Available online at www.sciencedirect.com
ScienceDirect
Expert Systems with Applications 35 (2008) 1422–1429
Expert Systems with Applications
www.elsevier.com/locate/eswa

Towards personalized recommendation by two-step modified Apriori data mining algorithm

Enrique Lazcorreta [a,b], Federico Botella [a,b], Antonio Fernández-Caballero [c,d,*]

[a] Instituto Universitario Centro de Investigación Operativa (CIO), Universidad Miguel Hernández de Elche, 03202 Elche, Spain
[b] Departamento de Estadística, Matemáticas e Informática, Universidad Miguel Hernández de Elche, 03202 Elche, Spain
[c] Instituto de Investigación en Informática de Albacete (I3A), Escuela Politécnica Superior de Albacete, Universidad de Castilla-La Mancha, 02071 Albacete, Spain
[d] Departamento de Sistemas Informáticos, Escuela Politécnica Superior de Albacete, Universidad de Castilla-La Mancha, 02071 Albacete, Spain

Abstract

In this paper a new method towards automatic personalized recommendation based on the behavior of a single user in accordance with all other users in web-based information systems is introduced. The proposal applies a modified version of the well-known Apriori data mining algorithm to the log files of a web site (primarily, an e-commerce or an e-learning site) to help the users to the selection of the best user-tailored links. The paper mainly analyzes the process of discovering association rules in this kind of big repositories and of transforming them into user-adapted recommendations by the two-step modified Apriori technique, which may be described as follows. A first pass of the modified Apriori algorithm verifies the existence of association rules in order to obtain a new repository of transactions that reflect the observed rules. A second pass of the proposed Apriori mechanism aims in discovering the rules that are really inter-associated. This way the behavior of a user is not determined by "what he does" but by "how he does". Furthermore, an efficient implementation has been performed to obtain results in real-time. As soon as a user closes his session in the web system, all data are recalculated to take the recent interaction into account for the next recommendations. Early results have shown that it is possible to run this model in web sites of medium size.

Keywords: Personalization; Data mining; Apriori-like algorithm; Recommendation

1. Introduction

In recent years, companies have concentrated on understanding the needs and expectations of their customers and grouping the existing and potential customers into classes with the purpose of improving the efficiency of their marketing strategies and increasing their market share (Saglam, Salman, Sayin, & Türkay, 2006). Personalization has become a reality and is possible by using efficient methods of data mining and knowledge discovery (Kim & Cho, 2007). To date, a variety of recommendation techniques has been developed (Cho, Kim, & Kim, 2002). Through analyzing user related information, it is possible to make a more accurate analysis of customer's interest or preference. In most cases, recommendation can be classified according to (1) whether customers for whom we want recommendations are all customers or selective customers, (2) whether the objective of recommendation is to predict how much a particular customer will like a particular product, or to identify a list of products that will be of interest to a given customer (top-*N* recommendation problem), and (3) whether the recommendation is accomplished at a specific time or persistently.

The abundance of large data collections and the need to extract hidden knowledge within them has triggered the

* Corresponding author. Address: Departamento de Sistemas Informáticos, Escuela Politécnica Superior de Albacete, Universidad de Castilla-La Mancha, 02071 Albacete, Spain. Tel.: +34 967599200; fax: +34 967 599224.
E-mail addresses: enrique@umh.es (E. Lazcorreta), federico@umh.es (F. Botella), caballer@dsi.uclm.es (A. Fernández-Caballero).

0957-4174/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2007.08.048

development of algorithms to detect unknown patterns in data sets (Han & Kamber, 2001). A paper (Ozmutlu, Spink, & Ozmutlu, 2002) reports results from a study using Poisson sampling to develop a sampling strategy to demonstrate how sample sets selected by Poisson sampling statistically effectively represent the characteristics of the entire data set. Moreover, clustering analysis is a data mining technique developed for the purpose of identifying groups of entities that are similar to each other with respect to certain similarity measures. In the past, different ways to discover groups using clustering techniques have been proposed (Schafer, Konstan, & Riedl, 2001). Very often, they are based on different definitions of similarity measure to represent the closeness between users. Users can also be grouped based on the transactions they perform (Wang, Lim, & Hwang, 2006). In Perkowitz and Etzioni (2000) a cluster mining algorithm – an unsupervised algorithm for efficiently identifying a small set of high-quality (and possibly overlapping) clusters with limited coverage – is introduced.

