association between each other. The manual feeding of the rules to the systems is not always feasible and effective especially when the text collections are large and from diverse subject domains. In other words, they suffer from limitation of scalability. Automatic knowledge-based models were introduced to overcome this limitation [17,22,32].

Automatic knowledge-based models typically employ a two step procedure for automatically generating rules [18]. In step one, the set of candidate rules are generated. In step two, the rules are evaluated using evaluation criteria and then thresholds are used to select important rules based on the evaluations. In regards to step one, word pairs are generated from vocabulary using topic clusters, topic transactions and from within topics [17,18,32]. In regards to step two, the evaluation criteria are typically based on association measures. In particular, the association measures compute the strength of relationship between word pairs which provide hints for forming potential rules. Some important association measures may be found in [13,56,60]. There is a lack of consideration in existing studies with regards to selection of using suitable thresholds. The generally used approach is to fix them for all the domain corpuses. For instance, fixed thresholds are used on Chen2014 dataset in [16,32] and predetermined thresholds are used on electronic product domains consisting of real users' reviews on Amazon.com in [17,18]. Since each domain has a specific vocabulary, document samples and word organizations. The thresholds are therefore to be adjusted for each domain and should not be fixed for all the domains. Moreover, the threshold values are defined based on expert intuition, thereby lack in providing an automated mechanism for their determination. In order to overcome these limitations, we propose a three-way approach for selecting important rules.

2.2. Three-way decisions

The notion of three-way decisions is primarily motivated and developed in the framework of rough sets [36–38,66,69]. In particular, the region based interpretation of approximations of an undefinable set by a definable set leads to the development of the theory of three-way decisions [15,69]. It is however, important to note that the essential ideas of three-way decisions are not new. Similar ideas may be found in medical decision making under the threshold approach [51], in psychology under the deferred decision making approach [55,61], in machine learning under the partial classification approach [8], in fuzzy set under the shadowed set approach [53], in statistics under three-way hypothesis testing [62] and in logic under the three-valued or Kleene's logic approach [33].

The fundamental notion in the theory of three-way decisions adopted from rough sets is the partitioning of the universal set into three pairwise disjoint regions [71]. It is recently argued that an equally important consideration is the construction of effective and efficient strategies for processing the three regions [69]. The realization of these two essential components has lead to the trisecting and acting framework of three-way decisions [42,69]. According to the framework, three-way decisions may be interpreted as a two step process. In step one, we seek tripartition or trisection of the universe and in step two, strategies are designed for effectively processing the three regions in the aim to obtain three-way decisions [35]. The three regions may be referred to as POS, NEG, BND [68], right, left, middle [69] or simply R-region, L-region, M-region [70]. We use the POS, NEG and BND notations.

The partitioning of the universe into three regions is typically based on an evaluation function and a pair of thresholds. Let $e(x) : U \longmapsto \mathbb{V}$ be an evaluation function which assigns and maps to every object $x \in U$ an evaluation value from a totally ordered set $(\mathbb{V}, \preceq)$. We partition U based on e by considering a threshold pair $(\alpha, \beta) \in \mathbb{V} \times \mathbb{V}$ with $\alpha \succ \beta$ as,

$$\mathrm{POS}_{(\alpha,\beta)} = \{x \in U \,|\, e(x) \succeq \alpha\}, \tag{1}$$

$$\mathrm{NEG}_{(\alpha,\beta)} = \{x \in U \,|\, e(x) \preceq \beta\}, \tag{2}$$

$$\mathrm{BND}_{(\alpha,\beta)} = \{x \in U \,|\, \beta \prec e(x) \prec \alpha\}. \tag{3}$$

This means that all objects with evaluation greater than α are assigned to POS region and all objects with evaluation smaller than β are assigned to the NEG region. In any other case, the objects are assigned to the BND region. The three-way decision

Table 1
Words occurrence and co-occurrence in domain corpus 1.

	Gene	DNA	Genetic	Life	Evolve	Organism	Brain	Neuron	Nerve	Data	Number	Computer
Gene	100	70	60	90	70	80	60	70	80	30	10	20
Dna		120	100	80	90	100	80	70	90	40	20	50
Genetic			130	110	80	90	80	100	90	40	50	20
Life				120	80	80	70	60	80	10	20	20
Evolve					130	70	100	110	120	20	30	50
Organism						110	100	80	70	30	50	30
Brain							100	80	80	30	40	10
Neuron								150	70	30	50	10
Nerve									140	30	40	20
Data										120	80	90
Number											150	120
Computer												130

Table 2
Words occurrence and co-occurrence in domain corpus 2.

	Disk	Port	Cable	Screen	Shape	Control	Ram	Script	Stage	Role
Disk	100	90	70	80	70	90	60	20	10	10
Port		100	80	80	90	70	80	30	30	40
Cable			110	90	80	80	90	40	30	20
Screen				130	100	80	80	30	10	20
Shape					120	80	90	50	60	30
Control						90	30	60	50	60
Ram							100	50	10	20
Script								120	100	110
Stage									110	110
Role										120

framework presented in Equations (1)–(3) is referred to as evaluation based framework for three-way decisions [68]. Some of the fundamental issues with regards to the above framework are construction and interpretation of evaluation function, determination of pair of thresholds, and measurement of the quality of tripartition [25,68]. We will return to these issues in Section 3.2.

3. Determination of thresholds for selecting rules

In this section, we highlight the limitations in existing models for selecting rules based on the thresholds. A demonstrative example is used for this purpose. Later in the same section, we will provide a solution on how to automatically adjust the thresholds based on the data itself.

3.1. Limitations in selecting rules with fixed and expert based thresholds

Consider two domain corpuses denoted as corpus 1 and corpus 2 consisting of 500 and 1,000 documents, respectively. The possible words or vocabulary in corpus 1 are *gene, dna, genetic, life, evolve, organism, brain, neuron, nerve, data, number, computer* and in corpus 2 are *disk, port, cable, screen, shape, control, ram, script, role*. Tables 1 and 2 show these words and their co-occurrences in the documents belonging to the two corpuses. A particular entry of these tables represents the number of documents in which the respective words co-occur. It may be noted that the two corpuses have different documents, vocabulary and organization of words into documents. We consider two types of rules based on the above vocabulary. A positive rule containing word pairs, such as *genetic* and *evolve*, which suggests for the presence of the two words under the same topic. A negative rule containing word pairs, such as *data* and *nerve*, which suggests that these words were hardly discussed together in a document and should belong to separate topics (see Table 3).

To evaluate the effectiveness of rules, we consider the measure of normalized point-wise mutual information (NPMI) [32]. The importance of this measure is that in contrast to other measures, it can be used to evaluate both positive and negative rules. It normalizes the value of Point-wise Mutual Information (PMI) into a range of $[-1, +1]$ to facilitate in applying thresholds. PMI shows the measure of association between the two words. However, the value of PMI could not be in a specific range. The NPMI for a candidate rule containing words pairs (w_1, w_2) is computed as [32],

$$\text{NPMI}(w_1, w_2) = -\frac{\text{PMI}(w_1, w_2)}{log(P(w_1, w_2))}, \tag{4}$$

where,

$$\text{PMI}(w_1, w_2) = log\frac{P(w_1, w_2)}{P(w_1)P(w_2)}. \tag{5}$$

Table 3
Topics in the two domain corpuses.

Topics in corpus 1			Topics in corpus 2		
Topic 1	Topic 2	Topic 3	Topic 1	Topic 2	Topic 3
organism	gene	computer	disk	screen	srcipt
evolve	dna	data	port	shape	stage
life	nerve	number	cable	control	role
genetic	neuron	evolve	ram		
brain	life	life			

A positive value indicates that the two words may form a positive rule and a negative value indicates that the two words may form a negative rule. The value of the measure reflect the strength or degree of a rule. A higher positive value indicates that a rule has a higher chance for being selected as a positive rule and a higher negative value indicates that the rule has a higher chance for being selected as a negative rule. The NPMI score may be computed based on the data in Table 1 and Table 2. For instance, for a candidate rule consisting of a word pair *genetic* and *life*, we have, $P(genetic) = 120/500$, $P(life) = 130/500$ and $P(genetic, life) = 110/500$. Substituting these values in Equation (5), we have PMI(genetic, life) = 1.26 which leads to NPMI(genetic,life) = 0.83. The degree of the rules vary between the two extremes of -1 and $+1$, thereby making some rules more important compared to others. Using all the rules is typically not very feasible due to computational and performance bottlenecks. A pair of thresholds are therefore being used to select important rules. The rules whose NPMI values are below t_1 are selected as negative rules and the rules with NPMI values above t_2 are selected as positive rules. An important issue in this context is how to determine the thresholds. A plausible way will be to consider thresholds based on improvement or enhancement in some performance aspects.

We consider two important aspects for highlighting the effectiveness of the rules, namely, the quality and quantity of rules. To measure and represent the quality of rules, we use the measure called topic coherence (TC) [48]. To measure the quantity of rules, we consider the measure called relative number of rules. We now define each of these measure.

