Hindawi Publishing Corporation
ISRN Artificial Intelligence
Volume 2013, Article ID 316913, 10 pages
http://dx.doi.org/10.1155/2013/316913

Research Article

A Novel Web Classification Algorithm Using Fuzzy Weighted Association Rules

Binu Thomas[1] and G. Raju[2]

[1] *Department of BCA, Marian College, Kuttikkanam, Kerala, India*
[2] *Department of Information Technology, Kannur University, Kannur, Kerala, India*

Correspondence should be addressed to Binu Thomas; binumarian@gmail.com

Received 30 September 2013; Accepted 20 October 2013

Academic Editors: J. García, J. A. Hernandez, and L. Mikhailov

In associative classification method, the rules generated from association rule mining are converted into classification rules. The concept of association rule mining can be extended in web mining environment to find associations between web pages visited together by the internet users in their browsing sessions. The weighted fuzzy association rule mining techniques are capable of finding natural associations between items by considering the significance of their presence in a transaction. The significance of an item in a transaction is usually referred as the weight of an item in the transaction and finding associations between such weighted items is called fuzzy weighted association rule mining. In this paper, we are presenting a novel web classification algorithm using the principles of fuzzy association rule mining to classify the web pages into different web categories, depending on the manner in which they appear in user sessions. The results are finally represented in the form of classification rules and these rules are compared with the result generated using famous Boolean Apriori association rule mining algorithm.

1. Introduction

Classification is a Data Mining function that assigns items in a collection to target categories or classes. The goal of classification is to accurately predict the target class for each case in the data. For example, a classification model could be used to identify loan applicants in a bank as low, medium, or high credit risks. A classification task begins with a data set in which the class assignments are known. A classification model that predicts credit risk could be developed based on observed data for many loan applicants over a period of time. In addition to the historical credit rating, the data might track employment history, home ownership or rental, years of residence, number and type of investments, and so on. Credit rating would be the target, the other attributes would be the predictors, and the data for each customer would constitute a case. Classification techniques include decision trees, association rules, fuzzy systems, and neural networks. Classification has many applications in customer segmentation, business modeling, marketing, credit analysis, web mining and biomedical, and drug response modeling.

Classification models include decision trees, Bayesian models, association rules, and neural nets. Although association rules have been predominantly used for data exploration and description, the interest in using them for prediction has rapidly increased in the Data Mining community. When classification models are constructed from rules, often they are represented as a decision list (a list of rules where the order of rules corresponds to the significance of the rules). Classification rules are of the form $P \rightarrow c$, where P is a pattern in the training data and c is a predefined class label (target) [1]. Association rule based classification is introduced by Liu et al. [2]. Association rule mining algorithm like Apriori can be used for generating rules and a second algorithm is used for building the classifier. The rules generated by association rules are called classification association rules (CARs), as they have a predefined class label or target. From the generated CARs, a subset is selected based on the heuristic criterion that the subset of rules can classify the training set accurately.

Servers register a Web log entry for every single access they get, in which important pieces of information about accessing are recorded, including the URL requested, the IP address from which the request originated, and a time-stamp. Applying Data Mining techniques on this web log data can reveal many interesting knowledge about the web users [3]. These web log data shows information accessed by the users and give their surfing pattern. When Data Mining techniques are implemented on these logs to extract hidden patterns between the URLs requested by the users [4], it is commonly known as Web Usage Mining. In recent years there has been an increasing interest and a growing body of work in Web usage mining [5] as an underlying approach in capturing and modeling Web user behavioral patterns and for deriving e-business intelligence. Web usage mining techniques rely on offline pattern discovery from user transactions. These techniques can be used to improve Web personalization based on historic browsing patters. Association rule mining can bring out precise information about user's navigational behavior. When we apply the association rule mining techniques with web log file, the result will be of the form $X \rightarrow Y$ where X and Y are URLs [6]. It means if a user accesses URL X then he would be accessing URL Y most likely. The user's navigational pattern information can be used in predictive prefetching of pages and web personalization. Development of such recommendation systems has become an active research area. Some recent studies have considered the use of association rule mining [7] in recommender systems [8, 9].

In this work, the association rule mining techniques are used for web classification based on the navigational patterns. A novel web classification algorithm is presented here, which is developed on the foundations of fuzzy association rule mining techniques. The concepts of weighted fuzzy transactions and fuzzy support and confidence framework are used to derive this algorithm. This associative classification algorithm finds longest possible access sequence patterns which lead to a web category. Here, each web category is considered as a class label. These identified classification rules can be later used for web personalization and predictive prefetching. The Boolean Apriori algorithm also used in the same framework to find access sequences which lead to a particular web category as the consequent. The results are compared and it is found that the new algorithm identifies more natural patterns.

2. Background and Related Work

In Data Mining area, general classification algorithms were designed to deal with transaction-like data. Such data has a different format from the sequential data, where the concept of an attribute has to be carefully considered. The association-rule representation is an extensively studied topic in Data Mining. Association rules were proposed to capture the co-occurrence of buying different items in a supermarket shopping. It is natural to use association rule generation to relate pages that are most often referenced together in a single server session [6]. In the association rule mining literature, weights of items are treated as insignificant until recently and a common weight of one (1) is assigned as a common practice. Some of the very recent approaches generalize this and give item weights to reflect their significance to the user. In weighted association rule mining, the weights may be as a result of particular promotions for the items or their profitability, and so forth, [10]. Fuzzy weighted support, confidence, and transactions are also defined in a fuzzy association rule framework [11, 12]. The concepts and methods used in weighted association rule mining can be extended to web mining [13].

Muyeba et al. [12] presented a novel approach for effectively mining weighted fuzzy association rules [14]. The authors address the issue of weight of each item according to its significance with respect to some user defined criteria. Most works on weighted association rule mining do not address the downward closure property while some make assumptions to validate the property. This paper generalizes the weighted association rule mining problem with binary and fuzzy attributes with weighted settings. This methodology follows an Apriori approach but employs T-tree data structure to improve efficiency of counting item sets. The authors' approach avoids preprocessing and postprocessing as opposed to most weighted association rule mining algorithms, thus eliminating the extra steps during rules generation. The paper also presents experimental results on both synthetic and real-data sets and a discussion on evaluating the proposed approach.

