Semantic Network Based Common Sense Measure for Association Rule Pruning

Ingi Lee[1] and Hwan-Seung Yong[1]

[1] *Computer Science Engineering, Ewha Womans University,*
lig@ewhain.net, hsyong@ewha.ac.kr

Abstract

Association rule mining is now widely used in many fields such as commerce, telecom, insurance, and bioinformatics. Although association rule mining has improved in performance, the real commerce database has also grown in its size and dimension to a point of creating millions of association rules. One of the biggest problems of association rule mining is that it frequently produces large numbers of rules, and this makes it difficult for users to select those that are of interest. We proposed the Common Sense Measure (CSM) so that only interesting knowledge can be selected in order to resolve the problem resulting from a large quantity of rules. The CSM is an interestingness measure that evaluates how closely rules match the common sense knowledge. We developed an algorithm of rule matching method with common sense knowledge using the common sense network (CSN).

Keywords: *Data Mining, Interestingness Measures, Common Sense Knowledge, Similarity, Semantic Network, Knowledge Representation*

1. Introduction

Association rule mining is now widely used in many fields such as commerce, telecom, insurance, and bioinformatics. However, with improved performance of association rule mining, the real commerce database has grown greatly in its size and dimension to a point of creating millions of association rules [1]. One of the biggest problems of association rule mining is that it frequently produces large numbers of rules, and this makes it difficult for users to select those that are of interest. Moreover, many of the discovered rules will be obvious, already known, or not relevant. Therefore, recent research has concentrated on various forms of interestingness measures. The users end up spending a lot of time and effort in finding interesting knowledge in rules that are useless [2, 3]. For this reason, we have developed an intelligent post processing technique in data mining that generates and evaluates association rules by interestingness measures [4].

The goal of measures for evaluating interestingness of knowledge is to show results that are most similar to knowledge that the user decides to be interesting. It always depends on the user's prior knowledge about the domain. Computers do not currently have the basic knowledge about the world that we consider "common sense". Open Mind Common Sense (OMCS) is an artificial intelligence project based at the Massachusetts Institute of Technology (MIT) Media Lab whose goal is to build and utilize a large common sense knowledge base from the contributions of many thousands of people across the web [5]. ConceptNet is a semantic network based on the information in the OMCS database and is expressed as a hypergraph whose nodes are concepts and whose edges are assertions of common sense about these concepts [6, 7]. Recently, ConceptNet 5 has been announced. Its new data sources include Wikipedia, Wiktionary, WordNet, and DBPedia. They include extensively the knowledge needed to the people [8]. ConceptNet 5 can be applied to many

different applications and has been actively studied. At the intersection of these two research fields, we propose interestingness measures by using the semantic approach of common sense network for finding interesting rules.

The remainder of this paper is organized as follows. In Section 2, we introduce related works, and in Section 3, we propose and define the Common Sense Measure (CSM). In Section 4, we implement the CSM and the core algorithm. We describe the experiment and results in Section 5 and conclude in Section 6.

2. Related Works

The post-processing process in data mining provides users with interesting knowledge by reducing a large number of rules and patterns produced after the data mining process. It touches on various practical issues, such as the pruning, interestingness, and redundancy-removal [9]. The rules filtered by applying the redundancy-removal technique and various pruning techniques still contain knowledge that is not interesting for the users [1].

Interestingness measures are divided into two categories by the objective and subjective aspects of interestingness. Objective measures are based on the statistical strength or properties of the rules. Subjective measures take into account both the data and the user's background knowledge about the data [1, 2]. Objective measures include lift, conviction, and leverage based on the theory of probability and various measures based on communication and information theory [3]. In recent years, numerous objective measures have been categorized and defined, and some measures have been applied to other fields [10-12]. However, objective measures cost a lot but do not provide reliable criteria. Therefore, more recent works related to subjective measures have been undertaken along with studies in the field of artificial intelligence and belief system [3].

