# A Method for Classification based on Association Rules using Ontology in Web Data

R.Hubert Rajan
Assit.Professor
Department of Information Technology
Noorul Islam University, Kumaracoil, India.

Julia Punitha Malar Dhas
Phd. Professor
Department of Computer Science and Engineering
Noorul Islam University, Kumaracoil, India.

## ABSTRACT
This paper shows a new method based on association rule mining and ontology for the classification of web pages. This work is pruning of association rules, generated by mining process. The main complexity arises due to the fact that there are various number of text documents that are considered for generating the association rules using the A-priori algorithm. But these rules that were generated are not based on the semantic knowledge. In order to obtain the most accurate rules we gone for the construction of the ontology, based on the domain knowledge. With this domain knowledge we design various operators which are helpful in reducing the rules generated. Thus the various rules that we get are semantically correct with regards to the domain selected. We use the high confidence value based classifier for classifying the given text document to that particular domain. Association rules are mined from this matrix using A-priori algorithm. Based on the high confidence value, a new text document is classified into one of the predefined classes. In general, from association rule mining, a huge amount of association rules are mined. All the association rules generated may not be useful for the classification purpose. So, In order to reduce the irrelevant association rules, we need semantic knowledge. For this purpose, propose new domain specific ontology to overcome this drawback of association rule mining method.

## Keywords
Data Mining, Association Rule, Web Data, Web Mining, Classification, Ontology.

## 1. INTRODUCTION
The association rules are divided into five categories: trivial, known and correct, unknown and correct, known and incorrect, unknown and incorrect. When applying association mining to real datasets, a major obstacle is that often a huge number of rules are generated even with very reasonable support and confidence. According to Ping Chen et al [10], post-processing can be efficiently integrated with existing rule reduction techniques to construct a concise, high-quality, and user-specific association rule set. Also Claudia et al. [1] have proposed the usefulness of association rules by overcoming an interactive approach to prune and filter discovered rules. This approach has produced sets of rules, and number of ways to reduce the rules and has integrated domain expert knowledge in the post processing step. The quality of the filtered rules was also validated by the domain expert at various points in the interactive process. Alaa Al Deen et al[12], has used the association rule mining in classification approach and experimental study against 13 UCI data sets is presented to evaluate and compare traditional and association rule based classification techniques with regards to classification accuracy, number of derived rules, rules features and processing time.

## 2. ONTOLOGY AND RULE SCHEMAS
Faten et al. [4], has proposed an algorithm for building ontology via set of rules generated by rule based learning system. This algorithm has extracted the rules generated from the original dataset in developing ontology elements. Domain ontology enhances the mining results of Association Rules, which also reduce the number of generated association rules. The adopted model is based on generalization and specialization processes in which the rules are filtered by metrics based on the coverage and confidence indicators. Hongyu Zhang et al. [2] have proposed the graphical representations of ontology's to help define some complexity measures intuitively. They have classified these metrics into two sets: one for measuring the overall design complexity of an ontology (ontology- level metrics), and the other for measuring the complexity of internal structure (class-level metrics). Claudia Marinica et al [1], has used Domain Ontology's, the integration of user knowledge in the post-processing task is strengthened. Furthermore, an interactive and iterative framework is designed to assist the user along the analyzing task. On the one hand, user domain knowledge is represented and on the other a novel technique is suggested to prune and to filter discovered rules.

## 3. DETAILED DESIGN
The WEBKB Dataset is a common benchmark used for testing classification algorithms. The dataset contains approximately 4000 web page contents, partitioned (nearly) evenly across 4 different groups. The 4 topics that are organized into broader categories as: faculty, student, project and course. This dataset has been used for many learning tasks. The above 4 topics under the category WEBKB (student, faculty, project and course), each containing 1000 documents has been used for this work.