In Boolean Apriori algorithm, all the products are treated uniformly, and all the rules are mined based on the occurrences of the products. However, in the social science research, the analysts may want to mine the rules based on the importance of the products, items or attributes. For example, total income attribute is more interesting than the height of a person in a household. Based on this generalized idea [15, 16], the items are given weights to reflect the importance to the users. The downward closure property of the support measure in the mining of association rules no longer exists in this approach. Here, they make use of a metric, called support bounds, in the mining of weighted fuzzy association rules. Furthermore, the authors introduce a simple sample method and the data maintenance method, based on the statistical approach, to mine the rules.

Mobasher et al. [4] proposed an effective and scalable technique for Web personalization based on association rule discovery from usage data. Here, the association rules are used for the development of a recommender system. In this work they proposed a scalable framework for recommender systems using association rule mining from click stream data. The recommendation algorithm utilizes a special data structure to produce recommendations efficiently in real-time, without the need to generate all association rules from frequent item sets. This method can overcome some of the limitations of low coverage resulting from high support thresholds or larger user histories and reduced accuracy due to the sparse nature of the data.

Suneetha and Krishnamoorti [17] suggested an improved version of Apriori algorithm to extracts interesting correlations, frequent patterns, and associations among web pages visited by users in their browsing sessions. In order to reduce repetitive disk read, a novel method of top down approach is proposed in this paper. The improved version of Apriori algorithm greatly reduces the data base scans and avoids generation of unnecessary patterns which reduces data base scan, time and space consumption. Kumar and Rukmani [18] used Apriori algorithm for web usage mining and in particular focuses on discovering the web usage patterns of websites from the server log files. In this work the memory usage and time usage of Apriori algorithm are compared with frequent pattern growth algorithm.

Ramli generates the university E Learning (UUM Educare) portal usage patterns using basic association rules algorithm called Apriori algorithm [19]. Server log files are used with Apriori algorithm to produce the final results. Here, web usage mining, approach has been combined with the basic association rule, Apriori algorithm to optimize the content of the university E Learning portal. The authors have identified several Web access pattern by applying the well known Apriori algorithm to the access log file data of this educational portal. This includes descriptive statistic and association Rules for the portal including support and confidence to represent the Web usage and user behavior for UUM Educare. The results and findings for this experimental analysis can be used by the Web administration and content developers in order to plan the upgrading and enhancement to the portal presentation.

Mei-Ling Shyu and Shu-Ching Chen proposed a new approach for mining user access patterns. The approach aims at predicting Web page requests on the website in order to reduce the access time and to assist the users in browsing within the website [20]. To capture the user access behavior on the website, an alternative structure of the Web is constructed from user access sequences obtained from the server logs, as opposed to static structural hyperlinks. Their approach consists of two major steps. First, the shortest path algorithm in graph theory is applied to find the distances between Web pages. In order to capture user access behavior on the Web, the distances are derived from user access sequences, as opposed to static structural hyperlinks. They refer to these distances as minimum reaching distance (MRD) information. The association rule mining (ARM) technique is then applied to form a set of predictive rules which are further refined and pruned by using the MRD information. In this paper, finally they propose a new method for mining user access patterns that allows the prediction of multiple nonconsecutive Web pages, that is, any pages within the website.

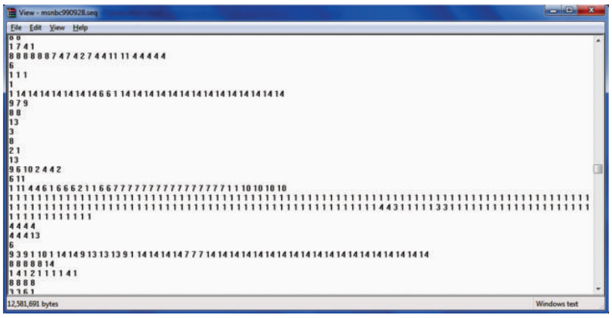
Srivastava et al. [21], proposed a data mining technique for finding frequently used web pages. These pages may be kept in a server's cache to speed up web access. Existing techniques of selecting pages to be cached do not capture a user's surfing patterns correctly. Here, they use a weighted association rule (WAR) mining technique that finds pages of the user's current interest and cache them to give faster net access [5]. This approach captures both user's habit and interest as compared to other approaches where emphasis is only on habit. If user A logs on to Internet every day for reading news and checking emails. He visits googlenews.com and gmail.com in any order. In this case, association rule would give rules (User A, googlenews.com) → (User A, gmail.com) and gmail can be pre-fetched to the cache to reduce the access time.

Among these classification methods in Data Mining, Association rule mining is simple and effective in classification. In fact rules generated from association rule mining can be easily converted to classification rules so it becomes a natural choice for classification in Data Mining. This technique is known as associative classification [1]. In the research work [22], the author focused on the construction of classification models based on association rules. In order to mine only rules that can be used for classification, the well-known association rule mining Apriori algorithm is modified to handle user-defined input constraints. Using this characterization, a classification system is implemented based on association rules. In this work, the performance of this classification method is compared with the performance of several model construction methods, including CBA (classification based on association). This classification algorithm mines for the best possible rules above a user-defined minimum confidence and within a desired range for the number of rules.

3. Finding Weighted Associations from Web Logs

We model the pieces of Web logs as sequences of events to find the associations between web pages on the basis of sequential patterns over a period of time. Each sequence is represented as an ordered list of discrete symbols and each symbol represents one of several possible categories of web pages requested by the user. Let E be a set of events. A Web log piece or (Web) access sequence $S = e_1, e_2, \ldots, e_n$ ($e_i \in E$) for ($1 \leq i \leq n$) is a sequence of events, while n is called the length of the access sequence. An access sequence with length n is also called an n-sequence. In an access sequence S, repetition is allowed. Duplicate references to a page in a web access sequence imply back traversals, refreshes or reloads [6, 23]. For example, 1, 1, 2 and 1, 2 are two different access sequences, in which 1 and 2 are two events. Figure 1 shows a sample of such sequence. The Data we used for the experiment comes from Internet Information server (IIS) logs for msbc.com and news related portions of msn.com for one entire day. Each sequence in the data set corresponds to page views of a user during that day. There are 1 million records and we selected 64,000 samples. Each event in the sequence corresponds to a request for a page. Requests are recorded only at the level of page category. There are 16 categories of pages and these categories are given numeric codes from 1 to 16. The pages are included into one of these categories based on their content. These categories are front page(1), news(2), technology(3), local(4), opinion(5), on air(6), miscellaneous(7), weather(8), health(9), living(10), business(11), sports(12), summary(13), bbs(14), and travel(15), msn news(16). Although other information pertaining to the web access is available, we model only the categories of page requests.