The TC shows the association or correlation between the words in a topic [48]. Topics with contextually correlated words have high TC and are termed as coherent topics. The TC for the jth topic in the ith domain corpus (where the topic is extracted based on thresholds (t_1, t_2) is given by,

$$TC_{T_{(i,j)}}(t_1, t_2) = \sum_{w, w' \in T_{(i,j)}} log(P(w|w')) \tag{6}$$

where $T_{(i,j)}$ represents the jth topic in the documents belonging to the ith domain corpus and w and w' are two distinct words in the topic $T_{(i,j)}$. Topic coherence is the double sum across all the words in the topic. It gives the measure of similarity between words representing a topic. The words with highest probability in a topic are chosen to represent it [48]. Usually top 30 words are used in the literature to represent a topic. The TC for a domain corpus is determined by averaging the TC over all the topics in that domain. Assuming there are K distinct topics in ith domain corpus, the overall TC for the ith domain corpus is given by,

$$TopicCoherece(t_1, t_2) = \frac{\sum_{j=1}^{K} TC_{T_{(i,j)}}(t_1, t_2)}{K} \tag{7}$$

Equation (7) means that the TC of a domain is the average TC of all its topics. Please be noted that the TC will results in a negative value with values closer to zero representing strongly coherent topics. The measure of relative rules based on thresholds (t_1, t_2) is computed as,

$$RelativeRules(t_1, t_2) = \frac{\text{Selected rules based on } (t_1, t_2)}{\text{Total rules}} \times 100 \tag{8}$$

It reflects the quantitative aspect of rules with respect to the total possible rules. One may aim at higher quality and lesser quantity of rules. It should be noted that in some domains, it may be more useful to assign different weights to topics while evaluating the overall TC i.e., some topics may be considered as more important than the others. For instance, when topics are considered as features of a product, we may assign more weights to some features based on customer preferences. However, the automatic knowledge-based technique does not require any human intervention. Moreover, it does not involve any user specified preferences about the topics. The probabilities of topics are therefore not considered and the topics are treated equally. Therefore, the probabilities of topics are ignored and the topics are treated equally.

We now return to the two limitations discussed in Section 2. Table 4 is constructed for this purpose. The table contains different thresholds and the resulting values of the measures of TC and relative rules based on Equations (7)–(8). Recall the first limitation, i.e., using fixed thresholds for all domain corpuses. We may note from Table 4, that changes in thresholds affect the measures differently when computed for the two corpuses. For instance, when thresholds are changed from $(-0.8, 1.0)$ to $(-0.8, 0.9)$ the TC for corpus 2 increases and for corpus 1 it decreases. We select rules based on per domain basis not on per topic basis. This is a more reasonable choice since we do not have any prior knowledge about the topics that may be present. However, in some supervised scenarios, rules based on topics may be considered. This means that

Table 4
TC and relative rules at different thresholds for the two corpuses.

(t_1, t_2)	Corpus 1		Corpus 2	
	Relative..(t_1, t_2)	*Topic*..(t_1, t_2)	*Relative*..(t_1, t_2)	*Topic*..(t_1, t_2)
$(1.0, -1.0)$	4.5	−880	2.22	−782
$(0.8, -1.0)$	10.6	−868	35.55	−763
$(0.9, -0.9)$	9.1	−854	13.33	−778
$(0.8, -0.9)$	12.12	−841	37.77	−761
$(1.0, -0.8)$	10.6	−859	11.11	−780
$(0.9, -0.8)$	13.63	−873	20	−771
$(0.8, -0.8)$	16.66	−878	44.44	−752
...	...	...	...	...
$(0.0, 0.0)$	100	−897	100	−805

when we aim to configure and determine the thresholds for one corpus (based on some performance measures), the same thresholds may not necessarily provide effective results on another corpus. This is due to different organization of words into documents, in each domain corpus. This calls for domain specific thresholds instead of using fix thresholds for processing all domains.

The second limitation discussed in Section 2 is the lack of automated mechanisms for justifying the selection of thresholds. In the above example, suppose that the user or expert based on his intuition selects thresholds $(-0.9, 0.8)$ for corpus 1. Although, we know from our example that this selection is reasonable as it provides the maximum TC with not many rules. However, the user may not be able to provide an acceptable justification such as a procedure, a process, a mechanism or some mathematical description for his choice. This puts questions on the reliability and repeatability of always selecting effective thresholds. Moreover, the cost of consulting experts for providing thresholds for each corpus may not be always effective. In the next section, we discuss on how these limitations may be addressed by considering a three-way approach.

3.2. Addressing the limitations with a three-way approach

We noted from previous section that the configuration of the thresholds affect two measures differently. At one extreme configuration of thresholds, i.e., $(t_1, t_2) = (-1, +1)$, we select the minimum number of rules with highest association or quality. However, the TC of the model with such rules may not be very effective. For instance, in the previous example, we have $TC(-1, 1) = -880$ and $TC(-1, 1) = -782$ for the two corpuses. In an another extreme setting of thresholds, i.e., $(t_1, t_2) = (0, 0)$, we select all the rules. We note again the TC for such thresholds may not be necessarily effective. Again, for instance in the example, we have $TC(0, 0) = -805$ for the two corpuses. In general, the aspects of relative rule (RR) and TC are affected differently when thresholds are being modified. The determination of suitable and effective thresholds may be considered by realizing a tradeoff between the two aspects. We consider this tradeoff based on a three-way interpretation of rules.

The three-way interpretation of rules is based on dividing the set of all possible rules based on thresholds (t_1, t_2) into three regions, namely, rules having strong positive associations, rules having strong negative associations and rules having weak association. Rules in the first two regions are selected while the rules in the third region are ignored. All the rules are evaluated individually by employing an evaluation criterion such as NPMI in this case. The rules with evaluation score smaller than t_1 are considered as negative rules and with greater than t_2 are considered as positive rule. The rules with evaluation score between t_1 and t_2 are considered weak rules. Fig. 1 shows the three-way interpretation of rules based on the trisecting and acting framework. In the trisecting step, we partition the rules into three regions and in the acting step, we take actions of selecting and ignoring rules. It is important to note that in contrast to existing models, where the positive and negative rules are selected separately, the three-way interpretation of rules provides a single framework for selecting the rules. In this article, we consider a game-theoretic rough sets model to determine three-way partitioning of rules. In particular, it implements a tradeoff between the measures of TC and RR. For the sake of completeness, we review the main ideas of the GTRS.

4. Three-way division using game-theoretic rough sets

The GTRS formulates a game in the aim to determine thresholds of three-way decisions [24,67]. Typically a game in GTRS has three important components i.e. game players, strategies and their payoffs or the utility functions. The components in a game are represented as tuple $\{P, S, u\}$. Each of these components are explained in detail.

Game players: The game players are denoted by set P having n number of players. However, to keep it simple, a two player game is commonly preferred in GTRS. The nature of game players depend on the overall game objectives and goals. For instance, region uncertainty is analyzed with players defined as the uncertainty of the immediate and deferred decision regions [5]. In another scenario of seeking for a balanced rough set model, accuracy and generality were used as players [6]. The players are selected to highlight the overall purpose of the game. GTRS define these players on aspects and properties of rough sets based classification and decision making such as accuracy, generality, precision and uncertainty.

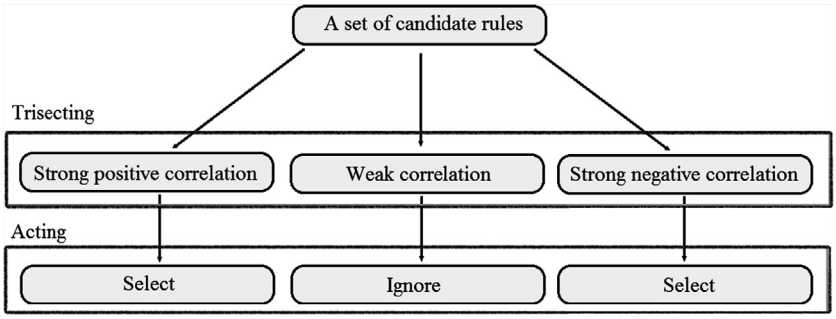

Fig. 1. Three-way interpretation of rules based on trisecting and acting framework.

Strategies: A player participates in the game by playing different strategies. The set of strategies available to player i are s_i. The Cartesian product of the possible strategy sets is denoted by $S = S_1 \times S_2 \times ... \times S_n$, where S contains ordered pairs of the form $(s_1, s_2, ..., s_n)$ such that $s_1 \in S_1$, $s_2 \in S_2$ and $s_n \in S_n$. Each order pair in S is called a strategy profile which represents a certain situation encountered in a game.

The strategies in GTRS are considered as modifications in the (t_1, t_2) thresholds. Based on the initial values of these thresholds, different types of strategies can be defined. For instance, when the initial values of (t_1, t_2) are set to $(0, 1)$ then the strategies are formulated such that t_1 and t_2 are decreased. Alternatively, when the initial values are set of (t_1, t_2) are set to $(-1, +1)$, then the strategies are formulated to increase t_1 and decrease t_2. Please note that in order to keep the regions disjoint, it is assumed that $-1 \leq t_1 < t_2 <\leq 1$. The players in a game devise strategies to modify thresholds until the final configuration of the thresholds is achieved.

