An Analysis and Knowledge Representation System to attain the genuine web user usage behavior

V.V.R. Maheswara Rao
Professor, Department of Computer Applications
Shri Vishnu Engineering College for Women
Bhimavaram, Andhra Pradesh, India
mahesh_vvr@yahoo.com

Dr. V. Valli Kumari
Professor, Department of Computer Science & Systems Engineering
College of Engineering, Andhra University
Visakhapatnam, Andhra Pradesh, India
vallikumari@gmail.com

***Abstract*—With the explosive growth of WWW, the web mining techniques are densely concentrated to discover the relevant behaviors of the web user from the web log data. In fact the pattern discovery techniques generate many hundreds, often thousands, of patterns, that are unwanted, unexpected, disputable and unbelievable in nature. The success of representing the real knowledge out of such patterns is highly reliant on the pattern analysis stage in investigating the web user usage behavior. To retain most genuine and interesting patterns it is necessary to filter out unqualified patterns and use more sophisticated visualization techniques to present the knowledge of web user usage effectively.**

The authors in the present paper propose an Analysis and Knowledge Representation System (AKRS) that equally concentrates on both knowledge identification and representation. The key measures are combinedly used for the knowledge identification as a three phase filtering system, to determine the interestingness of patterns in the proposed AKRS. Initially, the objective measures applied on the patterns discovered by pattern discovery techniques to filter out the patterns that do not meet statistical strengths with the frame work of interest factor. Later, subjective measures are applied to identify the patterns that are of most genuine interestingness based on web knowledge. Finally, the heuristic measures evaluate the semantics of patterns based on both user specific objectives and utility of mined patterns. The measures of AKRS efficiently determine the correlation among the most interesting patterns. In addition, to meet the challenges in knowledge representation, like identifying relevant information, finding the depth of information and achieving the visualization competency, the proposed AKRS also designates the recent knowledge visualization techniques like multidimensional and specialized hierarchical. The AKRS amplifies the truthfulness and rate of success in representing final knowledge of web user behavior.

Keywords-Pattern Analysis; objective meassures; subjective meassures; heuristic meassures; visulization techniques.

I. INTRODUCTION

The rapid advancement in web technology and declined costs of storage media, has led business to store enormous amounts of information in huge weblogs. Mining useful information and helpful knowledge from these weblogs has evolved as solid base for researchers and creates a scope for further research. Web mining is the application of data mining techniques to identify and represent useful information and patterns from the weblog data. The survey conducted by various authors and their research contributions identified three broad categories of web mining, namely Web Content Mining (WCM), Web Structure Mining (WSM) and Web Usage Mining (WUM).

Web Content Mining is a mining technique which can extract the knowledge from the content published on internet, usually as semi-structured (HTML), Unstructured (Plain text) and structured (XML) Documents. The content of a Web page may be varied, like text, images, HTML, tables or forms. Web Structure Mining is a mining technique which can extract the knowledge from the World Wide Web and links between references in the Web. Mining the structure of the Web involves extracting knowledge from the interconnections of the Hypertext documents in the WWW. Web Usage Mining, also known as web log mining, is the process of automatic discovery and investigation of patterns in click streams and associated data collected or generated as a result of user interactions with web resources on web sites.

The main goal of web usage mining is to capture, model and analyze the behavioural patterns and profiles of users interacting with web sites. The discovered patterns are usually represented as collection of pages, objects or resources that are frequently accessed by groups of users with common needs or interests. The primary data resources used in web usage mining are log files generated by web and application servers. The overall Web usage mining process can mainly be divided into three interdependent stages: Pre-Processing, Pattern Discovery & Pattern Analysis as shown in figure 1.

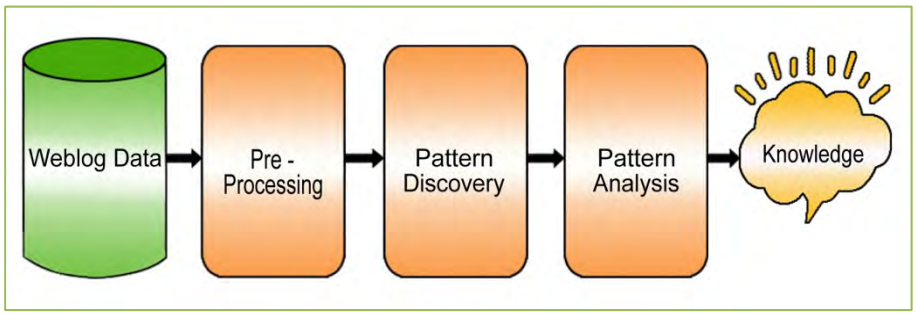

Figure 1. Stages of web usage mining

Pre-processing is the initial and very important stage in web usage mining applications in the creation of suitable target data set to which mining algorithms can be applied. Pattern discovery is the second stage of web usage mining process which can take the output generated by pre-processing stage. The goal of pattern discovery is the stage of learning certain general concepts from the pre-processed data. Pattern analysis is the final stage of usage mining which can extract interested patterns from the output of pattern discovery.

The goal of pattern analysis is the task of understanding, visualizing, and interpreting the discovered patterns and statistics. The patterns have been recognized by pattern discovery technique must be analyzed to determine patterns that can be used in investigating the web user usage behavior. In this process some patterns need to be deleted as they are identified with less interest. The evaluation of the interesting pattern has become the central issue in the analysis phase.

Thus, the usage of key measures is essential for notifying their degree of interestingness. The key measures are objective, subjective and heuristic. The objective measures discard the patterns on the basis of statistical strengths that include generality and reliability. The subjective measures that recognize the patterns on the basis of user belief factors include unexpectedness and actionability. The heuristic measures consider the semantics and explanation of the patterns take into utility. In addition, the significance of this stage depends on the presentation of the real knowledge using visualization techniques.

The remaining paper is organized as follows. In section II, related work is described. In the next section III, proposed work is presented in detail. In the subsequent section IV, the experimental analysis of the proposed work is shown. Finally in section V conclusions are mentioned.

II. RELATED WORK

In the next year 2000, Colton S. and Bundy A [4] focused on the concept of estimating the interestingness and extracted some common notations about interestingness using automated mathematical discovery. Hilderman R., Hamilton H [10] described a two step process for ranking the interestingness of discovered patterns. They have also shown how this 2 – step process can be applied to ranking generalized association rules. In the same year, Liu, B., Hsu, W., Chen, S., Ma, Y. [19] proposed a new approach to assist the user in finding interesting rules in particular unexpected rules from a set of discovered association rules. Ludwig, J. and Livingstone, G. [20] developed a definition of novelty concept to determine interestingness of discovered rules. Padmanabhan, B. and Tuzhilin [24] focused on discovering unexpected patterns and proposed new method for discovering a minimal set of unexpected patterns that measures the interestingness. They also demonstrated strength of this approach experimentally. Tan, P. and Kumar, V. [33] examined various interestingness measures proposed in statistics and data mining literature.

During 2001, Hilderman, R. J. and Hamilton, H. J [12] focused on classifying interestingness measure and provided general overview of more successful and widely used interestingness measures from the literature that have been employed in data mining applications. Hilderman, R.J., Hamilton, H.J. [13] explored diversity of measures for ranking the interestingness summaries. Jaroszewicz, S. and Simovici [14] presented a new general measure of rule interestingness.