In related studies, there are the method of reflecting knowledge learned from the domain expert in order to acquire knowledge from the belief system and the method of reflecting techniques such as inductive learning and machine learning [2, 3]. Other approaches include Sahar's method that eliminates useless rules to find patterns of "interestingness" [13]. This method proved to be useful in that it was effective in reducing time spent on reflecting and processing all the knowledge of the domain; however, it not only was hard to apply to other domains but also needed a knowledge acquiring process even for a small amount of knowledge. The biggest problem in subjective techniques has been how to reduce the bottleneck of the knowledge acquiring process. Hybrid type approaches combine objective criteria with subjective criteria [3, 14].

3. Common Sense Knowledge Based Interestingness Measure

In subjective interestingness measures, the user's knowledge related to data is important. Patterns and rules are more exciting and interesting when they cannot be predicted by the knowledge a user possesses, and for this reason, the user's knowledge representation is at the core.

In this paper, we use common sense knowledge for the knowledge representation of the user. Common sense knowledge is basic knowledge necessary in common social communication and interaction and can be regarded as uninteresting knowledge since it is mostly predictable and known. Common sense knowledge includes knowledge about the spatial, physical, social, temporal, and psychological aspects of everyday life. For example, "A lemon is sour", "A husband is man", and "To open a door, you must usually first turn the doorknob". As shown in Table 1, the CSM can be effectively used in finding interesting rules

since Common Sense Knowledge shows opposite characteristics of the interesting knowledge that we are hoping to discover.

Table 1. Comparison of Interestingness Criteria and Common Sense Knowledge

Interestingness criteria	Common Sense Knowledge
Novel	Known
Unexpected	Expected
Surprising	Common

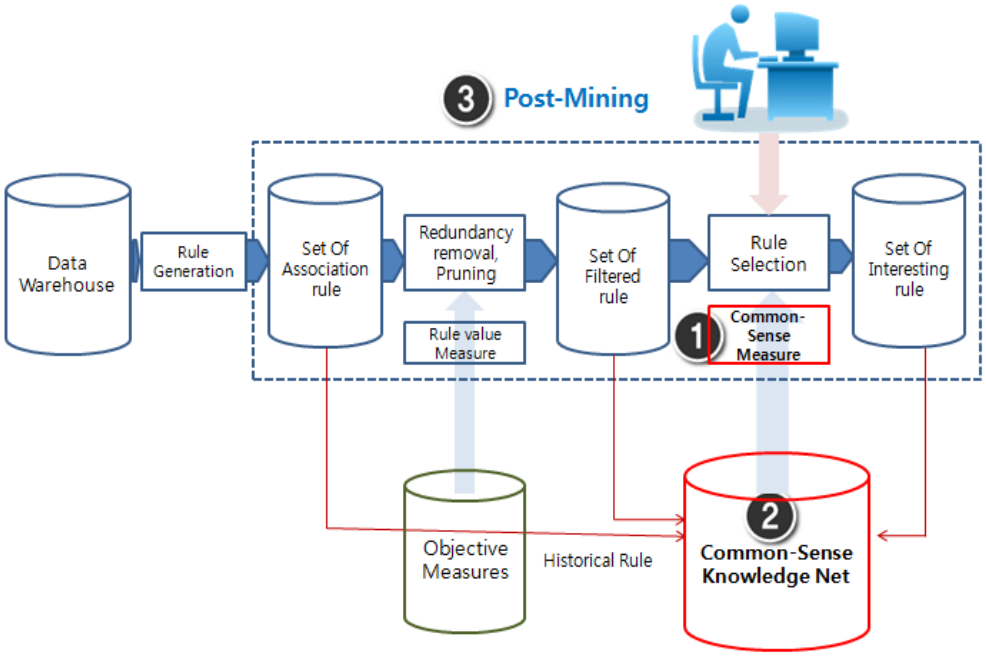


Figure 1. Post-processing in Data Mining

Figure 1 shows the post-processing technique of data mining. The CSM evaluates rules selected during the data mining post-processing step based on the knowledge collected in the common sense network(CSN) and ultimately delivers rules without any uninteresting knowledge. The CSN includes the domain knowledge of ConceptNet5.