Payoff functions: The payoff functions available to a player are shown by a set $u = (u_1, ..., u_n)$. Each u_i holds a real value as a utility function for player i and is represented as $u_i : S \longmapsto \Re$. The payoff values reflect the appropriateness or utility of performing or selecting a specific strategy. As the game players in GTRS possess aspects or properties that would facilitate three-way partitioning, thus the payoff function for a player depends on a measure that is employed to evaluate the properties of that player. Please be noted that the tuples represent the game and the ordered pairs represent the strategy profiles within the game.

Every player in the game wants to perform a strategy that would maximize their payoff. However, the strategies selected by a player has an impact on the payoff of opponent players as well. A better game solution is the one that offers a balance and tradeoff point based on the utilities of all the players. Such a solution in GTRS is commonly verified using Nash equilibrium.

A strategy profile without a specific player i is shown as $s_{-i} = (s_1, s_2, ..., s_{i-1}, s_{i+1}, ..., s_n)$. Similarly, the strategy profile $(s_1, s_2, ..., s_n)$ can be shown with the revised notation in the form of (s_i, s_{-i}). The strategy profile $(s_1, s_2, ..., s_n) = (s_i, s_{-i})$ satisfies Nash equilibrium as [34],

$$\forall i, \forall s_i' \in S_i, \quad u_i(s_i, s_{-i}) \geq u_i(s_i', s_{-i}), \quad \text{where } (s_i' \neq s_i) \tag{9}$$

It can be interpreted that for a player i with available strategies as s_i, is their best response to s_{-i}. Thus, a Nash equilibrium reached when none of the players are benefited with the new strategy.

5. A GTRS based approach for selecting rules

In this section, we introduce a GTRS based approach for determining the three-way thresholds for selecting important rules.

5.1. Game formulation based on a tradeoff between qualitative and quantitative aspects

From the explanation of GTRS in Section 4, we need to identify at least three game components in order to formulate and analyze problems with GTRS. This includes description of game players, the strategies that each player can perform and the utilities or payoff functions for each player. We now discuss each of them in detail.

The players are expected to reflect the overall purpose of the game. The objective in this game is to obtain effective three-way classification of rules by determining suitable thresholds. Since we are interested in evaluating the overall effectiveness of the rules based on the properties of RR and TC. Therefore, we consider these two properties as game players. The set of players in this case is given by $P = \{p_1, p_2\}$ with player p_1 representing percentage of rules and is denoted by P_{RR} and the p_2 representing the TC and is denoted by P_{TC}.

The players are affected differently when different thresholds (t_1, t_2) are considered. The strategies which represent possible moves of a player in a game setting are therefore represented in terms of different changes in thresholds. Each player is more interested in selecting a strategy that will maximize its benefits or utilities. By employing different strategies

Table 5
Payoff table for the game.

		P_{TC}		
		$s_1 = t_{1\downarrow}$	$s_2 = t_{2\uparrow}$	$s_3 = t_{1\downarrow}t_{2\uparrow}$
P_{RR}	$s_1 = t_{1\downarrow}$	$u_{RR}(s_1, s_1), u_{TC}(s_1, s_1)$	$u_{RR}(s_1, s_2), u_{TC}(s_1, s_2)$	$u_{RR}(s_1, s_3), u_{TC}(s_1, s_3)$
	$s_2 = t_{2\uparrow}$	$u_{RR}(s_2, s_1), u_{TC}(s_2, s_1)$	$u_{RR}(s_2, s_2), u_{TC}(s_2, s_2)$	$u_{RR}(s_2, s_3), u_{TC}(s_2, s_3)$
	$s_3 = t_{1\downarrow}t_{2\uparrow}$	$u_{RR}(s_3, s_1), u_{TC}(s_3, s_1)$	$u_{RR}(s_3, s_2), u_{TC}(s_3, s_2)$	$u_{RR}(s_3, s_3), u_{TC}(s_3, s_3)$

the players will compete with each other. The formulation of strategies based on the thresholds was studied in [4]. The study suggested different approaches such as the two ends approach, the middle start approach, the random approach and the range approach [4]. These approaches essentially differs in the initial setting of thresholds. We consider the two ends approach in this study as it provides useful results in the previous studies [5,24]. According to the two ends approach, the strategies are formulated based on the initial thresholds of $(t_1, t_2) = (-1, +1)$. As such, we consider three types of strategies for each player, namely, $s_1 = t_1\uparrow$ (increase of t_1), $s_2 = t_2\downarrow$ (decrease of t_2) and $s_3 = t_1\uparrow t_2\downarrow$ (increase of t_1 and decrease of t_2).

The utility or payoff functions measure the consequences of selecting different strategies by a player. The two players P_{TC} and P_{RR} representing the properties related to quality of rules and percentage of rules are affected based on different threshold values which in turn imply possible strategies. The utilities of two players are therefore determined as values of $TC(t_1, t_2)$ and $RR(t_1, t_2)$ as defined in Equations (7) and (8). From player TC's perspective, a lower negative value represents a higher gain or payoff. From player RR's perspective, a lower values represents a more desirable situation and a higher payoff. Both the players will choose the strategies in the aim to increase their respective payoffs. Please be noted that although we considered a two player game in GTRS in this study. The proposed game formulations may be extended to multiple criteria based games as suggested in [7].

For a certain strategy profile, say (s_m, s_n) which configures and leads to thresholds (t_1, t_2), the resulting utility of the players are given by,

$$u_{TC}(s_m, s_n) = TC(t_1, t_2), \tag{10}$$

$$u_{RR}(s_m, s_n) = RR(t_1, t_2), \tag{11}$$

where u_{RR} and u_{TC} are the payoff functions for the players P_{RR} and P_{TC}, respectively.

We consider the game in a payoff table in order to analyze and investigate how different strategies and payoff functions are related. Table 5 shows the payoff table for the considered game. The rows represent the strategies of player P_{RR} and the columns represent the strategies of player P_{TC}. Each cell or table entry corresponds to a particular strategy profile of the form (s_m, s_n) with corresponding payoffs given by $u_{RR}(s_m, s_n)$ and $u_{TC}(s_m, s_n)$ for players P_{RR} and P_{TC}, respectively.

The game solution such as Nash equilibrium can be used to determine a balance between the two player. For a two player game, a certain strategy profile such as (s_m, s_n) would be the Nash equilibrium if the following conditions based on Equation (9) hold,

$$\text{For player} P_{TC}: \ \forall s'_m \in S_1, u_{TC}(s_m, s_n) \geq u_{TC}(s'_m, s_n), \text{with } (s'_m \neq s_m), \tag{12}$$

$$\text{For player} P_{RR}: \ \forall s'_n \in S_2, u_{RR}(s_m, s_n) \leq u_{RR}(s_m, s'_n), \text{with } (s'_n \neq s_n). \tag{13}$$

This suggests that neither of the players are benefited by changing to a different strategy other than the one in profile (s_m, s_n). Please note that we use less than inequality in Equation (13). It is due to the fact that the player P_{RR} is aiming for a smaller value of its utility function as it represents a higher payoff.

An important issue in the above game formulation is the determination of threshold changes based on a particular strategy. In Table 5, we may note four types of changes in the two thresholds, i.e.,

$$t_1^{+} = \text{single player increase } t_1, \tag{14}$$

$$t_1^{++} = \text{both the players increase } t_1, \tag{15}$$

$$t_2^{-} = \text{single player decrease } t_2, \tag{16}$$

$$t_2^{--} = \text{both the players decrease } t_2 \tag{17}$$

The above definitions are used to associate a threshold pair with each strategy profile. For instance, thresholds based on a strategy profile, say, $(s_1, s_1) = (t_1\uparrow, t_1\uparrow)$ is computed as (t_1^{++}, t_2), since both the players suggest to increase threshold t_1 (Equation (14)). As another example, the strategy profile $(s_2, s_2) = (t_2\downarrow, t_2\downarrow)$ is computed as (t_1, t_2^{--}) using Equation (17). In the next section, we elaborate in detail on how to determine the values for the variables in Equations (14)–(17) based on a repetitive game mechanism.

Table 6
Payoff table for the game.

		P_{TC}		
		$s_1 = t_{1\downarrow}$	$s_2 = t_{2\uparrow}$	$s_3 = t_{1\downarrow}t_{2\uparrow}$
P_{RR}	$s_1 = t_{1\downarrow}$	$(-868, 6.1)$	$(-854, 4.5)$	$(-841, 7.6)$
	$s_2 = t_{2\uparrow}$	$(-854, 4.5)$	$(-859, 6.1)$	$(-873, 9.1)$
	$s_3 = t_{1\downarrow}t_{2\uparrow}$	$(\mathbf{-841, 7.6})$	$(-873, 9.1)$	$(-878, 12.1)$

5.2. An example of selecting rules with GTRS

We now explain the GTRS based approach for determining thresholds based on the example discussed in Section 3. Considering the game between the players P_{TC} and P_{RR} as discussed above in Section 5.1. Each player has three strategies based on an increase or decrease of 0.1 in the thresholds. Moreover, the game is being played with an initial threshold setting of $(t_1, t_2) = (-1, +1)$. A particular strategy, such as s_1, is interpreted as 0.1 decrease in t_2. The other strategies may be similarly interpreted. The thresholds corresponding to a particular strategy profile, say (s_1, s_2) is given by (t_1^+, t_2^-) which is equal to $(-0.9, 0.9)$. This means 0.1 increase in t_1 and 0.1 decrease in t_2. The utility or payoff functions are calculated based on Equations (10) and (11).

