Data mining as a tool in Personalized web sites

Ihsan Salman Jasim
University of Dyala / Internet and Computer Center

Abstract: this paper studies the problem of putting data mining techniques in personalized web services as assistance tool for collaborative filtering tool of personalization. Collaborative filtering collect information for the customer from the opinion of other people in some fields, by matching the favorites of customer with others like, matching the same data and meta data, here in this article will consider the algorithm of A-prior from association rule method of data Mining as assistance tool to get collaborative filtering from the customer baskets to find out the associate things normally other customers likes to buy within their baskets.

Introduction

In the growth of the world wide web and the deep using of this international network, and ingoing of the users experience, we can not consider the wide web is just a complex list of millions of life things the users have to mine about what they want, actually the modern technology give this net some live and become more than an advertise material work through our computers, it become like learning machines through the ideas of semantic web and employee a new technologies in that purpose.

"The web has become a borderless marketplace for purchasing and exchanging goods and services. While web users search for inspect and occasionally purchase products and services on the web, companies compete bitterly for each potential customer. The key to winning this competitive race is *knowledge* about the needs of potential customers and the ability to establish personalized services that satisfy these needs"[11].

One of these technologies is *personalization* which is ongoing process to deliver a suitable content for the user, and customers of the web sites by anticipate there needs [5]. This technology became more familiar in use for the customers by using their tools like (filling profiles, collaborative filtering, click stream, …and others) or using voting bills in sites or automatically giving the customers statistic about the visitors and their opinions, favorites, choices, which help them in getting what they want.

Personalization have tools techniques can be improved by using other techniques like what we trying in this article, using *association rules data mining* process to assist the tool of collaborative filtering in a huge database and have a lot of different items like (any store in marketing , or libraries) in sites employing personal services.

Personalization

Personalization is a process to learn about customers and their buying behaviors preferences, and then using these data to customize offers [10].

In marketing environment, the purposes of applying information technology to provide personalization are expressed by the personalization consortium [4] as to:

- Better serve the customer by anticipating needs.
- Make the interaction efficient and satisfying for both parties.
- Build a relationship that encourages the customer to return for subsequent purchases.

In fact the software that carryout personalization techniques is not just a program done for once but it is consider an ongoing process because of the user needs and favorites is changing during time, so it "involves process of gathering user-information during interaction with the user"[4].

Absolutely this effort of process, spent of time, code space must be important to be given from the owners and designers of the business web sites, that they get some important personal information about the customer without effecting their privacy [13], so they can better serve them and finally getting loyal customers, encourage to return for shopping transactions "building customer loyalty by building meaningful one-to-one relationship"[7], in the many to many world wide web.

Personalization techniques

There are growing number of techniques for personalization, starting from "filling profile" which can be found in different styles , and click stream , targeted E-mail, personal toolbar, collaborative filtering, cookies, … etc[1],[3],[4].

The technique will consider is collaborative filtering as a technique assist by association rules using Apriori-algorithm.

Collaborative filtering: This technique compares a user’s tastes with those of other users in order to build up a picture of like-minded people.[4] The choice of content is then based on the assumption that this particular user will value that which the like-minded people also enjoyed. The preferences of the community of like-minded people are used to predict appropriate content. The user’s tastes are either inferred from their previous actions (for example buying a book, or viewing a product is assumed to show an interest (or taste) for that product) or else measured directly by asking the user to rate products.

The reliance on a ‘critical mass’ of users can be a problem for collaborative filtering[8]:

- A small sample population may lead to lower-quality recommendations.
- The quality of recommendations increases with the size of the user population.

Another potential limitation is the inability to make a recommendation for an unusual user if a match with a like-minded set cannot be found. Collaborative filtering may be less important as a technique when categories of users and preferences are already well-known and well-defined.

Data Mining

When people engage in (information-seeking behavior), it's usually because they are hope to resolve some problem, or achieve some goal, for which there current state of knowledge is inadequate "not enough"[2].