4. CSM Implementation

We proposed the CSM in common sense knowledge based interestingness measure for data mining. We implement the CSM by using a new similarity technique that measures the similarity between association rules and common sense knowledge. The new similarity technique is based on the cosine distance similarity technique in the vector space model and takes into consideration the fact that association rules are formed in an itemset of the database and the fact that a concept of common sense knowledge may be a higher-order compound and not a simple lexical item. Furthermore, we need a semantic approach when we take into consideration the various problems that occur when human knowledge is expressed in a textual format.

4.1. The Knowledge Component Vector

In the vector space model, we define association rules as vectors and define attributes and values of the dataset as dimensions. The set $AV = \{av_1, av_2, av_3, \ldots, av_i\}$ is composed of attributes and values of the dataset. The definition of terms depends on the application. In this paper, terms are attributes and values of the dataset. They are the dimensionality of the vector.

Rule : *age = old* → *skin = wrinkle*
antecedent relation consequent

$$rule\ \vec{r} = \{\vec{r}_a, \vec{r}_c, \vec{r}_r\}. \quad (1)$$

$$\vec{r}_a = \left(w_{a,av_1}, w_{a,av_2}, w_{a,av_3} \ldots w_{a,av_i}\right). \quad (2)$$

$$w_r(a, av_i) = tf(av_i, r_a) \times ln\frac{av_i.count}{av_i.supp}, \quad \text{when } av_i.supp,\ av_i.count > 0 \quad (3)$$

where $av_i.supp$ is the support of av_i and $av_i.count$ is the distinct value count of attribute.

The dimensional weight of the rule r, $w_r(a, av_i)$ is based on the keyword importance that shows how important av_i is in the specific rule r. The association rule is generated by mathematical notions called "support" and "confidence". They are important objective measures to users. This importance should be reflected in the rule representation. Rules with very low support values are uncommon and probably represent outliers or very small numbers of transactions that are unlikely to be interesting or profitable. Items that have transactions in a skewed support distribution can make a cross-support pattern. It is obstacle to select interesting rules for a user.

Common Sense Knowledge: *Growing old has the effect of wrinkling the skin.*
concept1 relation concept2

$$Common\ Sense\ \overrightarrow{cs} = \{\overrightarrow{cs}_a, \overrightarrow{cs}_c, \overrightarrow{cs}_r\}. \quad (4)$$

$$\overrightarrow{cs}_a = \left(w_{a,av_1}, w_{a,av_2}, w_{a,av_3} \ldots w_{a,av_i}\right). \quad (5)$$

$$w_{cs}(a, av_i) = tf(av_i, cs_a) \times ln\frac{av_i.count}{av_i.supp} \times CN\text{-}SIM(av_i, cs_a), \quad (6)$$

when $av_i.supp$, $av_i.count > 0$,
where $av_i.supp$ is the support of av_i and $av_i.count$ is the distinct value count of attribute.

The dimensional weight of the common sense entry cs, $w_{cs}(a, av_i)$ is influenced by the common sense network similarity(CN-SIM). This is in order to resolve the problem of keyword matching of the vector space model and to measure the semantic similarity. We find the most similar concept and reflect the CN-SIM in the weight. CN-SIM(av_i, cs_a) is the similarity between the rule and concepts in CSN.

4.2. Semantic Network Based Rule Matching Method with Common Sense Knowledge

The set AV described in the previous subsection also contains dimensions for defining common sense knowledge into the vector space model. However, to correspond items in the rules with concepts of common sense knowledge, we apply context-based inference. This is in order to resolve the problem of keyword matching of the vector space model and to measure the semantic similarity. We propose a semantic matching technique using the CSN.

```
Algorithm RuleMatching()
Input :   An association rule rule, Common Sense Network CSN
Output :  Matched Common Sense Knowledge
          k={sentence, similarity, sentence_type, concept, relation}
Begin
1:  rule_cv = converted rule into logic by implication conversion law
2:  rule_sim[max] = decomposed simple rules
3:  max = the number of attribute
4:  n=decomposed simple rule's number
5:  structure knowledge k[max+1] = set of matched common sense knowledge
6:  if (a rule is simple){
7:          k[0]:=SearchCSN(rule);
8:          OutputK(k);}
9:  else (a rule is complex){
10:         k[0]:=SearchCSN(rule);
11:         rule_cv: = converted rule into logic by implication conversion law
12:         if (a type of rule is 1:M)
13:                 decompose rule_cv into a conjunction
14:         else if (a type of rule is N:1)
15:                 decompose rule_cv into a disjunction
16:         else (a type of rule is N:M)
17:                 decompose rule_cv into a conjunction of disjunction
18:         rule_sim[max]:= decomposed simple rules
19:         For(i=1 to n)
20:                 k[i]:=SearchCSN(rule_sim[i]);
21:         OutputK(k);}
end
```

