International Journal of Multimedia and Ubiquitous Engineering
Vol.10, No.11 (2015), pp.11-20
http://dx.doi.org/10.14257/ijmue.2015.10.11.02

Text Mining: Extraction of Interesting Association Rule with Frequent Itemsets Mining for Korean Language from Unstructured Data

Irfan Ajmal Khan, Junghyun Woo, Ji-Hoon Seo and Jin-Tak Choi

Department of computer Engineering
INU (Incheon National Univeristy)
Incheon, South Korea
manikhan@nate.com, sunhwangje@hanmail.net, sserz@inu.ac.kr, choi@inu.ac.kr

Abstract

Text mining is a specific method to extract knowledge from structured and unstructured data. This extracted knowledge from text mining process can be used for further usage and discovery. This paper presents the method for extraction information from unstructured text data and the importance of Association Rules Mining, specifically for of Korean language (text) and also, NLP (Natural Language Processing) tools are explained. Association Rules Mining (ARM) can also be used for mining association between itemsets from unstructured data with some modifications. Which can then, help for generating statistical thesaurus, to mine grammatical rules and to search large data efficiently. Although various association rules mining techniques have successfully used for market basket analysis but very few has applied on Korean text. A proposed Korean language mining method calculates and extracts meaningful patterns (association rules) between words and presents the hidden knowledge. First it cleans and integrates data, select relevant data then transform into transactional database. Then data mining techniques are used on data source to extract hidden patterns. These patterns are evaluated by specific rules until we get the valid and satisfactory result. We have tested on Korean news corpus and results have shown that it has worked well, and the results were adequate enough to research further.

Keywords: *Association Rules Mining, text mining, Frequent Item sets, Classification*

1. Introduction

The key purpose of data mining is to create a system that efficiently search the information from the data complexity and find the information user need. Data mining can be used for many purposes *i.e.*, Sales marketing, by knowing the shopping pattern of customer. This pattern/information can be used to as sales strategy and help increase revenue. Fraud Detection, Banks may wish to determine if any of their credit cards are being used for fraudulent purposes. Unusual spikes in a customer's spending pattern may indicate fraud. With the help of data mining techniques we can logically search through a large amount of data in order to extract important data. This is normally accomplished by finding the interesting hidden patterns. The relation between these patterns can be later process or analyze to solve number of problems. Mining data sometimes called data KDD (knowledge discovery in database). KDD is the process of analyzing data from different viewpoints and summarizing it into valuable information. Typically it is processed in three different steps Preprocessing, Mining and Evaluating. Preprocessing is where data cleaning, transformation, integration and selection of data happen. Next step is the main step, where different algorithms are used to find hidden knowledge from preprocessed data. After that is the evaluation of the mined data which is also called post processing.

ISSN: 1975-0080 IJMUE

These processes are done until we get the required result. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases. We have used same technique to extract meaningful association rules out of Korean text. First step is to pre-process data (getting rid of un-necessary data), then we transformed process data into transactional database. Grouping this transactional database with other hand-built databases we mined some interesting and useful patterns in given Korean text data collection (corpus).

Rest of the paper is structured as follows: Section 2 covers related work; Section 3 describes the association rules mining for Korean language. Experiments and results are in Sections 4, and Sections 5 are conclusion and future work.

1.1. Association Rules

Association rule mining was first proposed in [14]. The main goal of this research is to calculate and find hidden knowledge frequent patterns, correlations, associations or casual structures among sets of items in the transaction databases or other data repositories. Following is the original definition from [14]. The problem of association rule mining is defined as: Let I= {i1, i2, i3,…… in} be a set of n binary attributes called items. Let D= {t1, t2, t3,…… tm} be a set of transactions called the database. Each transaction in D has a unique transaction ID and contains a subset of the items in I. A rule is defined as an implication of the form $X \Rightarrow Y$ where X, $Y \subseteq I$ and $X \cap Y = \varnothing$. The sets of items (for short itemsets) X and Y is called antecedent (left-hand-side or LHS) and consequent (right-hand-side or RHS) of the rule respectively.

