International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

Design and Implementation of Association Rules Based System for Evaluating WSD

Samit Kumar[1], Dr. S. Niranjan[2]

[1]Ph.D. Scholar, Computer Science & Engineering, Mewar University, Chittorgarh, Rajasthan, India

[2]Professor, PDM College of Engineering, Women, Bahadurgarh, Haryana, India

Abstract: ***In this paper we presentnew method that usedassociation rules mining techniques in the field of natural language processing. Word sense disambiguation is persistently a central and challenging problem as increasing usage of internet in daily life. Every user has some queries that have to be searched on the internet. Transactional Database is created after preprocessing the text files. Ambiguous words along with its context and Senses are stored in the SQL database to create transactional database. Apriori algorithm applied on this transactional database to mine the rules. Strong rules are generated using frequent itemsets. Sense of ambiguous word is deduced by using the strong association rules generated.***

Keywords: Association rules, Apriori, Data mining, WordNet, Fuzzy Rules

1. Introduction

Word Sense Disambiguation (WSD) is a dynamic area which is very useful in today's world. Many WSD algorithms are available in literature. Words can have more than one dissimilar meaning. Words may be polysemous, but in actual text there is very little real ambiguity - to a person. Lexical disambiguation in its broadest definition is nothing less than finding the meaning of every word in context, which appears to be a largely unconscious process in people. As a computational problem it is often described as "AI-complete", that is, a problem whose solution presupposes a solution to complete natural-language understanding or common-sense reasoning.

In the field of computational linguistics, the problem is generally called word sense disambiguation (WSD), and is defined as the process of identifying which sense of a meaning is used in any given sentence, when the word has a number of distinct senses [1]. WSD is essentially a task of classification: word senses are the classes, the context provides the evidence, and each occurrence of a word is assigned to one or more of its possible classes based on the evidence. Words are assumed to have a finite and discrete set of senses from a dictionary, a lexical knowledge base, or ontology. Application-specific inventories can also be used. For instance, in a machine translation (MT) setting, one can treat word translations as word senses, an approach that is becoming increasingly feasible because of the availability of large multilingual parallel corpora that can serve as training data [15].

There are two main types of approach for WSD in natural language processing called as deep approachesand shallow approaches.

Deep approaches: These approaches involve the intention to understand and create meaning from whatis being learned, Interact vigorously with the content, make use of evidence, inquiry and evaluation, Take abroad view and relate ideas to one another and Relate concepts to every time experience. These approachesare not very successful in practice, mainly because such a body of knowledge does not exist in a computerreadableformat, outside of very limited domains. There is a long tradition in computational linguistics, of trying such approaches in terms of coded knowledge and in some cases; it is hard to say clearlywhether the knowledge involved is linguistic or world knowledge. The first attempt was that by MargaretMaster-man, at the Cambridge Language Research Unit in England, in the 1950s and Yarowsky's machine learning optimization of a thesaurus method in the 1990s.

Shallow approaches: These approaches are not concerned of learning the text instead they deal with thesurrounding words of the ambiguous word and try to identify only parts of interest for a particularapplication. They just consider the surrounding words, using a training corpus of words tagged with their wordsenses the rules can be automatically derived by the computer[14]. This approach, whiletheoretically not as powerful as deep approaches, gives superior results in practice, due to the computer'slimited word knowledge.

Association rules mining is an effective data mining technique to extract interesting patterns from transactional databases. This technique is usually used for market basket analysis i.e. to find out that which items are purchased together, so that management will be able to make effective decisions. Association rule mining can also be used for miningassociation rules from textual data with few changes.Association rules mining for textual data can use to createstatistical thesaurus, to extract grammatical rules and tosearch large online data efficiently.

2. Related Work

Research has progressed steadily to the point where WSD systems achieve sufficiently high levels of accuracy on a variety of word types and ambiguities. A rich variety of techniques have been researched, from dictionary-based methods that use the knowledge encoded in lexical resources, to supervised machine learning methods in which a classifier is trained for each distinct word on a corpus of manually sense-annotated examples, to completely

unsupervised methods that cluster occurrences of words, thereby inducing word senses. Among these, supervised learning approaches have been the most successful algorithms to date.