Figure 2. The Algorithm of the Rule Matching Method

As we explained in the previous section, CSN is conceptually represented as semantic network. It is built from nodes representing concepts, in the form of words or short phrases of natural language, and labeled relationships between them in ConceptNet5. To measure the semantic distance between nodes in CSN, we use an association method in ConceptNet5 to find a concept that is most similar to a rule.

As shown in Figure 2, to find the common sense knowledge to match a rule, we search the CSN. If the rule is simple, we can find the knowledge easily. However, complex rules might be matched by compound sentence through the several steps like the conversion process and decomposition process. We propose the following matching method to find the common sense knowledge by rule types.

1:1 Rules The rules have a single item in their antecedent and consequent. They have an implicative formula such as $X \rightarrow Y$. We search the CSN to find the common sense knowledge associated with items in the rule. If an assertion exists, we measure the similarity between items in the rule and the assertion. Let us consider the example of rule R_1: *husband* → *male*. We can find the knowledge "*A husband is a male spouse*" and the CN-SIM value 0.924.

1:M Rules The rules have a single item in their antecedent and more than one item in their consequent. They have an implicative formula such as $X \rightarrow Y_1, Y_2, Y_3, \ldots, Y_m (m \geq 2)$. In propositional logic, conditional implication is a valid rule of replacement that states that "*A implies B*" is logically equivalent to "*not-A or B*". The conditional statement $X \rightarrow Y_1, Y_2, Y_3$ is converted into $\neg X \vee (Y_1 \wedge Y_2 \wedge Y_3)$ by the implication conversion law [15]. R_2 : *husband* →

male, spouse, sports is decomposed into a conjunction of simple rules: *(husband→male) ∧ (husband→spouse) ∧ (husband→sports).* We measure the similarity between simple rules and assertions. We select the minimum value of similarity in conjunction of simple rules. We match the common sense knowledge with the maximum value of the similarity to the rule.

N:1 Rules The rules have more than one item in their antecedent and a single item in their consequent. They have an implicative formula such as $X_1, X_2, X_3, \ldots, X_n \rightarrow Y$ ($n \geq 2$). The conditional statement $X_1 X_2, X_3 \rightarrow Y$ is converted into $\neg(X_1 \wedge X_2 \wedge X_3) \vee Y$ by the implication conversion law. R_3 : *husband, white, USA → male* is decomposed into a disjunction of simple rules: *(husband→male) ∨ (white→male) ∨ (USA→male).* We measure the similarity between the simple rules and assertions. We select the maximum value of similarity in disjunction of simple rules. We match the common sense knowledge with the maximum value of the similarity to the rule.

N:M Rules The rules have more than one item in their antecedent and consequent. They have an implicative formula such as $X_1, X_2, X_3, \ldots, X_n \rightarrow Y_1, Y_2, Y_3, \ldots, Y_m$ ($n,m \geq 2$). The conditional statement $X_1, X_2 \rightarrow Y_1, Y_2$ is converted into $\neg(X_1 \wedge X_2) \vee (Y_1 \wedge Y_2)$ by the implication conversion law. R_4 : *husband, wrinkle → male, old* is decomposed into a conjunction of disjunction of simple rules: *{(husband → male) ∨ (wrinkle → male)} ∧ {(husband→old) ∨ (wrinkle→old)}.* We measure the similarity between the simple rules and assertions. We can find two assertions related to the rule. We select the minimum value of maximum values of similarity in disjunction of simple rules. We match the common sense knowledge with the maximum value of the similarity to the rule.