Document pre-processing is a prerequisite for any Natural Language Processing application. It is usually the most time consuming part of the entire process. The various tasks performed during this phase are,

- Retrieval of required information from the dataset.
- Parsing
- Tokenization
- Stop word Removal
- Stemming

## 3.1 Parsing

Parsing of text document involves removing of all the HTML tags. The web pages will contain lot of HTML tags for alignment purpose. They do not provide any useful information for classification. All the text content between the angle braces '<' and '>' are removed in this module. The tag information between them will not be useful for mining purpose. They will occupy more space and it should be removed. This step will reduce lot of processing complexity.

## 3.2 Tokenization

Tokenization is actually an important pre-processing step for any text mining task. Tokenization is the process of breaking a stream of text up into words, phrases, symbols, or other meaningful elements called tokens. The list of tokens becomes input for further processing such as parsing or text mining. Tokenization usually occurs at the word level. Often a tokenize relies on simple heuristics, for example:

- All contiguous strings of alphabetic characters are part of one token, likewise with numbers.
- Tokens are separated by whitespace characters, such as a space or line break, or by punctuation characters.
- Punctuation and whitespace may or may not be included in the resulting list of tokens.

## 3.3 Stop word Removal

Stop word removal removes the high frequent terms that do not depict the context of any document. These words are considered unnecessary and irrelevant for the process of classification. Words like 'a', 'an', 'the', 'of'','and' etc. that occur in almost every text are some of the examples for stop words. These words have low discrimination values for the categories. Using a list of almost 500 words, all stop words are removed from the documents.

## 3.4 Stemming

Stemming removes the morphological component from the term, thus reducing the word to the base form. This base form doesn't even need to be a word in the language. It is normally achieved by using rule based approach, usually based on suffix stripping. The stemming algorithm used here is the Porter Stemmer algorithm, which is the standard stemming algorithm for English language.
Eg:Playing, Plays, Played,Play→ Play

## 3.5 Vector-Space Modelling

All the documents are represented in the form of a vector called Term Frequency-Inverse Document Frequency vector (TF_IDF vector). The Term Frequency and Inverse Document Frequency are calculated as follows:
Term Frequency,

$$TFdt = freq(d,t)$$

Inverse Document Frequency,

$$IDFt = \log(|D| / |Dt|)$$

Where freq (d,t) is the number of occurrences of term t in document d
|D| is the total number of documents
|Dt| is the number of documents containing the term t
Now, the TF_IDF of a term is calculated by,

$$TF\_IDFdt = TFdt \times IDFt$$

The TF_IDF vector of a document will be represented as,
<TF_IDFterm1, TF_IDFterm2, . . . , TF_IDFtermn>