When we try to find the hidden patterns in web access sequence using the Boolean association, we can consider



```
8 8
1 7 4 1
8 8 8 8 8 8 7 4 7 4 2 7 4 4 11 11 4 4 4 4 4
6
1 1 1
1
1 14 14 14 14 14 14 14 6 6 1 14 14 14 14 14 14 14 14 14 14 14 14 14 14
9 7 9
8 8
13
3
8
2 1
13
9 6 10 2 4 4 2
6 11
1 11 4 4 6 1 6 6 6 2 1 1 6 6 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 1 1 10 10 10 10
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 4 4 3 1 1 1 1 1 3 3 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1
4 4 4 4
4 4 4 13
6
9 3 9 1 10 1 14 14 9 13 13 13 9 1 14 14 14 14 7 7 7 14 14 14 14 14 14 14 14 14 14 14 14 14 14 14 14 14 14 14
8 8 8 8 8 14
1 4 1 2 1 1 1 1 4 1
8 8 8 8
3 3 6 1
```

Figure 1: Web access sequences where numeric symbols are used to represent web pages.

only the presence or absence of pages in an access sequence. We do not give importance for the number of occurrences of a category of web pages in a sequence. If a particular category of web page is appearing together continuously then such occurrences also have to be processed with more significance. Instead of Apriori algorithm, if we use weighted fuzzy association rule mining algorithms, we will be able model the web sequences which reveal more natural patterns by considering all the above mentioned facts.

From Figure 1 it is clear that the weight of a category of web page in a browsing session of a user can be directly associated with the number of times the user visits that particular category of pages in his session. Again if the user is continuously visiting the same category of pages then more weight has to be given for such continuous accessing of same category of pages. Considering the above facts we define the following concepts for the development of the new web classification algorithm.

3.1. Definition 1. Fuzzy Weight of Web Page. The fuzzy weight of a web page is defined by considering the number of co-occurrences of a web category. The following expression is used for weight calculation:

$$\mu_i = \frac{n_i * \prod_{j=1}^{m} k_j}{n}, \qquad (1)$$

where μ_i is the fuzzy weight of the ith category web pages in a session. Here, we assume that there are m subgroups in a sequence which contains the ith category web pages. Then k_j is the number of successive ith category web pages appearing in the kth group and n_i is the total number of ith pages appearing in the session under consideration. Finally n is the total number of pages in the browsing session. The weight thus generated will be more than *one* in some cases and the values are normalized by dividing each weight with the maximum weight in a session. A portion of the values are given in Table 1.

3.2. Definition 2. Web Class of a Session. The class of a web sequence is defined as the web category in a sequence with maximum weight. It is assumed that in a web access sequence, the remaining page visits are leading to this web category with maximum weight. In an association rule framework, the page category with maximum weight can be considered as the consequent and the remaining pages as the antecedents. A web class WC_i for an access sequence S_i is defined as

$$\mathrm{WC}_i = \mathrm{Max}\left(S_i\left(\mu\right)\right). \qquad (2)$$

So, this concept work like a data mining classification problem where the available sequence patterns are classified into groups leading to a particular class and this information can be later used to predict the user behavior in browsing sessions. Using the above mentioned equation, all the web pages in the access patterns are converted into their corresponding weights. In this new approach, the web pages visited in a session are given weights, after considering the number of visits in that session and the extend of continuous visits of the same category of pages. The weights obtained in each session are normalized so that all the weights appear within the range of 0 and 1. Since we have sixteen categories of web pages, a new database table is created with sixteen attributes such that each attribute corresponds to a web category. All the sequences are converted into this fixed database table format with matching weights for each category. All the weights are normalized as shown in Table 1 so that they appear within a range of zero and one.

4. The Fuzzy Web Classification Algorithm (FWCA)

With the concept of web page weight and web access class, now the new algorithm for web classification can be derived. The algorithm has the following steps.

Step 1. Convert all the web pages in access sequences to corresponding fuzzy weights in a fixed database table format.

Step 2. Sort each web sequence in the descending order of weights and select the web page with maximum weight as the class (consequent) of a sequence.

Step 3. For each access sequence, the remaining pages (other than the consequent) are included into a classification rule sequence as long as the product of the weights is greater than a given support threshold.

Step 4. Select only those rules having the confidence value (associated with the number of times such rule sequences exist in the entire set of web access sequences) greater than the user specified threshold. The confidence for the jth rule for the ith web category C_{ji} is defined as

$$C_{ji} = \frac{s_{ji} * n_{ji}}{n_i}, \qquad (3)$$

where s_{ji} is support of jth rule for ith web category n_{ji} is number of jth rules identified for the ith web category n_i is total number of rules with i as the web class.

In Algorithm 1, we have the detailed pseudo code for the algorithm. In the algorithm, $W[n][p]$ is the weights of p page categories for the n browsing sessions. Weight is calculated

Table 1: Web groups and corresponding weights in browsing sessions.

Seq no.	Web categories and corresponding weights															
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	0.00	0.03	0.00	0.05	0.00	1.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.03	0.02
2	1.00	0.00	0.00	0.22	0.06	0.56	0.00	0.34	0.00	0.00	0.00	0.30	0.00	0.06	0.00	0.00
3	0.81	0.00	0.00	0.03	0.00	1.00	0.00	0.00	0.05	0.00	0.00	0.02	0.00	0.63	0.00	0.21
4	0.00	0.00	1.00	0.00	0.00	0.60	0.90	0.16	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5	0.00	0.00	0.08	1.00	0.08	0.08	0.00	0.00	0.08	0.23	0.00	0.00	0.00	0.26	0.00	0.00
6	0.71	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.14	0.00	0.04	0.02	0.00	0.00
7	1.00	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.04	0.00	0.00	0.00	0.00
8	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.36	0.00	0.00	0.00	0.00	0.00	0.00
9	0.00	0.00	0.22	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.24	0.00	0.00	0.21
10	0.34	0.00	0.00	0.05	0.00	1.00	0.00	0.00	0.00	0.00	0.30	0.00	0.00	0.48	0.00	0.00
11	0.00	1.00	0.00	0.00	0.00	0.27	0.00	0.00	0.00	0.46	0.00	0.00	0.00	0.00	0.47	0.00
12	0.00	0.00	0.00	1.00	0.00	0.00	0.35	0.00	0.50	0.00	0.00	0.50	0.00	0.00	0.00	0.00
13	0.22	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.28	0.00	0.00	0.54	0.00	0.00
14	0.00	0.00	0.25	0.00	0.00	0.00	0.74	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.85	0.00
15	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.55	0.00	0.00	0.32	0.06	0.00	0.00	0.00	0.00
16	0.05	0.08	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.00	0.00	1.00	0.00	0.31	0.00	0.00
17	0.00	0.00	0.13	0.00	1.00	0.00	0.00	0.40	0.00	0.00	0.00	0.00	0.00	0.10	0.00	0.00
18	0.05	0.00	0.00	0.04	0.21	0.00	0.00	0.00	0.15	0.00	0.00	0.00	1.00	0.00	0.03	0.00
19	0.00	0.15	0.00	0.00	0.23	0.14	0.00	0.00	5.14	0.00	0.06	0.00	0.00	0.00	0.00	0.03
20	0.00	0.50	0.00	0.00	0.00	0.50	0.00	0.00	1.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00
21	0.02	0.00	0.00	0.00	0.05	0.00	0.03	0.00	0.00	0.00	0.32	1.00	0.00	0.00	0.12	0.00