All the range in 2002, Keim, D.A. [17] proposed a classification of visualization techniques which is based on the data type to be visualized. Grinstein, G., Hoffman, P., & Pickett, R [8] described a set of benchmarking for visualization approaches. Collier, K., Medidi, M., & Sautter, D [3] based on their survey, expressed that the current visualization techniques have progressed dramatically in the past decade. Tan, P., Kumar, V., and

Srivastava J [34] presented an overview of various measures in the statistics, machine learning and data mining literature. Sahar S [30] introduced an approach that requires very little domain knowledge to eliminate the majority of the rules that are subjectively not interesting. Webb, G. I. and Brain D [38] addressed shortage of formal analysis about how to select useful interesting measures. Padmanabhan B. and Tuzhilin [25] addressed the problem of incorporating discovered contradicts into the belief system based on formal logic approach.

In the year 2003, Brijs T., Vanhoof K. and Wets G [1] provided an overview of existing measures of interestingness and divided them into objective and subjective measures. Dykes, J. and Mountain, D [5] described a novel representation designed for interactive graphical data analysis. E R Omiecinski [6] discussed interest measures for associations by applying downward closure property and described its importance.Robert Redpath and Bala Srinivasan [29] established suitable criteria for comparison of visualization. Wang, K., Jiang, Y. and Lakshmanan, L.V.S [37] studied important issues that are used in mining unexpected rules using dynamic entity. Hilderman, R. and Hamilton, H [11] evaluated many diversified measures used as heuristic measure of interestingness.

In the next year 2004, Jaroszewicz, S. and Simovici, D A [16] proposed a new definition of interestingness as the absolute difference between its support estimated from the data and from the Bayesian network. Lenca, P., Meyer, P., Vaillant, B., and Lallich, S. [18] expressed using probabilistic models and chain graphs instead of Bayesian is another research area. McGarry, K. and Malone, J [22] described goal driven method for subjective measures that are applied to measure the rules extracted from neural networks. X. Jin, Y. Zhou, and B. Mobasher [39] developed a unified framework that analyzes semantic relationships among the users and pages using probabilistic latent semantic analysis-PLSA.

During 2005, Furnkranz, J. and Flach, P. A [7] provided analysis of behavior of covering rule algorithms by visualizing their evaluation metrics and their dynamics and coverage space. Jaroszewicz, S. and Scheffer, T [15] studied the problem of discovering unexpected patterns in a databases. They formulated most interesting attribute sets and developed an algorithm. MCGarry, K. [21] presented a review of the available literature on the various measures devised for evaluating and ranking the discovered patterns produced by data mining process. Suzuki, E. and Zytkow, J.M [32] formalized the discovery of interesting exception rules as rule triplet discovery and categorized with literals.

All the range in 2006, Padmanabhan, B. and Tuzhilin [26] presented a new method for discovering a minimal set of unexpected patterns by combining two independent concepts of minimality and unexpectedness. Vasudha Bhatnagar, Ahmed Sultan Al-Hegami, and Naveen Kumar [35] studied novelty interestingness measure for discovered rules. They proposed a hybrid approach based on both objective and subjective measures. The qualification is performed objectively and a user involvement is sought for categorization of rules based on novelty index. Yao, Y. Y., Chen, Y. H., and Yang, X. D [40] suggested a critical review of rule evaluation for interestingness measures.

In the next year 2007, Heng-Soon Gan and Andrew [9] defined rescheduling stability quantitatively and have provided analytical mean for various heuristics. Rescheduling stability of heuristics is also important apart from effectiveness and efficiency. They extended empirical and analytical work on heuristic robustness. In their future research work, considering Spearman's foot rule, a measure of permutation disarray may shed some further light on heuristics.

During 2008, Vitaly Friedman [36] proposed USER approach that finds unexpected sequences and implication rules from data with user defined beliefs for mining unexpected behaviors from weblogs. They implemented by considering constraints both occurrences and semantics. Their experimental results with the belief basis have shown this approach is more suitable for web usage mining. As the unexpected behaviors impact the web usage analysis and many of the times the identification of unexpectedness depends on semantics and user beliefs, measure of unexpectedness elevated as an open research problem.

In the next year 2009, Michael Friendly [23] surveyed the visualization techniques from the deep roots to the current fruit. Rana Malhas and Zaher Al Aghbari [27] presented a new robust sensitivity measure of interestingness to discover interesting patterns based on background knowledge represented by a Bayesian network. To capture the sensitivity they defined a new method. Patterns that attain the highest sensitivity score are deemed interesting. Their experimental results triggered interesting future research paths towards automation of the process, quality, scalability and intelligence of the sensitivity measure.

The literature is evident that extending theory of interestingness for diversity measures is an open future research work. And also, the works in the literature triggered the present research towards combination of objective, subjective and heuristic measures as a unified measure for retaining the genuine interesting associated patterns. Some more authors articulated that the integration of visualization techniques along with interesting measures is another solid basis for future research.

III. ANALYSIS AND KNOWLEDGE REPRESENTATION SYSTEM (AKRS)

The associated patterns have been recognized by pattern discovery techniques must be analyzed to determine how those patterns can be used in investigating the web user usage behavior. In this process some of the patterns need to be deleted as they identified with less interest. To evaluate the interestingness of a pattern requires a set of popular measures and it becomes one of the central problems in the analysis phase. Besides, the representation of final knowledge understandable to every user is also an essential problem in knowledge representation. With these objectives the authors propose an Analysis and Knowledge Representation System (AKRS), the architecture of AKRS is as shown in figure 2.