4.3. CSM Calculation

By calculating the cosine distance similarity that compares the angles between vectors of each component of a rule and matched common sense knowledge, the CSM can be calculated. The cosine is a measure of the angle θ formed by two vectors $\vec{r}$ and $\overrightarrow{cs}$. Two vectors are the most similar when their coordinates are the same or positively proportional. In that case, the angle they form is 0° and the value of the cosine is one. If the value of the cosine is zero, it indicates that the rule and common sense knowledge vectors form a right angle and that there is no similarity between the two.

$$CSM(\vec{r}_a, \overrightarrow{cs}_a) = \frac{\sum_{k=1}^{i} w_r(a,av_k) w_{cs}(a,av_k)}{\sqrt{w_r(a,av_k)^2}\sqrt{w_{cs}(a,av_k)^2}} . \quad (1)$$

We measure the relation similarity between rules and common sense knowledge with values between zero and one by analyzing the similarity in 3000 detailed relationships of knowledge and the co-occurrence relationship in the rule.

Table 2. Example of Common Sense Knowledge

ID	Common Sense Knowledge	Antecedent-concept	Consequent-concept	Relation
1	Sports cars are used to impress young women	Sports cars	impress young women	{1} are used to {2}
2	Sport stars can only play while they are young.	Sport stars	play while they are young	{1} can only {2}
3	Rock climbing is a favorite sport of the young	Rock climbing	a favorite sport of the young	{1} is {2}
4	The young men can like sports	The young men	like sports	{1} can {2}

Table 3. Example of CSM

ID	$CSM(\vec{r}_a, \overrightarrow{cs}_a)$	$CSM(\vec{r}_c, \overrightarrow{cs}_c)$	$CSM(\vec{r}_r, \overrightarrow{cs}_r)$	$CSM(\vec{r}, \overrightarrow{cs})$
1	0.191	0.705289	0.5	0.45839
2	0.251	1	0.7	0.640291
3	0	0.2939	0.5	0.21756
4	1	0.524889	0.7	0.749955

5. Experiment

For the CSM experiment for finding interesting rules, the Internet shopping mall data from the Knowledge Discovery and Data Mining Data set were used. The Apriori algorithm was applied to 10,268 individual data entries with nine characteristics. After applying objective criteria, such as confidence, lift, conviction, and leverage, 500 best rules were selected. We defined these rules as vectors and defined attributes and values of the Internet shopping mall data set as dimensions. We searched the CSN in order to find the common sense knowledge matching the rules.

Table 4. Rule Matching with Common Sense Knowledge

Number of rules	%	Knowledge type	Knowledge structure	Model
212	42%	Common sense K	simple sentence	Vector space model
63	13%	Common sense K	complex sentence	Vector space model
45	9%	Common sense K	linked sentence	Graph model
55	11%	Common sense K	compound sentence	Graph model Vector space model
125	25%	Not match	-	-
500	**100%**	**Total**		

The number and percentage of the rules matching the common sense knowledge were 375 and 75%, respectively. These rules are regarded as the knowledge that is similar to common sense. They can be also determined as the knowledge that is not interesting according to the similarity. Table 4 shows that a rule can match the common sense knowledge with a compound or linked sentence structure when it cannot match the common senses knowledge with the simple sentence structure.

Table 5. Comparison of CSM and Expert Evaluation

CSM User evaluation	cs=0	0<cs<0.5	0.5<cs<1	cs=1
u=1	12%	0%	0%	0%
0.5<u<1	7%	6%	9%	0%
0<u<0.5	5%	1%	53%	0%
u=0	1%	0%	1%	5%