Association rules are usually required to satisfy a user-specified minimum support and a user-specified minimum confidence at the same time. Association rule generation is usually split up into two separate steps. First, minimum support is applied to find all frequent itemsets in a database. Support is the rule holds with support sup in D if sup % of transactions contain $X \cup Y$, $sup = Pr(X \cup Y)$. Second, these frequent itemsets and the minimum confidence constraint are used to form rules. Confidence is the rule holds in D with confidence conf if conf % of transactions that contain X also contain Y, $conf = Pr(Y \mid X)$. An association rule is a pattern that states when X occurs, Y occurs with certain probability.

1.2. Measures Association Rules

Basically association mining is about discovering a set of rules that is shared among a large percentage of the data. Association rules mining tend to produce a large amount of rules. The objective is to find the rules that are useful to users. There are two ways of computing usefulness, being subjectively and objectively. Objective measures involve statistical analysis of the data, such as support and confidence [14]. Support The rule $X \Rightarrow Y$ holds with support s if s% of transactions in D contains $X \cup Y$. Rules that have as greater than a user-specified support is said to have minimum support. Confidence- The rule $X \Rightarrow Y$ holds with confidence c if c% of the transactions in D that contain X also contain Y .Rules that have a c greater than a user-specified confidence is said to have minimum confidence.

2. Related Work

Association rule mining, a data mining technique was first introduced in [14]. The objective of the research was to extract frequent patterns, hidden relationships among sets of items in the transaction databases. Nowadays, frequent itemset mining is a main interest of research and already a lot of algorithms have already been proposed [1], simplest and most common ARM algorithm. The main focus of ARM is on the sub-problem of efficient frequent rule generation. Some of other researches have shown great

interest in infrequent associations for mining the rare association rules. In [2] extraction of rare association is done by setting the level of support low enough to find rare association rules. The number of itemsets is much larger than the text database, in [15] the research was done by cutting off the infrequent itemsets and dividing database to improve the algorithm efficiency. This technique has shown greater efficiency as compared to Apriori algorithms. In [7], a partition algorithm is proposed which can minimize the scans on database to only two by further dividing the database into small partitions to fit them into main memory. Chao Tang in [8], proposed a model for mining grammatical rules from Chinese text utilizing. Model was based on three main steps, preprocessing, mining association rules and then verification of those rules. The results showed that the algorithm works better on smaller length sentences. In [9], Solution based on Association Rules mining and Apriori algorithm for Word Sense Disambiguation problem was proposed. The result has shown a promising result extracting association rules between sense of ambiguous words and context. Dough Won Choi proposed an improved version of Apriori algorithm for candidate and frequent itemsets in [10], this idea was to eliminate the candidate itemsets on the basis of 'minimum and minimum relative support' values to find transitive relations. The results show that this method generated more items. Recent studies on reducing the large amount of mined association rules has been done in, normally two different ways, either by increasing the minimum confidence parameter or by increasing the minimum support parameter. In [11-12], proposed the work on reducing the amount of mined association rules by adjusting the rule induction or pruning the rule set.

3. Association Rules Mining for Korean Language

Association rules mining for Korean language is a complex task because, it is a mixture of Hanja, Chinese and Japanese languages and it is written as a group of blocks and each block transcribes a syllable Modern Hangul syllable blocks can be expressed with either two or three jamo, either in the form consonant + vowel or in the form consonant + vowel + consonant. There are 19 possible leading (initial) consonants (choseong), 21 vowels (jungseong), and 27 trailing (final) consonants (jongseong). Thus there are 399 possible two-jamo syllable blocks and 10,773 possible three-jamo syllable blocks, giving a total of 11,172 modern Hangul syllable blocks [5]. Followings are the objectives of our research project for text mining.

- Pre-processing of data: Data in the text form is multifaceted to extract something meaningful data needs to be preprocessed. Major task to measure the quality of data is done by data cleaning, data integration, data transformation, data reduction, data discretization is done
- Management of words transactional database: In text databases, distribution of words varies from the conventional transactional databases. Numbers of unique words are significantly larger than the number of unique items in a transactional database.