Knowledge-based: In this category, disambiguation is carried out by using information from an explicit lexicon or knowledge base. The lexicon may be a machine readable dictionary, thesaurus or hand-crafted. [2, 3] use WordNet as the knowledge base to disambiguate word senses, and [5] uses Roget's International Thesaurus.

Corpus-based: This category of approaches attempt to disambiguate words by using information gained from raining on some corpus, rather than taking it directly from an explicit knowledge source [4]. Training can be carried out either on a disambiguated corpus or a raw corpus. In a disambiguated corpus, the semantics of each polysemous lexical item has been marked, while in a raw corpus, the semantics has not been marked yet.

Hybrid Approaches: A good example is Luk's system [7] which uses the textual definitions of senses from a machine readable dictionary to identify relations between senses. It then uses a corpus to calculate mutual information scores between the related senses in order to discover the most useful information. In this way, theamount of text needed in the training corpus is reduced.

In [8], John and Soon have given the idea of Parallel Multipass Inverted Hashing and Pruning for text databases. Proposed algorithm used hash tables to avoid several passes on the database during the mining process. This algorithm adopted the pruning strategy to cut off the infrequent itemsets on the occurrence of items in transactions that were stored in hash tables. It divided the database among various partitions to improve the efficiency of algorithm as compare to Apriori and Count Distribution algorithms. Greater efficiency was achieved by using this technique.

In [9], Yong-le Sun and Ke-liangJia utilized the Association rules mining based upon Apriori algorithm to solve the Word Sense Disambiguation problem. Apriori successfully extracted association rules between the sense of the ambiguous words and contexts and generated very precise association rules.

In [10], Chao Tang and Chen Liu utilized Apriori algorithm to find out grammatical rules from Chinese text. A new model was proposed for grammatical rules mining which had three major steps: pre-processing, association rules mining and verification of association rules to get real rules. Four different corpuses had been chosen for testing purpose and results have indicated the interesting fact about the length of sentences and effectiveness of Apriori algorithm i.e. For small sentences the algorithm worked well as compare to large sentences because the large sentences have contained combined smaller rules.

In [11], Zhou targeted the engineering documents for association rules. The documents mining procedures distributed in two sub-processes: one was document structure generation and other was document content generation. Apriori algorithm was used for mining interesting patterns in engineering documents. This algorithm filtered out structure-structure association rules, structure-item association rules and item-item association rules.

In [12], Wu Gongxing devised a new distributed algorithm to extract association rules for the XML data. This algorithm created the DOM tree at the beginning and then had used this tree to extract association rules. The distributed algorithm had worked on multiple web sites. Each website had executed the FreqTree algorithm to compute local support count and sent it to global website. The global site then determined the global frequent items (FI) on the basis of sum of support counts which were gathered from all local sites. At the end, a verification process had applied to filter out the valid XML rules from the global frequent items.

In [13], Al-Zoghby, A., Eldin, A.S., Ismail, N.A. and Hamza, T. introduced a new system based upon Apriori and CHARM algorithm to determine soft-matching association rules for Arabic language. Frequent Closed Itemsets were tested along with Frequent Itemsets. Proposed system has converted the Arabic corpora in transactional databases, then performed cleaning and morphological analysis on the database and finally used Apriori and CHARM algorithms to find out association rules mining. Results has shown that Frequent Closed Itemsets worked well as compare to Frequent Itemsets because Frequent Closed Itemsets reduced redundancy up to a significant level which was present in Frequent Itemsets.

3. WSDBased on Association Rules

Association rule learning is a popular and well researched method for discovering interesting relations between variables in large databases. It is intended to identify strong rules discovered in databases using different measures of interestingness. Association rule mining is user-centric because its objective is the elicitation of interesting rules from which knowledge can be derived. Interesting ness of rules means that they are novel, externally significant, unexpected, nontrivial, and actionable. An association mining system aids the process in order to facilitate the process, filter and present the rules for further interpretation by the user.