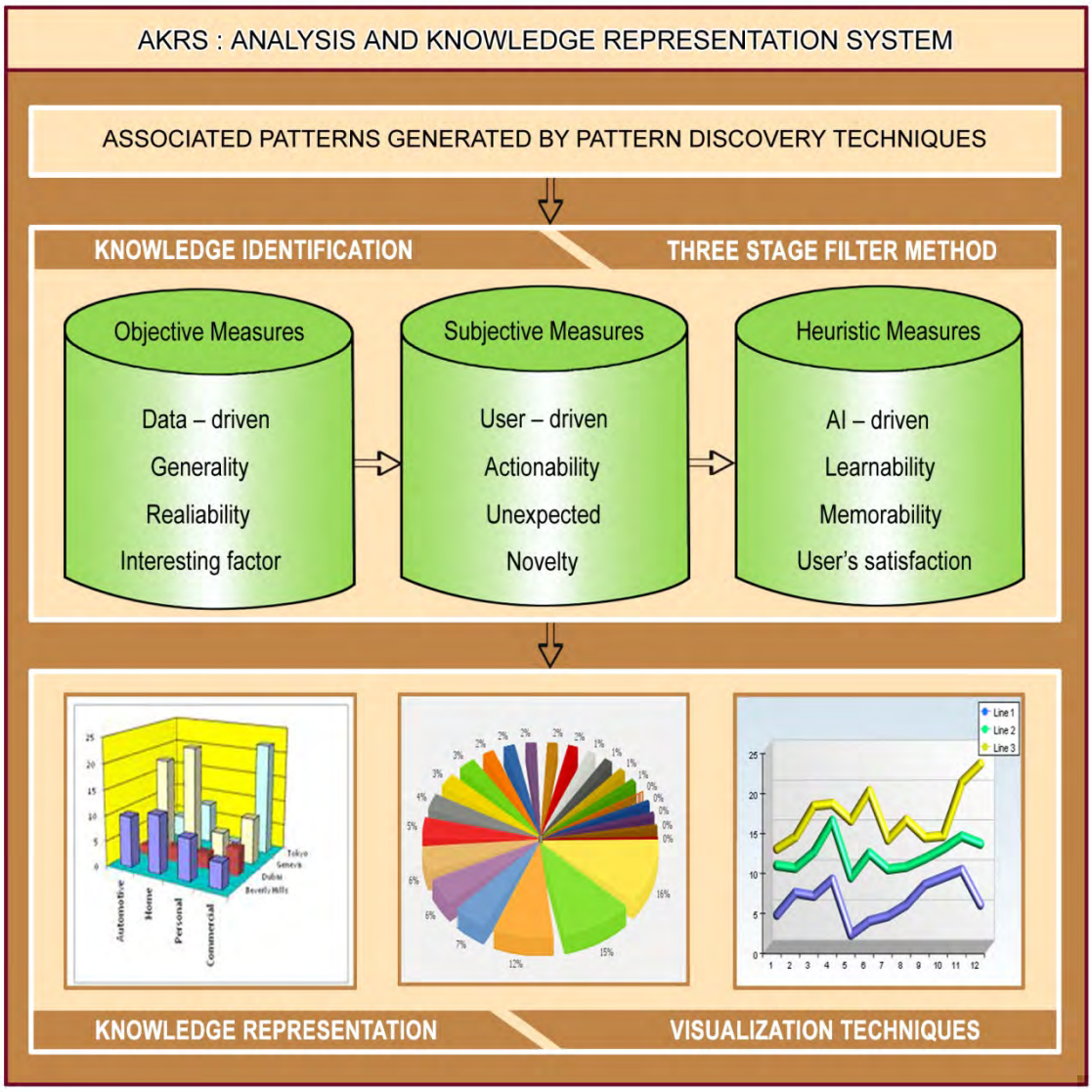

Figure 2. Architecture of Analysis and knowledge representation system

The AKRS, initially, identifies the knowledge inherited in associated patterns by using interestingness measures. These measures collectively identify the knowledge and helps in deleting less interest patterns. Later, it represents the knowledge by interpreting the interesting patterns using data visualization techniques to the level of desires and intentions of user.

A. *Knowledge Identification*

The knowledge identification is a process that evaluates the degree of interestingness of the associated patterns. The interestingness depends on quality of pattern, especially, cardinality, relativity and domain context. To integrate these qualities, selection of a number of good measures plays a vital role in identifying the actual knowledge. Towards this the proposed AKRS employs a three stage filter method which consists of objective, subjective and heuristic measures. It exploits the advantages of each measure to improve the identification of associated pattern with high interestingness.

The input for three stage filter AKRS is the associated set of patterns generated by the pattern discovery techniques, a sample output is as shown in figure 3.

{P5}, {P4, P5}, {P1, P4, P5}, {P3, P5}, {P1, P5}
{P4}, {P3, P4}, {P2, P3, P4}, {P1, P3, P4}, {P2, P4}, {P1, P2, P4}, {P1, P4}
{P3}, {P2, P3}, {P1, P2, P3}, {P1, P3}
{P2}, {P1, P2}

Figure 3. Set of associated patterns

1) Objective measures :

Objective measures serve as a first stage filter to evaluate the interestingness of associated patterns based on data driven method and they take only data cardinalities into account. These measures are basically application independent and require additional input from the user apart from the specified threshold to eliminate unrelated patterns. These measures extract meaningful patterns depending on the structure of the pattern.

The objective measures consider the probability and statistical values in filtering the patterns that do not meet the statistical strengths and retain the potential interesting patterns. The appropriate selection of objective measures in the web environment is made based on properties of the measures. The principles and three key properties are as shown in figure 4.

Property 1: M = 0 if P1 and P2 are statistically independent;

Property 2: M monotonically increases with P(P1P2) when P(P1) and P(P2) remain the same.

Property 3: M monotonically decreases with P(P1) (or P(P2)) when the rest of the parameters (P(P1P2) and P(P2) or P(P1)) remain unchanged.

Where M is a good measure, P1 & P2 are the pages of an associated pattern.

Figure 4. Set of key properties

Property1 is practically rigid as it states that an associated pattern which takes place by chance has zero interestingness value. In Property2, the higher support of P1P2 gives higher interestingness value when the support for P1 and P2 is fixed. Similarly in Property3, smaller support for P1 (or P2) shows more interestingness value when the supports for P1P2 and P2 (or P1) are fixed.

A good objective measure is able to distinguish between direct associated pattern and sub associated pattern. To identify these objective measures, properties of the objects are the key differentiators. These properties play a vibrant role in choosing good object measures as a significant number of objective measures provide conflicting information about interestingness of an associated pattern. However, a set of measures are also available to provide consistent and common statistical strengths about interestingness of an associated pattern.

In web environment and considering the properties of good measure, the following properties are considered by AKRS for validating the associated patterns.

- Generality
- Reliability

More so, the AKRS employs the Interest Factor (IF) that adjoins the potential interestingness patterns involving low support and also pay attention to the result cause by support.

A common methodology behind all these measures is developing a contingency table for set of associated patterns. This table tabulates the frequency counts of pages in associated patterns. The objective measures are derived based on the probability values represented in contingency table. A more practical approach is to determine the most appropriate measure by providing smaller set of contingency tables.

An example of 2X2 contingency table for an associated pattern {P1, P2} is as shown in table I.

TABLE I. 2 X 2 CONTINGENCY TABLE

	$P2$	$\overline{P2}$	
$P1$	f_{11}	f_{10}	f_{1+}
$\overline{P1}$	f_{01}	f_{00}	f_{0+}
	f_{+1}	f_{+0}	N

where,

P1, P2 denote the web pages in an associated pattern {P1, P2}

N denotes the total number of associated patterns

f1+ denotes the row sum of support count satisfying P1

f0+ denotes the row sum of support count not satisfying P1

f+1 denotes the column sum of support count satisfying P2

f+0 denotes the row sum of support count not satisfying P2

f11 denotes the number of associated patterns satisfying both P1 and P2

f10 denoted the number of associated patterns satisfying P1 but not P2

f01 denotes the number of associated patterns satisfying P2 but not P1

f00 denoted the number of associated patterns not satisfying both P1 and P2

a) Generality:

An associated pattern is general and it covers relatively large subset of a given weblog. It measures comprehensiveness of an associated pattern, that is, the fraction of all patterns in the weblog that matches the associated pattern. The associated pattern interestingness is directly proportionate to the coverage in weblog. Generality normally coincides with sub associated patterns since they tend to have greater coverage.

The AKRS identifies the below probability values to represent the generality of an associated pattern from the contingency table I.

- The coverage of P1 is derived as probability of P1, P(P1) = n(P1) / N.
- The coverage of P2 is derived as probability of P2, P(P2) = n(P2) / N.
- Then, the support, P(P1P2) represents the generality of the associated pattern {P1,P2}.

b) Reliability:

An associated pattern is reliable if the relationships described between the pages occur highly applicable for many cases. For example, an associated pattern is reliable if its predictions are highly accurate, and association is reliable if it has high confidence. Many measures from probability, statistics, and information retrieval have been proposed to measure the reliability of associated pattern.

