Sādhanā Vol. 30, Parts 2 & 3, April/June 2005, pp. 291–309. © Printed in India

Interestingness of association rules in data mining: Issues relevant to e-commerce

RAJESH NATARAJAN[1] and B SHEKAR[2]

[1]IT & Systems Group, Indian Institute of Management Lucknow (IIML),
Prabandh Nagar, Off Sitapur Road, Lucknow 226 013, India
[2]Quantitative Methods and Information Systems (QMIS) Area, Indian Institute of Management Bangalore (IIMB), Bannerghatta Road, Bangalore 560 076, India
e-mail: rajeshn@iiml.ac.in; shek@iimb.ernet.in

Abstract. The ubiquitous low-cost connectivity synonymous with the internet has changed the competitive business environment by dissolving traditional sources of competitive advantage based on size, location and the like. In this level playing field, firms are forced to compete on the basis of knowledge. Data mining tools and techniques provide e-commerce applications with novel and significant knowledge. This knowledge can be leveraged to gain competitive advantage. However, the automated nature of data mining algorithms may result in a glut of patterns – the sheer numbers of which contribute to incomprehensibility. Importance of automated methods that address this immensity problem, particularly with respect to practical application of data mining results, cannot be overstated. We first examine different approaches to address this problem citing their applicability to e-commerce whenever appropriate. We then provide a detailed survey of one important approach, namely interestingness measure, and discuss its relevance in e-commerce applications such as personalization in recommender systems. Study of current literature brings out important issues that reveal many promising avenues for future research. We conclude by reiterating the importance of post-processing methods in data mining for effective and efficient deployment of e-commerce solutions.

Keywords. Data mining; association rules; post-processing; interestingness measures; subjective/objective interestingness.

1. Introduction

Emergence of the internet as a global information superhighway has contributed to fundamental changes in many spheres of our daily life. One area that has been profoundly affected by the internet revolution is ‘business.’ Electronic commerce (Kalakota & Whinston 1999), the buying and selling of information, products and services via computer networks and more specifically the internet, has revolutionized the way business transactions are conducted across the world. e-Commerce has cut across traditional competitive advantages based on physical features such as firm size, location, employee strength, etc. All firms that have a presence on the internet can now compete on equal terms. This has created a level-playing field among firms (Burt & Sparks 2003). Inexpensive connectivity, accessibility and low entry/exit barriers characteristic to e-commerce, are some factors instrumental in the creation of this level-playing field. Today, businesses consider presence on the internet as a prime necessity – something that is taken for granted.

e-Commerce systems provide a firm with novel opportunities for both, selling its products and understanding its customers' behaviour. Virtual market places offer a cheap and effective avenue for showcasing a firm's products in accordance with its customers' requirements. Any e-commerce system generates large amounts of customer behaviour data. This is a by-product of its ease of use and its unrestricted accessibility to customers. In addition, certain categories of data, previously considered inaccessible, can now be recorded easily in an unobtrusive manner. For example, an e-commerce system can record all actions of customers visiting a virtual store right from the point of entry to exit. Use of "cookies", user-identification logins and other innovative methods can track the activities of the same customer over a period of time. Burt & Sparks (2003) stress the utilization of customer data for improving the "stickiness" of web stores. Repeat visits and long stays are desirable in the virtual world. The importance of leveraging "data" for creating a 'personal' environment in e-commerce systems (Geoffrin & Krishnan 2003) cannot be overemphasized, since there is no direct interaction with customers. Here we are mainly concerned with e-commerce systems that undertake B-to-C transactions. Hence our discussions will be limited to the same.

Any firm can innovatively use the large amount of transactional data generated by an e-commerce system, to create a personalized environment for its customers. This is because hidden in the operational data is knowledge about a firm – its business processes, customers and environment. Data mining aims to uncover trends and patterns, that would otherwise remain buried and therefore of no consequence. Kohavi & Provost (2001) have identified some key factors that may lead to a high success rate of e-commerce-related data mining applications. These include accessibility of data with rich descriptions, inexpensive collection of large volumes of data, controlled and reliable data collection, direct evaluation of data mining results and ease of integration with e-commerce systems. According to them (Kohavi & Provost 2001), any data mining application will rarely support all these factors simultaneously. For example, in a retail outlet, it might be easy to record customer purchase transactions while getting access to personal details of the customer might be more difficult. Thus, linking purchase transactions with customer profiles may not be a straightforward task. This might have to be done on an off-line basis. Thus, the store might miss an opportunity to cross-sell and up-sell products. On the other hand, the hurdles in e-commerce applications are significantly lower. To take the same example, a virtual store will find it very easy to link customers with their purchase transactions. This is especially true if the store insists on the user logging on the store using a pre-defined identification. Data collected by e-commerce systems are orders of magnitude larger than data collected by traditional means. It is not possible to manually analyse such magnitudes of data and even semi-automated methods become unwieldy after the size of data crossing a threshold. This calls for automated methods and thus data mining applications have increasingly proliferated e-commerce systems in novel ways. In some e-commerce applications like web merchandising, it has become essential to use data mining in order to increase system effectiveness for targeting customers.

Data mining results are expressed in the form of 'patterns' (Fayyad *et al* 1996). These patterns represent novel, hidden and previously implicit knowledge that may be used for increasing sales revenue, etc. in a profitable manner. However, effective application of knowledge represented by data mining patterns is impeded by the 'glut' of patterns generated by data mining algorithms (Silberschatz & Tuzhilin 2001). The numbers of these generated patterns are so large that manual inspection and analysis is impractical if not impossible. In addition, most of these mined patterns represent strong domain facts. Such facts are obvious to a domain expert since they represent common place knowledge. Researchers in the data mining community have acknowledged the importance of addressing this understandability problem (Kohavi & Provost 2001). This paper presents a survey of current approaches to addressing the understandability problem with appropriate references to e-commerce applications. In particular, we focus on one important approach, namely the use of interestingness measures to rank association rules. Association rules are implication rules that inform the user about items most likely to be purchased by a customer during a visit to the retail store. We concentrate on association rule mining since it features as one of the main data mining techniques used in e-commerce applications such as personalization applications, collaborative filtering and recommender systems. Geyer-Schulz & Hahsler (2002) describe a recommender algorithm that uses ARs derived from past purchases, for making recommendations to new anonymous customers. The main advantages of ARs are simplicity, intuitiveness and freedom from model-based assumptions. In addition, ARs are more general than other patterns like decision trees and classification trees. This is due to lack of constraint on the cardinality of attributes that an antecedent/consequent of an AR can contain.

This survey is organized as follows. After introducing of the post-processing problem, we present a short overview of the various post-processing techniques present in data mining and their relevance in e-commerce applications. We then concentrate on an important post-processing method namely, usage of “interestingness measures” for ranking ARs. After classifying interestingness measures, we allude to their various advantages and limitations including those in the e-commerce context. In particular, we discuss the utility of interestingness measures in recommender systems for personalization applications, an important e-commerce application. This naturally leads us to a discussion of, some important issues relevant to e-commerce applications, and future research issues. Finally we conclude the survey with a summary.