Table 6 represents the game in the payoff table. The values in the payoff table are based on the data for corpus 1 and are represented in Table 1. The cell with bold values represents the Nash equilibrium or game solution. The game solution in this case corresponds to strategy profile (s_3, s_1) which leads in thresholds of $(-0.9, 0.8)$ and the resulting payoffs of $(-841, 7.6\%)$. The game solution suggests that neither of the two players can achieve a higher payoff, given the other player chosen action. The resulting topic coherence in this case is better than both extreme cases of $TC(-1, 1) = -880$ (with minimum number of rules, i.e., 4.54%) and $TC(0, 0) = -897$ (with maximum number of rules, i.e., 100%) as discussed in Section 3. The computed thresholds $(t_1, t_2) = (-0.9, 0.8)$ have more rules (by 3.06%) but improve topic coherence (by 39 points) as compared to extreme point $(t_1, t_2) = (-1, 1)$. Moreover, it has lower relative rules (by 92.4%) and improved topic coherence (by 47 points) as compared to another extreme point $(t_1, t_2) = (0, 0)$.

From the above example, we may note that the GTRS based approach can overcome the two limitations discussed in Section 3. Firstly, the limitation of fixed thresholds is addressed by implementing games based on the data of each domain corpuses. Secondly, the game solution employed in the GTRS provides a useful justification for a the selection of suitable thresholds.

5.3. Algorithms for implementing the GTRS based approach

In this section, we provide algorithmic details for implementing and incorporating the proposed model in the automatic knowledge-based topic models. Algorithm 1 shows working of the main model that loads dataset consisting of N domain corpuses. The topics of all domains in the dataset are represented as T. In lines 5 to 9, the algorithm processes each individual domain corpus D_i. As pointed out in Section 2, the automatic knowledge-based topic models work in two steps. In step one, the candidate rules are generated and in step two an evaluation criteria and thresholds are used to select important rules. To execute step one, the algorithm in line 6 uses the basic LDA model (presented as Algorithm 2) to extract initial topics T_i for the current domain D_i. The topics extracted with baseline LDA are used as intuition to generate candidate rules in Algorithm 3 (which is called in line 7 of Algorithm 1). Highest probability words in same topic form positive candidate rules while highest probability words in different topics from negative candidate rules. It helps to limit the sample space from all possible combinations in the vocabulary. Step two is executed in line 8, the generated rules are passed to the GTRS based rule selection algorithm, i.e., Algorithm 4 which selects important rules (based on thresholds of three-way approach) and returns topics based on the selected rules. Finally, the domain specific topics are merged to obtain overall topics for the dataset. We now explain each of the sub algorithms used in the main algorithm.

Algorithm 1 Automatic knowledge based-topic model with three-way rule selection.

1: **Input:** Multiple corpuses D_1, D_2, D_3, ... D_N
2: **Output:** Collective topics T of all domains
3: T_i: Topics in domain corpus D_i
4: CR_i: Candidate rules for domain D_i
5: **for** domain corpus D_1 to D_N **do**
6: $T_i \leftarrow$ Initial topics without rules using LDA(D_i)//**Algorithm 2**
7: $CR_i \leftarrow$ Generate Candidate Rules (T_i, D_i) //**Algorithm 3**
8: $T_i \leftarrow$ GTRS based rule selection (CR_i, D_i) // **Algorithm 4**
9: **end for**
10: **return** $\bigcup_{i=1}^{N} T_i$

Algorithm 2 explains the topic extraction mechanism using LDA model [12] in lines 1 and 3–6. The LDA model is used when no *relevantRules* are found in line 2. In order to process a domain corpus, the model initially assigns topics to each

Algorithm 2 Topic extraction Algorithm(D_i) without rules adopted from [12,9] and with rules adopted from [16].

1: Assign words to topics at random
2: **if** $relevantRules = \emptyset$ **then**
3: **for** Each word in the document **do**
4: Randomly choose a topic t from the distribution over topics (i.e. per-topic word distribution and per-document topic distribution)
5: Randomly choose a word w from the corresponding distribution over vocabulary
6: **end for**
7: **else**
8: **for** Each word in the document **do**
9: Randomly choose a topic t from the distribution over topics (i.e. per-topic word distribution and per-document topic distribution)
10: Randomly choose a word w from the corresponding distribution over vocabulary
11: **for** RelevantRules having word w **do**
12: Positive $rule(w, w')$ has probability of w' increased for topic t
13: Negative $rule(w, w')$ has probability of w' decreased for topic t
14: **end for**
15: **end for**
16: **end if**
17: $T_i \leftarrow$ Topics in Domain D_i
18: **return** T_i

Algorithm 3 Generate Candidate Rules (T_i, D_i).

1: **for** each topic $T_{(i,j)} \in T_i$ **do**
2: make positive rules based on distinct word pairs $(w, w') \in T_{(i,j)}$
3: $CR_i = CR_i \cup w \longrightarrow w'$
4: **end for**
5: **for** word pairs (w, w') where $w \in T_{(i,j)}$ and $w' \in T_{(i,j')}$ **do**
6: make negative rules based on distinct word pairs (w, w')
7: $CR_i = CR_i \cup w \longrightarrow \neg w'$
8: **end for**
9: **return:** CR_i

word in each document at random in line 1. Each word has a different probability in each topic. In lines 3–6, collapsed Gibbs Sampling technique is used to extract hidden topic structures from the observed documents [10]. For each document in the collection, the words are generated in two steps. In step one, we randomly choose a distribution over topics such that each document exhibit all topics in different proportion. In step two, each word in each document is drawn from its current topic and is switched to a newly selected topic. The selected topic is chosen from the document-topic distribution [9]. The procedure stops when in successive iterations the words in particular topics are not assigned to other topics. The topic extraction with rules is also initialized in line 1. Lines 8–15 extract topics with the help of rules. However, each time a word w is switched to the newly selected topic; the relevant rules containing word w are employed in line 11. The probability of words paired with w in positive rules is increased for the new topic, while it is decreased for the words paired with w in negative rules in lines 12 and 13. Thus, the impact of rules is added by increasing or decreasing the probabilities of the accompanying words in rules by using Generalized Polya Urn (GPU) model [45]. This way the rules enforce positive rule words to co-occur under the same topic and vice versa.

Algorithm 3 presents the semantics of generating candidate rules CR_i for a domain D_i. As discussed in Section 3, the rules are based on word pairs and are primarily of two types, i.e., positive rules and negative rules. However, based on the evaluation of the rules and thresholds, we select a few positive and negative rules whose NPMI value is below t_1 or above t_2. To generate candidate positive rules, the algorithm considers rules based on distinct word pairs (w, w') in each topic $T_{(i,j)} \in T_i$. These rules are expected to have a positive value, (such as positive NPMI value). To generate candidate negative rules, the algorithm considers distinct words pairs (w, w') from different topics such that $w \in T_{(i,j)}$ and $w' \in T_{(i,j)}$. These rules are expected to have a negative value, (such as negative NPMI value). These candidate rules are passed to the GTRS based rule selection algorithm for selecting important rules.

Algorithm 4 represents the GTRS based rule selection algorithm. The algorithm accepts a particular domain D_i and the candidate rules CR_i generated for that domain based on Algorithm 3. The algorithm applies the game-theoretic analysis to determine thresholds which are used to select important rules. Finally, the rules are used to learn topics for the domain. The algorithm starts with initial values of thresholds $(t_1, t_2) = (-1, +1)$ and parameters $t_1^+, t_1^{++}, t_2^-, t_2^{--}$. With a one time non-repeated game, it may not be possible to achieve effective thresholds. A repeated game is therefore considered which iteratively modifies and improves the quality of the thresholds. In particular, the relationship between the modification in the thresholds and their impact on the utilities of the players are considered. We use this relationship to define the variables $t_1^+, t_1^{++}, t_2^-, t_2^{--}$ as follows.

$$t_1^+ = t_1 - k_1(t_1 \times RR(t_1', t_2') - RR(t_1, t_2)), \tag{18}$$

$$t_1^{++} = t_1 - k_2(t_1 \times RR(t_1', t_2') - RR(t_1, t_2)), \tag{19}$$

$$t_2^- = t_2 - k_1(t_2 \times RR(t_1', t_2') - RR(t_1, t_2)), \tag{20}$$

Algorithm 4 GTRS rules selection (CR_i, D_i).