The AKRS identifies the below probability values to represent the reliability of a associated pattern from the contingency table I.

- The coverage of P1 is derived as probability of P1, P(P1) = n(P1) / N.
- The coverage of P2 is derived as probability of P2, P(P2) = n(P2) / N.
- Then, the confidence, P(P1P2) represents the reliability of the associated pattern {P1, P2}.

c) Interest Factor:

The Interest Factor (I) facilitates a general and practical approach to automatically identifying interesting patterns. It compares the support of an associated pattern with a baseline support calculated under support-confidence frame work. This factor is defined as the proportion between the joint probability of two pages with respective to their expected probabilities under the individual page assumption.

$$I(P1, P2) = \frac{P(P1, P2)}{P(P1)P(P2)} \quad (1)$$

The interest factor is a non negative real number. It is 1 when both the pages in the associated pattern are statistically independent. This metric is desirable since it satisfies all three fundamental principles of measures. Additionally, this factor is closely related to the coefficient. For 2X2 contingency table:

$$I\ (P1, P2) = \frac{f_{11}\ N}{f_{1+}\ f_{+1}} \quad (2)$$

The correlation coefficient calculates the degree of linearity between two random pages. Mathematically, it is defined as covariance between two pages in an associated pattern, divided by their standard deviation:

$$\rho_{A\,B} = \frac{Cov\ (P1,\ P2)}{\sigma_{P1}\ \sigma_{P2}} \quad (3)$$

The range of $\rho_{A\,B}$ is located between -1 and +1.

For binary variables $\sigma_{P1} = \sqrt{p(1-p)}$, where $P \equiv P\ (P1) = f_{1+}/N$

For finite samples of entire population, the phi-coefficient and correlation coefficient $\rho_{A\,B}$ are same. Thus phi-coefficient :

$$\emptyset = \frac{f_{11}f_{00} - f_{10}f_{01}}{\sqrt{f_{1+}f_{0+}f_{+1}f_{+0}}} \quad (4)$$

The numerator equation further simplified as below:

$$= \frac{(f_{1+}f_{+1})I - f_{1+}f_{+1}}{\sqrt{f_{1+}f_{0+}f_{+1}f_{+0}}}$$

$$= \frac{(I-1)\sqrt{f_{1+}f_{+1}}}{\sqrt{f_{0+}f_{+0}}} \quad (5)$$

For high volume web log, the support count of a particular associated pattern is very low, $\frac{f_{1+}}{N} \ll 1$ and $\frac{f_{+1}}{N} \ll 1$. At this juncture, both $\frac{f_{0+}}{N}$ and $\frac{f_{+0}}{N}$ are close to 1. In addition highly correlated associated patterns have $I \gg 1$. Based on the approximation, the equation 5 re-written as:

$$\emptyset \approx I\sqrt{\frac{f_{1+}f_{+1}}{N^2}} \quad (6)$$

$$= \frac{Nf_{11}}{f_{1+}\ f_{+1}}\sqrt{\frac{f_{1+}f_{+1}}{N^2}}$$

$$= \sqrt{\frac{Nf_{11}}{f_{1+}\ f_{+1}} \cdot \frac{f_{11}}{N}}$$

$$\emptyset = \sqrt{I \times \frac{f_{11}}{N}} \quad (7)$$

From equation 6 it is evident that a better interesting measure shown from statistical correlation, in the area of low support and high interest value is:

$$IS = \sqrt{I \times \frac{f_{11}}{N}} \quad (8)$$

IS has many desirable properties despite violating the first principle of objective measure. First of all, it contains a product of two important quantities, interest factor and support. In other words, this measure takes into account both interestingness and significance of a pattern. In addition, the confidence is also taken enough importance.

2) Subjective measures:

Subjective measures serve as a second stage filter to evaluate the interestingness of associated patterns based on user-driven method and it takes only users domain information into account. These measures basically depend on application context and require user understandability of the domain to select desired interesting patterns. These measures operate by comparing the user's beliefs against the patterns discovered by the mining algorithm. These measures do not depend on the structure of the pattern but depend on class of users who examine the pattern.

The subjective measures cannot be represented by simple mathematical formulae as the user knowledge needs to represent in various forms. The subjective measures impart formal specification of user knowledge, user interactiveness and desires of the user to identify the degree of interestingness. To attain this subjective measures apply the process of knowledge acquisition, inductive learning and conditional probabilities. In the web context the AKRS adopts the following measures to find out the interesting associated patterns.

- Actionability
- Unexpectedness

a) Actionability:

A pattern is actionable in a domain if it enables decision making about future actions in the domain. Actionability is an approach to define interestingness of pattern based on the action of the user. It is an important subjective measure since users are mostly interested in the knowledge that permits them to do their jobs better by taking specific actions in response to the newly discovered knowledge.

The actionable associated patterns allow the user to do favorable actions. In the web environment, most of the websites deal with a number of categories along with their activities. The actions of the user in these categories resemble the interestingness of the user. Thus, actionability helps in identification of degree of interestingness.

The subjective measures discover the interestingness of an associated pattern by considering the domain knowledge and become a central problem in the analysis. To arrive this, the AKRS present an approach to identify the actionability based on the domain knowledge.

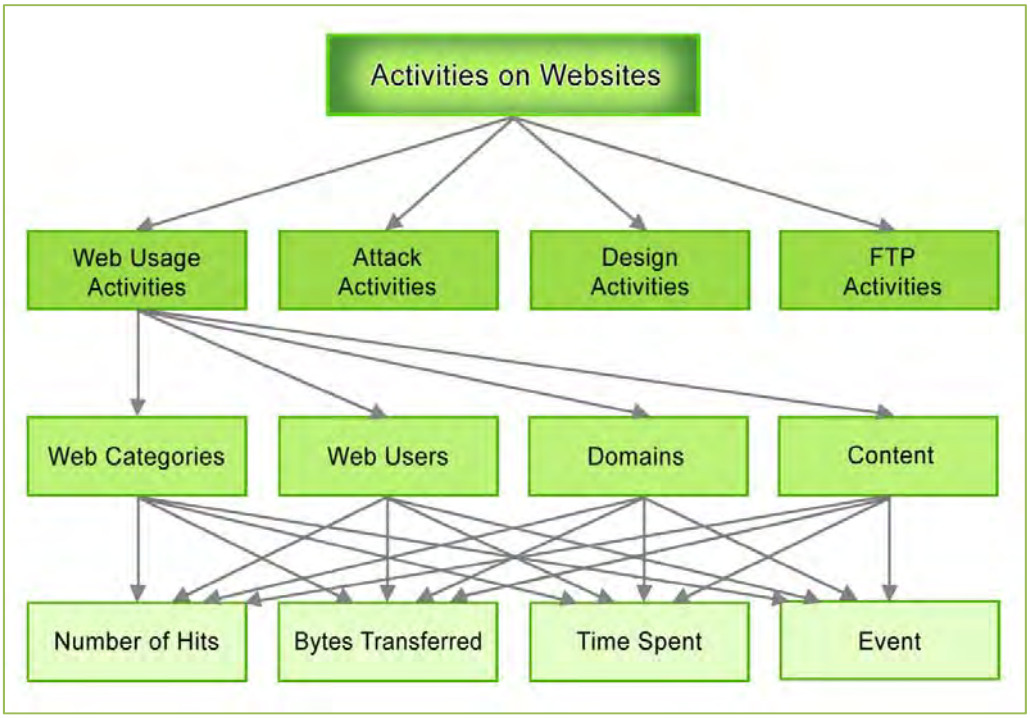

Figure 5. Web knowledge tree (WK – Tree)

The WK – tree consists of root node, non leaf nodes and leaf nodes. At root node all actions of the websites are denoted. Top level non leaf nodes represents general web activities like advertising, user related actions, category wise actions and so on. Similarly in the next level specific activities are denoted. At leaf nodes represents statistics like time spent, download bytes etc. WK – Tree is shown in figure 5.