# 4. ONTOLOGY CONSTRUCTION

In the early 1990s, ontology was defined by a formal, explicit specification of a shared conceptualization. By conceptualization, we understand here an abstract model of some phenomenon described by its important concepts. The formal notion denotes the idea that machines should be able to interpret ontology. Moreover, explicit refers to the transparent definition of ontology elements. Several other definitions are proposed in the literature. For instance, ontology is viewed as a logical theory accounting for the intended meaning of a formal vocabulary. There exists a more artificial-intelligence-oriented definition. Thus, ontology's are described as (Meta) data schemas, providing a controlled vocabulary of concepts, each with an explicitly defined and machine process able semantics.
Ontology introduced in data mining for the first time in early 2000, can be used in several ways: Domain and Background Knowledge Ontology, Ontology for Data Mining Process, or Metadata Ontology's. Background Knowledge Ontology organizes domain knowledge and play important roles at several levels of the knowledge discovery process. Ontology for Data Mining Process codify mining process description and choose the most appropriate task according to the given problem; while Metadata Ontology describe the construction process of items. In this project, we focus on Domain and Background Knowledge Ontology.
Formally, an ontology is a quintuple, O = {C, R, I, H, A}. C = {C1, C2 . . . Cn} is a set of concepts and R = {R1, R2, . . .,Rm} is a set of relations defined over concepts. I is a set of instances of concepts and H is a Directed Acyclic Graph (DAG) defined by the subsumption relation (is-a relation) between concepts. We say that C2 is-a C1, C1 ≥ C2, if the concept C1 subsumes the concept C2. A is a set of axioms bringing additional constraints on the ontology.
Domain knowledge is defined as the user information concerning the database, is described in our framework using ontology. Compared to taxonomies used in the specification language, ontology offer a more complex knowledge representation model by extending the only is-a relation presented in taxonomy with the set R of relations. In addition, the axioms bring important improvements permitting concept definition starting from existing information in the ontology. In this scenario, it is fundamental to connect ontology concepts C of { C,R, I,H,A } to the database, each one of them being connected to one by several items of I. To this end, we consider three types of concepts: leaf-concepts, generalized concepts from the assumption relation (≥) in H of O, and restriction concepts proposed only by ontology. In order to proceed with the definition of each type of concepts, let us remind that a set of items in a database is defined as I = {i1, i2, . . . , in}.The leaf-concepts (C0) are defined as C0 = { c0 ε C | c`ε C , c` ≥ c0}.They are connected in the easiest way to database—each concept from C0 is associated to one item in the database:f0 : C0 → I , ¥ c0 ε C0 ; i ε I; i = f0(c0)
Generalized concepts (C1) are described as the concepts that subsume other concepts in the ontology. A generalized concept is connected to the database through its subsumed concepts. This means that, recursively, only the leaf-concepts subsumed by the generalized concept contribute to its database connection. Restriction concepts are described using logical expressions defined over items and are organized in the C2 subset. In a first attempt, we base the description of the concepts on restrictions over properties available in description logics. Thus, the restriction concept defined could be connected to a disjunction of items.

# 5. ASSOCIATIVE TEXT CLASSIFIER ENGINE

ATCE (Classifier based on High Confidence Association Rule Agreements) is a new special classifier able to return class

name when processing a test data. The following definition is required before detailing the HiCARe algorithm. A *match* occurs when the class name satisfy the body part of a rule. The ATCE algorithm stores all item sets (set of keywords) belonging to the head of the rules in a data structure.

An item set is returned in the suggested class *if* the condition stated here is satisfied

$$nM(h) / nM(h)+nN(h) \geq \beta$$

Where nM(h) is the number of matches of the item set and nN(h) is the number of not-matches. A threshold is employed to limit the minimal number of matches required to return an item set in the classification process. As we will show in the

section of experiments, our proposed method is well-suited to suggest class names for documents, enhancing and bringing more confidence to the classification process.

**ATCE Algorithm: Pseudo code**

**Input:** Feature Vector F of the test image, β

**Output:** Set of Keywords K

```
for each rule s ε S of the form body→head do
  for each item set h ε head do
      if body matches F
          then   nM(h)++
        else   nN(h)++
          end if
    end for
  end for
  for each rule s ε S of the form body→head do
if nM(h) / nM(h)+nN(h) ≥ β then
    if h € K then Add h in K
        end if
    end if
  end for
end for
Return K
```

Associative Classifier assigns a known class label to the new unknown text document based on matching and non-matching criteria with high confidence value.

# 6. IMPLEMENTATION AND RESULTS

This chapter describes the design of An Association Rule Based Method for Classifying the WebPages using Ontology**.** It explains the overall implementation and results obtained for the proposed system.

## 6.1 Document Pre-Processing

The underlying benchmark dataset considered is an unstructured dataset such that the document should be pre processed for the text mining techniques to be applied. Steps involved in document pre-processing are:

**Pseudo Code**

Step 1: Tokenize the file into individual tokens using space as the delimiter.

Step 2: Use porter stemmer algorithm to stem the words with common root word.

Step 3: Removing the stop words which does not convey any meaning.

## 6.2 Ontology Evaluation Metrics

The metric values obtained for the ontology is shown in Table 6.1

**Table 6.1 Metrics for evaluating Ontology**

| S.NO | METRICS | VALUES |
|---|---|---|
| 1 | SOV | 142 |
| 2 | ENR | 3.66 |
| 3 | TIP | 367 |

## 6.3 Defining Rule Schemas

The rule schemas are defined over four operators: Conforming, Pruning, Unexpected and Exceptions. The rules are generated from the ontology which is constructed and validated. Rule schemas are defined for each and every operator as follows:

Let C(RS) be the conforming rule schema, C(RS): X and Y → Z.

