A Novel Approach and Comparative Study of Association Rule Algorithms in Validation of Semantics of Sentences

Yamuna Devi. N
Assistant Professor(Senior Grade)
Department of MCA
Coimbatore Institute of Technology
Coimbatore, India

Devi Shree J, PhD.
Assistant Professor(Senior Grade)
Department of EEE
Coimbatore Institute of Technology
Coimbatore, India

ABSTRACT
Efficient Human Computer Interaction (HCI) is an absolute necessary for many applications these days. Computational Linguistics supports HCI to make computers to understand human languages. Advanced Computational models can be built using many technologies to provide easy communication between human and computers. Data mining has emerged to address problems of understanding ever-growing volumes of information for structured data. Data mining is a process to extract hidden knowledge from huge amount of data which can be used to build computational model. The usage of Association Rules (AR), one of the data mining techniques, to build an effective communication between human and computers is elucidated in this paper. The comparative performance of two different Association rule algorithms is illuminated in building a model to legalize semantics of sentences in linguistics domain. The sequence of operations to build the model is explored with necessary constraints at each stage. The model is to verify the meaning of English sentences which are syntactically correct using Apriori and Frequent-pattern tree growth algorithm in a limited domain. As a prerequisite, syntax verification of the sentence is also done and as a follow up, it also provides an interface which can be used for interaction between human and computer. The association rules, a data mining concept is employed in semantic analysis in a distinct way. Since the natural language understanding is an endless process, this work opens the door for the usage of association rules in semantic analysis of natural language sentences in a defined domain.

General Terms
Association Rules, Human Computer Interaction

Keywords
Syntax Analysis, Semantic analysis, Apriori algorithm, Question Answering System.

1. INTRODUCTION
Data mining is one of fervent field in which research is handled for various application domains. A large amount of data can be the input for data mining task to extract knowledge from it. Linguistics is a domain with vast data which is the study of natural languages that people use for communication. Computational linguistics is related to linguistics and computer science in building computational models of linguistic theories. Building computational models for linguistic analysis is a useful and necessary mission for human-machine communication. It can be achieved to a greater extent by analyzing the syntax and semantic of the sentences pertaining to the natural language. As a new approach, data mining techniques are applied in natural language analysis to find meaning of a sentence as knowledge. There are various disciplines in natural languages, like phonetics, syntax, semantic, pragmatic, morphology, utterance etc [1] [2]. Among these disciplines syntax and semantic analysis are used in a range of applications like machine learning, word sense disambiguation, voice recognition systems and information retrieval etc. The natural languages are also analyzed in computational aspects via Natural Language Processing (NLP), Natural Language Understanding (NLU), etc.

2. PROBLEM DESCRIPTION
The syntax analysis in NLP defines the process of analyzing the structure of a sentence. It demonstrates that how the words are related to each other in a sentence [3]. The semantic analysis in NLU defines the process of capturing and understanding the meaning of a sentence in a context. It needs focus today as it helps the people to interact with computers through natural languages. Example, the information about the trains, train times, etc can be obtained by posting a natural language query through an interface [4]. Since, data mining techniques capable of handling huge amount of data, Apriori and FP tree growth algorithms of Association Rules, are applied in verifying the meaning of an English sentence. The performance of both algorithms is compared.

One of the popular applications of the semantic analysis is Question Answering System (QAS) [5]. Generally the queries are posted in predefined formats or through menus. It will be easier and useful if the queries are entered as natural language sentences. A natural language sentence which is meaningful can be converted to a formatted SQL query which can be executed to retrieve the information from a database. Thus by providing natural language interface the human-machine interaction is improved with non-computer people.

Though data mining has touched greater heights of application domains, it is an endless process to search newer heights in different domains. In this paper, the focus is given to apply Apriori and FP-tree growth algorithms in verifying the meaning of a sentence. As the meaning can be verified for syntactically valid sentences, the syntax analysis is also carried out. By applying the algorithms, the association rules as valid combinations of constituents are generated for verification of meaning and stored in semantic database for future use. The semantically valid sentences are considered for formal query generation which is executed to produce results for end users [5]. An interface is used to post a query in natural language.

3. SYNTAX ANALYSIS

Syntax shows the role of words in a sentence and computing the structure of the sentence like how words are related in a sentence. Given a sentence, the system assigns to it a syntactic structure approved by the English grammar rules. A domain specific lexicon is constructed that shows which terminal symbol a word in the language belongs to. The lexicon can be considered as language dictionary.