2. Post-processing in data mining

Personalization or one-to-one marketing is the delivery of a targeted solution to a customer by using the customer’s information such as likes, dislikes, preferences, etc (Murthi & Sarkar 2003). In the e-commerce domain personalization takes various forms such as dynamic web-content presentation, purchase recommendations and targeting of advertisements. Utilization of customer information for effective and efficient personalization has become an important business problem in e-commerce applications. One important task in personalization applications is related to the construction of profiles for individual customers or customer segments. This can be done by rule discovery methods like AR mining methods (*A priori* and related algorithms) (Agrawal *et al* 1993), classification algorithms like CART (Brieman *et al* 1984) C 4·5 (Quinlan 1993), and the like. A set of relevant rules pertaining to a customer then constitutes his/her profile. One serious problem with many of the rule mining methods is the generation of a large number of patterns.

This immensity in the number of generated patterns results in two kinds of related problems. The sheer numbers of mined rules render manual inspection practically infeasible. It also increases the difficulty in interpreting the results and obtaining a holistic picture of the domain. An equally or possibly more important issue concerns the ‘quality’ of the mined rules. Rules such as “age $= 5 \rightarrow$ unemployed” and “Bread $\rightarrow$ Butter”, while being statistically valid in a database are obvious since they are commonplace knowledge. In addition, such facts form a core component of the user's domain knowledge due to repeated observation and application. Examination of these patterns is a waste of time since they do not further a user's knowledge base. In addition, some of them may not be relevant to the task at hand. For example, Major & Mangano (1995) mined 529 rules from a hurricane database of which only 19 were found to be actually novel, useful and relevant. Information about the 'quality' of a rule can be used to either retain or prune it. Pruning rules reduces the 'quantity' of the final rule-set. In addition, the overall quality of the final set is improved since only the relevant rules are retained. Post-processing of mined rules is an important task in data mining, relevant to improving the efficacy of mined results. Researchers have adopted various strategies to address the rule immensity problem. Accordingly, redundancy reduction, rule templates, incorporation of additional constraints, ranking, grouping and visualization are some of the important post-processing techniques. We briefly discuss them and highlight some of their advantages and limitations.

2.1 *Incorporation of additional constraints*

In AR mining, additional constraints in conjunction with support and confidence thresholds (Agrawal R *et al* 1993), can reveal specific relationships between items. 'Rule templates' (Klemettinen *et al* 1994) helps the expert specify the structure of interesting and uninteresting class of rules in the form of inclusive and restrictive templates respectively. Rules matching an inclusive template are interesting and (or) relevant to the user. Such templates are typical post-processing filters. They select the relevant rules only after the entire set of rules has been extracted. Adomavicius & Tuzhilin (2001) have applied template-based rule filtering for validating rules. Validated rules are then used for constructing user-profiles.

Constraint-based mining (Bayardo *et al* 2000) directly embeds user-specified rule-constraints in the mining process itself. These constraints eliminate any rule that can be simplified to yield a rule of equal or higher predictive ability. Association patterns like negative ARs (Savasere *et al* 1998; Subramanian *et al* 2003), cyclic ARs (Ozden *et al* 1998), inter-transactional ARs (Lu *et al* 2000), ratio rules (Korn *et al* 1998), and substitution rules (Teng *et al* 2002) bring out particular relationships that are characteristic to the set of items being considered. In the market-basket context, negative ARs reveal the set of items a customer is unlikely to purchase with another set. This implies that in addition to preferences, user dislikes can also be modelled. Thus, web page contents can be dynamically modified, by removing those displays that may not be viewed by users. Cyclic association rules reveal purchases that display periodicity over time. Thus, they introduce a temporal element that might be used to customize web offerings. While customizing a web page, if the customer is unlikely to utilize the offer during the period of reference, a web site may not display a particular advertisement. A customer may not look at an offer if he/she has utilized the same offer in the recent past. In such cases some other offer(s) may be displayed. This may increase the impact of the web site.

Substitution rules reveal items that behave as substitutes. For example, e-bookshops like Amazon.com may not be able to stock all titles at any given point of time. However, some of the titles pertaining to the same subject may serve as equally effective substitutes. Thus, unavailability of a title can activate a user-tailored dynamic display of equivalent titles. Imposition of additional constraints into rule mining may offer insight into the domain by discovering focused and tighter relationships. As in rule templates, the constraints themselves may be a consequence of domain-expert knowledge. Each method discovers a specific kind of behaviour relevant for a particular aspect of application. A large number of mined patterns might necessitate the usage of other pruning methods.

Methods that enforce constraints are characterized by low user-involvement, except in the case of rule templates. From the business perspective, reluctance to apply unverified business models from data mining results to business decisions increases the importance of end-user involvement. Some of the patterns described are still in the realm of research. Hence, they are yet to find widespread application in e-commerce. We believe that the knowledge afforded by incorporating methods that enforce additional constraints may enrich profile descriptions by revealing relationships not revealed by plain ARs. In addition, some patterns in conjunction with others might throw light on interesting customer behaviour.

2.2 *Redundancy reduction*

'Redundancy reduction' refers to a class of techniques specifically aimed at pruning out patterns that do not convey new information. They address the quantity problem and the associated understandability issue by succinct characterization of the domain. If a set of rules refer to the same feature of the data, then the most general rule may be retained. 'Rule covers' (Toivonen *et al* 1995) is a method that retains a subset of the original set of rules. This subset refers to all rows (in a relational database) that the original rule set covered. Another strategy (Zaki 2000) in AR mining is to determine a subset of frequently occurring closed itemsets from their supersets. Though, the subset's cardinality is much lower than that of the superset, there is no loss of information. Sometimes, one rule can be generated from another using a certain inference system. Retaining only the basic rules may reduce the cardinality (Cristofor & Simovici 2002). In addition, the basic rules may provide users with a bird's eye-view of the domain. Such inference systems can also recover the original rule set using reversible mechanisms. Thus, information content of the basic un-pruned set is retained. Redundancy reduction methods may not provide a holistic picture if the size of the pruned rule-set is large. Specific knowledge with respect to certain customers may be lost, especially if the process has a bias towards generalization. Hence, in e-commerce applications, such methods may help in the concise characterization of a population of customers but not in generating descriptions of individual customers. A subset of the customer population may be described in general terms with new customers being assigned to these population segments. A method arriving at generalizations might remove interesting exceptions. Thus the important issue of identification of interesting patterns is left unaddressed.

2.3 *Visualization*

Visualization techniques take advantage of the intuitive appeal of visual depiction (Hilderman *et al* 2002). In the realm of e-commerce, visualization techniques have been extensively used in applications such as web merchandising, analysis of click streams in online stores (Lee *et al* 2001), profile building and description and other applications. Various features such as graphs, colour and charts help in improved visualization. In addition, innovative ways of depicting the results of data mining in two or three dimensions helps in compressing details given the restrictions of display space. Visual depiction of data mining results also helps easy iterative interaction. Rules depicted in a visual form can be easily navigated to various levels of detail by iteratively and interactively changing the thresholds of rule parameters. This means that many 'What-If' scenarios can be simultaneously analysed and compared. Groups of rules can be validated simultaneously on the basis of their visual depiction. This may reduce the load on the expert to a certain extent. However, the main drawback in visualization-based approaches is the difficulty of depicting a large rule/attribute space. As the number of rules increases the following two related problems become apparent. Visual depiction of the rules themselves becomes difficult. Second, it becomes difficult to perceive interconnections between rules constituting the rule-space. Thus it becomes more difficult to obtain a holistic picture of the domain. An increase in the number of dimensions is usually accompanied by a reduction in understandability of the resulting visual depiction. Hence if it is inlaid in a myriad of mundane facts, a user might fail to detect an interesting phenomenon. However, for browsing a limited rule space, visualization techniques do provide an intuitive summarization and overview of a domain. In addition, facts presented in a visual mode, are more often than not easier to comprehend than facts presented in non-visual forms such as tables, lists and the like.