Based on the web user's desires the associated pattern is assigned to a node of WK – Tree. This process is stopped on completing assignment of associated patterns to the nodes. Later the web miner traverses the WK – Tree to take a decision on identifying the interestingness of associated pattern.

The WK – Tree yields high modularity as it follows top down approach. It allows building the set of sub activities in several stages as it is a directed acyclic graph. The maintenance of WK – Tree is easy even the activities are changed over a time.

b) Unexpectedness:

An associated pattern is unexpected if it intuitively contradict an existing user's belief and surprises the user's expectations. The user belief notices the unexpected pattern from the associated patterns that are authenticated by objective measures. The set of prior user beliefs is an essential component to characterize the unexpectedness as it is inherently biased. The degree of unexpectedness is estimated in terms of deviations between discovered associated patterns and user belief system, including changes to the degrees of these beliefs. Thus, the prior user's beliefs are mandatory to update in the light of new evidences and facts.

The AKRS constitutes its belief system with a set of web user specifications and interactive feedback in the web domain where each belief is defined as a logical statement. Further, it considers the concept of logical contradiction to find unexpectedness by estimating the deviations between an associated pattern and belief system. This concept utilizes the aggregation of web knowledge in its process. Finally, it captures unexpected patterns with high degree by taking into account the term "intersection of associated and belief patterns".

Both the belief and associated patterns carries almost equal statistical strength as the associated patterns is validated by objective measures. Thus, the intersection of associated and belief patterns derives the logical contradiction that identifies the characterization of unexpected pattern. The large intersection value yields high degree of unexpected patterns.

For a given associated pattern {P1, P2} where P1 and P2 are atomic pages and {X, Y} is a belief pattern from the belief system. The pattern {P1, P2} is unexpected with respect to the belief {X, Y} on the web knowledge W then the AKRS derives the following forms:

- P1 AND X holds on a statistically large subset of patterns in W. The intersection of an associated pattern with respect to user belief pattern defines the subset of patterns in W such that the user belief pattern and the associated pattern are true.
- P2 AND Y $\neq$ FALSE. This condition enforces the limitation that P2 and Y logically contradict each other.
- The associated pattern {P1, {X , P2}} holds, since previous condition, P2 and Y are logically contradict each other, thus, {P1, {X , $\neg$ Y}} holds.

The concept of logical contradiction of AKRS initially, starts with a set of initial beliefs to begin the discovery of all associated patterns in the form of {X, {P1, P2}} such that P2 contradicts the belief {X, Y}. Subsequently, for each belief it incrementally adds all large associated patterns that facilitate to bring out the unexpected patterns. Later, it identifies sophisticated associated patterns based on the intersection value of the associated and belief patterns. The comparison property of logical contradiction helps in finding the high degree of unexpected patterns in the form of {X1, {P1, P2}}, where X1 $\subset$ X. The logical contradiction of AKRS extracts valuable domain knowledge as it considers aggregation of web knowledge. The amalgamation of user belief and logical contradiction drastically reduces the low degree of unexpected patterns efficiently.

In the web knowledge scenario, a few associated unexpected patterns are actionable and non-actionable at a time. Although the two types of interestingness are independent of each other, the greater part of actionable patterns is unexpected and the greater part of unexpected patterns is actionable. Therefore, unexpectedness is a good approximation for actionability and actionability is a good approximation for unexpectedness. Thus, AKRS believe that, both actionability and unexpectedness are important subjective measures in the web scenario.

3) Heuristic measures:

The heuristic measures are the final stage filters of AKRS to evaluate the interestingness of associated patterns based on artificial intelligence that mimic the experts. Basically, these measures are self learning techniques, build on knowledge drawn from experience. This evolution is the most informal and plays a vital role in validating the associated patterns quickly. In fact, it is a proven usability engineering method with high cost benefit. Initially, heuristic techniques obtain overall understanding and general scope of the associated patterns.

Subsequently, they focus on web usage analysis to determine the interestingness by validating the associated patterns.

The heuristic measures of AKRS works on utility principles like learnability, memorability, users' satisfaction and so on from associated patterns. The principle of learnability makes the system to learn the functionality and behavior in a simplest manner. Memorability crafts the system to remember the functionality with no effort so as to avoid repeated learning. Users' satisfaction quantifies the system through which the user gets agreeable to utilize. This approach captures the inherent relations among the associated patterns based on probability inference method.

The utility is identified through defining L heuristic factors and correlate them with each page in an associated pattern, where L is the length of the pattern. The heuristic factors acquire the learning capability and memorability by estimating the conditional probability distribution on associated patterns that are validated by previous filters of AKRS. Moreover, these heuristics identify inherent knowledge and helps in validating associated patterns competently.

For example, for a given pattern {P1, P3, P4} it is necessary to define three heuristic factors L1, L2, L3 as the length of the pattern is three. The heuristic factor L1 represents specification of a product details in page P1, L2 represents details of the prices of product in page P3 and L3 represents shipping details of product in page P4. These heuristic factors combinedly represent the overall interestingness of the associated pattern.

Before applying the heuristic measures on associated patterns it is necessary to bring in the mathematical model to estimate the conditional probability distribution. The probability definitions are shown below,

- $P(EP_i)$ represents the probability of associated patterns ending with a particular page EP_i
- $P(L_k \mid EP_i)$ represents probability distribution of an associated pattern ending with a specific page with a heuristic factor L_k
- $P(p_j \mid L_k)$ represents the conditional probability distribution of pages on a specific heuristic factor L_k

Based on these probability distributions the heuristic measure can be articulated as below,

- Pick the associated pattern ending with a particular page EP_i by a probability $P(EP_i)$
- Choose heuristic factor L_k by a probability $P(L_k \mid EP_i)$
- Create a page p_j by a probability $P(p_j \mid L_k)$

Finally, the probability model translates the occurrence of probability of an observed pair (EP_i, p_j) by taking on the heuristic factor L_k results as below,

$$P(EP_i, p_j) = P(EP_i).P(p_j \mid EP_i) \quad (9)$$

Where,

$$P(p_j/EP_i) = \sum_{l \in L} P(p_j/L) \ . P(L/EP_i) \quad (10)$$

These expressions characterize a set of heuristic factors based on length of the associated pattern. The factors acquire the knowledge, and the degree is determined by probability inference process. The high degree infers associated patterns with genuine interestingness.