In general, to each syntactic rule combining some sequence of child constituents into a parent constituent, there will be some corresponding semantic operation combining the meaning of the children to produce the meaning of the parent. The Context-Free-Grammar (CFG) rules are developed which defines the structure of an English sentence [6]. A small portion of a CFG rules used is shown in figure 1. These rules are also called as Re-write rules. Various strategies are available for finding the structure of an English sentence. This system uses Bottom-up, Breath-first and Left-right strategy to parse the sentence. Parsing process proves the syntactic structure of any sentence. As a result of parsing process, a parse tree can be generated [7].

<Decl. sentence>→ <subject><predicate>
<subject> → <simple subject>| <compound subject>
<simple subject>→ <noun phrase> | <nominative personal pronoun>
<noun phrase> → <proper noun> | <art>[<adverb>* <adj>]<noun> …
<predicate> → <verb>|<verb phrase> [<complement>
<verb phrase> → <aux> <verb> | <link verb> [<vpastp>|<ving>] |….
<complement> → <subject> | [<adverb>* <adjective>]|..

Figure 1: Grammar to Construct the Parse Tree.

A parser is constructed using shift-reduce technique. The parser expects just one sentence. The parser will act as a grammar checker, simply rejecting sentences that it considers ungrammatical. The parser will produce a parse tree for grammatically valid sentences. The parser uses the grammar and lexicon to find the structure in a language which is syntactically correct. The parser produces the constituents of the given sentence as noun, verb, adjective, adverb etc which can be used by semantic analyzer. As an example, the sample sentence and its parse tree is given in figure 2.

4. SEMANTIC ANALYSIS

Semantic analysis is the process to devise the meaning of a sentence based on the domain knowledge. The meaning can be inferred using the way in which the words combine at the sentence level [8]. For an example, consider a sentence, "The boy went to the park with a dog" which is syntactically and semantically valid. But the sentence "The park went to the dog with a boy" which is syntactically valid but semantically invalid. The meaning of a sentence can be analyzed based on the context in which the words are used in the sentence. Natural languages such as English, German and French are having many ambiguities in words usage based on the context in which they are used [8][9]. Example, the author, F.R. Palmer explained the different usage of the word "bank" with different meanings as 'sloping ground on each side of river', 'financial institution', 'deposit money at a bank" [2]. So, it is necessary to analyze and stay tuned where to use a word with appropriate meaning. The semantic analyzer considers syntactically valid sentences to check whether it is meaningful sentence. The semantic analyzer is constructed using Apriori algorithm and FP-tree growth algorithm of Association rules [10]. A semantic database is created and used to store the sentence fragments which can be used by the semantic analyzer.

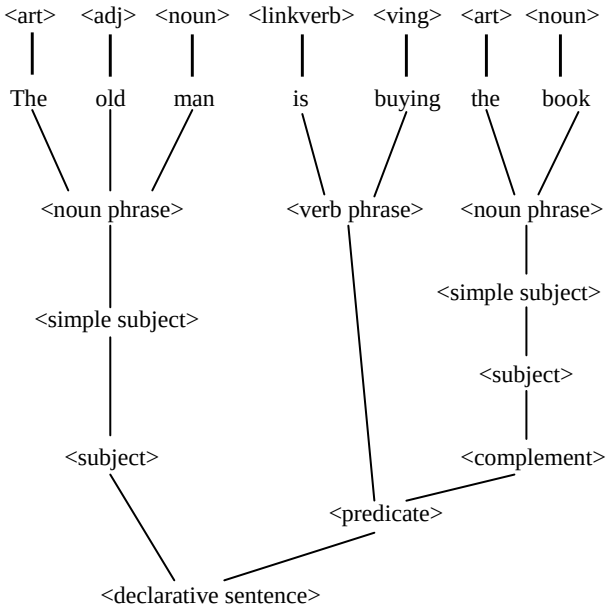


Figure 2: Parse Tree

4.1 Association Rules in Semantic Analysis

The Association Rule, which is a data mining concept, is used for defining the association between any two given data sets [11]. The two different techniques of Association Rules namely Apriori and Frequent-Pattern Tree Growth are used in semantic validation of sentences in this work, the performance is compared.

4.1.1Association Rules

For a given data set D, an association rule is an expression of the form X=>Y, where X and Y are subsets of D and X=>Y holds with confidence T, if t% of data set D that support X also support Y. The rule X=>Y has support S in the data set D, if S% of data set in D supports X ∪ Y [10]. Here, the verbs and nouns are considered as two different data subsets of lexical categories set (D). Let the verbs subset is denoted by X and the nouns subset is denoted by Y. Here, the support S is assigned as logical value and is calculated using a predefined table. The confidence is calculated as, Confidence (X => Y) = P(Y / X) = Support(X U Y) / support(X). Since the support values are logical, confidence value is either 100 % (for valid associations) or 0% (for invalid associations).