Where X, Y, Z are generalized concepts from the ontology. They are denoted by the symbol *f* ().

The Rule Schema's which are to be defined for each operator from the ontology is as follows:

## 6.5 Conforming Rule Schemas

RS: *f(*designation) → *f(*department)

*f(*designation) → {professor, assistant professor, lecturer, visiting faculty}

*f(*department) → { cse, it, ece, eee, textile, mech, eie}

R1: professor, teaching, research, computer → cse, it

R2: lecturer, classroom, schedule → mech, project

The association rules similar to the above pattern will be confirmed.

## 6.6 Pruned Rules

The rules rejected by the conforming operator are considered as the pruning rules. There is no Rule Schema defined for the pruning operator. The rules that are pruned by the conforming rule schema are:

R3: book, study, table → primary, education, receive

R4: software, perform → current, numeric

## 6.7 Unexpected Rule Schemas

Certain concepts which are not supposed to occur in the antecedent and consequent side are said to unexpected

RS: *f*(electronic_gadgets) → f(grade)

*f*(electronic_gadgets) → {mobile, computer, laptop, bluetooth, pendrive}

*f*(grade) → { gradeA, gradeB, gradeS, gradeU}

R5: mobile → gradeA

R6: table, pendrive → gradeU, professor

## 6.8 Exceptional Rule Schemas

Any rule opposing C(RS) i.e., X and Y → ! Z is considered to be exceptions.

RS: department → designation

*f*(designation) → {professor, assistant professor, lecturer, visiting faculty}

*f*(department) → { cse, it, ece, eee, textile, mech, eie}

R7: cse, it → professor, teaching, research, computer

R8: mech → lecturer

## 6.9 Reducing the rules over operator

The association rules are by taking input from the rule schemas of ontology. The similarly other operators are also executed.

With various values of support values and a constant confidence values, the system is executed and the results are mentioned in Table 6.2.

**Table 6.2 Output Table**

| | Min_ Support =30 | Min_ Support =25 | Min_ Support =18 |
|---|---|---|---|
| Total Rules | 202 | 256 | 762 |
| Pruned Rules | 192 | 239 | 723 |
| Unexpected Rules | 200 | 252 | 750 |
| Exceptional Rules | 198 | 250 | 748 |

## 6.10 Associative Classifier

| | Before Pruning | After Pruning |
|---|---|---|
| Accuracy | 90 | 93 |
| Sensitivity | 97 | 98 |
| Specificity | 70 | 72 |

ATCE, a new Associative Classifier is used, which assigns known class labels to unknown data based on association rules with high confidence matching and non-matching values with high values of accuracy.

## 6.11 Confusion Matrix

With 100 sample documents evenly taken from each category is tested with this classifier and the obtained confusion matrix is shown.

**Table 6.3 Confusion matrix**

| | Faculty | Student | Project | Course |
|---|---|---|---|---|
| Faculty | 22 | 1 | 2 | 0 |
| Student | 2 | 21 | 0 | 1 |
| Project | 1 | 1 | 22 | 1 |
| Course | 0 | 1 | 1 | 23 |

## 6.12 Performance Comparison

The performance of the proposed system is compared with the system before pruning the rules which is shown in the below Fig 6.4.