One of the most important tasks for mining association rules is to perform pre-processing on Korean text files and convert them in a transactional database. Step required to deal with complexity of preprocessing of Korean language are discussed as below in detail.

3.1. Mining Association Rules from Korean Text

Our proposed mining association rule method (model) provides one solution to specific natural language processing tasks for Korean language, from the first step, preprocessing of data up to extraction of association rules. Figure1 illustrates a Korean language model. It consists of following major steps: pre-processing of Korean text files (corpus), then, is to create a transactional database from the preprocessed data and finally applying

algorithm to search and find meaningful patterns within processed data from Korean text and collection of databases, using association rules mining technique. It is very important to Pre-process data because it contains some unnecessary and meaningless data, numeric and punctuation data. These unnecessary data causes the higher number of meaningless frequent itemsets also wastage of time and resources.

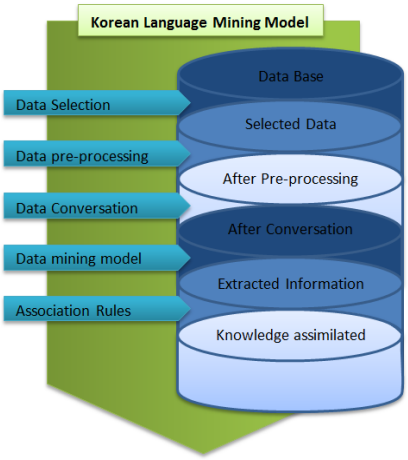


Figure 1. Korean Text Mining Model

3.2. Normalization/Unification

Performing normalization is the first step and it further has two main functions. Han unification is an effort by the authors of Unicode and the Universal Character Set to map multiple character sets of the so-called CJK languages into a single set of unified characters. Han characters are a common feature of written Chinese (hanzi), Japanese (kanji), and Korean (hanja). First is to unifying different Unicode, which means that characters of identical shape but with different encoding or meaning in other language. Table1 and Figure 2 show an example of this situation. Figure4 on left is the dictionary of words borrowed from other languages. Second, if text documents contain any Romanized (konglish) words, then to remove diacritical so latter it can be processed for better understanding and mining (Korean government officially revised the Romanization of the Korean language in July 2000 to eliminate diacritic).

Hanja	Korean	Chinese	Unicode
月	월	달	U+6708
音	음	소리	U+97F3
木	목	나무	U+6728
毛	모	털	U+6BDB
水	수	물	U+6C34
家	가	집	U+5BB6
小	소	적다	U+5C0F
不	부	아니다	U+4E0D
苦	고	괴롭다	U+82E6

English Word	Korean Words
website	웨브사잇, 웹사이트
homepage	홈페이즈, 홈페이지
computer	컴퓨터, 컴퓨러
condition	콘디션, 컨디션
service	서비스, 써비스
video	비디오, 비데오
Fusion	퓨션, 휴션

Figure 2. Hanja and its Meaning in Korean and Chinese

Table 1. Unicode Standards & Number of Characters

Standard	*Number of Characters*
ANSI Z39.64-1989 (EACC)	13053
Big Five	13481
CCCII, Level 1	4808
CNS 11643-1986	13051
CNS 11643-1986 User Characters	3418
GB 2312-80 (GB0)	6763
GB 12345-90 (GB1)1	2176
GB 7589-87 (GB3)	7327
GB 7590-87 (GB5)	7039
General Use Characters for Modern Chinese (GB7)2	41
GB 8565-89 (GB8)3	287
GB 12052-89 (Korean)	94
JEF (Fujitsu)	3149
JIS X 0208-1990	6355
JIS X 0212-1990	5801
KS C 5601-1989	4888
KS C 5657-1991	2856
PRC Telegraph Code	~8000
Taiwan Telegraph Code	9040
Xerox Chinese	9776
Total characters covered	~121403
Total unique characters	21001