2.4 *Organization and summarization*

A user might be able to get a good overview of the domain by examining a few general rules that describe the domain's essentials. Mining generalized association rules using product/attribute taxonomies is one such approach (Srikant & Agrawal 1995). In a product-taxonomy, lower levels might depict 'items' while higher levels may depict the item's categories of membership. One approach to representing general relationships using product taxonomy for summarization is as follows. If all items belonging to one category exhibit the same relationship with all items belonging to another category, then the set of rules describing them may be replaced by a general rule that directly relates product categories. Thus, the relationship gets described at a meta-level. In the process exceptional behaviour exhibited by individual items might get lost. GSE (general rules, summaries and exceptions) patterns introduced by Liu *et al* (2000) try to address this limitation. In their approach to summarization, general rules along with summaries convey an overview while exceptions point to cases differing from general cases. An expert may examine groups of rules by first examining the general rule that describes them. Particular rules from the group may be explored further depending on the generated interest. Summarization techniques are very helpful for understanding a domain. This is particularly true if the summarization highlights essential features of the domain. However, summarization at a very high level might not say anything new. This is because summaries might pertain to commonplace knowledge that might form a core component of a user's domain knowledge. Thus, the wider and important issue of identifying truly novel and interesting "nuggets" of knowledge is left unaddressed.

From the e-commerce perspective, summarization might be useful for applications that involve projection of inferences from a group of customers to a particular customer. Suppose a firm detects a new visitor on its web site. Then, instead of displaying a standard web page with common contents, some degree of personalization can be incorporated. This may be done by tracking customer behaviour on the website. Summarization techniques may be used to capture the essential characteristics of a group of customers. On identification of the possible group-membership of a new visitor, personalization can be effected by the use of group characteristics. For such applications, effectiveness of summarization techniques depends on the quality of summaries and timeliness of actions taken on the basis of these summaries. For dynamically changing groups, effectiveness may lie in the responsiveness of summarization techniques.

2.5 *Rule grouping and clustering*

A group of randomly selected rules may not make much holistic sense. This may be due to rules describing different and unrelated aspects of the domain. However, groups of related rules may provide some holistic perspective by revealing interconnections at the individual and at the group levels. There may be many 'similar' rules among the discovered rule set. A combined evaluation and validation of 'similar' rules might be manually easier than separately evaluating (Adomavicius & Tuzhilin 2001) individual rules. In addition, at a meta-level a user might gain insights such as patterns, within and across product categories. 'Grouping' of rules according to some criteria aids in the understandability of mined rules. The user now has a structure that may form the basis of rule evaluation.

One straightforward approach is to group ARs on the basis of extraneous characteristics of items, such as, economic assessment, profit margin, period of purchase, etc. (Baesens *et al* 2000). Users may consider them important from financial or other economic perspectives. However, these criteria may not get directly reflected in the transactions and may have to be imposed by the incorporation of additional knowledge. For example, users might wish to see rules describing items having similar profitability, storage costs etc. At other times, it might be advisable to have temporal comparisons. Consider the Christmas period. It might be more appropriate to examine rules describing purchase patterns during the corresponding period of the previous year than rules pertaining to a few months prior to Christmas. However, such groupings have elements of artificiality creeping in due to the use of extraneous factors.

Adomavicius & Tuzhilin (2001) have adopted a similarity-based grouping approach using attribute hierarchies in the context of personalization applications in e-commerce. They define a grouping operator that allows users to group similar rules. Similarity is defined (by an expert) by selecting a specific 'cut' of the attribute hierarchy. The expert defines aggregation granularity of the rules by specifying the 'cut.' Rules that have the same 'aggregated rule' are brought together in a group. However the basis of grouping may be devoid of any semantic meaning. Effectiveness of the grouping strategy is entirely dependent on the expert's knowledge and ability to define the 'cut.' Another limitation is the method not giving any consideration to natural groupings that might reflect a customer's purchasing behaviour.

Another approach is to 'let the rules speak for themselves' by means of transactions. Cluster analysis is a class of techniques used to classify objects or cases into relatively homogenous groups called clusters (Anderberg 1973; Jain *et al* 1999; Kaufman & Rousseeuw 1990). Objects in each cluster tend to be similar to each other and dissimilar to objects in other clusters. Application of clustering techniques might improve the understandability of mined rules by bringing together 'similar' rules into the same cluster. It may be easier to infer item-behaviour from rule clusters than from a rule list. This is because consecutive rules in a rule list may not have any relationship to each other. This can confound the user thus making interpretation difficult.

Clustering differs from grouping in that there is no preconceived notion of the structure or the number of groups that may exist in the data (Anderberg 1973). The idea here is to look for a 'natural' structure in the data on the basis of which clusters are evolved. Researchers have used clustering and grouping as strategies to improve the understandability of rules. Lent *et al* (1997) have introduced the notion of a 'clustered' AR. A clustered AR is a rule that is formed by combining similar, 'adjacent' association rules to form a few general rules. Clustering is defined as the merging of rules with adjacent attribute values or bins (intervals) of attribute values to form one rule that can represent the group. For example, the clustered rule $(40 \leq \text{age} < 42) \Rightarrow (\text{own\_home} = yes)$ could be formed from the two association rules $(\text{age} = 40) \Rightarrow (\text{own\_home} = \text{yes})$ and $(\text{age} = 41) \Rightarrow (\text{own\_home} = \text{yes})$. Wang *et al* (1998), on the other hand, allow any combination of numeric and categorical attributes in the antecedent and one or more categorical attributes in the consequent. Both approaches merge adjacent values of numeric attributes in a bottom-up fashion. Lent *et al* (1997) utilize a clustering approach to merging while the study by Wang *et al* (1998) maximizes certain interestingness criteria during the merging process. Both approaches are limited. While constructing clusters, they neither use relationships between items nor use behaviour of items in transactions. In addition, their approach being syntactic does not throw any fresh light on relationships among rules.

Toivonen *et al* (1995) proposed another approach (more akin to classical clustering). Distance between two rules is defined as the number of transactions in which the two rules with the same consequents differ. Gupta *et al* (1999) have proposed a normalized distance function called Conditional market-basket probability (CMPB) distance. This distance function tends to group all those rules that 'cover' the same set of transactions. Gupta *et al* (1999) state "rules involving different items but serving equal purposes were found to be close good neighbors." (Gupta *et al* 1999) Thus, their approach is able to capture some amount of customer purchasing behaviour. One of the limitations of both the schemes is the arbitrariness of the distance measures used for rule clustering (Adomavicius & Tuzhilin 2001). Moreover, they do not develop any framework to concisely describe the generated rule clusters.

Sahar (2002) has proposed a distance measure called d_{SC}. This distance measure combines the approaches in Dong & Li (1998) and Toivonen *et al* (1995). This measure utilizes the five characteristics that fully define an AR, namely the antecedent set, consequent set, rule support, antecedent support, and consequent support. This approach has utilized both syntactic matching of item sets and rule coverage of data. However it fails to consider the items constituting a rule, at an individual level. In addition, ARs arising from different sub-domains but showing identical behaviour in terms of characteristics of their constituent items may not be brought together. Hence, in some applications such clustering schemes may not be relevant and thus ineffective. Jorge (2004) has studied hierarchical clustering in the context of thematic browsing and summarization of large sets of ARs. The binary distance between two rules, defined in terms of set-theoretic operations, is proportional to the number of items occurring in only one of the two rules. Rules having a large overlap in their item sets might be brought together in a cluster. However, this again does not cover identical behaviour in transactions. Usage of clustering to alleviate the understandability problem in rule mining is still in its infancy. Not withstanding that, clustering has provided the base for many e-commerce applications such as partitioning and identification of relevant customer segments for product targeting, coupon design, cross-selling, up-selling etc.