B. Knowledge Representation

The knowledge representation is a process that presents the discovered knowledge graphically through powerful visualization techniques. These techniques visualize the meaning of knowledge which is in the abstract form. The insight of knowledge visualization is to address every stakeholder intentions in detail at different levels. In addition, knowledge representation improves the transfer of knowledge from the mined results to understanding level of common user. To transfer the knowledge effectively, both presentation and functionality necessarily go hand in hand.

The AKRS chooses a set of visualization techniques that includes standard and specialized tools, to transfer the knowledge effectively. These tools give the importance to attractive, elegant and descriptive presentation along with functionality. The AKRS uses standard visualization tools such as line, column, bar and pie graphs to demonstrate the experimental results. In addition, to meet the challenges in knowledge representation like identifying relevant information, finding the depth of information and achieving the visualization competency, the AKRS also designates the recent knowledge visualization techniques as below,

- Multidimensional visualizations
- Specialized hierarchical visualizations

Multidimensional visualization techniques enable visual comparison and contrast the knowledge of one category patterns to another category pattern. They are also used to discover intrinsic relations between them. Specialized hierarchical visualization techniques facilitate enhancement and exploitation of the inherent structure of the associated patterns. These can also be used to explore the relationships between the hierarchies of associated patterns.

The representation techniques employed by AKRS are able to present the knowledge that is inherited in the discovered associated patterns effectively. Through these techniques a common man can understand the underlying knowledge. Thus, these techniques allow users to analyze the results in simple manner. Further, they visually present complete web usage statistics that helps in taking the right decisions about web user usage easily.

IV. EXPERIMENTAL ANALYSIS

The proposed AKRS is evaluated by conducting a series of experiments on three different sizes of weblogs over a period of six months. The universal parameters of sessions are set to have hours from 00 hrs to 23 hrs, day from Monday to Sunday. Due to privacy issues of the user, other sensible information is not included. All the three weblog files are subjected through each phase of web usage mining. Thereby, the associated patterns are generated by pattern discovery techniques for each weblog and given as input for proposed AKRS. The three phase filtering system AKRS validated the associated patterns with objective, subjective and heuristic measures. The performance results of AKRS measures on three webs are recorded and shown below.

The Interest factor is a key objective measure and its performance is evaluated based on cumulative probability of support and confidence. A graph is plotted between accuracy rate of interest factor and cumulative probability; results are shown in figure 6. The results are evident with the theoretical analysis of cumulative probability.

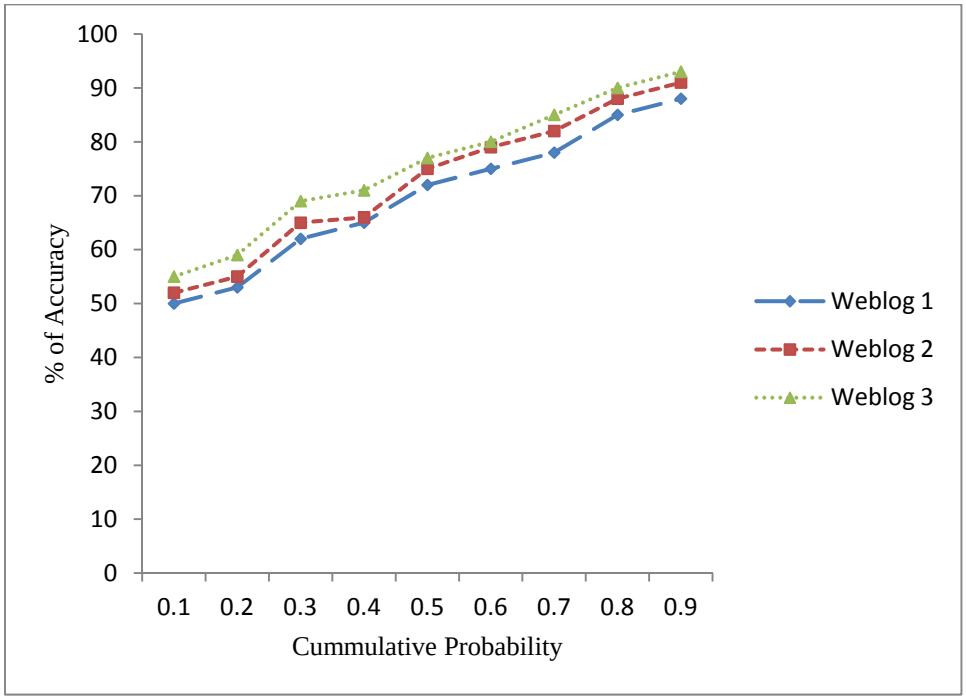

Figure 6. Accuracy performance of Interest factor

A) The subjective measures are evaluated by considering 20 beliefs for each weblog. The actionable associated patterns that allow the user to do favorable actions are identified based on WK – Tree. A graph is plotted between actionability of associated patterns with number of beliefs as shown in figure7.

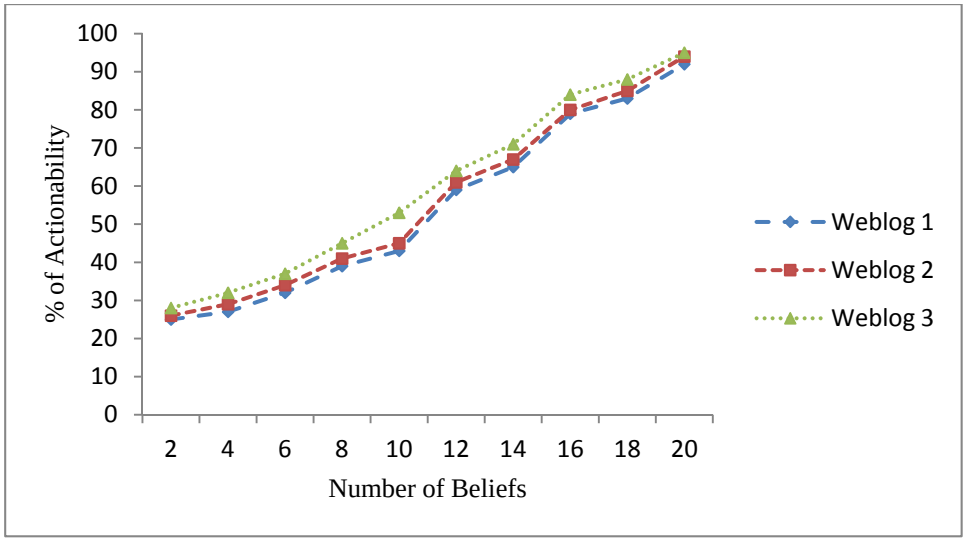

Figure 7. Rate of actionability against number of beliefs

B) The beliefs are also utilized for extracting unexpected associated patterns based on logical contradiction. A graph is plotted between unexpectedness of associated patterns against to beliefs the results are shown in figure 8.