3.3. Tokenization

The next step is, word segmentation also known as tokenization. The main focus of tokenization is to recognize word boundaries exploiting orthographic word boundary delimiters, punctuation marks, written forms of alphabet and affixes. We have combined rule based methods and dictionary based methods and converts various forms of writing to a unique standard one. Tokens from Korean language text as shown in Figure 3

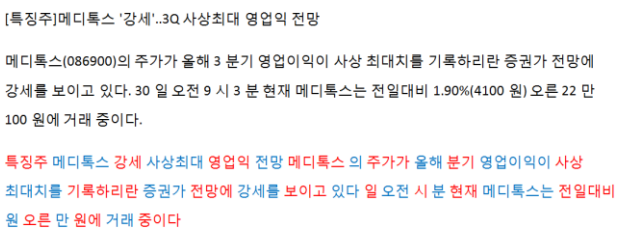


Figure 3. Korean Text Tokens

Tokenization strategies

- Whitespace
- Word, boundaries
- Unicode category-based
- Linguistic
- Whitespace/newline-preserving

3.4. Handling Word Joiner and Zero Width Space

Text that is encoded with both a DBCS and SBCS is typically displayed such that the glyphs representing DBCS characters occupy two display cells—where a display cell is defined in terms of the glyphs used to display the SBCS (ASCII) characters. In these systems, the two-display-cell width is known as the fullwidth and the one-display-cell width is known as the halfwidth form [5-11]. When Hangul compatibility jamo are transformed with a compatibility normalization form, NFKD or NFKC, the characters are

converted to the corresponding conjoining jamo characters. Where the characters are intended to remain in separate syllables after such transformation, they may require separation from adjacent characters. This separation can be achieved by inserting any non-Korean character.

- U+200B zero width space is recommended where the characters are to allow a line break.
- U+2060 word joiner can be used where the characters are not to break across lines.

Original	NFKD	NFKC	Display
ㄱ ㅏ 3131 314F	ㄱ ㅏ 1100 1161	가 AC00	가
ㄱ ZW SP ㅏ 3131 200B 314F	ㄱ ZW SP ㅏ 1100 200B 1161	ㄱ ZW SP ㅏ 1100 200B 1161	ㄱㅏ

Figure 4. Separating Jamo Characters

3.5. Tokenization of Clitics

Clitics are elements having some properties of affixes and words but they are not different from affixes and word. When contrasted with independent words, clitics have some of the properties of affixes and when contrasted with clitics, words have some of the properties of syntactic phrases. Sometimes clitic acts as a verb, sometimes as an object and so on. Both nouns and pronouns take case clitics. Pronouns are somewhat irregular. As with many clitics and suffixes in Korean, for many case clitics different forms are used with nouns ending in consonants and nouns ending in vowels. The most extreme example of this is in the nominative (subject), where the historical clitic I ᄋᆡ is now restricted to appearing after consonants, and a completely unrelated (suppletive) form –ka (pronounced -ga) appears after vowels as show in Figure 5 [6].

Case	After V	After C
Nominative	**ka** 가 *-ga*	**-i** 이
Accusative	**lul** 를 *-reul*	**ul** 을 *-eul*
Genitive	**-uy** 의 *-ui*[1]	
Dative (also destination)	**-ey** 에 *-e* (inanimate) **-ey key** 에게 *-ege* (animate)	
Locative (place of event, also source)	**-ey se** 에서 *-eseo* (inanimate) **-ey key se** 에게서 *-egeseo* (animate)	
Instrumental	**-lo** 로 *-ro*[2]	**-ulo** 으로 *-euro*
Comitative (also *and*)	**-hako** 하고 *-hago*	
	-wa 와	**kwa** 과 *-gwa*
	lang 랑 *-rang*	**-i lang** 이랑 *-irang*

Figure 5. Case Clitics

3.6. Stemming

Stemming is a simple experimental process which is often used to chops off the affixes of words to get the concrete and precise sense of the word. To obtain information using morphological techniques Stemming has been widely adopted for morphological technique for extracting information [13]. We have used five different databases for stemming our Korean text files, out of five databases 3 were individually built by our group members. The First database just contains some irregular nouns and their corresponding stems. Second database consists of Korean words with their POS tags along with frequencies. Third database is collection of Korean verbs stems linked to both

present and past stem. Fourth dictionary contains words borrowed from other languages and its written variations. The last database stores the entire pattern collection which for each syntactic category keeps their acceptable morphological rules and valid prefixes and affixes. These patterns are later classified according to different syntactic categories and then for each category every possible and valid rules, prefix and postfixes were collected.