3. Interestingness measures

One of the more popular approaches to alleviating the understandability problem in data mining is rule ranking using 'interestingness measures' (Silberschatz & Tuzhilin 1996). In recent years, a lot of work has been done in defining and quantifying 'interestingness.' As a result, several measures that view interestingness from different perspectives have been proposed, developed and applied. Interestingness measures attempt to capture the amount of 'interest' any pattern is expected to evoke on inspection. Merriam Webster's collegiate dictionary defines 'interest' as "a feeling that accompanies or causes special attention to an object or class of objects, or something that arouses such attention." Interesting patterns are supposed to arouse strong attention from users. A user might find a pattern interesting due to various reasons – some of which may be difficult to articulate. It has been found that 'interestingness' is an elusive concept (Silberschatz & Tuzhilin 1996; Fayyad & Uthurusamy 2002) consisting of many facets that are difficult to operationalize and therefore difficult to capture. In some cases, a particular behaviour in a domain might be interesting. The same behaviour exhibited in another domain may not be interesting. Thus, interestingness may be domain and user-dependent. In some other cases the same features may be domain and user-independent. Capturing all features of interestingness in one single measure simultaneously is an arduous if not an impossible task. Therefore, researchers typically focus on capturing only those features that may be important and relevant for a particular application. The discovered patterns may then be assigned scores and those that rank high on these measures may be ultimately displayed to the user.

Ranking the 'relevance' of a particular document or website with respect to a query is an important problem in the field of information retrieval (Page *et al* 1998; Baeza-Yates & Ribeiro-Neto 1999; Jeh & Widom 2002). Although related, relevance ranking is quite different from interestingness in the following respects. While the purpose of relevance ranking is to bring out the relevance and therefore the implied usefulness of a particular object with respect to a query, interestingness is supposed to identify new and previously implicit knowledge. Although relevance contributes to interestingness, a relevant pattern may not be interesting if it is commonplace knowledge. Interplay among various facets such as relevance, novelty, unexpectedness, surprisingness and user-knowledge determines the interestingness of an AR. A complete and composite characterization of interestingness is impossible.

An important and useful classification scheme of interestingness measures may be based on user-involvement. This results in two categories - objective and subjective measures (Silberschatz & Tuzhilin 1996; Freitas 1998). Objective measures usually deal with data-related aspects such as its distribution, structure of the rule and others while subjective measures are more user-driven.

3.1 *Objective measures of interestingness*

Objective measures quantify a pattern's interestingness in terms of the pattern's structure and the underlying data used in the discovery process. Researchers have used measures developed in diverse fields such as statistics, social sciences, information theory, and artificial intelligence in order to measure specific data characteristics. Typical objective measures of interestingness include statistical measures like confidence, support (Agrawal *et al* 1993), lift (Piatetsky-Shapiro *et al* 2000), conviction (Brin *et al* 1997a), rule interest (Brin *et al* 1997a) and collective strength (Aggarwal & Yu 2001). Information theoretic measures such as entropy, amount of information, Kullback and Hellinger measures have also been used in other studies (Hilderman & Hamilton 1999). Occurrence of unusual phenomenon like Simpson's paradox in data is deemed interesting by Freitas (1998). Freitas (1999) has adopted a multi-criteria approach for objective evaluation of a rule's interestingness. Incorporation of rule-quality factors such as disjunct size, imbalance of class distributions, misclassification costs, and asymmetry, helps in characterizing a rule's 'surprisingness' to a greater extent. Exception rule mining (Hussain *et al* 2000) is another approach that reveals interesting rules (exceptions) that differ from expected ones. The expert pre-specifies the structure of interesting exceptions. Patterns that satisfy these conditions are reported as exceptions.

Hilderman & Hamilton (1999) have surveyed seventeen interestingness measures (mostly objective) that have been successfully employed in data mining applications. Tan & Kumar (2000) have examined various measures proposed in statistics, machine learning and data mining literature. Based on experimental results they show the similarity between several of the measures and the correlation coefficient. Jaroszewicz & Simovici (2001) have proposed an objective measure that is a generalization of many conditional and unconditional classical measures. Omiecinski (2003) has come up with three alternative interestingness measures for associations: any-confidence, all-confidence and bond.

Objective measures are strongly domain and user-independent. They reveal data characteristics that are not tied to domain/user specific definitions. This increases their applicability in different situations. However, this property of independence may limit their power of discrimination. Since any objective measure has to hold across all domains, it considers interestingness from a limited perspective that is common across domains. Hence, objective measures cannot capture all complexities of the discovery process (Silberschatz & Tuzhilin 1996). Many objective measures are based on the strength of dependence relationships between items (Meo 2000; Shekar & Natarajan 2004b), and interestingness is regarded as being directly proportional to this strength. However, this view may lead to erroneous results (Brin *et al* 1997b). Take 'support' as an example. While it is useful in measuring the statistical significance of a rule, rules that are most obvious to the user have high support values. Thus, bringing out rules based on support values may not add additional knowledge. Similarly, other objective measures have their own limitations and biases.

It is also common for different objective measures to convey contradictory evaluation or conceal certain domain-related facts. Therefore, it is not only important to select the appropriate measure(s) for each domain, but it is also important to specify the correct order of application (Tan *et al* 2004). Only then, truly interesting rules would get revealed. Objective measures when used as initial filters to remove definitely uninteresting or unprofitable rules. Rules that are statistically insignificant may be removed since they do not warrant further attention. Further, considerations spanning diverse domains and that are not user dependent may be modelled by objective measures.

3.2 *Subjective measures of interestingness*

Domain experts (users) play an important role in the interpretation and subsequent application of data mining results. Hence the need to incorporate user-views, in addition to data-related aspects. Generally, users differ in their beliefs and interests since they may possess varied experience, knowledge and psychological make-up. They may be interested in different aspects of the domain depending on their area of work. In addition, they may also have varying goals and difference of opinions about the applicability and usefulness of KDD results. This variation in interest enhances the importance of injecting subjectivity into interestingness evaluation (Silberschatz & Tuzhilin 1996). 'Actionability' and 'unexpectedness' are two facets that determine subjective interestingness (Silberschatz & Tuzhilin 1996). Rules are interesting if they are unexpected (surprising to the user) or actionable (if the user can act advantageously).

Actionable patterns are interesting since they offer opportunity for direct action that may translate into profitable results. However, operationalization of 'actionability' has proved to be an extremely difficult task due to the inherent difficulty in associating rules with actions. A large rule space and the possibility of a single rule implying multiple actions may increase the difficulty of associating rules with actions. Hence, studies oriented towards actionability tend to be extremely domain specific (Silberschatz & Tuzhilin 1996). In particular, they demand that patterns to be related to actions like the KEFIR system (Matheus *et al* 1996). This is a difficult task in all but the smallest domains where actions are clearly defined. Adomavicius & Tuzhilin (1997) have proposed an approach to defining actionability, using 'action hierarchy'. An action hierarchy is a 'tree' of actions with patterns and pattern templates (KDD queries) assigned to its nodes. This approach provides a framework for operationalizing actionability with some domain independence.