Nevertheless, the existing researches could not afford to give a formal way for capturing individual customer's preference or associations among products through web usage mining. Given a set of transactions where each transaction is a set of items (itemset), an association rule implies the form $X \Rightarrow Y$, where X and Y are itemsets; X and Y are called the body and the head, respectively. The support for the association rule $X \Rightarrow Y$ is the percentage of transactions that contain both itemset X and Y among all transactions. The confidence for the rule $X \Rightarrow Y$ is the percentage of transactions that contain itemset Y among transaction that contain itemset X. The support represents the usefulness of the discovered rule and the confidence represents certainty of the rule. Association rule mining is the discovery of all association rules that are above a user-specified minimum support *minsup* and minimum confidence *minconf* (Tseng & Lin, 2007). Apriori algorithm is one of the prevalent techniques used to find association rules (Agrawal, Imielinski, & Swami, 1993; Agrawal & Srikant, 1994). Apriori operates in two phases. In the first phase, all itemsets with minimum support (frequent itemsets) are generated. This phase utilizes the downward closure property of support. In other words, if an itemset of size k is a frequent itemset, then all the itemsets below $(k-1)$ size must also be frequent itemsets. Using this property, candidate itemsets of size k are generated from the set of frequent itemsets of size $(k-1)$ by imposing the constraint that all subsets of size $(k-1)$ of any candidate itemset must be present in the set of frequent itemsets of size $(k-1)$. The second phase of the algorithm generates rules from the set of all frequent itemsets.

Association rule mining, as originally proposed in Agrawal et al. (1993) with its Apriori algorithm, has developed into an active research area. Association rule discovery and classification are analogous tasks in data mining, with the exception that classification main aim is the prediction of class labels, while association rule mining discovers associations between attribute values in a data set (Thabtah, Cowling, & Hammoud, 2006). Many additional algorithms have been proposed for association rule mining (e.g. Pujari, 2001; Lin & Kedem, 2002). End users of association rule mining tools encounter several well-known problems in practice. First, the algorithms do not always return the results in a reasonable time. A fuzzy mining algorithm based on the AprioriTid approach to find fuzzy association rules from given quantitative transactions has been proposed for reduced time complexity (Hong, Kuo, & Wang, 2004). Further, the association rules sets are sometimes very large. In Palshikar, Kale, and Apte (2007) a concept called a heavy itemset is proposed to compactly represent the association rules. An algorithm named as BitTableFI (Dong & Han, 2007) has significant difference from the Apriori and all other algorithms extended from Apriori. It compresses the database into BitTable, and with the special data structure, candidate itemsets generation and support count can be performed quickly. Also, mining association rules with multiple minimum supports is an important generalization of the association rule mining problem. Instead of setting a single minimum support threshold for all items, (Liu, Hsu, & Ma, 1999) allow users to specify multiple minimum supports to reflect the natures of the items, and an Apriori-based algorithm, named MSapriori, is developed to mine all frequent itemsets. In a recent paper (Hu & Chen, 2006), the same problem is suited but with two additional improvements. Another approach is the cluster-based association rule (CBAR) method (Tsay & Chiang, 2005), aimed to create cluster tables by scanning the database once, and then clustering the transaction records to the kth cluster table, where the length of a record is k. Moreover, the large itemsets are generated by contrasts with the partial cluster tables. This not only prunes considerable amounts of data reducing the time needed to perform data scans and requiring less contrast, but also ensures the correctness of the mined results. In Lee, Hong, and Lin (2005) another point of view about defining the minimum supports of itemsets when items have different minimum supports is provided. The maximum constraint is used, and then a simple algorithm based on the Apriori approach to find the large-itemsets and association rules under this constraint is introduced. Another interesting proposal is to utilize methods and techniques from Information Retrieval (IR) in order to assist data mining functions (Kouris, Makris, & Tsakalidis, 2005).

In this paper a new method towards automatic personalized recommendation based on the behavior of a single user in accordance with all other users in web-based information systems is introduced. The proposal applies a modified version of the well-known Apriori data mining algorithm to the log files of a web site to guide the users to the selection of the best user-tailored links. The paper mainly analyzes the process of discovering association rules in this kind of big repositories and of transforming them into user-adapted recommendations.