4.2 Semantic Database

The semantic database is constructed by storing the fragments of sentences as rules which give meaningful associations of nouns (as subject or object) and verbs. These meaningfully valid association rules are automatically generated with the help of Association Rules algorithm and stored in semantic database for every new verb appended to the syntactic lexicon. These rules are used for verification of

meaning of the input sentence.

4.3 The Apriori Algorithm

The Apriori algorithm of association rules can be used to validate the meaning of the sentences. Apriori property [10] is used to generate the rules of each and every subsequent level. Apriori property implies that, "all nonempty subsets of a valid set must also be valid". This property helps to eliminate the useless rules at each level and in turn save memory space used to store the semantic rules. The association rules are generated using Apriori algorithm in two steps. In the first step, the valid associations are generated. In second step, these associations are converted into rules and stored in the semantic database.

4.3.1. Generation of Valid Associations

The tokens, except the articles in a sentence are considered as set of items. The associations between these tokens according to the rules of the language are considered as valid associations. The language rules are defined and stored in a dictionary. This work deals with all types of sentences that consist of one verb. The verbs are also categorized as action verbs, regular verbs etc. The valid associations are generated in level wise. At each level, the complexity of sentences can be increased as more than one noun can occur with the verb. The combination of other tokens with the verb is validated for subject place and object place of a sentence using the dictionary. The voice patterns such as active voice and passive voice based on the verb is also analyzed for prediction of accurate meaning of the sentence. The validness of the combination is achieved with the help of support count which is a logical value. Considering as example, a verb 'write', which is an action verb can be combined with various categories of other tokens in a sentence. The various levels in which one or more tokens that are combined with the verb are shown in figure 3 at various levels.

L1 = {(write), Place-hold}
{(write), Human-being}
{(write), Instruments}
L2 = {(write), Place-hold, Instruments}
{(write), Place-hold, Human-being}
{(write), Instruments, Human-being}
L3 = {(write), Place-hold, Instruments, Human-being}

Figure 3: Valid Associations of Constituents

4.3.2 Generation of Association Rules and Semantic Validation

The strong associations are categorized as the rules that satisfy minimum confidence threshold value. The confidence can be defined as, Confidence(X=>Y) = P(Y/X) = Support (XUY) / Support(X). Here, strong association rules can be generated from the valid combinations of all levels. The sample strong rules are listed in figure 4 with confidence value of 100%. The strong rules are stored in semantic database. These rules are used to validate any given input sentence for its meaning. If the sentence does not satisfy the rules constraints, then the erroneous part of the sentence is identified.

V → Place-Hold
V → Human-being
V → Instruments
V → Place-Hold ^ Instruments
V → Place-hold ^ Human-being
V → Place-hold ^ Human-being ^ Instruments

where, V is the unique identity code for 'write'.

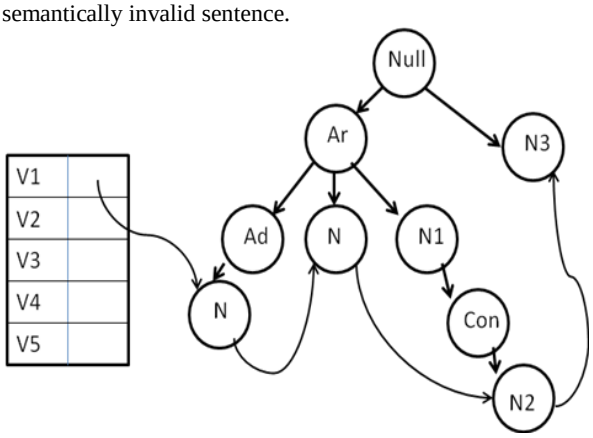
Figure 4 :Sample Association Rules

4.4 Frequent-Pattern Tree Growth Algorithm

The FP-tree growth algorithm is a technique to generate association rules without generating large set of candidate item sets which is a major drawback of Apriori algorithm. In addition, the number of scans of the database is reduced as two. This algorithm needs to create a frequent-pattern tree which is to be mined for association rules. The FP-tree is constructed using the semantic database as input which consists of valid combinations of verbs and nouns according to the language rules. The header table is constructed with verbs. The prefix nodes of the branches of the tree contain the permissible combination of nouns and other constituents for a verb which is in leaf node. The database is created which contains the entries of all possible combinations of all constituents [10].