Roddick & Rice (2001) have brought out the temporal and dynamic nature of interestingness. Using events in a sporting arena as a running example, they show how 'anticipation' has a critical effect on both, selection of interesting events and variation of interestingness threshold as the events unfold. This brings out the dynamic nature of interestingness. Interestingness with respect to a particular event or object may vary with availability of further information. However, the concept of anticipation needs to be explored further. This is because anticipation may be useful in domains characterized by sequences of real-time transactions taking place in a quick succession. Typical examples would be electronic auctions and internet shopping where each purchase is considered as an event.

Another feature of interestingness pertains to prior knowledge of the user. Prior knowledge increases the interestingness about a subject. In addition, current knowledge about a subject becomes interesting if this knowledge is directly relevant to user-goals (Ram 1990). Accordingly, interestingness may be defined as a heuristic that measures relevance of the input to a person's knowledge goals. "Knowledge goals" (Ram 1990) is related to acquiring some piece of information required for a reasoning task. If a piece of information is relevant to "knowledge goals", interest towards it increases.

Silberschatz & Tuzhilin (1996, p. 971) have argued that "the majority of actionable patterns are unexpected and that the majority of unexpected patterns are actionable." Hence, they hypothesize that unexpectedness is a good approximation for actionability and vice-versa. Since unexpectedness is easier to operationalize than actionability, most studies concerning subjective interestingness employ 'unexpectedness' as the main subjective criterion. Approaches to the determination of subjective interestingness using 'unexpectedness', adopt the following scheme.

- Eliciting user-views
- Representing user-views in a form suitable for computation
- Mining the database to extract rules about the domain
- Comparing mined rules with user-views to determine the degree of conflict
- Presenting and labelling, rules that conflict user-views (on attributes such as relationship, strength, and direction), as interesting

However, these methods differ in details such as representation schema, method of comparison and interestingness measures (Padmanabhan & Tuzhilin 1999; Liu *et al* 1999; Liu *et al* 2000; Shekar & Natarajan 2004a). Sahar (1999) has proposed a scheme to identify interesting rules indirectly by the elimination of uninteresting rules. Effectiveness of this approach lies in its ability to quickly eliminate large families of rules that are not interesting, while limiting user interactions to a few, simple classification questions.

One limitation with respect to utilizing user-beliefs as the basis for interestingness is the 'knowledge acquisition' needed for evaluation. Eliciting views from users is difficult in practice and acquired knowledge could also be incomplete. Users may not be able to specify each and every belief about a domain completely. Partial specification of user-beliefs may result in incorrect assignment of interestingness scores to a large number of rules. Many of these incorrectly scored rules might pertain to attributes, relationships and beliefs, the user had failed to specify. In addition user-beliefs can also change with time. Continuously maintaining and updating the knowledge base (user-beliefs) is an important issue that requires attention. Other limitations are more approach-specific; such as specification of a priori probabilities in the Bayesian approach (Silberschatz & Tuzhilin 1996) and fuzzy membership functions (Liu *et al* 1999).

Figure 1 displays a partial classification of the approaches towards interestingness in data mining. Pruning and ranking of patterns on the basis of interestingness measures is an intuitive approach to rule quality and rule quantity problems.

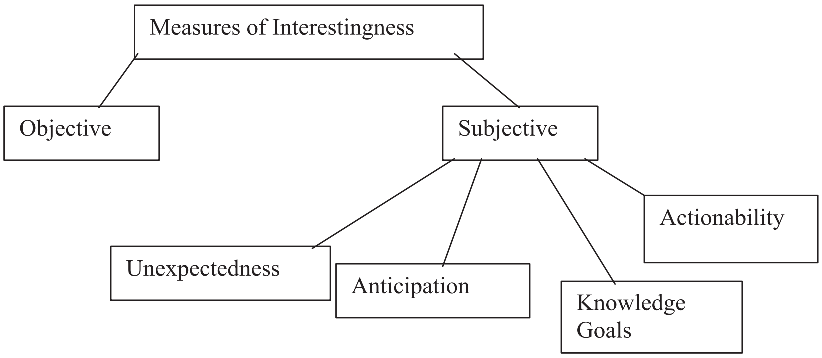

Figure 1. Partial classification of interestingness measures (based on Silberschatz & Tuzhilin 1996).

Interestingness measures can play an important role in the identification of novel, relevant, implicit and understandable patterns from the multitude of mined patterns. They help in automating most of the post-processing work in data mining. Integration of interestingness evaluation schemes into the data mining process can help the firm in gaining new knowledge about its customers. e-Commerce applications like personalization in web merchandising and others require a high degree of customization with respect to specific customer requirements. One of the possible areas where interestingness measures can play a significant role in building accurate customer profiles. Interesting rules need not have high statistical significance. Hence, approaches biased towards stronger rules might disregard 'interesting' rules characteristic to a particular customer. However, a high degree of personalization implies catering to the idiosyncratic beliefs and preferences of a customer. Precise identification of 'interesting' rules characteristic to a particular customer might complete the characterization of a customer's profile. Applications like recommender systems may use this knowledge to make accurate, relevant product recommendations ultimately furthering the sales. We explore related issues in the next section wherein we discuss personalization applications and the relevance of interestingness measures for increasing their effectiveness.

4. Utility of interestingness rankings in personalization applications for e-commerce

Personalization is the use of a customer's information for delivering a customized solution to that customer thus catering to personal needs (Murthi & Sarkar 2003; Schafer *et al* 2001). The importance of personalization in e-commerce applications has increased due to various reasons. If a firm is able to make an accurate estimate of its customer's needs and deliver personalized products and services based on the needs, then this can become an important source of competitive advantage in some areas. This is because of the increase in search costs with increase in the available choices. Mechanisms to provide relevant information through personalization applications can aid in the customer's decision-making process. Drastic reduction in costs pertaining to information technology (IT) and database has enabled firms to collect, process and store large amounts of customer data. e-commerce and other internet-based applications are typical examples, wherein large amounts of customer-related data can be easily collected.

A common way to incorporate personalization in a firm's interactions with a customer is through the use of recommender systems (Resnick & Varian 1997). Recommender systems make product recommendations based on prior interactions of the customer with the firm. For example, Amazon.com and other web-based booksellers may use diverse techniques such as collaborative filtering and AR mining to make book recommendations. In addition to product recommendations, personalization may also imply communication of appropriate messages to the 'right' customer. This may be done on the basis of customer profiles. Here we are concerned with the applicability of interestingness measures and other typical post-processing mechanisms in internet-based personalization applications.

The personalization process itself may be viewed (Murthi & Sarkar 2003) as consisting of three main stages: learning, matching and evaluation. In the learning stage a firm collects data about the customer's preferences and tastes. Sometimes demographic data may also be collected. The customer may provide data in an explicit manner at the account creation time at the website. Alternatively, tracking a customer's interactions on the firm's website may provide a wealth of information about his/her preferences. User registrations and/or cookies may help in identifying a visitor as an existing customer. Similarly, transaction data/point-of-sale data, and web-server logs may also track a customer's interaction with the firm. Thus, e-commerce applications in an online environment aid in fast, accurate and unobtrusive data collection.