We state the problem of mining association rules as follows: $I = \{i1, i2, \ldots, im\}$ is a set of items, $T = \{t1, t2, \ldots, tn\}$ is a set of transactions, each of which contains items of the itemset I . Thus, each transaction ti is a set of items such that $ti \subseteq I$. An association rule is an implication of the form: $X \rightarrow Y$, where $X \subset I$, $Y \subset I$ and $X \cap Y = \theta$. X (or Y) is a set of items, called itemset.
An example for a simple association rule would be {bread }→{butter} . This rule says that if bread was in a transaction, butter was in most cases in that transaction too. In other words, people who buy bread often buy butter as well. Such a rule is based on observations of the customer behavior and is a result from the data stored in transaction databases.

Looking at an association rule of the form X=>Y, X would be called the antecedent, Y the consequent. It is obvious that the value of the antecedent implies the value of the consequent. The antecedent, also called the "left hand side" of a rule, can consist either of a single item or of a whole set of items. This applies for the consequent, also called the "right hand side", as well.

The most complex task of the whole association rule mining process is the generation of frequent itemsets. Many different combinations of items have to be explored which can be a very computation-intensive task, especially in large databases. As most of the business databases are very large, the need for efficient algorithms that can extract itemsets in a reasonable amount of time is high. Often, a compromise has to be made between discovering all itemsets and computation time. Generally, only those itemsets that fulfill a certain support requirement are taken into consideration. Support and confidence are the two most important quality measures for evaluating the interestingness of a rule.
. The support of a rule is represented by the formula

$$\text{supp}[X \rightarrow Y] = \left[\frac{|X \cap Y|}{n}\right]$$

where$|X \cap Y|$ is the number of transactions that contain all the items of the rule and n is the total number of transactions.

Confidence: The confidence of a rule describes the percentage of transactions containing X which also contain Y.

$$\text{conf}[X \rightarrow Y] = \left[\frac{|X \cap Y|}{|X|}\right]$$

We look for rules that exceed pre-defined support (minimum support) and have high confidence.

3.1 Fuzzy Association Rules

Kuok et al. describe fuzzy association rules as follows: "Mining fuzzy association rule is the discovery of association rules using fuzzy set concepts such that the quantitative attribute can be handled" As in classical association rules, $I=\{i_1,i_2, \ldots ,i_m\}$ represents all the attributes appearing in the trans action database $T=\{t_1,t_2,\ldots,t_n\}$ .The set*I*contains all the possible items of a database, different combinations of those items are called itemsets. Each attribute $i_k$will associate with several fuzzy sets.
As an example, the attribute salary could look as follows: $F_{Salary}=\{high,medium,low\}$. Fuzzy sets and their corresponding membership functions have to be defined by domain experts. Each of the fuzzy sets can be viewed as a [0,1] valued attribute, called fuzzy attribute.

A fuzzy association rule has the following form:

If X is AthenY isB

In this case, X ={x1, x2, ... , x p} and Y ={y1, y2,... , yq} are itemsets which are subsets of I . It is important to notice that those two sets must be disjoint and thus do not have any attributes in common. A={ f_{x1}, f_{x2},... , f_{xp}} and B={f_{y1}, f_{y2},... , f_{yq}} contain the fuzzy sets that are associated with X and Y . Known from classical association rules, X is A is the antecedent; Y is B is the consequent. If a sufficient amount of records approves this rule, we will call it satisfied.

3.2 Basic Idea of the WSD Algorithm Based on Mining Association Rules

The context of an ambiguous word is regarded as a transaction record, the words in the context and the senses of the ambiguous word areregarded as items. Given an ambiguous word, Y presents the set of the senses of the word, X means the set of its context words. The item sets I includes all senses of the ambiguous word and all words of its context. The database D is the context document sets of the word, each record T is a context of the word. The association rule X=>Y means that the sense of the word can be determined by its context.

The basic idea of the WSD algorithm based on mining association rules is: to discover the frequent item sets composed of the sense of the ambiguous word and its context by scanning its context database, which support degree is no less than the threshold of support degree; to produce the association rules X=>Y which confidence degree is no less than the threshold of the confidence degree from maximum frequent item sets; at last to determine the sense of the ambiguous word by choosing the sense which the most association rules deduced.In the producing process of the frequent item sets, there are too many words in the item sets which leads to produce too many N-frequent item sets. In order to avoid to loss these frequent item sets, the paper gives different min-support degree according to different N-frequent item sets. N is larger, the min-support degree is smaller. N is smaller, the degree is larger. The association rule X=>Y , which confidence degree is no less than the threshold of the confidence degree, is produced.