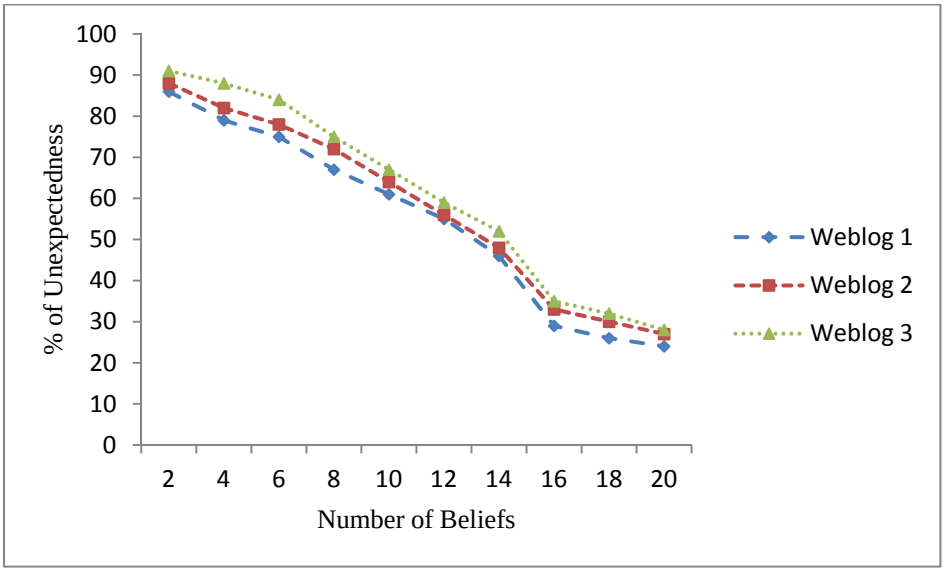

Figure 8. Rate of unexpectedness against number of beliefs

C) Heuristic factor performance is evaluated based on learnability with respect to number of iterations and a graph is plotted as shown in figure 9. The experimental study indicates that learnability increases with number of iterations for all the three weblogs.

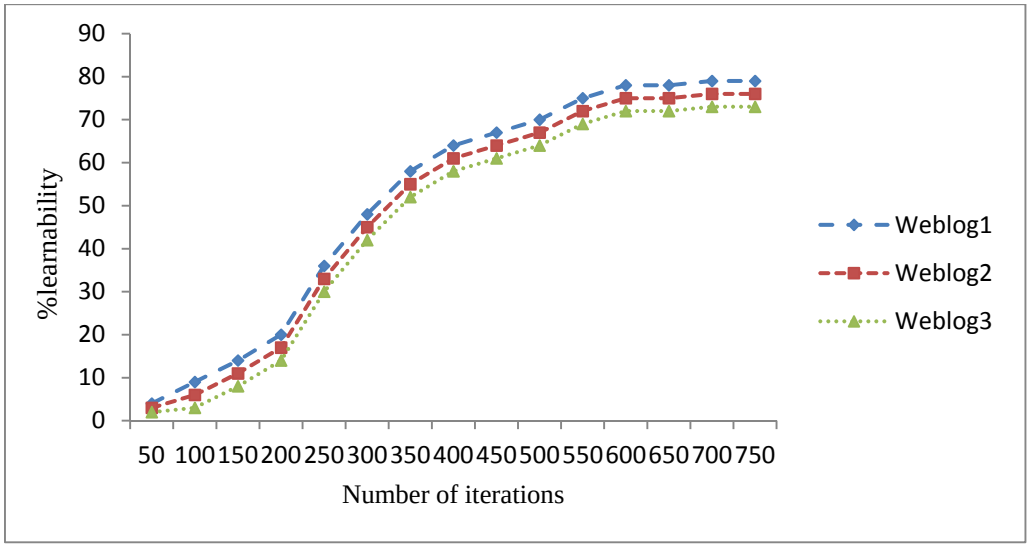

Figure 9. Heuristic factor learnability

D) All the three weblogs are mined and the associated patterns are validated using AKRS three phase filtering system. Initially, associated patterns are validated by Interest factor along with statistical strengths. The validation is continued based on the user beliefs. Finally, heuristic functions are used in finalizing the associated patterns, combinedly all the results are tabulated as shown in table II.

TABLE II. FINALIZED ASSOCIATED PATTERNS FOR THREE WEBLOGS

APAS / WEBLOG	Time	All Patterns	Associated patterns	Objective measures	Subjective measures	Heuristic Measures	Finalized patterns
Weblog1	T1	18234	1556	125	144	59	1228
	T2	17811	1489	112	145	56	1176
	T3	17262	1405	102	123	62	1118
Weblog2	T1	150435	1298	112	135	47	1004
	T2	14045	1274	124	119	54	977
	T3	14100	1156	134	117	61	844
Weblog2	T1	12651	1096	149	105	40	802
	T2	12569	1010	158	89	46	717
	T3	12032	952	139	77	37	699

V. SUMMARY

In this paper the proposed AKRS is designed for knowledge identification and knowledge representation. The objective, subjective and heuristic measures combindly used for knowledge identification. The interest factor of objective measures is able to eliminate poorly correlated associated patterns effectively. The formulation of user beliefs in the subjective measures evaluates the associated patterns and extracts useful and effective unexpected actionable patterns. The utility principle of heuristic measure takeout the semantic meaning inherited in the associated patterns and enhances learnability along with memorability. The visualization techniques make the common man to understand the final knowledge clearly and help in the analysis of the associated patterns.

ACKNOWLEDGMENT

The authors record their acknowledgements to the authorities of Shri Vishnu Engineering College for Women, Bhimavaram; Andhra University, Visakhapatnam and Acharya Nagarjuna University, Guntur for their constant support and cooperation.

REFERENCES

[1] Brijs, T., Vanhoof, K., Wets, G., “Defining interestingness for association rules”, International Journal of Information Theories and Applications, 10(4), pp: 370–376, 2003.
[2] Bayardo, R. J., Agrawal, R., “Mining the most interesting rules”, In Proceedings of the 5th International Conference on Knowledge Discovery and Data Mining (KDD-99). San Diego, CA., pp: 145–154, 1999.