Table 2. Some of the Used Tags

Original Tag	*Equivalent tag*	*Description*
A0, A1, A2	ADJ	Adjective
Ad	ADV	Adverb
N1, N2, N3, N4	NN	Noun
V1	VVP	Present Verb
V2	VVT	Past Verb

Name	Frequency	Description	
가격	101	명사	NOUN
가격상한폭까지	5	부사	ADVERB
가격의	3	형용사	ADJECTIVE
가격제한폭까지하락했다	1	과거동사	PAST VERB
가까워지자	26	부사	ADVERB
가깝게	8	부사	ADVERB
가능한	24	형용사	ADJECTIVE
가능한만큼	1	부사	ADVERB
가능했다	7	과거동사	PAST VERB
가동	6	명사	NOUN
가동률	9	명사	NOUN
나설지	3	부사	ADVERB
나섰고	2	과거동사	PAST VERB
나스닥	14	명사	NOUN
나온다	19	현재동사	PRESENT VERB
나왔습니다	8	과거동사	PAST VERB
나왔을	11	형용사	ADJECTIVE
나타난다	6	현재동사	PRESENT VERB
다가온	9	형용사	ADJECTIVE
다가왔다	1	과거동사	PAST VERB
다각적으로	2	부사	ADVERB
다각화	1	명사	NOUN
다르다	3	현재동사	PRESENT VERB
다문화가정	1	명사	NOUN

Figure 6. Part of Word Database

Every single Korean sentence ends in either a verb or an adjective syllable 다 100% of the time, also several verbs and adjectives end with the two syllables 하다 which means ‘do’. 하다 is a very important verb, because you can simply eliminate the 하다 to make the noun form of that verb/adjective. Table3 shows some of the collected patterns from the pattern database.

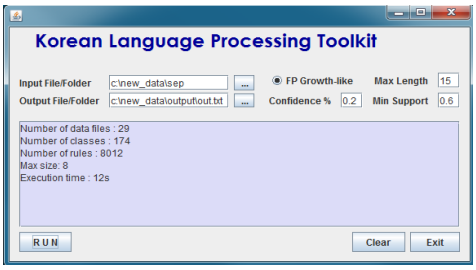
Table 3. Unicode Standards & Number of Characters

Verb Stem 하다	Formal	Standard	Intimate
Present	합니다	해요	해
Past	했습니다	했어요	했어

Stemming process as follow
Input: data
Step1: look up for the word in first dictionary
IF found then RETURN matching result
ELSEIF search in words lexicon by eliminating affix RETURN matching result
Step2: Crosschecks the word after affix elimination with pattern database IF found THEN RETURN as valid stem.
Step3: IF it matches any verb patterns THEN RETUEN as a valid verb stem.
Step4: IF needed eliminate one or more affix to find the stem and REPATE until matches
Output: Valid Stem

4. Experiment and Results

For testing and experiment we have run test on Korean news corpus and performance result of different processes of our Korean language mining model. Our Korean language mining model has extracted some interesting association rules. Initially Korean text files have contained 98,693 words, after preprocessing the number of words reduced to 43,002 words as shown in Figure 10. Each word has maximum length of 15 characters and minimum. A transactional database was created on the basis of these token extracted from corpus and assigning each word a unique transactional ID. Testing is done by using monthly news data of 30 news casters with different number of words. Figure 7 (right) show some of the extracted frequent term sets from the collection of documents. A graphical view of associations between words is shown in Figure 8 and Figure 9 below, Result show some interesting relations/patterns among words and it would produce better result with little bit better techniques while pruning and pre-processing data. During processing of data it has observed that the number of association rules are increased as the number of words are increased, validity of Korean language are greatly affected by the number of words being processed.