1: Initial thresholds $(t_1, t_2) = (-1, +1)$
2: SR_i: Selected rules based on thresholds (t_1, t_2)
3: Initial values of $t_1^+, t_1^{++}, t_2^-, t_2^{--}$
4: **repeat**
5: Calculate the utilities of players based on Equation (10) and (11)
6: Populate the payoff table with calculated values
7: Calculate equilibrium in a payoff table using Equations (12) and (13)
8: Determine the selected strategies and the respective thresholds (t_1', t_2')
9: Calculate $t_1^+, t_1^{++}, t_2^-, t_2^{--}$ based on Equations (18)–(21)
10: $(t_1, t_2) = (t_1', t_2')$
11: **until** $TC(t_1', t_2') < TC(t_1, t_2)$ or $RR(t_1', t_2') > c$
12: **for** each rule containing word pair (w, w') in CR_i **do**
13: **if** NPMI$(w, w') \leq t_1$ or NPMI$(w, w') \geq t_2$ **then**
14: Select rule based on (w, w')
15: $SR_i = SR_i \cup$ rule from (w, w')
16: **else**
17: Ignore rules from (w, w')
18: **end if**
19: **end for**
20: $T_i \leftarrow$ Generate topics based on SR_i **Algorithm 2**
21: **return** T_i

$$t_2^{--} = t_2 - k_2(t_2 \times RR(t_1', t_2') - RR(t_1, t_2)), \tag{21}$$

where (t_1, t_2) are the initial thresholds at a certain iteration and the (t_1', t_2') are the updated thresholds based on game analysis. The variables k_1 and k_2 controls the change in thresholds where k_1 brings lower change as compared k_2 and $RR(t_1, t_2)$ and $RR(t_1', t_2')$ are the utilities of the player P_{RR} according Equation (11). The repeated game is shown in line 4–11 of the Algorithm 4. It stops when the subsequent iterations either do not improve topic coherence or reaches maximum number of available rules c.

An important issue with regards to Equations (18)–(21), is how to set parameter k_1 and k_2. We note from the game description that a strategy profile corresponds to a t_1 value form the set $\{t_1, t_1^+, t_1^{++}\}$ and a t_2 value from the set $\{t_2, t_2^-, t_2^{--}\}$, (where t_1, t_2 implies no change). From Equations (18)–(21), we know that t_1^- is the result of applying k_1 and t_1^{--} is the result of applying k_2. In the same way t_2^+ is the result of applying k_1 and t_2^{++} is the result of applying k_2. We determine an average k value that is based on the variables used in calculating the two thresholds. For instance, the profile (s_3, s_3) which corresponds to thresholds (t_1^{++}, t_2^{--}) are the result of applying k_2. In this case it will be $(k_2 + k_2)/2 = k_2$. The new values of k_1 and k_2 for the next iteration are now calculated based on the average value. Denoting an average value as k, the new value of k_1 is determined as a unit value lesser than k, i.e., $k_1 = k - 1$ and the new value of k_2 is determined as a unit value greater than k, i.e., $k_2 = k + 1$. For the thresholds which represent no change, we consider the variable with the lower value, i.e., k_1 as the corresponding variable, since it is closer to no change compared to k_2.

The determined thresholds with GTRS are next used by the Algorithm 4 to select important rules. This component is represented in lines 13 to 20 of the algorithm. All the word pairs in candidate rules are evaluated based on the NPMI measure. If the evaluation of the rule is below t_1 or above t_2 the rule is selected. In any other case, the rule is ignored and not selected. Finally, the rules are used to extract topics for the domain. Lastly, we may note that the complexity of the GTRS based algorithm, i.e., Algorithm 4, is $log(n)$. This may be verified by examining that there are two independent simple loops in this algorithm. Where n represents the sum of iterations of two loops, i.e. number of games played in lines 5–10, and rules evaluated in lines 13–18. It is interesting to note how the above approach take cares of polysemy, i.e., words with different contextual meaning. In this regards it may be noted that the suggested approach also saves the rule context by preserving its vocabulary. The context is matched to ensure that the rule is transferred in the same context in which it was extracted.

6. Experimental results and discussion

We used the most recently reported Chen2014 dataset in our experiments. It should be noted that there are many datasets used for topic modeling. However, majority of these do not fit into the context of automatic knowledge based topic modeling. In particular, they do not contain data from multiple domain corpuses. The Chen2014 dataset has 100 domains corpuses with 50 each from electronic and non-electronic commercial products. Each domain on average contains 5,000 documents that reflect real user reviews from Amazon.com. We employ the GTRS based approach to learn thresholds for each domain. In all experiments, the results are evaluated using ten fold cross validation. To simplify computations, we apply the restrictions suggested in [16–19], i.e., 15 topics are extracted for each domain where each topic is represented with only 30 words in that topic. The thresholds (t_1, t_2) are initialized to $(-1, +1)$ and are repeatedly modified until one of the stop conditions in Algorithm 4 is reached.

Tables 7 and 8 show the detailed results on the 50 electronic product domains. Each row of these tables represents the results for one domain corpus. The second column shows the topic coherence of the topics extracted with the LDA without

Table 7
Results on electronic product domains from Chen2014 dataset (Part 1).

Domain	Topic coherence			Relative rules		Thresholds
	TC	$TC(-1,1)$	$TC(t_1,t_2)$	$RR(-1,1)$	$RR(t_1,t_2)$	(t_1,t_2)
Alarm clock	−820	−818	−787	0.14	24.72	(−0.66, 0.81)
Amplifier	−866	−841	−803	0.13	23.88	(−0.68, 0.80)
Battery	−813	−801	−776	0.6	42.04	(−0.28, 0.32)
Blu-ray	−990	−982	−888	0.12	11.10	(−0.89, 0.77)
Cable	−908	−905	−850	0.37	10.94	(−0.88, 0.42)
Camcorder	−928	−918	−835	0.17	3.23	(−0.82, 0.75)
Camera	−955	−953	−859	0.24	12.52	(−0.84, 0.59)
Car stereo	−892	−887	−855	0.06	15.3	(−0.89, 0.61)
CD player	−798	−785	−721	0.11	30.95	(−0.68, 0.62)
Cell phone	−836	−822	−792	0.11	29.23	(−0.68, 1.0)
Computer	−885	−880	−819	0.04	23.72	(−0.67, 0.80)
DVD player	−848	−836	−776	0.12	23.28	(−0.89, 0.81)
Fan	−821	−819	−787	0.23	29.36	(−0.62, 0.43)
GPS	−896	−893	−824	0.13	15.3	(−0.81, 0.75)
Graphics	−894	−836	−774	0.12	19.56	(−0.79, 0.72)
Hard drive	−840	−932	−854	0.15	11.61	(−0.68, 0.81)
Head phone	−908	−870	−817	0.13	11.86	(−0.90, 0.57)
Home	−913	−882	−828	0.08	18.08	(−0.74, 0.47)
Iron	−878	−854	−836	0.12	14.4	(−0.82, 0.95)
Keyboard	−859	−856	−798	0.04	24.27	(−0.68, 0.81)
Kindle	−820	−815	−786	0.26	32.96	(−0.47, 0.56)
Lamp	−788	−778	−763	0.16	27.25	(−0.63, 0.53)
Laptop	−784	−771	−738	0.10	27.19	(−0.89, 0.81)
Media player	−1025	−977	−918	0.11	10.16	(−0.68, 0.81)
Memory	−825	−822	−790	0.20	23.51	(−0.65, 0.78)

Table 8
Results on electronic product domains from Chen2014 dataset (Part 2).

Domain	Topic coherence			Relative rules		Thresholds
	TC	$TC(-1,1)$	$TC(t_1,t_2)$	$RR(-1,1)$	$RR(t_1,t_2)$	(t_1,t_2)
Microphone	−845	−824	−803	0.14	17.35	(−0.89, 0.81)
Microwave	−846	−837	−793	0.19	15.44	(−0.79, 0.76)
Monitor	−897	−870	−826	0.33	12.46	(−0.88, 0.55)
Mouse	−862	−857	−827	0.22	23.11	(−0.67, 0.71)
MP3 player	−922	−900	−847	0.19	17.11	(−0.79, 0.52)
Network	−877	−862	−838	0.07	17.11	(−0.6, 0.8)
Printer	−950	−941	−878	0.23	29.95	(−0.82, 0.57)
Projector	−892	−888	−841	0.11	13.37	(−0.77, 0.66)
Radar	−924	−916	−841	0.23	9.44	(−0.89, 0.80)
Remote	−928	−924	−847	0.06	14.63	(−0.88, 0.67)
Rice cooker	−838	−837	−772	0.26	19.49	(−0.74, 0.57)
Scanner	−917	−906	−850	0.12	11.89	(−0.86, 0.76)
Speaker	−872	−865	−815	0.06	24.79	(−0.78, 0.62)
Sub woofer	−878	−856	−809	0.20	23.96	(−0.68, 0.81)
Tablet	−919	−912	−858	0.13	6.47	(−0.95, 0.62)
Telephone	−901	−899	−830	0.15	16.74	(−0.79, 0.64)
TV	−863	−850	−791	0.07	15.48	(−0.89, 0.80)
Vacuum	−963	−953	−879	0.13	8.6	(−0.89, 0.80)
Video player	−951	−908	−847	0.08	16.48	(−0.82, 0.75)
Video	−933	−880	−840	0.19	10.46	(−0.89, 0.78)
Voice	−888	−883	−824	0.04	18.41	(−0.89, 0.41)
Watch	−828	−826	−767	0.10	21.75	(−0.69, 0.81)
Web cam	−870	−867	−811	0.04	18.96	(−0.89, 0.43)
Wireless	−964	−937	−889	0.21	8.19	(−0.93, 0.82)
Xbox	−867	−857	−813	0.22	15.58	(−0.89, 0.81)

any rules. This is included as a baseline model for the sake of comparison. The third and fourth columns represent the values of the topic coherence based on the initial thresholds $(-1,+1)$ and the obtained thresholds (t_1,t_2). The fifth and sixth column represent the values of the relative rules based on the initial thresholds $(-1,+1)$ and the obtained thresholds (t_1,t_2). The last column shows the values of the determined thresholds.