At the matching stage (Murthi & Sarkar 2003), the firm uses the knowledge of customer preferences obtained from the learning stage, to customize its offerings in accordance with the customer's tastes and needs. As we have seen, this can be performed by recommender systems. These systems use rule-based, collaborative filtering and content filtering techniques. Other data mining methods such as clustering may also find application in tasks such as customer segmentation, assignment of a new customer to a segment etc. The evaluation stage involves development of appropriate metrics for assessing the effectiveness of the personalization schemes employed by the firm. Figure 2 provides a pictorial representation of the personalization process. It should be noted that the entire personalization process is iterative. Interactions with the customer constantly provide input to the learning stage. In addition, evaluation of the offering provided by the firm and also the degree of match between customer requirements and the offering constitute another set of inputs to the learning process. Sometimes, the matching algorithms might be modified based on the feedback from the evaluation phase. In a nutshell, we may observe that the personalization process is a dynamic one with customer interactions constantly defining changes in the offerings.

As we have seen, ARs and other data mining methods may be used to match the firm offerings with customer requirements. Consider the use of ARs in recommender applications. ARs bring out co-occurrence based affinity between sets of items/products. Essentially, ARs

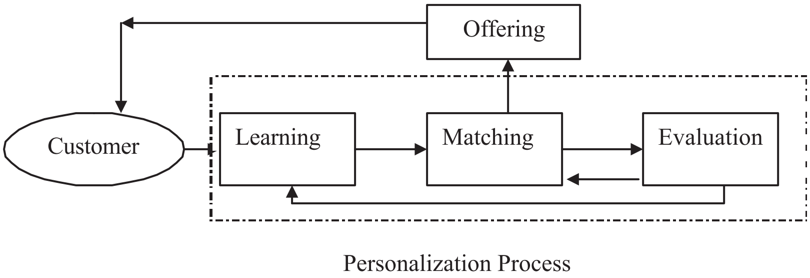

Figure 2. Typical stages in personalization.

discover associations between two sets of items/products, such that the presence of one set in a particular transaction implies the presence of another in the same transaction. A typical approach towards using ARs in recommender systems is as follows (Sarwar *et al* 2000).

- For each customer, a transaction containing all items purchased by the customer in the past can be constructed
- AR discovery algorithm is used to discover sets of items that are frequently purchased by customers
- For a particular customer, the relevant set of ARs are those ARs whose antecedents contain items purchased by the customer in the past
- Finally, the recommender system recommends items that are present in the consequent of the ARs in the set. The confidence measure (Agrawal *et al* 1993) might be used to discern which of the ARs from the set may be the most effective for recommendation

Interestingness measures may be used in different ways for improving recommendations in personalization applications. In some cases, the discovered ARs might be too numerous. It might then be necessary to select the most appropriate ARs for the given situation and context. Ranking rules on the basis of their interestingness, aid in selecting the most relevant rules for a given application.

As seen in the previous sections, post-processing methods may be based on generalizations or specifications. It might be advantageous to use generalized ARs and interestingness measures that preserve generalizations when we do not have information pertaining to a customer. Subsequently, preferences of the customer segment can be extrapolated while personalizing the offering.

Subjective measures of interestingness are based on user-beliefs pertaining to customer behaviour. ARs that deviate from these user-beliefs are deemed interesting. In personalization applications, ARs ranking high on subjective measures (based on unexpectedness) are indicative of novel and unknown knowledge. These represent opportunities for making recommendations with an aim to cross-sell or up-sell products. This can be accomplished by extrapolating the novel behaviour of a customer segment to other customer segments that exhibit either similar or identical behaviour. Thus, we can say that interestingness measures and other related post-processing mechanisms are useful in increasing the effectiveness of recommender systems and other personalization applications.

5. Research issues

A study of the different strategies aimed at mitigating the pattern immensity problem in data mining reveals some interesting research issues relevant to e-commerce. Each approach comes with its own set of advantages and limitations. One main difference between a traditional data mining environment and an e-commerce application like recommender system is that the end-user of a recommender system is the consumer. This key difference leads to several desirable properties required by recommender systems and other e-commerce applications. e-Commerce takes place in a dynamic environment, wherein there can be thousands of users simultaneously accessing a website. Each user interaction is a source of information about its customer for the firm. The firm can use this information in a profitable manner. For example, a user must feel that with each interaction, the degree of personalization has increased. At the same time the system deployed by the firm must be fast enough to respond immediately to the user with least latency. This calls for a balance between techniques that are accurate but overhead-intensive and techniques that are not so accurate but more responsive.

Interestingness may consist of many facets intrinsic to a particular domain. A combination of objective and subjective measures may be necessary to reveal interesting patterns. Some issues concerning application of interestingness measures may be generic while others may be domain-specific. One important issue is the genesis of interestingness and its constituent features. Unexpectedness and actionability do not characterize interestingness in totality. There might be other domain-dependent features worthy of consideration and incorporation into interestingness evaluation schemes. These features could then form the basis for more comprehensive interestingness evaluation.

Another issue concerns the joint application of objective and subjective measures. Objective measures could be used as a first filter to remove rules that are definitely uninteresting. This should be based on certain requirements, such as significance and predictive ability. Subjective measures can then bring in user-biases and beliefs into interestingness evaluations. An important issue with respect to subjective interestingness concerns maintenance of the knowledge base and its continuous updation thus preserving its relevance. Interaction between objective and subjective measures has not been sufficiently explored. In addition, few studies have considered the appropriateness of applying a specific interestingness measure across domains. The effect of changing the order of application of interestingness measures and the interaction among them are other issues worthy of future study. Some relationships are a logical consequence of a firm's operational business rules. Such knowledge being intuitive and tacit may not get specified during the knowledge elicitation phase. Incorporating such inferences in subjective measures is another issue for future research.

An important consideration with respect to e-commerce applications is context inclusion. Without contextual information, interestingness evaluation and other methods may not be effective. This is because interestingness may be strongly domain and user-dependent and hence highly subjective. On the other hand, if objective measures can be made context-dependent by infusing domain-related data definitions into them, then user-dependence may be reduced. This could lead to sharing of data mining results across different e-commerce applications.

Application of interestingness measures during the various phases of data mining has its own advantages and disadvantages. If the dataset is large, then it may be advantageous to mine rules and then apply interestingness measures. On the other hand, for a small one, it may be preferable to apply interestingness measures during the mining phase itself. Ideally, a data mining system should contain a repository of interestingness measures, both objective and subjective. Choice of measures could be based on the mined patterns, application and purpose of the user. Here, the importance of integrating mining methods with interestingness evaluation in the context of e-commerce cannot be over-emphasized (Ansari *et al* 2000).

Another problem with rule ranking concerns the possible lack of relationship among interesting rules. Thus, two consecutive interesting rules could pertain to different domains/sub-domains. Hence, it may be difficult for a user to connect them and obtain an overview of the domain. Combining methods that address the rule quantity problem along with interestingness, might partially address this problem. Clustering of similar rules could possibly be studied as a possible pre-processing step followed by a ranking scheme based on interestingness. Thus, a user might be able to obtain an overview of the domain and also discover implicit hidden knowledge brought out by interesting rules. Visualization techniques that display interestingness evaluations in an intuitive and understandable manner may also be helpful here.

6. Summary

e-Commerce applications typically generate huge amounts of operational and customer behavioural data. Automated methods that make use of data mining tools and techniques are necessary to analyse and unearth hidden knowledge from this data. In this paper, we have examined the understandability problem in data mining, and its relevance to e-commerce applications. The genesis of this problem can be traced to the glut of patterns generated by data mining applications. Large numbers of mined patterns render manual inspection impractical and infeasible. In addition, the commonplace and obvious nature of knowledge revealed by most of these patterns necessitates usage of automated methods for selecting the most interesting, novel, relevant and significant patterns for further action. We have surveyed some of the available post-processing methods in data mining. These include redundancy reduction, incorporation of additional constraints, visualization, organization and summarization, rule grouping and clustering. This was done with respect to association rule mining since ARs find application in many e-commerce applications such as recommender systems. Each post-processing problem uses a different approach to mitigating the rule immensity problem. We brought out some of the advantages and limitations of each method. We examined the relevance and applicability of each method in e-commerce context.