Data mining is the process of extracting or "mining" knowledge from large amount of data [6],also "called machine learning or knowledge discovery"[10]. Data mining essentially is a step of knowledge discovery which consist of the steps:

Data cleaning, data integration, data selection, data transformation, data mining, pattern evaluation, knowledge representation.

Data mining work of different kinds of data repositories like relational data base, data warehouses, transactional database, object-oriented, … and others, here transactional web data base will be consider which simply a table represent the customer basket consist of two field at least transaction number, and the list items been deal with like list of items customer buy.

Data mining functionalities are used to specify the kind of patterns to be found in data mining tasks. In general, classify into two categories: (1) descriptive and (2) predictive. Descriptive mining tasks characterize the general properties of the data in the data base. Predictive mining tasks perform inference on the current data in order to make predictions.[6] *In this article will consider the predictive category*.

Data mining functionalities include the discovery of concept/class description, association, classification, prediction, clustering,[6] …and other functions. *In this article will consider the association analysis*.

Association analyses: is the discovery of association rules, showing attributes-value conditions that occur frequently together in a given set of data. This analysis widely used for market basket or transaction data analysis. The association rules can discover buying patterns and link the presence of one item in a transaction to another item in the transaction [9].

Association Rules mining finds interesting association or correlation relationships among a large set of data items, as example: "one may discover a set of symptoms frequently occurring together with certain kinds of diseases and further study the reasons behind it".

There are algorithms to perform association rules, *here will consider the Apriori algorithm* because it's results match the need of collaborative filtering technique

<u>The Apriori algorithm</u> [6]: finding frequent itemsets using candidate generation.

Apriori is an influential algorithm for mining frequent item sets for Boolean association rules. The name of the algorithm is based on the fact that the algorithm use prior knowledge of frequent item set properties, as we shall see below. Apriori employs an iterative approach known as a level-wise search, where k-itemsets are used to explore (k+1)-itemsets. first, the set of frequent 1-itemsets is found .this set is denoted L1.L1 is used to find L2, the set of frequent 2-itemsets, which is used to find L3, and so on, until no more frequent k-itemsets can be found. The finding of each L_k requires one full scan of the database.

The proposal work

We can put these tables in any database platform like MS-ACCESS employing SQL queries to get the above (simple and important) results.

As have been seen above we can employ the result of Apriori algorithm to give recommendations to customers from behaviors of others registered at the website by analyzing the data using data mining techniques, and that can be done through three steps or phases:

1- Data gathered in web site and arranged from personal web site, (gathering data, personalization phase).
2- Process a data mining association rule, to get of association items, (data mining phase).
3- Then: the results used by the personal site to be used in collaborative filtering tool, (final phase).

Experimental and results

Lets D be a database of transactional basket in shopping session, (figure1-a), and let this transaction in a personalized web service, if we deploy a collaborative filtering process only, we will get an offer given to the visitor or costumer about what the other costumers bay table in (figure 1-b).

Transaction No.	**Items**
1	I1, I2, I5
2	I2, I4
3	I2, I3
4	I1, I2, I4
5	I1, I3
6	I2, I3
7	I1, I3
8	I1, I2, I3, I5
9	I1, I2, I3

(A) transaction table

Itemset	**Sup. count**
I1	6
I2	7
I3	6
I4	2
I5	2

(B) item frequencies can be used as offers by collaborative filtering

Figure 1: implementing collaborative filtering on transaction table

While if we used Apriori process with it we can get the almost shared items appear together in most transactions in the site database, the result of Apriori can be given to the costumers as recommendation offers they can take it, and this can be done through the following steps.

1- In the first iteration of the algorithm, each item is a member of the set of candidate 1-itemsets; C_1.The algorithm simply scans all of the transactions in order to count the number of occurrences of each item(figure 2-a).

2- Suppose that the minimum transaction support count required is 2 (i.e.,min_ sup=2/9=22%). The set if frequent 1-itemsets, L_1, can then be determined. It consists of the candidate 1-itemsets satisfying minimum support (figure 2-a).