We may note from Tables 7 and 8 that the GTRS determines different thresholds for each domain. The determined thresholds increases the relative rules and topic coherence compared to initial thresholds. It is noted that the topic coherence is improved for every domain compared to the baseline. In comparison to the initial thresholds, we may note that the

Table 9
Results on non-electronic product domains from Chen2014 dataset (Part 1).

Domain	Topic coherence			Relative rules		Thresholds
	TC	$TC(-1,1)$	$TC(t_1,t_2)$	$RR(-1,1)$	$RR(t_1,t_2)$	(t_1,t_2)
Android App	−812	−805	−781	0.18	9.94	(−0.87, 0.62)
Appliances	−858	−856	−811	0.14	12.2	(−0.86, 0.75)
Arts crafts	−800	−795	−764	0.4	14.09	(−0.88, 0.65)
Automotive	−968	−962	−781	0.14	12.1	(−0.95, 0.66)
Baby	−903	−893	−838	0.11	12.4	(−0.88, 0.73)
Bag	−918	−883	−825	0.01	7.2	(−0.95, 0.62)
Beauty	−938	−903	−875	0.36	6.4	(−0.94, 0.91)
Bike	−887	−876	−831	0.14	15.2	(−0.67, 0.82)
Books	−807	−803	−772	0.2	8.3	(−0.91, 0.64)
Cable	−824	−791	−774	0.34	13.8	(−0.89, 0.81)
Care	−886	−869	−820	0.14	11.9	(−0.83, 0.78)
Clothing	−849	−836	−786	0.13	11.7	(−0.88, 0.74)
Conditioner	−813	−810	−778	0.33	13.5	(−0.86, 0.62)
Diaper	−882	−857	−817	0.32	8.3	(−0.93, 0.65)
Dining	−882	−875	−825	0.27	10.5	(−0.87, 0.60)
Dumbbell	−909	−887	−833	0.09	7.7	(−0.92, 0.64)
Flashlight	−914	−903	−827	0.14	9.7	(−0.68, 0.81)
Food	−910	−880	−824	0.27	10.4	(−0.61, 1.0)
Gloves	−875	−862	−815	0.16	11.01	(−0.87, 0.59)
Golf	−846	−841	−805	0.14	8.5	(−0.91, 0.63)
Home	−828	−763	−687	0.24	25.7	(−0.69, 0.42)
Industrial	−790	−778	−758	0.31	13.3	(−0.85, 0.59)
Jewelry	−781	−747	−713	0.18	9.3	(−0.9, 0.81)
Kindle	−1026	−957	−901	0.26	7.6	(−0.68, 0.62)
Kitchen	−825	−804	−781	0.09	17.1	(−0.79, 0.66)

domain corpus “Battery” holds the highest increase in relative rules of 41.44% and the domain corpus “Camcorder” has the lowest increase in relative rules of 3.06%. In the same way, the domain corpus “BluRay” has the highest increase in topic coherence of 94 points at the cost of 10.98% increase in relative rules and the domain corpus “Lamp” has the lowest rise of 15 points in topic coherence at the cost of 27.09% increase in relative rules. In general, corpuses having weak association between words under the same topic brings little improvement in topic coherence by increasing the relative rules. Table 11 shows a sample of some prominent rules that contributed in improving topic coherence of their respective corpuses.

Tables 9 and 10 are similar to Tables 7 and 8 and summarize the results for the non-electronic domains. Again the determined thresholds produce better topic coherence in comparison to baseline and initial thresholds. The domain corpus “Home” has the highest increase in relative rules of 25.7% and domain corpus “Water” has the lowest increase of 0.5%. Similarly, the domain corpus “Tent” has the highest increase of 102 points in topic coherence and “Pet Supp.” has the lowest increase of 8 points with respect to initial thresholds. It is achieved at the cost of 9.1% and 15.2% increase in relative rules for domain corpuses “Tent” and “Pet Supp.”, respectively.

We include Figs. 2 and 3 to visualize the results in Tables 7 and 8. Fig. 2 shows the increase in the relative rules for individual domains. Each bar in the figure corresponds to a rise in the relative rules for a particular domain. The domains on the x-axis, written as D1 to D50 represents the domains in Tables 7 and 8 are in the same alphabetical order as given in that table. For instance, D1 represents the domain “Alarm Clock” and D2 represents the domain “Amplifier”. All the increase in the relative rules are with respect to the initial thresholds of $(-1, 1)$. The increase in the relative rules are different for different domains. All the increases are less than 45%. An average increase in relative rules when the thresholds are being modified with GTRS is 17.93%. Fig. 3 shows the comparison of the topic coherence with the determined thresholds and the initial thresholds. Again on the x-axis we have 50 domains and on the y-axis we have topic coherence at initial and modified thresholds. It highlights the improvement in topic coherence for individual domains that averages to 52.82 at modified thresholds at a cost of 17.93% increase in the relative rules.

Figs. 4 and 5 are similar to Figs. 2 and 3 and highlights the results of non-electronic domains given in Table 9 and 10. It may be noted from Fig. 4 that all the increases in the relative rules are less than 20%. The average increase is noted as 10.19% for all the domains. From Fig. 5, it is noted that the topic coherence is increased for all domain corpuses with determined thresholds having an average increase of 43.82 points.

Finally, we look at the mean deviation in the thresholds for different domains. It may be computed as,

$$MD = \frac{\sum_{i=1}^{50} |(t_i - \bar{t})|}{50} \times 100, \tag{22}$$

where t_i is one of the two thresholds (t_1, t_2) of a specific domain and $\bar{t}$ is the mean of the thresholds across the 50 domains. A higher value of the MD will mean high deviation in the threshold across different domains. Due to high deviation, we should therefore discourage the use of global thresholds to be used across all the domains. In other words, a higher value for the MD advocates for the use of domain specific thresholds. The determined values of the MD for the thresholds t_1 and t_2

Table 10
Results on non-electronic product domains from Chen2014 dataset (Part 2).

Domain	Topic coherence			Relative rules		Thresholds
	TC	$TC(-1,1)$	$TC(t_1,t_2)$	$RR(-1,1)$	$RR(t_1,t_2)$	(t_1,t_2)
Knife	−857	−841	−799	0.1	13.3	$(-0.87, 0.60)$
Luggage	−838	−818	−791	0.27	11.8	$(-0.86, 0.61)$
Magazines	−891	−848	−824	0.19	4.8	$(-0.97, 0.74)$
Mat	−863	−832	−788	0.16	16.9	$(-0.88, 0.60)$
Mattress	−920	−888	−832	0.26	9.1	$(-0.91, 0.62)$
Movies TV	−884	−881	−856	0.19	5.1	$(-0.89, 0.96)$
Music	−947	−885	−859	0.31	6.0	$(-0.68, 0.62)$
Musical	−887	−887	−840	0.17	6.9	$(-0.83, 0.64)$
Office	−915	−893	−824	0.14	11.1	$(-0.86, 0.61)$
Patio lawn	−914	−909	−844	0.07	10.8	$(-0.87, 0.79)$
Pet supp.	−833	−831	−823	0.42	15.2	$(-0.81, 0.54)$
Pillows	−906	−866	−832	0.10	7.6	$(-0.92, 0.66)$
Sandal	−871	−815	−779	0.26	14.5	$(-0.68, 0.81)$
Scooter	−868	−841	−802	0.17	13.0	$(-0.85, 0.56)$
Shoe	−911	−867	−847	0.22	5.71	$(-0.96, 0.66)$
Software	−923	−899	−838	0.13	7.0	$(-0.94, 0.68)$
Sports	−871	−848	−814	0.11	11.1	$(-0.89, 0.58)$
Table	−952	−944	−879	0.17	8.3	$(-0.93, 0.64)$
Tent	−945	−928	−826	0.09	9.1	$(-0.91, 0.63)$
Tire	−945	−861	−821	0.15	6.0	$(-0.95, 0.66)$
Toys	−877	−873	−823	0.27	14.4	$(-0.68, 0.62)$
Video	−829	−817	−798	0.16	8.8	$(-0.89, 0.64)$
Vitamin	−865	−845	−807	0.17	10.3	$(-0.87, 0.61)$
Wall	−957	−929	−871	0.24	5.9	$(-0.95, 0.66)$
Water	−670	−533	−502	0.48	0.5	$(-1, 0.73)$

Table 11
Corpus domains and their sample word pairs forming rules.

Domain	Type	Rules
Laptop	positive	(longer, charge), (disk, file), (program, software)
	negative	(customer, drive), (easy, port), (port, software)
Fan	positive	(quiet, noise), (panel, control), (ac, adaptor)
	negative	(hour, small), (compact, control), (purchase, size)
Vacuum	positive	(floor, room), (compact, space), (long, cord)
	negative	(company, wall), (model, wall), (hand, purchase)

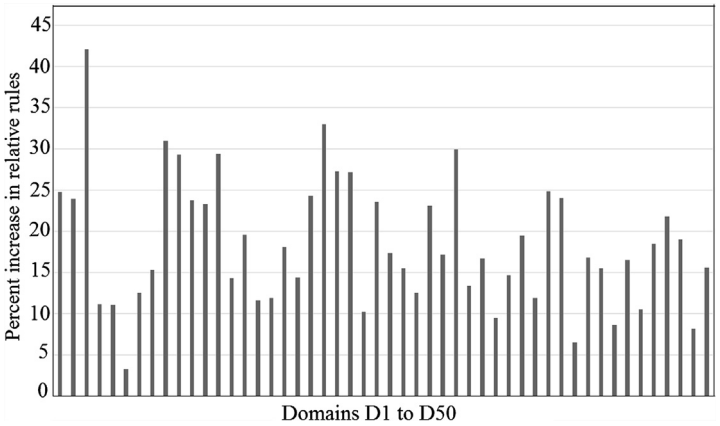

Fig. 2. Increase in relative rules for electronic domains.

in case of electronic domains are 18% and 13%, respectively. These values represent a noticeable deviation in the thresholds. Therefore, domain specific thresholds are favored against universal thresholds. In conclusion, it provides additional support for the use of automated thresholds for each domain.