using expression 1. *Sort*$(W[i][p])$ is used to sort the ith access sequence on the basis of the weights of web pages. The function *addwebtype*(Wm) is used to get the web page category whose weight is the maximum in an access sequence. Finally *Seq*$[n]$ stores the rules generated. The algorithm find all fuzzy weighted classification rules from the web access sequence for a user specified support (∞) and confidence (β) threshold values.

In the algorithm, the weights of each category of web pages are calculated and stored in the two dimensional array w for all available sequences. In the next steps the weights are sorted in the decreasing order for every sequence. So in each sequence, the web category with maximum weight will be placed first and it is considered as the consequent of that sequence. The then highest weight category will be placed in the second position and so on. In each sequence the web categories are arranged in the decreasing order of their importance. The web page categories are included in to the rules (as antecedents in the association rule) in the order they are arranged.

We have a variable support and its initial value is the weight of the consequent. When one web category is included in to the rule its weight is multiplied with the support value. When the support becomes smaller than the given threshold alpha (∞) the rule generation for that user session is stopped. The next user session sequence is considered for the same process and this is done for all the available user session sequences and finally there will be "n" rules. In the next step, the global *confidence* of rules is checked.

Since the web categories are included into the rules as antecedents in the descending order of their weights, only the most significant web categories in a user session will be included into the rule generated for that session. Once all rules are generated, their global confidence count is found (the number of times the same antecedent-consequent sequence appears in the entire sessions). All the rules with confidences lesser than the user specified threshold *beta* are removed from the set of rules. By applying the weighted association rule mining approach to the web mining problem we could identify fifty five rules from one 64000 sessions these rules are given in (Table 2).

5. Finding Web Page Associations Using Apriori Algorithm

To find the Boolean associations between the co-occurrences of the web pages, first we converted the entire web sequences into true or false values. Since there are sixteen categories of web pages, here also we designed a database with sixteen fields. The presence and absence of a web page in a sequence is represented with true or false values in the corresponding record representing the web sequence. In this approach, we are considering only the presence of a web page in the sequence and we do not give any significance for the number of occurrences of the page in the sequence. With this preprocessing, the web sequence database is converted in to a true Boolean database with only true or false values indicating the presence and absence of web pages in the user sessions. Table 3 shows a portion of such Boolean database generated.

To apply the Apriori algorithm to the new dataset we used the IBM SPSS Modeler 14.1 data mining software. The Table 4

```
Algorithm FWCA ($n$, $p$)
{
// $n$ is the number of access sequence
// $p$ is the number of web categories
// $W[n, p]$ is the weights of web pages for $n$ sessions
// seq[$n$, $p$] is the rules generated
for $i = 1$ to $n$
  {
    for $j = 1$ to $p$
    {
     $w[i, j]$ = weight($i$, $p$)
    }
  }
Sort ($w[n, p]$)
rule = 1
for $i = 1$ to $n$
    {
     $k = 2$
     Support = (Max($W[i, 1 \ldots p]$))
     while support≥ $\infty$
       {
        wm = (Max($W[i, k \ldots p]$))
        Seq[rule, $k$] = Addwebtype(wm)
        Support = support $*$ wm
        Delete(wm)
        $k = k + 1$
       }
    Rule = rule + 1
    }
for $i = 1$ to rule
    {
    if confidence(seq[$i$, $p$]) > $\beta$
    print(seq[$i$, $p$])
    }
}
```

Algorithm 1: The new algorithm used for web classification.

Table 2: The association rules generated using the FWCA method.

No.	Rules	Support	Confidence
	Rules 1–29		
1	2, 10, 4, 6, 7 → 1	4.88	0.762
2	11, 12 → 1	0.38	0.62
3	3, 7, 2, 15, 6 → 1	0.31	0.554
4	3, 11 → 1	0.17	0.409
5	6, 14 → 1	0.14	0.378
6	7, 6, 1 → 1	0.3	0.548
7	3, 4 → 2	0.09	0.296
8	6, 3, 12 → 2	0.02	0.125
9	3, 10, 1, 11 → 2	0.01	0.12
10	14, 1 → 2	0.01	0.119
11	7, 1, 6 → 3	0.04	0.198
12	1 → 3	0.02	0.139
13	12, 1, 11 → 3	0.01	0.1
14	4 → 3	0.01	0.98
15	11, 2 → 3	0.01	0.92
16	9 → 3	0.02	0.89
17	10, 3, 5, 6, 9 → 4	0.35	5.93
18	3, 9, 7, 12 → 4	0.07	0.26
19	7, 9, 1, 2 → 4	0.04	0.189
20	11, 2, 7, 8, 9 → 4	0.02	0.13
21	1 → 5	0.16	0.82
22	2, 6, 11, 1, 4 → 5	0.04	0.7
23	9 → 5	0.01	0.51
24	1 → 6	0.06	0.241
25	7, 3 → 6	0.05	0.233
26	15 → 6	0.05	0.23
27	2, 1, 10 → 6	0.03	0.176
28	9 → 6	0.02	0.152
29	12 → 6	0.01	0.106
	Rules 30–55		
30	1 → 7	0.04	0.202
31	3, 4, 6, 9 → 7	0.01	0.1
32	10, 2, 6, 1, 4 → 7	0.01	0.19
33	7, 6 → 8	0.02	0.132
34	2 → 8	0.01	0.12
35	9 → 8	0.01	0.118
36	1, 7 → 8	0.01	0.114
37	4, 7 → 8	0.01	0.103
38	3, 11 → 9	0.02	0.131
39	6, 7 → 9	0.02	0.13
40	4, 1, 3 → 9	0.01	0.12
41	12 → 9	0.01	0.114
42	6, 7, 2, 12, 1 → 10	0.02	0.135
43	1, 4, 15 → 10	0.02	0.127
44	2, 4, 7, 12, 1 → 10	0.01	0.105
45	1, 5 → 11	0.04	0.061
46	4, 9, 2 → 11	0.01	0.032
47	1 → 12	0.07	0.26
48	2, 1, 6, 15 → 12	0.01	0.114
49	2, 3, 4 → 12	0.01	0.106