3.3 Algorithm

Given an ambiguous word, firstly choose the association rules from the association rules database according to its context words, the selected rules may be more than one; they are all used in WSD. At last the sense of the ambiguous word is determined by the rules committee voting, the sense which is decided by the most rules is selected as the sense of the ambiguous word in its context. Association rule generation is usually split up into two separate steps:

1. First, minimum support is applied to find all frequent itemsets in a database.
2. Second, these frequent itemsets and the minimum confidenceconstraints are used to form rules.

Finding all frequent itemsets in a database is difficult since it involves searching all possible itemsets(item combinations). The set of possible itemsets is the power set over I and has size 2^n-1 (excluding the empty set which is not a valid itemset). Although the size of the powerset grows exponentially in the number of items n in *I*, efficient search is possible using the downward-closure property of support which guarantees that for a frequent itemset, all its subsets are also frequent and thus for an infrequent itemset, all its supersets must also be infrequent.

3.4 Apriori Algorithm

procedureApriori (T, minSupport) { //T is the database and inSupport is the minimum support
L1= {frequent items};
for (k= 2; L_{k-1} !=θ; k++) {
C_k= candidates generated from L_{k-1}
//that iscartesian product L_{k-1} x L_{k-1} and
eliminating any k-1 size itemset that is not
//frequent
for each transaction t in database do{
#increment the count of all candidates ink that are contained in t
L_k = candidates in C_k with minSupport
}//end for each
}//end for
return$U_k L_k$;
}

Given an ambiguous word, choose the association rules according to its context words.At last the sense of the ambiguous word is determined by the rules committee voting. The sense, which is decided by the most rules, is selected as the sense of the ambiguous word in its context.Apriori uses a "bottom up" approach, where frequent subsets are extended one item at a time (a step known as candidate generation), and groups of candidates are tested against the data. The algorithm terminates when no further successful extensions are found. Apriori uses breadth-first search and a tree structure to count candidate item sets efficiently. It generates candidate item sets of length k from item sets of length k − 1. Then it prunes the candidates which have an infrequent sub pattern. According to the downward closure lemma, the candidate set contains all frequent k-length item sets. After that, it scans the transaction database to determine frequent item sets among the candidates.

4. Experimental Setup

Problem is to gain an understanding of the senses of the word with its relative position in the entence.The mining model consists three major steps: 1. Pre-processing—splits the text into tokens(tokenization), POS tagging, chunking and parsing, 2.Creating transactional database from the preprocessed files and finallyapplying Apriori algorithm to get association rules on transactional database.

Pre-processing is a very important step because numeric data and punctuation marks increase the number of 1-itemsets that results in more higher order invalid frequent itemsets and mining of these frequent itemsets is awastage of precious resources.
Conversion of text files to transactional database is a fundamental requirement of Apriori algorithm.