Min Support (80)	Frequent term set
84	[현재,포인트,시가총액,SK 하이닉스]
86	[현재,포인트,시가총액,각각]
84	[있다,포인트,시가총액,중이다]
87	[억원,코스피,오른,시가총액]
	[…]
Resulting Cluster	현재, 있다,포인트,시가총액,SK 하이닉스

Figure 7. (Right) Front End User Interface, (Left) Frequent Itemsets from Out.Txt File

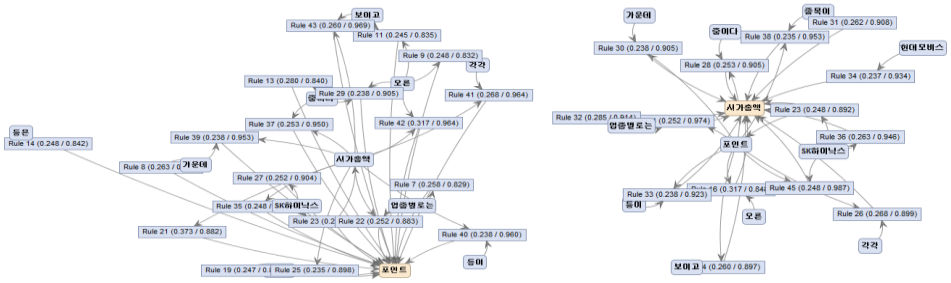


Figure 8 & 9. Associations of Word (시가총액 Market Capitalization & 포인트 Point)

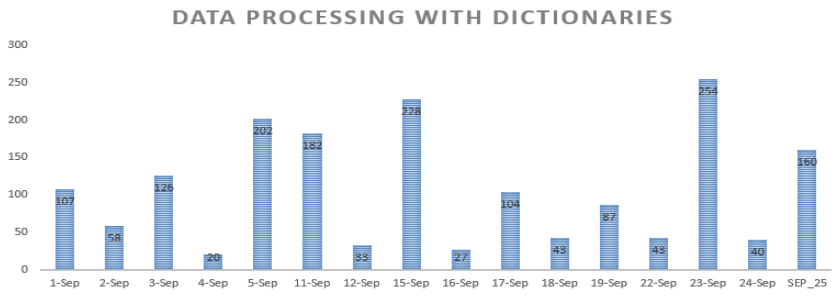


Figure 10. Data Preprocessing Using Dictionaries Individually Built by Our Group Members

5. Conclusion and Future Work

Data mining techniques can effectively apply for extracting association rules from unstructured Korean text. Corpus plays a vital rule for association rules mining as the association rules vary from corpus to corpus and are affected by the number of words, sentences, paragraphs and even text files. Proposed methods worked just fine for Korean text news documents. To extract more accurate rules, it is necessary that corpus is significantly large and contains logically related data. Another important consideration for Korean language is processing of unification and normalization with huge and accurate amount of data dictionaries and Unicode processing *i.e*. file pre-processing, conversion from text to transactional database and implementation are needed to be set according to Unicode coding scheme that varies among programming languages and database management system. Korean language requires research work on association rules not only among letters but on a broader spectrum *i.e*. among words, sentences and grammar. Korea text is needed to be digitized and more Korean databases are required for this purpose. So Korean scholars and users will have more automated tools such as grammar rules extractor, thesaurus and efficient web search in future. We have presented the method of text mining tasks to extract the information from collections of unstructured text data and the result showed great satisfaction. Further research on text mining will be carried out to explore for better and accurate association between words from unstructured text in large collections of data.