represents the rules identified using Apriori algorithm. By applying Apriori algorithm, 45 rules were identified with a support value of 5 and confidence of 20. Among the rules, only seven are having more than one antecedent.

By comparing the rules generated from both the methods, it is clear that the fuzzy weighted approach for associative classification using the new proposed algorithm is far superior to Boolean Apriori Algorithm. This is in terms of the coverage of the rules and inclusion of web categories into the classification rules. In the case of Apriori algorithm we got only two antecedents in five cases and only one in all the remaining cases but in the case of weighted approach many rules are having more than three web categories as antecedents.

6. Discussions

Web server log files contains repository of web browsing information by the internet users. Mining on this data collection can bring out valuable information about the web access patterns of users. When we apply classification and

Table 2: Continued.

No.	Rules	Support	Confidence
50	7, 14, 4 → 13	0.44	0.66
51	9 → 13	0.03	0.173
52	2, 1 → 14	0.38	0.616
53	13, 8 → 14	0.05	0.223
54	1, 10, 2, 12 → 14	0.02	0.131
55	6, 7, 2, 5, 10 → 15	0.01	0.137

Table 3: Boolean values representing presence of web categories in sessions.

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
F	T	T	T	F	F	F	F	F	F	F	F	F	F	F	F
F	F	F	F	F	T	T	T	F	F	F	F	F	F	F	F
F	F	T	T	T	T	F	F	T	T	F	F	F	F	F	F
T	F	F	F	F	T	F	F	F	F	T	F	F	T	F	F
T	F	F	F	F	T	F	F	F	T	F	F	F	F	F	T
F	F	F	F	F	F	F	T	F	F	F	F	F	F	T	F
F	F	F	F	F	T	F	F	F	F	F	T	T	F	F	F
F	F	F	F	F	F	F	F	T	F	F	T	F	F	F	F
F	F	T	F	F	F	F	F	T	T	F	F	F	F	F	F
F	F	F	F	F	F	F	F	T	F	F	F	F	F	F	F
F	F	T	F	F	F	F	F	F	F	F	F	F	F	F	F
F	F	F	F	T	F	F	F	F	T	F	F	T	F	F	F
F	F	F	F	F	F	F	F	F	F	F	F	T	F	F	F

prediction techniques in web usage mining environment, the access patterns web users can be predicted. The data used in this work contains *sixteen* web categories and 64,000 samples of web access sequences involving these web categories. Some of the web categories are highly popular among web users that these web categories appear in several access patterns. The importance of web categories is evident from the graphical representations (Figures 2 and 3) which are directly linked with the number of occurrences of web categories in access sequences.

Figure 2 shows the number of sequences in which the web categories appear. The Figure 3 is the total occurrences web categories in different sequences (a web category may appear many times in a sequence). From the figures it is clear that some of the web categories are more important in comparison with others. The concept of importance of web categories in access sequences is modeled using the concept of fuzzy weight of web categories (Table 1).

In this paper, the associations are found between the web categories using conventional Boolean Apriori algorithm and the FWCA. By using the new algorithm, *fifty-five* rules are identified and *forty-five* rules are identified using Apriori algorithm. It is found that the rules generated using the FWCA algorithm have more coverage (classification rules are identified for more web categories) and it identified classification rules leading to *fifteen* web categories. A comparison between the two techniques in terms of the number of rules identified from each web category is given in Figure 4.

Table 4: The rules identified using apriori algorithm.

	Using Boolean Apriori algorithm		
No.	Rules	Support	Confidence
1	11 → 1	5.97	57.39
2	10 → 1	5.00	53.21
3	7 → 1	8.01	46.56
4	2 → 1	17.76	42.99
5	12 → 1	11.18	39.35
6	14 → 1	11.96	34.09
7	4 → 1	12.43	33.05
8	3 → 1	12.38	27.48
9	6 → 1	21.84	18.88
10	15 → 1	21.84	18.88
11	10 → 2	5.00	34.53
12	11 → 2	5.97	30.88
13	7 → 2	8.01	24.28
14	6 → 2	32.19	23.72
15	4 → 2	12.43	22.23
16	12 → 2	11.18	21.04
17	3 → 2	12.38	20.23
18	14 → 2	11.96	16.14
19	1 → 3	7.64	18.96
20	12 → 3	5.00	18.36
21	7 → 4	8.01	33.40
22	3, 6 → 4	7.64	22.02
23	9 → 4	9.10	21.29
24	11 → 4	5.97	19.32
25	10 → 4	5.00	17.89
26	2 → 4	17.76	15.56
27	1, 7 → 6	8.01	40.31
28	15 → 6	5.00	27.53
29	1, 2 → 6	7.64	20.69
30	11 → 6	5.97	19.29
31	4 → 6	12.43	17.30
32	9 → 6	9.10	16.89
33	4 → 7	12.43	21.53
34	1, 2 → 7	7.64	16.94
35	6, 7 → 9	8.01	16.47
36	3 → 9	12.43	15.59
37	2, 7 → 10	7.64	15.83
38	12 → 11	7.64	18.07
39	2 → 12	7.64	20.30
40	15 → 12	5.97	19.09
41	10 → 12	5.00	18.03
42	14 → 13	11.96	15.25
43	13 → 14	7.78	23.44
44	2 → 14	7.64	16.48
45	11 → 14	5.97	15.93

The classification rules generated using the techniques show the associations between the web categories. The number of web categories involved in each rule shows the ability of the rule generation technique to find more inclusive rules

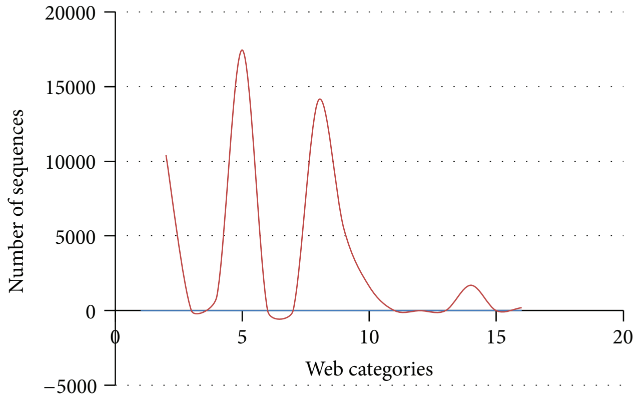


Figure 2: The web categories and the number of sequences they appear.