One important and widely used approach to tackle the rule immensity and the consequent understandability problem is the usage of interestingness measures. Interestingness measures attempt to capture the amount of ‘interest’ that a pattern is expected to evoke on inspection. Interestingness is an elusive concept that has both data-driven and user-driven aspects. Accordingly, interestingness measures may be classified as objective or subjective. Properties of objective and subjective measures and their usefulness have been discussed in light of online web-based applications.

Personalization in e-commerce applications is increasingly gaining importance. This is because: personalization enables a firm to cater to individualistic needs of its customers. Issues involved in the various stages of personalization were brought out. Interestingness measures, by identifying the most relevant rules for action, can be used to improve the effectiveness of personalization applications.

Interestingness is a perceptual concept whose different facets are difficult to capture and operationalize. Some of these facets are yet to be identified. The ensuing discussions identified some research issues relevant to e-commerce applications. The dynamic nature of e-commerce applications and the need for immediate response to customer interactions necessitates a substantially different approach towards interestingness evaluation and deployment of its results. Future research pertaining to interestingness and associated methods is expected to yield results with more complete interestingness characterizations.

e-Commerce creates a level-playing field by removing all location-based and other traditional advantages. Businesses must leverage on information advantages. With large volumes of data, data mining is expected to play a significant role in increasing the effectiveness of e-commerce applications. Merely mining of patterns is not enough. It is important to effectively employ the knowledge gained from these patterns. Interestingness measures and other methods, which address the problem of immensity of mined patterns, are vital contributors to the post-processing phase in data mining. It is not uncommon to find organizations struggling to make sense of data captured through automated processes. Frameworks and methodologies that aid not only in the selection of relevant and significant patterns but also in their effective and efficient deployment need to be researched as they may help firms in leveraging their information advantage.

References

Adomavicius G, Tuzhilin A 2001 Expert-driven validation of rule-based user models in personalization applications. *Data Mining Knowledge Discovery* 5(1/2): 33–58

Adomavicius G, Tuzhilin A 1997 Discovery of actionable patterns in databases: the action hierarchy approach. *Proc. Third Int. Conf. on Data Mining and Knowledge Discovery (KDD 1997)* (Meulo Park, CA: AAAI Press) pp 111–114

Aggarwal C C, Yu P S 2001 Mining associations with the collective strength approach. *IEEE Trans. Knowledge Data Eng.* 13: 863–873

Agrawal R, Imielinski T, Swami A 1993 Mining association rules between sets of items in large databases. *Proc. 1993 ACM SIGMOD Int. Conf. on Management of Data* (Washington, DC: ACM Press) pp 207–216

Anderberg M R 1973 *Cluster analysis for applications* (New York: Academic Press)

Ansari S, Kohavi R, Mason L, Zhang Z 2000 Integrating e-commerce and data mining: Architecture and challenges. *Proc. WEBKDD'2000 Workshop: Web mining for e-commerce Challenges and Opportunities*; http://ai.stanford.edu/∼ronnyk/WEBKDD2000/index.html

Baesens B, Viaene S, Vanthienen J 2000 Post-processing of association rules. *Proc. Workshop on Post-Processing in Machine Learning and Data Mining: Interpretation, visualization, integration, and related topics* with in *Sixth ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD 2000)* Boston, MA, pp 20–23, http://www.cas.mcmaster.ca/∼bruha/kdd2000/kddrep.html

Baeza-Yates R, Ribeiro-Neto B 1999 *Modern information retrieval* (Reading, MA: Addison Wesley)

Bayardo Jr R J, Agrawal R, Gunopulos D 2000 Constraint-based rule mining in large, dense databases. *Data Mining Knowledge Discovery* 4(2/3): 217–240

Brieman L, Friedman J H, Olshen R, Stone C 1984 *Classification and regression trees* (Belmont, CA: Wadsworth)

Brin S, Motwani R, Silverstein C 1997a Beyond market baskets: Generalizing association rules to correlations. *Proc. ACM SIGMOD Conf.* (New York: ACM Press) pp 265–276

Brin S, Motwani R, Ullman J D, Tsur S 1997b Dynamic itemset counting and implication rules for market basket data. *Proc. ACM SIGMOD Conf.* (New York: ACM Press) pp 255–264

Burt S, Sparks L 2003 e-Commerce and the retail process: A review. *J. Retailing Customer Services* 10: 275–286

Cristofer L, Simovici D 2002 Generating an informative cover for association rules. *Proc. 2002 IEEE Int. Conf. on Data Mining (ICDM 2002)* (Washington, DC: IEEE Comput. Soc. Press) pp 597–600

Deo N 1989 *Graph theory with applications to engineering and computer science.* (New Delhi: Prentice Hall of India)

Fayyad U, Uthurusamy R 2002 Evolving data mining into solutions for insights. *Commun. ACM* 45(8): 28–31

Fayyad U M, Piatetsky-Shapiro G, Smyth P 1996 From data mining to knowledge discovery: An overview. In *Advances in knowledge discovery and data mining* (eds) U M Fayyad, G Piatetsky-Shapiro, P Smyth, R Uthurusamy (AAAI/MIT Press) pp 1–34

Freitas A A 1998 On objective measures of rule surprisingness. Proc. *Second European Symposium on Principles of Data Mining and Knowledge Discovery, (PKDD-98),* Lecture Notes in Artificial Intelligence. (LNAI 1510), (Springer-Verlag) pp 1–9

Freitas A A 1999 On rule interestingness measures. *Knowledge-Based Syst.* 12: 309–315

Geoffrion A M, Krishnan R 2003a e-Business and management science: Mutual impacts (Part 1). *Manage. Sci.* 49: 1275–1286.

Geoffrion A M, Krishnan R 2003b e-Business and management science: Mutual impacts (Part 2). *Manage. Sci.* 49: 1445–1456

Geyer-Schulz A, Hahsler M 2002 Comparing two recommender algorithms with the help of recommendations by peers. In *WEBKDD 2002 - Mining web data for discovering usage patterns and profiles: 4th Int. Workshop,* (eds) O R Zaiane, J Srivastava, M Spiliopoulou, B Masand (Revised Papers) Lecture Notes in Computer Science LNAI 2703 (Berlin: Springer-Verlag) pp 137–158

Grabmeier J, Rudolph A 2002 Techniques of cluster algorithms in data mining. *Data Mining Knowledge Discovery* 6: 303–360

Gupta G K, Strehl A, Ghosh J 1999 Distance based clustering of association rules. *Proc. Intelligent Engineering Systems through Artificial Neural Networks (ANNIE 1999)* (St. Louis, MO: ASME Press) vol. 9, pp 759–764

Hilderman R J, Hamilton H J 1999 Knowledge discovery and interestingness measures: A survey. Technical Report, Department of Computer Science, University of Regina, Canada

Hilderman R J, Li L, Hamilton H J 2002 Visualizing data mining results with domain generalization graphs. In *Information visualization in data mining and knowledge discovery* (eds) U M Fayyad, G G Grinstein, A Wierse Andreas (San Franscisco, CA: Morgan Kaufman) pp 251–270