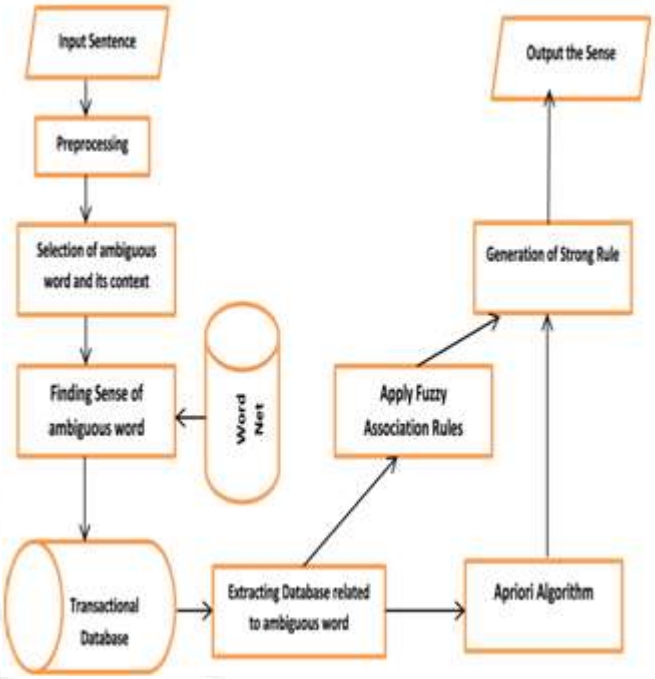


Figure 1: Steps of Experimental Setup

When only ambiguous word is selected and search for the meaning in the WordNet database, then its returns all the possible senses of ambiguous word. When we select it context words then it returns the meaning of the word that is related to its context. In above example if search bank in WordNet database then records are given above.

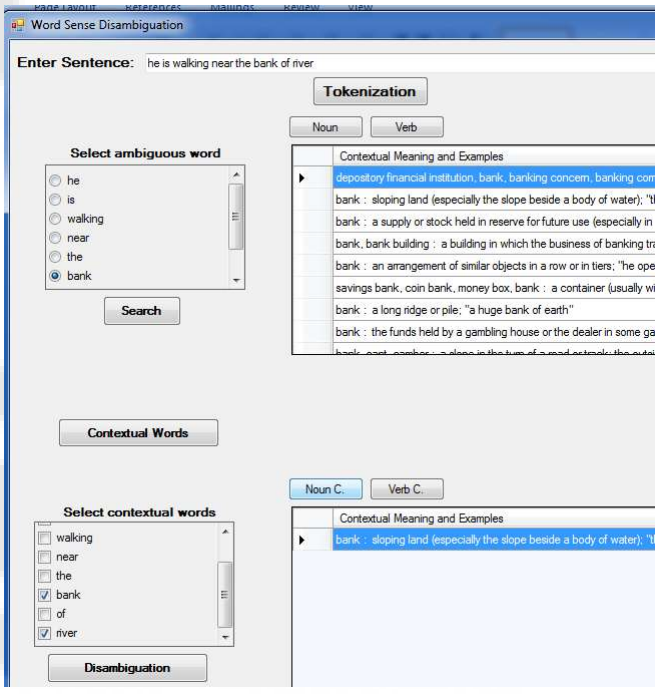


Figure 2: Finding the sense of ambiguous word using WordNet

This way, a database is created for the mining the association rules. A transactional file that contains the ambiguous word, its context and possible meaning related to its context.

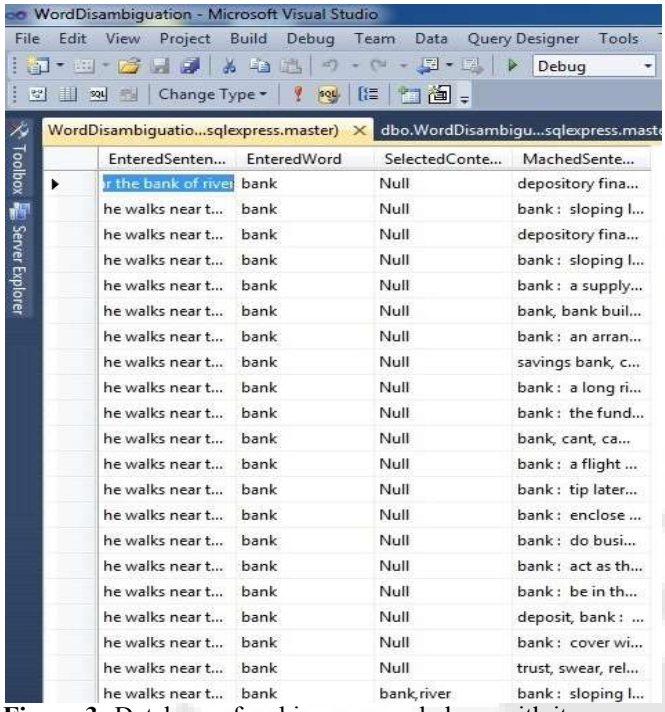
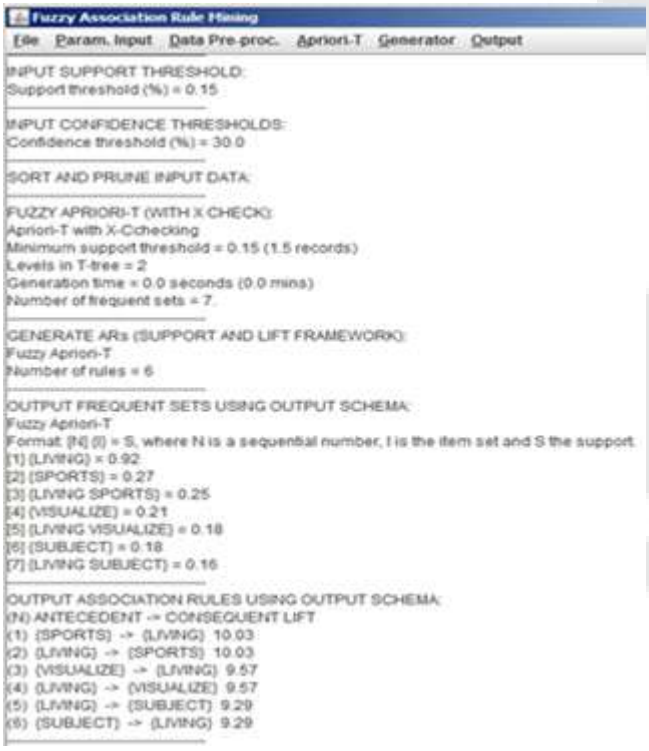