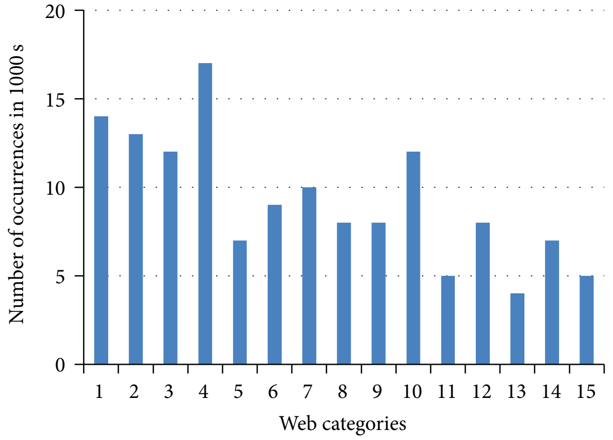


Figure 3: The total number of occurrences of web categories in all the sequences.

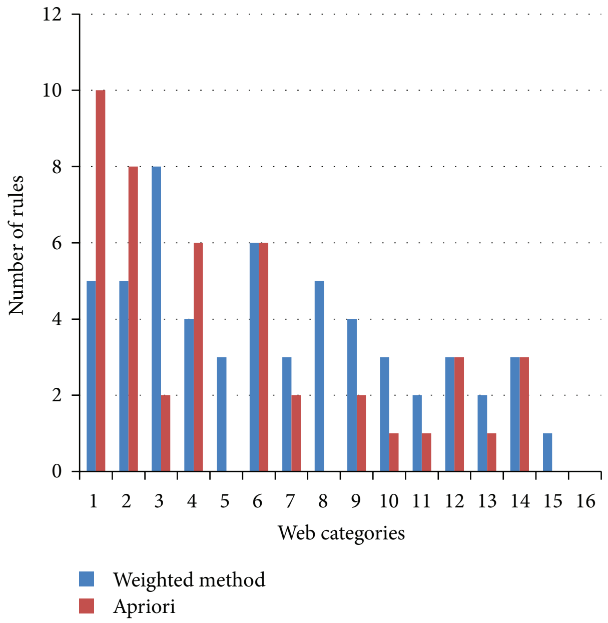


Figure 4: The number of rules identified from each web category using the two methods.

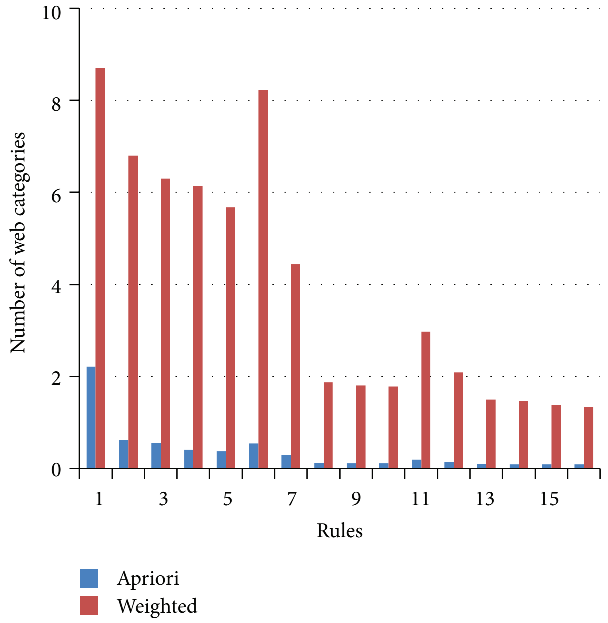


Figure 5: Average numbers of web categories (antecedents) involved in rules.

from the available web access sequences. The average number of web categories (antecedents) involved in each class of rules using the two rule generation techniques are given in Figure 5. It is evident from the figure that the FWCA is more inclusive (more web categories are included as antecedents in rules) while identifying the classification rules.

From the sample data, the number of access sequences in which the web categories appear similar to the antecedents of the web classification rules are also found. This is similar to the support threshold in association rule mining. Even though the rules generated using the fuzzy weighted algorithm have more antecedents, those rule patterns appear more in the access sequences than the Boolean rules (Figure 6). It shows that the new fuzzy based algorithm is more capable of identifying natural associations between web categories.

Finally the fuzzy weighted rules out perform the Boolean rules in terms of the number of access patterns which actually satisfy the rules. This is equivalent to the confidence measure of association rule mining technique. Actual validity and authority of the rules are analyzed by finding the number of occasions in which the access sequences from the sample data perfectly satisfy the classification rules. The advantage of fuzzy weighted rules over Boolean rules in terms of number of cases from the sample data, satisfying the rules is demonstrated in Figure 7 The graph shows the number of cases satisfying the *forty-five* Boolean rules and *fifty-five* fuzzy weighted rules.

From the above discussion, it follows that the FWCA algorithm presented here to classify the access sequences has

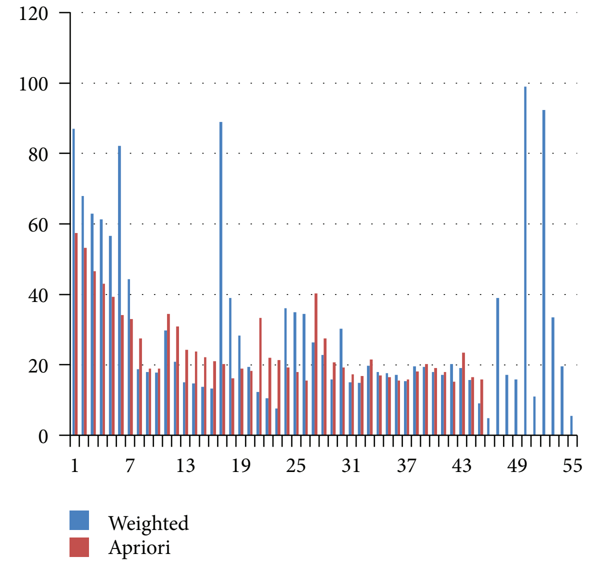


Figure 6: The rules (x axis) and support count (y axis): a comparison.