[3] Collier, K., Medidi, M., Sautter, D., "Visualization in the knowledge discovery process", In Information visualization in data mining and knowledge discovery, California, Morgan Kaufmann, 2002.
[4] Colton, S., Bundy, A., "On the notion of interestingness in automated mathematical discovery", International Journal of Human-Computer Studies 53, pp: 351–375, 2000.
[5] Dykes, J., Mountain, D., "Seeking structure in records of spatio-temporal behaviour: visualization issues, efforts and applications", Computational Statistics and Data Analysis 43, pp: 581–603, 2003.
[6] E. R. Omiecinski, "Alternative Interest Measures for Mining Associations in Databases", IEEE Trans. on Knowledge and Data Engineering 15, pp: 57–69, 2003.
[7] Furnkranz, J., Flach, P. A., "ROC 'n' rule learning: Towards a better understanding of covering algorithms" Mach. Learn. 58, (1), pp: 39–77, 2005.
[8] Grinstein, G., Hoffman, P., Pickett, R., "Benchmark development for the evaluation of visualization for data mining", In Information visualization in data mining and knowledge discovery, California, Morgan Kaufmann, 2002.
[9] Heng-Soon, Gan., Andrew, "Wirth Heuristic stability: A permutation disarray measure", Computers & Operations Research, Volume 34, Issue 11, pp: 3187-3208, 2007.
[10] Hilderman, R., Hamilton, H., "Applying objective interestingness measures in data mining systems". In Fourth European Symposium on Principles of Data Mining and Knowledge Discovery (PKDD 2000), Springer Verlag, pp: 432–439, 2000.
[11] Hilderman, R. J., Hamilton, H. J., "Measuring the interestingness of discovered knowledge: A principled approach", Intelligent Data Analysis 7(4), pp: 347–382, 2003.
[12] Hilderman, R. J., Hamilton, H. J., "Knowledge Discovery and Measures of Interest", Kluwer Academic, Boston, MA, 2001.
[13] Hilderman, R. J., Hamilton, H. J., "Evaluation of interestingness measures for ranking discovered knowledge", Lecture Notes in Computer Science 2035, pp: 247–259, 2001.
[14] Jaroszewicz, S., Simovici, D. A., "A general measure of rule interestingness", In Proceedings of the 5th European Conference on Principles of Data Mining and Knowledge Discovery (PKDD 2001). Freiburg, Germany, pp: 253–265, 2001.
[15] Jaroszewicz, S., Scheffer, T., "Fast discovery of unexpected patterns in data, relative to a bayesian network", Proc. 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Chicago, Illinois, USA, pp:118–127, 2005.
[16] Jaroszewicz, S., Simovici, D. A., "Interestingness of frequent itemsets using Bayesian networks as background knowledge", in Proceedings of the 2004 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, WA, USA, pp: 178–186, 2004.
[17] Keim, D.A., "Information visualization and visual data mining", IEEE Transactions On Visualization and Computer Graphics 7, pp: 100–107, 2002.
[18] Lenca, P., Meyer, P., Vaillant, B., Lallich, S., "A multicriteria decision aid for interestingness measure selection", Tech. Rep. LUSSI-TR-2004-01-EN, May 2004.
[19] Liu, B., Hsu, W., Chen, S., Ma, Y., "Analyzing the subjective interestingness of association rules", IEEE Intelligent Systems 15, pp: 47–55, 2000.
[20] Ludwig, J., Livingstone, G., "Whats new? using prior models as a measure of novelty in knowledge discovery", in 'Proceedings of the 12th IEEE Conference on Tools with Artificial Intelliegnce', pp: 86–89, 2000.
[21] MCGarry, K., "A survey of interestingness measures for knowledge discovery", Knowledge Engineering Review 20, (1), pp: 39–61, 2005.
[22] McGarry, K., Malone, J., "Analysis of rules discovered by the data mining process", in Applications and Science in Soft Computing Series: Advances in Soft Computing, Springer-Verlag, pp: 219–224, 2004.
[23] Michael Friendly, "Milestones in the history of thematic cartography, statistical graphics, and data visualization", 2009.
[24] Padmanabhan, B., Tuzhilin, A., "Small is beautiful: Discovering the minimal set of unexpected patterns", In Proceedings of the 6th International Conference on Knowledge Discovery and Data Mining (KDD 2000)., Boston, MA., pp: 54–63, 2000.
[25] Padmanabhan, B., Tuzhilin, A., "Knowledge refinement based on the discovery of unexpected patterns in data mining", Decision Support Systems 33, pp: 309–321, 2002.
[26] Padmanabhan, B., Tuzhilin, A., "On characterization and discovery of minimal unexpected patterns in rule discovery", IEEE Transactions on Knowledge and Data Engineering, Vol. 18, No. 2, pp: 202–216, 2006.
[27] Rana Malhas, Zaher Al Aghbari, "Interestingness filtering engine: Mining Bayesian networks for interesting patterns", Expert Systems with Applications, Volume 36, Issue 3, Part 1, pp: 5137-5145, April 2009.
[28] Robert J. Hilderman, Regina, Saskatchewan, "Assessing the Interestingness of Discovered Knowledge Using a Principled Objective Approach", Philadelphia, Pennsylvania, USA., ACM 1595934405/06/0008, 2006.
[29] Robert Redpath, Bala Srinivasan, "Criteria for a Comparative Study of Visualization Techniques in Data Mining", Proceedings of the IEEE 3rd International Conference On Intelligent Systems Design and Application, Tulsa, USA,(ISDA 2003), ISBN: 3540404260 Springer-Verlag, Berlin, 2003.
[30] Sahar, S., "On incorporating subjective interestingness into the mining process", in Proceedings of the 2002 IEEE International Conference on Data Mining (ICDM 2002), Maebashi City, Japan, pp: 681–684, 2002.
[31] Shen, Y. D., Zhang, Z., Yang, Q., "Objective-Oriented utility-based association mining", In Proceedings of the 2002 IEEE International Conference on Data Mining (ICDM '02)., Maebashi City, Japan, pp: 426–433, 2002.
[32] Suzuki, E., Zytkow, J.M., "Unified algorithm for undirected discovery of exception rules", International Journal of Intelligent Systems, Vol. 20, No. 7, pp: 673–691, 2005.
[33] Tan, P., Kumar, V., "Interestingness measures for association patterns: A perspective", Tech. Rep. 00-036, Department of Computer Science, University of Minnesota, 2000.
[34] Tan, P., Kumar, V., Srivastava J., "Selecting the right interestingness measure for association patterns", In Proceedings of the 8th International Conference on Knowledge Discovery and Data Mining (KDD 2002), Edmonton, Canada., pp:32–41, 2002.
[35] Vasudha Bhatnagar, Ahmed Sultan Al-Hegami, Naveen Kumar, "Novelty as a Measure of Interestingness in Knowledge Discovery", International Journal of Information and Communication Engineering, 2006, pp:320-325, 2006.
[36] Vitaly Friedman, "Data Visualization and Infographics" in: Graphics, Monday Inspiration, January 14th, 2008.
[37] Wang, K., Jiang, Y., Lakshmanan, L.V.S., "Mining unexpected rules by pushing user dynamics", Proc. 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington DC, USA, pp: 246–255, 2003.
[38] Webb, G. I., Brain, D., "Generality is predictive of prediction accuracy", In Proceedings of the 2002 Pacific Rim Knowledge Acquisition Workshop (PKAW 2002), Tokyo, pp: 117–130, 2002.
[39] X. Jin, Y. Zhou, B. Mobasher, "Web usage mining based on probabilistic latent semantic analysis", in KDD '04: Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining., New York, NY, USA: ACM Press, pp: 197-205, 2004.
[40] Yao, Y. Y., Chen, Y. H., Yang, X. D., "A measurement-theoretic foundation of rule interestingness evaluation", In Foundations and Novel Approaches in Data Mining, T. Y. Lin et al., Eds. Springer-Verlag, Berlin, pp: 41–59, 2006.

AUTHORS PROFILE

Prof. V.V.R. Maheswara Rao received his Master of Computer Applications degree from Osmania University, Hyderabad, India. He is working as Professor in the Dept of Computer Applications at SVECW, Bhimavaram, AP, India. He is currently pursuing his Ph.D. in Computer Science Engineering at Acharya Nagarjuna University, Guntur, India. His Research interests include Webmining, Artificial Intelligence. He is life member of CSI, CI and ISTE.

Dr. V. Valli Kumari holds a Ph.D. degree in Computer Science and Systems Engineering from Andhra University Engineering College, Visakhapatnam and is presently working as Professor in the same department. Her research interests include Security and privacy issues in Data Engineering, Network Security and E-Commerce. She is a member of IEEE and ACM.