- Let I = {I1, I2,…In} be the domain of literals called constituents, here nouns and verbs alone.
- A record called transaction, contains a set of constituents with permissible combinations as I1,I2, ……Ik ⊂ I.
- The input to the FP-growth algorithm is a set of transactions T.
- We call any set of constituents I1, I2, ……Im ⊂ I collectively as a constituents set or sentence.
- The constituent set has a measure called support and confidence.

The support count is assigned as logical value. If the logical value of the constituents and verb combination is 1, then a branch of a tree is constructed using those constituents and the verb. A sample FP tree is given in figure 5. The FP-tree growth algorithm adopts a divide-and-conquer strategy. The valid set of constituents for a verb is constructed and generated as a separate branch of the tree without candidate generation. Any input sentence can be traced for any of the branch of the tree based on the constituents in the sentence using pre-order traversal. If none of the branch follows the constituent sequence of the sentence, then it can be declared as semantically invalid sentence.



Where, V1,V2..V5 = verbs, Ar = Article, Ad = Adjective, N,N1,N2,N3=Nouns, Con=Conjugate

Figure 5: Sample FP Tree with Nouns and Verbs

4.5. Comparative Study

The two different algorithms of association rules namely, Apriori and FP-tree growth are used in the process of semantic validation of sentences. Each algorithm has its advantages and drawbacks of its own in this method. The FP tree growth algorithm generates frequent sets as valid constituents of a verb in short time compared to Apriori which takes more time to generate the same, since it generates candidate sets.. The Apriori algorithm does not require any additional memory space except to store semantic database. But FP-tree growth algorithm requires some extra memory space to store the Frequent Pattern tree which is to be mined. So the algorithm can be selected by making trade-off between time and memory according to the application in which they are applied.

5. QUERY GENERATION

As a simple application of semantic validation a Question Answering System (QAS) is developed with restricted domain [12]. Semantically valid sentences are converted to a SQL query. The conversion to valid SQL query is achieved with the help of some predefined templates. The keywords are extracted from the input sentences and are fed to the templates [13]. A query is formed for the sentence using a template which is suitable for the keywords in the input sentence. The generated query is executed by the database engine to retrieve the resulting records from the database. The templates for specific types of SQL queries such as Simple query, Queries with Boolean and special operators, Queries with aggregate functions and projection are created.

6. PERFORMANCE ANALYSIS

The algorithms are executed with same set of sample sentences. The sentences are selected such that it consists of constituents in more numbers. The execution time is calculated for two different algorithms to validate semantics of sentences and their performance is compared in Figure 6. The execution time, to validate the meaning is measured with the number of new nouns that are not available in the dictionary. The result is shown in Figure 7 and Figure 8.