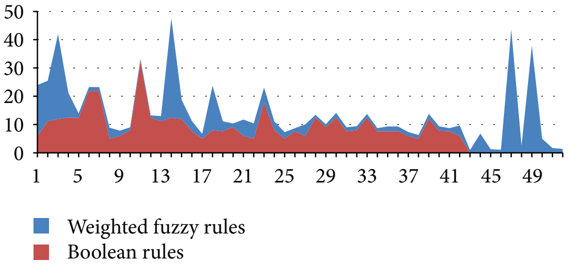


Figure 7: Number of cases (y axis in 100 s) satisfying the *forty six* Boolean rules and *fifty-five* fuzzy weighted rules (x axis).

noticeable advantage over the Boolean Apriori method. The benefits of this method are listed as follows.

(i) In a web access sequence the importance of web categories vary according to the user preferences. By using FWCA, we can assign more weight-age for frequently visited pages by uses. This consideration will help in evolving rule which represent natural access habits.

(ii) In the experiment, There are *sixteen* categories of web pages. A classification rule generation system for this sample data is efficient if it can identify rules which lead to most these web categories. FWCA identified rules which lead to *fifteen* web categories. Apriori algorithm identified rules for only *eleven* web categories.

(iii) These web classification rules can be used for prediction and selective prefetching of web pages. The rules will become useful if more number of web categories are involved in the rules. Using FWCA, more web categories are included as antecedents in the rules. It helps in identifying wide range of associations between web categories.

(iv) The number of sequences in which the antecedents of a rule appear together from the sample sequences (the support count) is more in the case of FWCA. It shows that the rules generated by the algorithm are exactly revealing the access patterns of users.

(v) The number of sequences which actually satisfy the rules (Confidence measure) from the sample is also more for FWCA. It proves that that the rules generated by the algorithm are correct, that is, the antecedent sequences identified by the rules are leading to the web category of the rule.

The main advantage of this algorithm is that it can identify the longest possible frequent patterns in a single step by using the concepts of fuzzy weighted association, while the Apriori algorithm requires many passes over the data to generate the rules.

7. Summary

The concept of a market basket can be extended as the pages visited by a user in one session in web mining. Association rule mining techniques are used here to find associations between web pages visited by users. Here the problem is redefined like which pages are most frequently visited simultaneously by web users? The Boolean Apriori algorithm for association rule mining is used to find the association between the web pages visited together by users. But by using Apriori algorithm, only the presence and absence of web pages in a browsing session is considered. But, we also have to consider other important factors like the number visits of a web category, the time spent on a web page, and so forth, This paper discussed about a novel web classification algorithm using the principles of fuzzy association rule mining to classify the web pages into different classes in a single step, depending on the manner in which they appear in user sessions. In this approach, page visits in a browsing sessions are converted into fuzzy weighted values and association rules are generated from this. These fuzzy rules are used to classify access patterns in the form of classification rules.

References

[1] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, Morgan Kaufmann, 2006.

[2] B. Liu, W. Hsu, and Y. Ma, "Integrating classification and association rule mining," in *Proceedings of the Knowledge Discovery and Data Mining (KDD '98)*, pp. 80–86, AAAI, 1999.

[3] W. Jicheng, H. Yuan, W. Gangshan, and Z. Fuyan, "Web mining: knowledge discovery on the web," in *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics 'Human Communication and Cybernetics'*, vol. 2, pp. 137–141, October 1999.

[4] B. Mobasher, R. Cooley, and J. Srivastava, "Automatic personalization based on web usage mining," *Communications of the ACM*, vol. 43, no. 8, pp. 142–151, 2000.

[5] A. Srivastava, A. Bhosale, and S. Sural, "Speeding up web access using weighted association rules," *Pattern Recognition and Machine Intelligence Lecture Notes in Computer Science*, vol. 3776, pp. 660–665, 2005.

[6] J. Srivastava, R. Cooley, and P.-N. Tan, "Web usage mining: discovery and applications of usage patterns from Web data," *ACM SIGKDD Explorations Newsletter*, vol. 1, no. 2, pp. 12–23, 2000.

[7] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," in *Proceedings of the 20th International Conference on Very Large Data Bases (VLDB '94)*, pp. 487–499, Morgan Kaufmann, 1994.

[8] R. Kosala and H. Blockeel, "Web mining research: a survey," *ACM SIGKDD Explorations Newsletter*, vol. 2, no. 1, pp. 1–15, 2000.

[9] A. Abraham, "Business intelligence from web usage mining," *Journal of Information & Knowledge Management*, vol. 2, no. 4, pp. 375–390, 2003.

[10] W. Wang, J. Yang, and S. Philip, "Efficient mining of weighted association rules (WAR)," in *Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '00)*, pp. 270–274, August 2000.

[11] M. S. Khan, M. Muyeba, C. Tjortjis, and F. Coenen, "An effective fuzzy healthy association rule mining algorithm (FHARM)," *databases*, vol. 4, no. 5, article 14, 2006.

[12] M. Muyeba, M. S. Khan, and F. Coenen, "Effective mining of weighted fuzzy association rules," *Computer*, vol. 90, 9 pages, 2010.

[13] S. K. Pal, V. Talwar, and P. Mitra, "Web mining in soft computing framework: relevance, state of the art and future directions," *IEEE Transactions on Neural Networks*, vol. 13, no. 5, pp. 1163–1177, 2002.

[14] M. Muyeba, M. S. Khan, and F. Coenen, "Fuzzy weighted association rule mining with weighted support and confidence framework," *New Frontiers in Applied Data Mining Lecture Notes in Computer Science*, vol. 5433, pp. 312–320, 2009.

[15] F. Karel, *Quantitative association rules mining—department of cybernetics [Ph.D. thesis]*, 2012, http://cyber.felk.cvut.cz/phd/completed/KAREL-phd%2012_2009.pdf.