References

[1] S. Kotsiantis and D. Kanellopoulus, "Association rules mining: A recent overview", International Transactions on Computer Science and Engineering Journal, vol. 32, no. 1, **(2006)**, pp. 71-82.
[2] Y. .Koh and N. Rountree, "Rare association rule mining and knowledge discovery", Information Science Reference, **(2009)**.
[3] D. Jurafsky and J. H. Martin, "Speech and Language Processing", **(2008)**.
[4] M. Konchady, "Text Mining Application Programming", **(2006)**.
[5] "The Unicode Standard Version", www.unicode.org/versions/Unicode7.0.0/ch18.pdf.
[6] "Korean Grammar", http://en.wikipedia.org/wiki/Korean_grammar
[7] A. Savasere, E. Omiecinsky and S. Navathe, "An Efficient Algorithm for Mining Association Rules in Large Databases", Proceedings of 21st International Conference on Very Large Databases. Zurich, Switzerland, **(1995)**, pp. 432-444.
[8] C. Tang and C. Liu, "Method of Chinese Grammar Rules Automatically Access Based on Association Rules", Proc. Computer Science and Computational Technology (ISCSCT 2008), vol. 1, **(2008)**, pp. 265-268.
[9] Y. Sun and K. Jia, "Research of Word Sense Disambiguation Based on Mining Association Rules", Intelligent Information Technology Application Workshops, **(2009)**, pp. 86-88.
[10] D. W. Choi and Y. J. Hyun, "Transitive Association Rule Discovery by Considering Strategic Importance Computer and Information Technology (CIT)", Proceedings 2010 IEEE 10th International Conference, **(2010)**, pp. 1654-1659.
[11] I. N. M. Shaharanee, F. Hadzic and T. S. Dillon, "Interestingness measures for association rules based on statistical validity", Knowledge-Based Systems, vol. 24, no. 3, **(2011)**, pp. 386–392.
[12] Y. Xu, Y. Li and G. Shaw, "Reliable representations for association rules", Data & Knowledge Engineering, vol. 70, no. 6, **(2011)**, pp. 555–575.
[13] A. K. Ingason, S. Helgadottir, H. Loftsson and E. Rognvaldsson, "A Mixed Method Lemmatization Algorithm Using a Hierarchy of Linguistic Identities (HOLl)", in: Adv. Nat. Lang. Process, Springer, **(2008)**, pp. 205-216.
[14] Agrawal R., Imielinski T. and Swami A. N., "Mining association rules between sets of items in large databases", In Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, P. Buneman and S. Jajodia, Eds. Washington, D.C., **(1993)**, pp. 207-216.
[15] J. D. Holt and S. M. Chung, "Parallel Mining of Association Rules from Text Databases on a Cluster of Workstations", in Proceedings of the 2004 18th international Parallel and Distributes Processing Symposium, **(2004)**.

Authors

Irfan Ajmal Khan, received his BSc degree in Pre Engineering (science) in 1998 from Government College of Faisalabad, Pakistan. In 2006 he received his M.S. Degree in Computer Engineering at Hoseo University where he finished his Ph.D. in Computer Engineering in 2013. His research interest includes Databases, Data Mining, Machine Learning, Algorithms, and Mobile Computing.

Jung Hyun Woo, received her B.S. degree in Applied physics from Inha University, Incheon, Korea, in 1990. She is currently pursuing her PhD degree in department of Computer Science at Incheon National University, where she finished her M.S degree in Computer Science (IT policy) in 2010. Her research interest includes Databases, Cryptography, Educational technology, Algorithms, and Mobile Computing.

Seo-Ji Hoon, received his Bachelor degree in 2008 at Seoul National University of Science and Technology from department of Safety Engineering. He finished his MS and Ph.D. at Incheon National University from Department of Computer Science and Engineering in 2010 and 2015 respectively. His research interest includes Data Mining, Database Management and Sensor Networking.

Jin Tak Choi, received his B.S. degree in Mathematics and his M.S. degree in Computer Science from Dongkuk University, Seoul, Korea, in 1977 and 1982, respectively. He received Ph.D. degree in Electronics from Kyunghee University, Seoul, Korea, in 1991.Since 1987, he has been a Faculty Member at the Department of Computer Science of Incheon National University. His research interests include cryptography, database systems, and mobile and distributed computing.