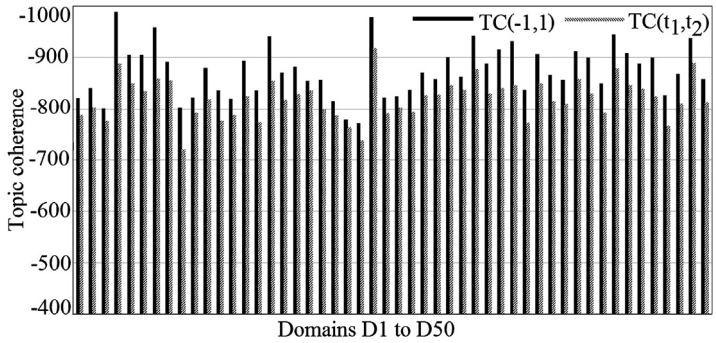

Fig. 3. Increase in topic Coherence for electronic domains.

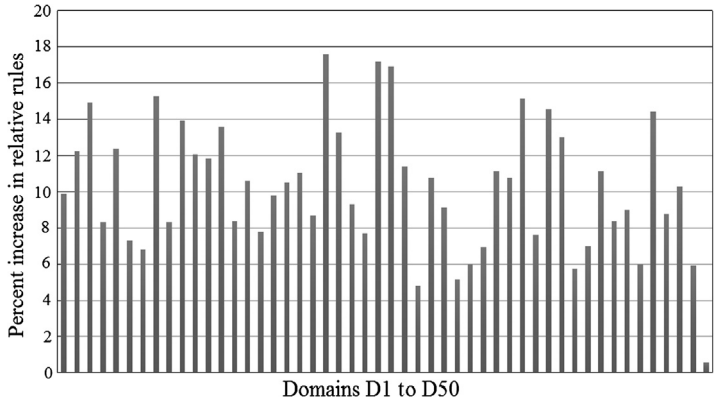

Fig. 4. Increase in relative rules for non-electronic domains.

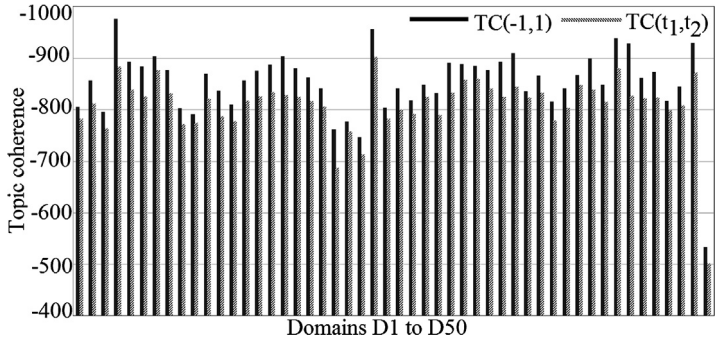

Fig. 5. Increase in topic Coherence for non-electronic domains.

7. Conclusion

Automatic knowledge-based topic models are recently introduced to meet with the demands of processing large scale text collections. They are based on automatic selection of rules for extracting topics. The rules vary in their respective levels of significance or importance and therefore only a few useful rules are selected by employing evaluation criteria and thresholds. There are two limitations in existing models for selecting rules. Firstly, they are based on fixed thresholds for extracting rules from all domain corpuses thereby ignoring domain specific characteristics. Secondly, the thresholds are based on expert opinions and not on some automated mechanisms. We address these limitations by considering a three-way approach for selecting rules where a pair of thresholds (t_1, t_2) partitions the rules into three groups or regions, namely, rules with strong positive association, strong negative association and weak association. To obtain domain specific and automated thresholds, we utilize the GTRS model to implement a game that considered a tradeoff between the quality and quantity of the rules. Algorithms for incorporating the proposed GTRS based approach into automatic knowledge-based topic modeling are introduced and discussed. Experimental results on Chen2014 dataset suggest an average improvement of 52.82 points in topic coherence by increasing the quantity of rules to 17.93%.

In future the three-way approach for selecting rules may be extended to online automatic knowledge-based topic models that has sequential flow of domain corpuses.

References

[1] R. Agrawal, R. Srikant, Fast algorithms for mining association rules, in: Proceedings of the 20th International Conference on Very Large Data Bases, VLDB'94, vol. 1215, 1994, pp. 487–499.
[2] D. Andrzejewski, X. Zhu, M. Craven, Incorporating domain knowledge into topic modeling via Dirichlet forest priors, in: Proceedings of the 26th Annual International Conference on Machine Learning, ICML'96, 2009, pp. 25–32.
[3] D. Andrzejewski, X. Zhu, M. Craven, B. Recht, A framework for incorporating general domain knowledge into latent Dirichlet allocation using first-order logic, in: Proceedings of 22nd International Joint Conference on Artificial Intelligence, vol. 22, 2011, pp. 1171–1177.
[4] N. Azam, J.T. Yao, Formulating game strategies in game-theoretic rough sets, in: Proceedings of 8th International Conference on Rough Sets and Knowledge Technology, RSKT'13, in: Lect. Notes Comput. Sci., vol. 8171, 2013, pp. 145–153.
[5] N. Azam, J.T. Yao, Analyzing uncertainties of probabilistic rough set regions with game-theoretic rough sets, Int. J. Approx. Reason. 55 (1) (2014) 142–155.
[6] N. Azam, J.T. Yao, Game-theoretic rough sets for recommender systems, Knowl.-Based Syst. 72 (2014) 96–107.
[7] N. Azam, J.T. Yao, Interpretation of equilibria in game-theoretic rough sets, Inf. Sci. 295 (2015) 586–599.
[8] Y. Baram, Partial classification: the benefit of deferred decision, IEEE Trans. Pattern Anal. Mach. Intell. 20 (8) (1998) 769–776.
[9] D.M. Blei, Probabilistic topic models, Commun. ACM 55 (4) (2012) 77–84.
[10] D.M. Blei, M.I. Jordan, et al., Variational inference for Dirichlet process mixtures, Bayesian Anal. 1 (1) (2006) 121–144.
[11] D.M. Blei, J.D. Lafferty, Topic models, in: A. Srivastava, M. Sahami (Eds.), Text Mining: Classification, Clustering, and Applications, vol. 10 (71), 2009.
[12] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent Dirichlet allocation, J. Mach. Learn. Res. 3 (2003) 993–1022.
[13] G. Bouma, Normalized (pointwise) mutual information in collocation extraction, in: Proceedings of the International Conference of the German Society for Computational Linguistics and Language Technology, 2009, pp. 31–40.
[14] S. Brody, N. Elhadad, An unsupervised aspect-sentiment model for online reviews, in: Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics, 2010, pp. 804–812.
[15] F. Cabitza, D. Ciucci, A. Locoro, Exploiting collective knowledge with three-way decision theory. Cases from the questionnaire-based research, Int. J. Approx. Reason. (2017), http://dx.doi.org/10.1016/j.ijar.2016.11.013.
[16] Z. Chen, B. Liu, Mining topics in documents: standing on the shoulders of big data, in: Proceedings of the 20th International Conference on Knowledge Discovery and Data Mining, 2014, pp. 1116–1125.
[17] Z. Chen, B. Liu, Topic modeling using topics from many domains, lifelong learning and big data, in: Proceedings of the 31st International Conference on Machine Learning, ICML'14, 2014, pp. 703–711.
[18] Z. Chen, A. Mukherjee, B. Liu, Aspect extraction with automated prior knowledge learning, in: 52nd Annual Meeting of the Association for Computational Linguistics, 2014, pp. 347–358.
[19] Z. Chen, A. Mukherjee, B. Liu, M. Hsu, M. Castellanos, R. Ghosh, Discovering coherent topics using general knowledge, in: Proceedings of the 22nd International Conference on Information & Knowledge Management, 2013, pp. 209–218.
[20] Z.B. Chen, Chen2014 dataset, https://www.cs.uic.edu/zchen/downloads/KDD2014-Chen-Dataset.zip, 2014.
[21] K.W. Church, P. Hanks, Word association norms, mutual information, and lexicography, Comput. Linguist. 16 (1) (1990) 22–29.
[22] E. Eaton, P.L. Ruvolo, Ella: an efficient lifelong learning algorithm, in: Proceedings of the 30th International Conference on Machine Learning, ICML'13, 2013, pp. 507–515.
[23] J. Han, H. Cheng, D. Xin, X. Yan, Frequent pattern mining: current status and future directions, Data Min. Knowl. Discov. 15 (1) (2007) 55–86.
[24] J.P. Herbert, J.T. Yao, Game-theoretic rough sets, Fundam. Inform. 108 (3–4) (2011) 267–286.
[25] B.Q. Hu, Three-way decisions space and three-way decisions, Inf. Sci. 281 (2014) 21–52.
[26] Y. Hu, J. Boyd-Graber, B. Satinoff, A. Smith, Interactive topic modeling, Mach. Learn. 95 (3) (2014) 423–469.
[27] J. Huang, M. Peng, H. Wang, Topic detection from large scale of microblog stream with high utility pattern clustering, in: Proceedings of the 8th Workshop on Ph. D. Workshop in Information and Knowledge Management, ACM, 2015, pp. 3–10.
[28] C. Jacobi, W. van Atteveldt, K. Welbers, Quantitative analysis of large amounts of journalistic texts using topic modelling, Digit. Journal. 4 (1) (2016) 89–106.
[29] J. Jagarlamudi, H. Daumé III, R. Udupa, Incorporating lexical priors into topic models, in: Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, 2012, pp. 204–213.
[30] Y. Jo, A.H. Oh, Aspect and sentiment unification model for online review analysis, in: Proceedings of the 4th International Conference on Web Search and Data Mining, ACM, 2011, pp. 815–824.
[31] J.-H. Kang, J. Ma, Y. Liu, Transfer topic modeling with ease and scalability, in: Proceedings of the International Conference on Data Mining, 2012, pp. 564–575.
[32] M.T. Khan, M. Durrani, S. Khalid, F. Aziz, Online knowledge-based model for big data topic extraction, Comput. Intell. Neurosci. 2016 (2016) 1–9.
[33] S.C. Kleene, Introduction to Metamathematics, Groningen, New York, 1952.
[34] K. Leyton-Brown, Y. Shoham, Essentials of Game Theory: A Concise Multidisciplinary Introduction, Morgan & Claypool Publishers, 2008.
[35] W. Li, Z. Huang, Q. Li, Three-way decisions based software defect prediction, Knowl.-Based Syst. 91 (2016) 263–274.
[36] D. Liang, D. Liu, Deriving three-way decisions from intuitionistic fuzzy decision-theoretic rough sets, Inf. Sci. 300 (2015) 28–48.
[37] D. Liang, W. Pedrycz, D. Liu, P. Hu, Three-way decisions based on decision-theoretic rough sets under linguistic assessment with the aid of group decision making, Appl. Soft Comput. 29 (2015) 256–269.
[38] D.C. Liang, D. Liu, A. Kobina, Three-way group decisions with decision-theoretic rough sets, Inf. Sci. 345 (2016) 46–64.
[39] K.W. Lim, W. Buntine, C. Chen, L. Du, Nonparametric Bayesian topic modelling with the hierarchical Pitman–Yor processes, Int. J. Approx. Reason. 78 (2016) 172–191.
[40] C. Lin, Y. He, R. Everson, S. Rüger, Weakly supervised joint sentiment-topic detection from text, IEEE Trans. Knowl. Data Eng. 24 (6) (2012) 1134–1145.
[41] B. Liu, W. Hsu, Y. Ma, Mining association rules with multiple minimum supports, in: Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1999, pp. 337–341.
[42] D. Liu, D. Liang, C. Wang, A novel three-way decision model based on incomplete information system, Knowl.-Based Syst. 91 (2016) 32–45.
[43] Y. Lu, C. Zhai, Opinion integration through semi-supervised topic modeling, in: Proceedings of the 17th International Conference on World Wide Web, 2008, pp. 121–130.
[44] Y. Lu, C. Zhai, N. Sundaresan, Rated aspect summarization of short comments, in: Proceedings of the 18th International Conference on World Wide Web, 2009, pp. 131–140.
[45] H. Mahmoud, Pólya Urn Models, CRC Press, 2008.