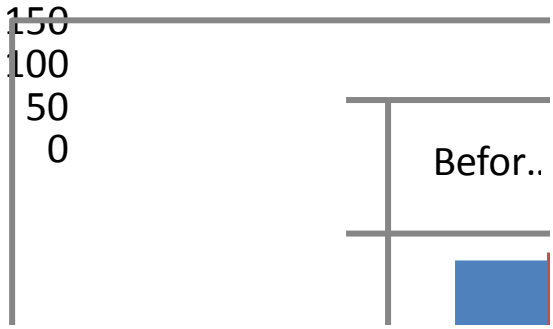


**Fig 6.4 Performance Comparison**

## 7. CONCLUSION

Automatically classifying the text documents is very important, in the field of Knowledge Discovery. This paper presents a new method based on association rule mining and ontology for the classification of web pages. Root words of the text documents are automatically extracted from the dataset after doing preprocessing and from that, a document term matrix is constructed. Association rules are mined from this matrix using Apriori algorithm. Based on the high confidence value, a new text document is classified into one of the predefined classes. In general, from association rule mining, a huge amount of association rules are mined. All the association rules generated may not be useful for the classification purpose. So, In order to reduce the irrelevant association rules, we need semantic knowledge. For this purpose, we propose new domain specific ontology to overcome this drawback of association rule mining method.

Ontology has been constructed with the available domain knowledge and validated with some metrics. In future a generalized ontology is to be constructed which will include knowledge from various domains.

## 8. ACKNOWLEDGMENTS

Our thanks to Dr.Julia Punitha Malar Dhas for his valud suggestions and guidance which helped us for successful completion.our deepest gratitude to Dr.Selva Kumar of Annamalai University for his motivation of this paper.

## 9. REFERENCES

[1] Claudia Marinica, Fabrice Guillet, "Knowledge-Based Interactive Postmining of Association Rules Using Ontologies", IEEE Transactions On Knowledge And Data Engineering, vol. 22, no. 6, pg 784-797, June 2010 .

[2] Hongyu Zhang, Yuan-Fang Li, Hee Beng Kuan Tan, "Measuring design complexity of semantic web ontologies", The Journal of Systems and Software 83 (2010) 803–814 at ElSEVIER.

[3] Marcela X. Ribeiro, Agma J. M. Traina, Caetano Traina, Paulo M. Azevedo-Marques, "An Association Rule-Based Method to Support Medical Image Diagnosis With Efficiency", IEEE Transactions On Multimedia, vol. 10, no. 2, pg.277-285, February 2008.

[4] Faten Kharbat, Haya Ghalayini, "New Algorithm for Building Ontology from Existing Rules: A Case Study", International Conference on Information Management and Engineering, pg no.12-16, 2009.

[5] Jorge J. Villalon, Rafael A. Calvo, "Concept Map Mining: A definition and a framework for its evaluation", IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, pg.357-360, 2008

[6] Supaporn Buddeewong, Worapoj Kreesuradej, "A New Association Rule-Based Text Classifier Algorithm", IEEE International Conference on Tools with Artificial Intelligence, 2005.

[7] Yuefeng Li, Ning Zhong, "Rough Association Rule Mining in Text Documents for Acquiring Web User Information Needs" ,Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence.

[8] Abe, H. Tsumoto, S., "Detection of trends of technical phrases in text mining", IEEE International Conference on Granular Computing, GRC '09, , 2009, GRC '09, Pages: 7 – 12, 2009

[9] Pak Chung Wong, Whitney, P. Thomas, J., "Visualizing association rules for text mining", Proceedings on Information Visualization (Info Vis '99), Pages: 120 - 123, 152, 1999

[10] Ping Chen, Rakesh Verma, Janet C. Meininger, Herman Pressler Dr.Houston, "Semantic Analysis of Association Rules", Association for the Advancement of Artificial Intelligence

[11] FABRIZIO SEBASTIANI, "Machine Learning in Automated Text Categorization", *Consiglio Nazionale delle Ricerche, Italy*

[12] Alaa Al Deen, Mustafa Nofal, Sulieman Bani-Ahmad, "Classification Based On Association-Rule Mining Techniques: A General Survey And Empirical Comparative Evaluation" .

[13] Inhauma Neves Ferraz, Ana Cristina Bicharra Garcia, "Ontology In Association Rules Pre-Processing And Post-Processing", Pages-87-91, IADIS European Conference Data Mining, 2008