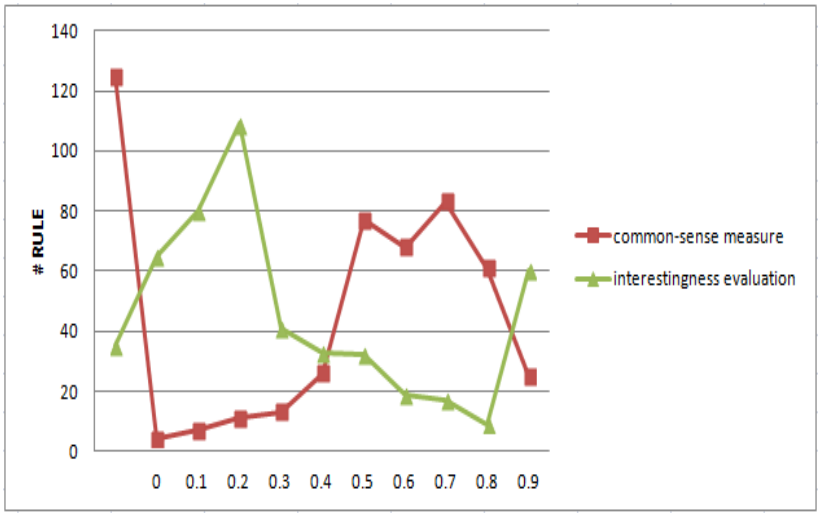


Figure 3. A Comparison of the Number of Rules in CSM and Experts' Evaluation

The CSM of rules was measured and evaluated by four domain users who evaluated interestingness of the rules based on the evaluation standards provided to them. The categories of the evaluation standards included: Expected/Unexpected, Actionable/ Unactionable, Noble, and Accuracy. The final interestingness was measured between zero and one. In Table 5, cs=1 represents the rule closest to the common sense knowledge and u=1 has been determined to be the most interesting rule. The interesting rule had a CSM value close to zero. Figure 3 shows a comparison of the number of rules by the CSM and experts' evaluation. It shows that the number of the rules according to the CSM is inversely proportional to the number of the rules according to the results of the evaluation of experts. As we have predicted earlier, this experimental result proves that the rules with large CSM values are the rules that are uninteresting according to experts' evaluation. Figure 4 shows the accuracy, precision, recall, and F-measure of the CSM. The experimental result of the CSM coincides with the evaluation result of the experts. and 76%. 87% of rules, Rules with the CSM value larger than 0.5 were deemed uninteresting knowledge.

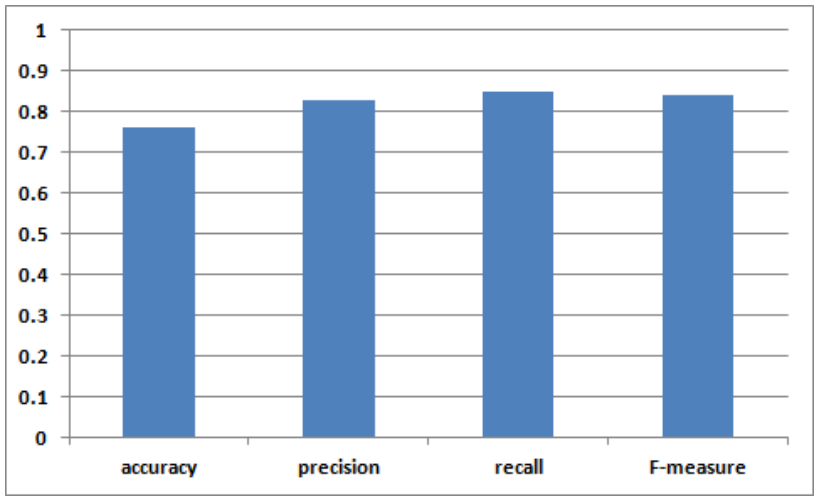


Figure 4. The Evaluation of CSM

6. Conclusion

In this paper, we proposed the CSM so that only interesting knowledge can be selected in order to resolve the problem resulting from a large quantity of rules and patterns. The CSM is an interestingness measure that evaluates how closely the rules produced from data mining match common sense knowledge and uses a hybrid approach that combines objective measures and subjective measures. We proposed the algorithm of rule matching method with common sense knowledge using the CSN. We implemented a new similarity technique in order to take into consideration the fact that association rules are formed in an itemset of the database and the fact that a concept of common sense knowledge may be a compound in higher order and not a simple lexical item. The result of the experiments shows that the CSM accurately finds similarities among structuralized knowledge and is efficient in eliminating uninteresting rules that are not filtered by previous objective measures.