Figure 3: Database of ambiguous word along with its senses and context word

Now we applied, the association rules using apriori algorithm and fuzzy association rules on the database created above. First, it generates the 1item frequent sets. Then using more itemset are generated using previous generated candidate sets. In the last, strong association rules are generated using the generated frequent sets. This rule will deduce the sense of ambiguous word.



5. Results and Discussion

We have taken the followings performance measures to evaluate our work.

Precision : p is the number of correct results divided by the number of all returned results

$$\text{Precision} = \frac{\text{number of true positive}}{\text{number of true positive} + \text{false positive}}$$

Recall: r is the number of correct results divided by the number of results that should have been returned

$$\text{Recall} = \frac{\text{number of true positive}}{\text{total expected results}}$$

The F score (also F-measure) is a measure of a test's accuracy. It considers both the precision p and the recall r of the test to compute the score :

$$F = 2.\frac{\text{Precision.Recall}}{\text{Precision} + \text{Recall}}$$

Accuracy is also used as a statistical measure of how well a binary classification test correctly identifies or excludes a condition.

We applied these two algorithms first on data without sampling. Then we apply the sampling on the data file and then association rules algorithms. Following Table shows the results of our experiments.

Table 1: Results without Sampling

Parameter	Apriori Algorithm	Fuzzy Association Rules
Precision	0.84	0.88
Recall	0.81	0.86
Accuracy	82.4%	86.4%

Table 2: Results after Sampling

Parameter	*Apriori Algorithm*	*Fuzzy Association Rules*
Precision	0.86	0.90
Recall	0.83	0.882
Accuracy	83%	87.6%

This is experimented that association rules techniques can also be used in the field of Natural Language Processing. The outcomes existing in above tables determine that high accuracy can be achieved using data mining techniques. The algorithm fuzzy association rules have high performance than traditional Apriori Algorithm. It can be predicted that above than 80% association rules shall deduce the correct sense of ambiguous word.

6. Conclusion and Future Scope

In this paper, we have implemented a database for WSD that can be used for mining the strong decision rules using association rules. Fuzzy association rules have greater performance than the crisp association rules.Ambiguous word, its context and its possible senses are stored in a database. This database is presented to the algorithms for mining the decision rules. The sense which is inferred by most association rules will be the exact sense of that ambiguous word. This work can be further extended to neurofuzzy rules. Automatic creation of database is main challenge of this research. One can further work to create the transactional database automatically. More advance algorithm like genetic algorithm can be also applied to find out the sense.