[16] W. Wang, J. Yang, and S. Philip, "WAR: weighted association rules for item intensities," *Knowledge and Information Systems*, vol. 6, no. 2, pp. 203–229, 2004.

[17] K. R. Suneetha and R. Krishnamoorti, "Web log mining using improved version of apriori algorithm," *International Journal of Computer Applications*, vol. 29, no. 6, pp. 23–27, 2011.

[18] B. S. Kumar and K. V. Rukmani, "Implementation of web usage mining using Apriori and FP Growth algorithms," *International Journal of Advanced Networking and Applications*, vol. 400, pp. 400–404, 2010.

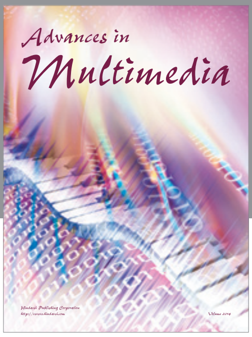
[19] A. A. Bin Ramli, "Web usage mining using apriori algorithm: UUM learning care portal case," in *Proceedings of the International Conference on Knowledge Management*, pp. 212–220, 2001.

[20] C. Haruechaiyasak, M.-L. Shyu, S.-C. Chen, and X. Li, "Web document classification based on fuzzy association," in *Proceedings of the 26th Annual International Computer Software and Applications Conference*, pp. 487–492, August 2002.

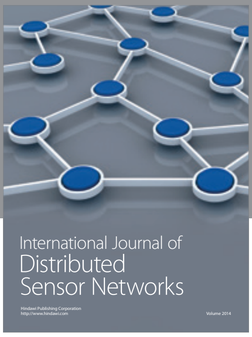
[21] A. Srivastava, A. Bhosale, and S. Sural, "Speeding up web access using weighted association rules," in *Pattern Recognition and Machine Intelligence*, pp. 660–665, Springer, Berlin, Germany, 2005.

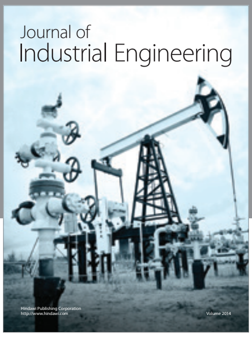
[22] S. K. Palanisamy, *Association rule based classification [Ph.D. thesis]*, Worcester Polytechnic Institute, 2006.

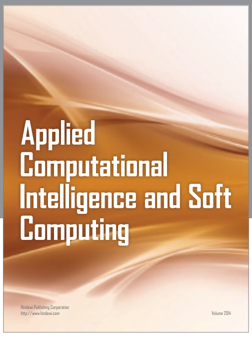
[23] J. Pei, J. Han, B. Mortazavi, and H. Zhu:, "Mining access patterns efficiently from web logs," in *Proceedings of the 4th PAKDD*, pp. 396–407, Kyoto, Japan, 2000.

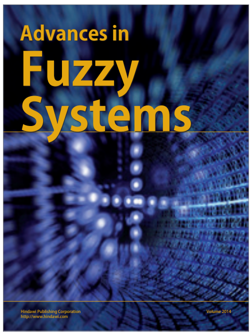

Advances in
Multimedia

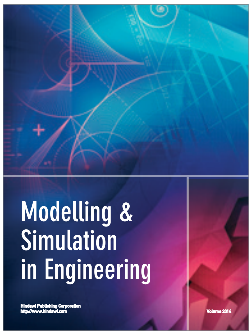
The Scientific
World Journal

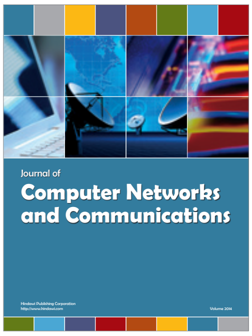

International Journal of
Distributed
Sensor Networks

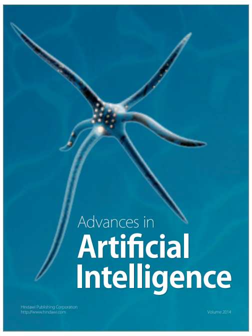

Journal of
Industrial Engineering

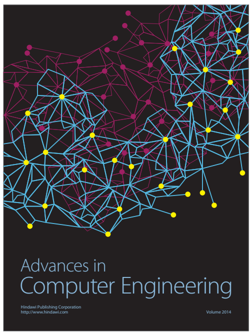

Applied
Computational
Intelligence and Soft
Computing


Advances in
Fuzzy
Systems

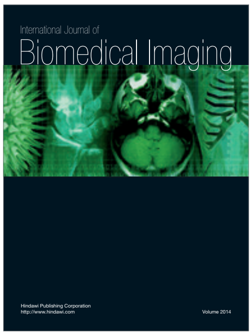

Modelling &
Simulation
in Engineering

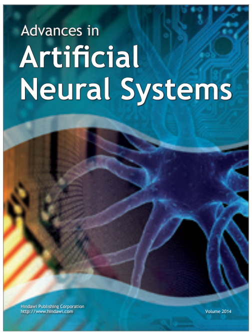

Journal of
Computer Networks
and Communications

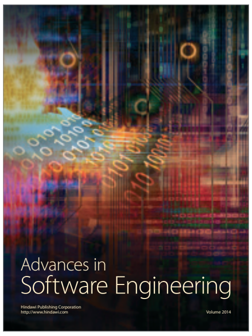

Advances in
Artificial
Intelligence


Advances in
Computer Engineering

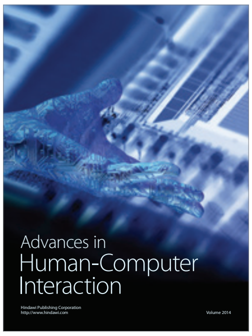
International Journal of
Computer Games
Technology

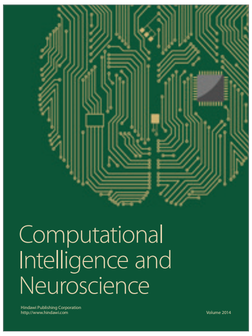

International Journal of
Biomedical Imaging


Advances in
Artificial
Neural Systems


Advances in
Software Engineering

Journal of
Robotics


Advances in
Human-Computer
Interaction


Computational
Intelligence and
Neuroscience

International Journal of
Reconfigurable
Computing

Journal of
Electrical and Computer
Engineering