Hussain F, Liu H, Suzuki E, Lu H 2000 Exception rule mining with a relative interestingness measure. *Proc. Pacific Asia Conf. on Knowledge Discovery in Databases (PAKDD 2000)* (London: Springer Verlag) pp 86–97

Jain A K, Murty M N, Flynn P J 1999 Data clustering: A review. *ACM Comput. Surv.* 31(3): 264–323

Jeh G, Widom J 2002 SimRank: A measure of structural-context similarity. *Proc. 8th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, KDD-2002*

Jorge A 2004 Hierarchical clustering for thematic browsing and summarization of large sets of association rules. *Proc. 2004 SIAM Int. Conf. on Data Mining (SDM 2004)* (SIAM Press)

Kalakota R, Whinston A B 1999 *Frontiers of electronic commerce* (Singapore: Addison Wesley/Longman)

Kaufman L, Rousseeuw P J 1990 *Finding groups in data: An introduction to cluster analysis* (New York: Wiley)

Klemettinen M, Mannila H, Ronkainen P, Toivonen H, Verkamo I A 1994 Finding interesting rules from large sets of discovered association rules. *Proc. Third Int. Conf. on Information and Knowledge Management (CIKM 1994)* (ACM Press) pp 401–407

Kohavi R, Provost F 2001 Applications of data mining to electronic commerce. *Data Mining Knowledge Discovery* 5(1/2): 5–10

Korn F, Labrinidis A, Kotidis Y, Faloutos C 1998 Ratio rules: A new paradigm for fast, quantifiable data mining. *Proc. 24th Int. Conf. on Very Large Databases (VLDB 1998)* (New York: Morgan Kaufmann) pp 582–593

Lee J, Podlaseck M, Schonberg E, Hoch R 2001 Visualization and analysis of clickstream data of online stores for understanding web merchandising. *Data Mining Knowledge Discovery* 5(1/2): 59–84

Liu B, Hsu W, Mun L, Lee H 1999 Finding interesting patterns using user expectations. *IEEE Trans. Knowledge Data Eng.* 11: 817–832

Liu B, Hu M, Hsu W 2000 Multi-level organization and summarization of the discovered rules. *Proc. ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD 2000)* (Boston: ACM Press) pp 208-217

Lu H, Feng L, Han J 2000 Beyond intra-transaction association analysis: Mining multi-dimensional inter-transaction association rules. *ACM Trans. Inf. Syst.* 18(4): 423–454

Major J A, Mangano J J 1995 Selecting among rules induced from a hurricane database. *J. Intell. Inf. Syst.* 4: 39–52

Matheus C, Piatetsky-Shapiro G, McNeill D 1996 Selecting and reporting what is interesting: The KEFIR application to healthcare data. In *Advances in knowledge discovery and data mining* (eds) U M Fayyad, G Piatetsky-Shapiro, P Smyth, R Uthurusamy (Menlo Park, CA: AAAI/MIT Press) pp 495–516

Meo R 2000 Theory of dependence values. *ACM Trans. Database Syst.* 25: 380–406

Murthi B P S, Sarkar S 2003 The role of the management sciences in research on personalization. *Manage. Sci.* 49: 1344–1362

Ozden B, Sridhar R, Silberschatz A 1998 Cyclic association rules. *Proc. Fourteenth Int. Conf. on Data Engineering (ICDE 1998)* (Washington, DC: IEEE Comput. Soc. Press) pp 412–421

Padmanabhan B, Tuzhilin A 1999 Unexpectedness as a measure of interestingness in knowledge discovery. *decision support syst.* 27: 303–318

Page L, Brin S, Motwani R, Winograd T 1998 The PageRank citation ranking: Bringing order to the web. Technical report, Stanford University Database Group. http//citeseer.nj.nec.com/368196.html

Piatetsky-Shapiro G, Steingold S 2000 Measuring lift quality in database marketing. *ACM SIGKDD Explorations Newslett.* 2(2): 76–80

Quinlan J 1993 *C4.5: Programs for machine learning* (New York: Morgan Kaufmann)

Ram A 1990 Knowledge goals: A theory of interestingness. *Proc. 12th Annual Conf. of the Cognitive Science Society*, Cambridge, MA

Resnick P, Varian H R 1997 Recommender systems. *Commun. ACM* 40(3): 56–58

Roddick J F, Rice S 2001 What's interesting about cricket? – On thresholds and anticipation in discovered rules. *ACM SIGKDD Explorations Newslett.* 3(1): 1–5

Sahar S 1999 Interestingness via what is not interesting. *Proc. ACM Conf. on Data Mining (KDD-99)* (San Diego, CA: ACM Press) pp 332–336

Sahar S 2002 Exploring interestingness through clustering: A framework. *Proc. IEEE Int. Conf. on Data Mining (ICDM 2002)* (Washington, DC: IEEE Comput. Soc. Press) pp 677–680

Sarwar B, Karypis G, Konstan J, Riedl J 2000 Analysis of recommendation algorithms for e-commerce. *Proc. e-Commerce 2000* (New York: ACM Press) pp 158–167

Savasere A, Omiecinski E, Navathe S 1998 Mining for strong negative associations in a large database of customer transactions. *Proc. Fourteenth Int. Conf. on Data Engineering (ICDE 1998)* (Washington, DC: IEEE Comput. Soc. Press) pp 494–502

Schafer J B, Konstan J A, Riedl J 2001 e-Commerce recommendation applications. *Data Mining Knowledge Discovery* 5(1/2): 115–153

Shekar B, Natarajan R 2004a A framework for evaluating knowledge-based interestingness of association rules. *Fuzzy Optim. Decision Making* 3(2): 157–185

Shekar B, Natarajan R 2004b A transaction-based neighbourhood-driven approach to quantifying interestingness of association rules. *Proc. Fourth IEEE Int. Conf. on Data Mining (ICDM 2004)* (Washington, DC: IEEE Comput. Soc. Press) pp 194–201

Silberschatz A, Tuzhilin A 1996 What makes patterns interesting in knowledge discovery systems. *IEEE Trans. Knowledge Data Eng.* 8: 970–974

Srikant R, Agrawal R 1995 Mining generalized association rules. *Proc. 21st Int. Conf. on Very Large Databases (VLDB 1995)* (New York: Morgan Kaufmann)

Subramanian D K, Ananthanarayana V S, Narasimha Murty M 2003 Knowledge-based association rule mining using AND-OR taxonomies. *Knowledge-Based Syst.* 16: 37–45

Tan P, Kumar V, Srivastava J 2004 Selecting the right interestingness measure for association patterns. *Inf. Syst.* 29(4): 293–331

Teng W, Hsieh M, Chen M 2002 On the mining of substitution rules for statistically dependent items. *Proc. IEEE Int. Conf. on Data Mining (ICDM 2002)* (Washington, DC: IEEE Comput. Soc. Press) pp 442–449

Toivonen H, Klemettinen M, Ronkainen P, Hatonen K, Mannila H 1995 Pruning and grouping discovered association rules. *Proc. Mlnet Workshop on Statistics, Machine Learning and Knowledge Discovery in Databases*, Herakhion, Crete, Greece

Wang K, Tay Soon H W, Liu B 1998 Interestingness-based interval merger for numeric association rules. *Proc. 4th Int. Conf. on Data Mining and Knowledge Discovery (KDD 98)* (New York: AAAI Press) pp 121–128

Zaki M J 2000 Generating non-redundant association rules. *Proc. 6th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD 2000)* (New York: ACM Press) pp 34–43