[46] J.D. Mcauliffe, D.M. Blei, Supervised topic models, in: Advances in Neural Information Processing Systems, 2008, pp. 121–128.
[47] Q. Mei, X. Ling, M. Wondra, H. Su, C. Zhai, Topic sentiment mixture: modeling facets and opinions in weblogs, in: Proceedings of the 16th International Conference on World Wide Web, ACM, 2007, pp. 171–180.
[48] D. Mimno, H.M. Wallach, E. Talley, M. Leenders, A. McCallum, Optimizing semantic coherence in topic models, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2011, pp. 262–272.
[49] A. Mukherjee, B. Liu, Aspect extraction through semi-supervised modeling, in: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers, vol. 1, 2012, pp. 339–348.
[50] B. Pang, L. Lee, Opinion mining and sentiment analysis, Found. Trends Inf. Retr. 2 (1–2) (2008) 1–135.
[51] S.G. Pauker, J.P. Kassirer, The threshold approach to clinical decision making, N. Engl. J. Med. 302 (20) (1980) 1109–1117.
[52] P. Pecina, A machine learning approach to multiword expression extraction, in: Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions, MWE'08, 2008, pp. 54–61.
[53] W. Pedrycz, Shadowed sets: representing and processing fuzzy sets, IEEE Trans. Syst. Man Cybern., Part B, Cybern. 28 (1) (1998) 103–109.
[54] D. Ramage, D. Hall, R. Nallapati, C.D. Manning, Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora, in: Proceedings of the Empirical Methods in Natural Language Processing, vol. 1, 2009, pp. 248–256.
[55] A. Rapoport, G.J. Burkheimer, Models for deferred decision making, J. Math. Psychol. 8 (4) (1971) 508–538.
[56] I.A. Sag, T. Baldwin, F. Bond, A. Copestake, D. Flickinger, Multiword expressions: a pain in the neck for NLP, in: International Conference on Intelligent Text Processing and Computational Linguistics, Springer, 2002, pp. 1–15.
[57] C. Sauper, A. Haghighi, R. Barzilay, Content models with attitude, in: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, vol. 1, Association for Computational Linguistics, 2011, pp. 350–358.
[58] K. Schouten, F. Frasincar, Survey on aspect-level sentiment analysis, IEEE Trans. Knowl. Data Eng. 28 (3) (2016) 813–830.
[59] L. Shen, L. Wu, Z. Li, Topic modelling for object-based classification of VHR satellite images based on multiscale segmentations, Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci. (2016) 359–363.
[60] Y. Tsvetkov, S. Wintner, Extraction of multi-word expressions from small parallel corpora, Nat. Lang. Eng. 18 (04) (2012) 549–573.
[61] A. Tversky, E. Shafir, Choice under conflict: the dynamics of deferred decision, Psychol. Sci. 3 (6) (1992) 358–361.
[62] A. Wald, Sequential tests of statistical hypotheses, Ann. Math. Stat. 16 (2) (1945) 117–186.
[63] H. Wang, Y. Lu, C. Zhai, Latent aspect rating analysis on review text data: a rating regression approach, in: Proceedings of the 16th International Conference on Knowledge Discovery and Data Mining, 2010, pp. 783–792.
[64] T. Wang, Y. Cai, H.-f. Leung, R.Y. Lau, Q. Li, H. Min, Product aspect extraction supervised with online domain knowledge, Knowl.-Based Syst. 71 (2014) 86–100.
[65] G.-R. Xue, W. Dai, Q. Yang, Y. Yu, Topic-bridged PLSA for cross-domain text classification, in: Proceedings of the 31st International Conference on Research and Development in Information Retrieval, 2008, pp. 627–634.
[66] X.P. Yang, J.T. Yao, Modelling multi-agent three-way decisions with decision-theoretic rough sets, Fundam. Inform. 115 (2–3) (2012) 157–171.
[67] J.T. Yao, J.P. Herbert, A game-theoretic perspective on rough set analysis, J. Chongqing Univ. Posts Telecommun. (Nat. Sci. Ed.) 20 (3) (2008) 291–298.
[68] Y.Y. Yao, An outline of a theory of three-way decisions, in: Proceedings of Rough Sets and Current Trends in Computing, RSCTC'12, in: Lect. Notes Comput. Sci., vol. 7413, 2012, pp. 1–17.
[69] Y.Y. Yao, Rough sets and three-way decisions, in: Proceedings of 10th International Conference on Rough Sets and Knowledge Technology, RSKT'15, in: Lect. Notes Comput. Sci., vol. 9436, 2015, pp. 62–73.
[70] Y.Y. Yao, H. Yu, An introduction to three-way decisions, in: H. Yu, G.Y. Wang, T.R. Li, J.Y. Liang, D.Q. Miao, Y.Y. Yao (Eds.), Three-Way Decisions: Methods and Practices for Complex Problem Solving, Science Press, Beijing, 2015, pp. 1–19 (in Chinese).
[71] Y. Zhang, J.T. Yao, Gini objective functions for three-way classifications, Int. J. Approx. Reason. 81 (2017) 103–114.
[72] W.X. Zhao, J. Jiang, H. Yan, X. Li, Jointly modeling aspects and opinions with a MaxEnt-LDA hybrid, in: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2010, pp. 56–65.
[73] X. Zheng, Z. Lin, X. Wang, K.-J. Lin, M. Song, Incorporating appraisal expression patterns into topic modeling for aspect and sentiment word identification, Knowl.-Based Syst. 61 (2014) 29–47.