References

[1] Carpuat, M. and D. Wu,. Word sensedisambiguation vs. statistical machine translation, Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, (ACL '05), Association for Computational Linguistics Stroudsburg, PA, USA, pp: 387-394. DOI: 10.3115/1219840.1219888,2005
[2] Voorhees, E.: Using WordNet to Disambiguate Word Senses for Text Retrieval. SIGIR (1993) 171-180
[3] Agirre, E., Rigau, G.: Word sense disambiguation using conceptual density. Proc. Of COLING (1996)
[4] Richardson, R., Smeaton, A.: Using wordnet in a knowledge-based approach to information retrieval. Proc. of the BCS-IRSG Colloquium, Crewe (1995)
[5] Yarowsky, D.: Word-sense disambiguation using statistical models of Roget's categories trained on large corpora. Proc. of the 14th International Conference on Computational Linguistics (COLING-92), Nantes, France (1992) 454-460
[6] Luk, A.: Statistical sense disambiguation with relatively small corpora using dictionary definitions. Proc. of the 33rd Meetings of the Association for Computational Linguistics (ACL-95), Cambridge, M.A. (1995) 181-188
[7] Holt, J.D., and Chung, S.M. Parallel Mining of Association Rules from Text Databases on a Cluster ofWorkstations. Proceedings of the 2004 18th international Parallel and Distributes Processing Symposium, (DigitalObject Identifier: 10.1109/IPDPS.2004.1303027).
[8] Yong-le Sun and Ke-liangJia. Research of Word Sense Disambiguation Based on Mining Association Rules.Intelligent Information Technology Application Workshops, 2009. Third International Symposium on (DigitalObject Identifier: 10.1109/IITAW.2009.85, Publication Year: 2009), pp. 86-88.
[9] Chao Tang and Chen Liu. Method of Chinese Grammar Rules Automatically Access Based on Association Rules.Computer Science and Computational Technology, 2008 (ISCSCT '08) International Symposium on Volume:1 (Digital Object Identifier: 10.1109/ISCSCT.2008.68) Publication Year: 2008, pp. 265 – 268.
[10] Zhou, J. Discovering Association Rules in Engineering Documents. Natural Language Processing and KnowledgeEngineering, 2003.Proceedings. 2003 International Conference on (Digital Object Identifier:10.1109/NLPKE.2003.1275927, Publication Year: 2003), pp. 339 – 344.
[11] Wu Gongxing. A Study on the Mining Algorithm of Fast Association Rules for the XML Data.Computer Scienceand Information Technology, 2008. (Digital Object Identifier: 10.1109/ICCSIT.2008.89 Publication Year: 2008),pp. 204 – 207.
[12] Al-Zoghby, A., Eldin, A.S., Ismail, N.A. and Hamza, T. Mining Arabic Text Using Soft-Matching AssociationRules. Computer Engineering & Systems, 2007.(ICCES '07, Digital Object Identifier:10.1109/ICCES.2007.4447080, Publication Year: 2007) 2007, pp. 421– 426.
[13] M.H.Margahny and A.A.Mitwaly "Fast Algorithm for Mining Association Rules" AIML 05 Conference, 19-21 December 2005, CICC, Cairo, Egypt.
[14] Zhang Zheng, Zhu Shu "A New Approach to Word Sense Disambiguation in MT System" World Congress on Computer Science and Information Engineering, 2009, Los Angeles, CA, P 407 – 411.

Author Profile

Samit Kumar, is a Ph.D. research scholar in Mewar University, Chittorgarh. He completed his BE (CSE) in 2003, and M.Tech (CSE) in 2007 and M. Phil (Computer Science) in 2010. He has 10.5 years of experience in total. He has presented three papers in National Conference and seven papers in International conferences sponsored by IEEE.He has also published six papers in reputed journals.

Prof. S. Niranjan, did his M. Tech (Computer Engineering) from IIT Kharagpur in 1987. Completed Ph.D. (CSE) in 2004andPh.D.(I&CT) in 2007. He has in total 28 years of teachingexperience. Presently working as Professor at PDM College of Engg for Women, Bahadurgarh(Haryana). He has published more than 20 research papers in reputed national and international journals. He also published 30papers in sponsored conferences. He guided more than 10 students for their Ph.D.