In future studies, we should focus not only on selecting interesting data for the user by post-processing common sense knowledge but also on showing effects of improved capabilities that can reduce time, space, and cost in the whole data mining process. This research can be expanded to remove the uninteresting rules and patterns and improve the performance of the mining algorithm in the various data mining research fields such as clustering and classification. In addition, the common sense knowledge research is applicable to the ontology and artificial intelligence field application where the computer processes human language and text.

Acknowledgements

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology(2012R1A1A2003764).

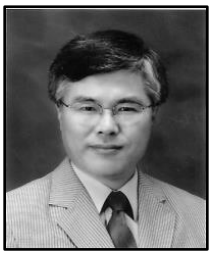
References

[1] H. Jiawei and K. Micheline, “Data Mining-Concept and Technique”, Morgan Kaufmann, USA, **(2006)**.
[2] K. McGarry, “A Survey of Interestingness Measures for Knowledge Discovery”, The Knowledge Engineering Review, vol. 20, no. 1, **(2005)**, pp. 39-61.
[3] G. Liqiang and J. Hamilton, “Interestingness Measures for Data Mining: A Survey”, ACM Computing Surveys, vol. 38, no. 3, **(2006)**, pp. 1-32.
[4] I. Lee and H. S. Yong, “Common sense knowledge based hybrid interestingness measures for data mining”, Proceedings of 6th International Conference on Convergence and Hybrid Information Technology (ICHIT), LNCS, vol. 7425, **(2012)**, pp. 146-154.
[5] R. Speer, C. Havasi and H. Lieberman, “AnalogySpace:Reducing the dimensionality of common sense knowledge”, Proceedings of the 23 Association for the Advancement of Artificial Intelligence(AAAI) Conference on Artificial Intelligence, **(2008)**, pp. 548-553.
[6] H. Liu and P. Singh, “ConceptNet: A Practical Commonsense Reasoning Toolkit”, BT Technology Journal, vol. 22, no. 4, **(2004)**, pp. 211-226.
[7] C. Havasi, R. Speer and J. Alonso, “ConceptNet 3: a Flexible, Multilingual Semantic Network for Common Sense Knowledge”, Proceedings of Recent Advances in Natural Languges Processing, **(2007)**, pp. 27-29.
[8] R. Speer, ConceptNet5, http://conceptnet5.media.mit.edu/.
[9] Z. Yanchang, “Post-Mining of Association Rules”, Information Science Reference, USA, **(2009)**.
[10] T. Wu, Y. Chen and J. Han, “Re-examination of interestingness measures in pattern mining: a unified framework”, Data Mining and Knowledge Discovery, vol. 21, no. 3, **(2010)**, pp. 371-397.
[11] J. David, F. Guillet, R. Gras and H. Briand, “Comparison of interestingness measures applied to textual taxonomies matching”, Revue des Nouvelles Technologies de l'Information, **(2008)**.
[12] E. Suzuki, “Compression-Based Measures for Mining Interesting Rules”, Next-Generation Applied Intelligence, **(2009)**, pp. 741-746.
[13] S. Sahar, “On incorporating subjective interestingness into the mining process”, Proceedings of the 2002 IEEE International Conference on Data Mining, Maebashi City, Japan, **(2002)**, pp. 681-684.

[14] C. Sengstock, M. Gertz and T. Van Canh, "Spatial interestingness measures for co-location pattern mining", Proceedings of 12th IEEE International Conference on Data Mining Workshops(ICDMW), IEEE, article.6406524, **(2012)**, pp. 821-826.
[15] K. H. Rosen, "Discrete Mathematics and Its Applications", 7th Edtion, McGrawHill, USA, **(2011)**.

Authors

Ingi Lee is a Ph.D. student in Department of Computer Science and Engineering at Ewha Womans University in Seoul, South Korea. She obtained her master degree in database system from Ewha Womans University in 2001. Her main research interests include data mining, ontology, and big data management.



Hwan-Seung Yong is a Professor in Department of Computer Science and Engineering at Ewha Womans University in Seoul, South Korea. He obtained his Ph.D degree in database system from Seoul National University in 1994. His main research interests include data mining, ubiquitous computing, and big data management.