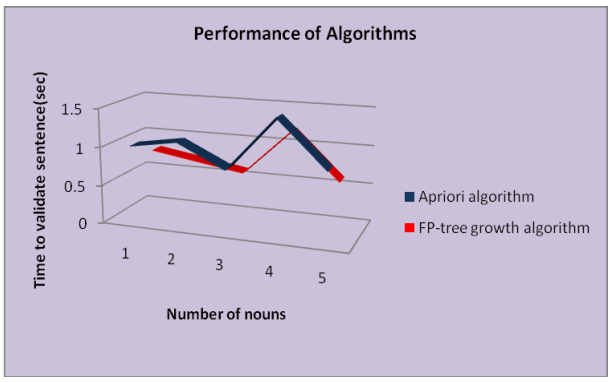


Figure 6: Performance of algorithms to validate sentences

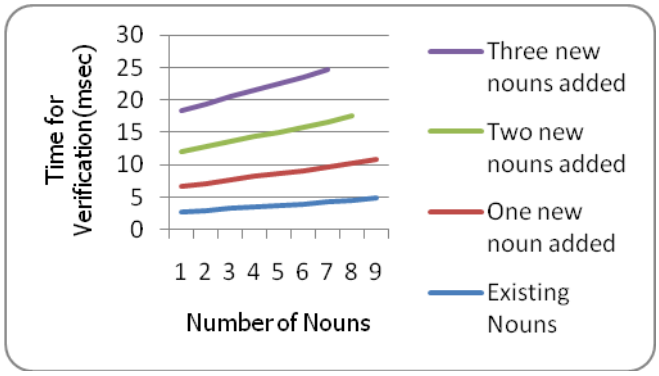


Figure 7: Performance of the system for new nouns + existing nouns

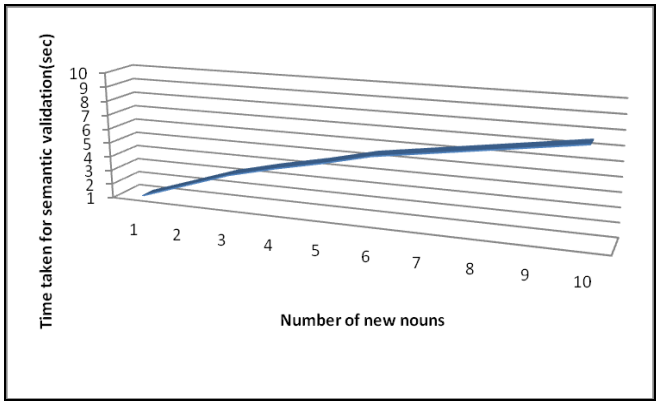


Figure 8: Performance of the system for complete set of new nouns

The standard increase in size of the semantic database according to the type of the new verb appended to the dictionary is analyzed. The performance is shown in Figure 9.

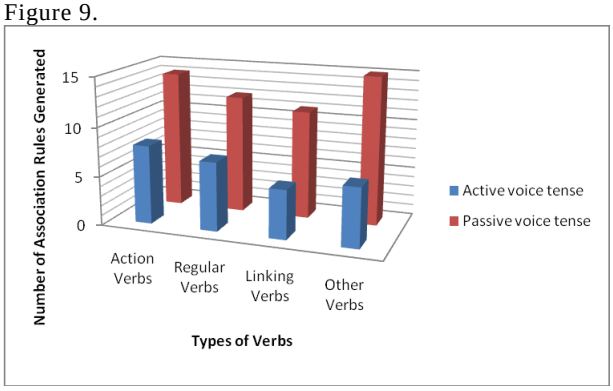


Figure 9: Performance of the system for new verbs

7. CONCLUSION

Though data mining techniques employ great participation in solving problems in various domains, the efficiency of data mining algorithms is proved through the application of the same in linguistics. The performance of two different algorithms of association rules in semantic validation of sentences is compared for efficiency. This system is developed and implemented for domain specific English language sentences. The capabilities of the system will increase with its usage. As future enhancements, by

removing the constraints on the structure and types of sentences, the system can be extended for more types of sentences in a more general way. Semantic analysis can be improved by eliminating the constraint on the specific domain and can be enhanced for open domain.

8. REFERENCES

[1] Hornby, A.S., The teaching of structural words and sentence patterns – stages three and four, The English Language Book Society and Oxford University Press.

[2] Palmer, F.R., Semantics, Cambridge University Press, Second Edition.

[3] Rutherford, W., "Principled Sentence Arrangement", Maxtes01 Journal, No.4, 43-48.

[4] Sunil Kopparapu, Akhilesh Srivastava and PVS Rao. 2006. Building a Natural Language Interface for a Railway Website. In proceedings of Second National Conference on Innovation in information and Communication Technology. 67-71.

[5] Paloma moreda, Hector liorens and Estela Saguete and Manuel Palomar, "Combining semantic information in question answering systems", Journal on Information Processing and Management, November 2011, Volume 47, Issue 6, 870-885.

[6] Alfred V. Aho and Jeffrey D. Ullman 1977. Principles of Compiler Design, Addison-Wesley Publishing Company.

[7] Jean-Paul Tremblay and Paul G. Sorenson. 1984. An Introduction to Data Structures with Applications. Tata-Mc-Graw Hill Company, Second Edition.

[8] Bollegala, D., Massuo, Y., Ishizuk, M. "A Web search engine-based approach to measure semantic similarity between words", IEEE Transactions on Knowledge and Data Engineering, July 2011, Volume 23, Issue 7, 977-990.

[9] Yuhua Li, Zuhair A. Bandar and David McLean. "An approach for measuring semantic similarity between words using multiple information sources", IEEE Transactions on Knowledge and Data Engineering, July/August 2003, Volume 15, No.4.

[10] Jiawei Han and Micheline Kamber. Data Mining-Concepts and Techniques. Morgan Kaufmann Publishers.

[11] Sam Y. Sung, Zhao Li, Chew L Tan and Peter A. Ng. "Forecasting Association Rules using Existing Datasets", IEEE Transactions on Knowledge and Data Engineering, Volume 15, No.6.

[12] Leila Kosseim and Jamileh Yousefi. "Improving the performance of question answering with semantically equivalent answer patterns". International Journal on Data and Knowledge Engineering, July 2008, Volume 66, Issue 1, 53-67.

[13] Demidova, E., Xuan Zhou and Neidl, W. "A Probabilistic scheme for keyword-based incremental query construction" IEEE Transactions on Knowledge and Data Engineering, March 2012, Volume 24, Issue 3, 426-439.