3- To discover the set of frequent 2-itemsets, L_2, the algorithm uses $L_1 \times L_1$ to generate the candidate of 2-itemsets, C_2 (figure 2-b).

4- Next, the transaction in D are scanned and the support count of each candidate item set in C_2 is accumulated ,as shown in the second table of the second row (figure 2-b).

5- The set of frequent 2-item sets, L_2, is then determined, consisting of those candidate 2-item sets in C_2 having minimum support .

6- The generation of the set of the candidate 3-itemsets, C_3, is detailed in figure. First, let $C_3=L_2 \times L_2$ = {{I1,I2,I3}, {I1,I2,I5}, {I1,I3,I5}, {I2,I3,I4}, {I2,I3,I5} ,{I2,I4,I5}}. Based on the Apriori property that all subset of a frequent itemset must be frequent, we can determine that the four latter candidates cannot possibly be frequent. We therefore remove them from C_3, thereby saving the effort of unnecessarily obtaining their counts during the subsequent scan of D to determine L_3. Note that when given a candidate k-itemset, we only need to check if its (k-1)-subset are frequent since the Apriori algorithm uses a level-wise search strategy (figure 2-c).

7- The transaction in D are scanned in order to determine L_3, consisting of those candidate 3-itemsets in C_3 having minimum support (figure 2-c).

8- The algorithm uses $L_3 \times L_3$ to generate a candidate set of 4-itemsets, C_4 although the join results in {{I1,I2,I3,I4}}, this itemset is pruned since its subset {{I2,I3,I4}} is not frequent. Thus $C_4 = \emptyset$, and the algorithm terminates, having found all of the frequent itemsets.

There is no candidate of C4 with acceptable support threshold … so the algorithm terminates with triple candidates in this example (figure2).

Transaction No.	Items
1	I1, I2, I5
2	I2, I4
3	I2, I3
4	I1, I2, I4
5	I1, I3
6	I2, I3
7	I1, I3
8	I1, I2, I3, I5
9	I1, I2, I3

Scan D for count of each candidate C1:

Itemset	**Sup. count**
I1	6
I2	7
I3	6
I4	2
I5	2

Compute candidate support with minimum support count L1:

Itemset	**Sup. count**
I1	6
I2	7
I3	6
I4	2
I5	2

(A)

Compute C2 candidates from L1:

C2

Itemset
(I1, I2)
(I1 , I3)
(I1 , I4)
(I1 , I5)
(I2, I3)
(I2, I4)
(I2,I5)
(I3,I4)
(I3,I5)
(I4,I5)

scan D for count of each candidate:

C2

Itemset	**Sup. count**
(I1,I2)	4
(I1,I3)	4
(I1,I4)	1
(I1,I5)	2
(I2,I3)	4
(I2,I4)	2
(I2,I5)	2
(I3,I4)	0
(I3,I5)	1
(I4,I5)	0

Compute candidate support count with minimum support count:

Itemset	**Sup. count**
(I1,I2)	4
(I1,I3)	4
(I1,I5)	2
(I2,I3)	4
(I2,I4)	2
(I2,I5)	2

L2

(B)

Generate C3 candidates from C2:

C3

Itemset
(I1, I2, I3)
(I1, I2, I4)
(I1, I2, I5)

Scan D for count of each candidate:

C3

Itemset	**Sup. count**
(I1, I2, I3)	2
(I1, I2, I4)	1
(I1, I2, I5)	2

Compute candidate count with minimum support count:

L3

Itemset	**Sup. count**
(I1, I2, I3)	
(I1, I2, I5)	

(c)

Figure1 Tables shows the process of Apriori algorithm

(A) Candidates with 1-itemsets

(B) Candidates with 2-itemsets

(C) Candidates with 3-itemsets

Measures

Here in the above example has been used the support measure to test the items while pass the threshold or not, the support give as follow [6]:

Support(set of items)=(tuples containing he set of items) / (total of tuples).

conclusions

1. In talking about personalization tool "collaborative filtering" the collecting information will be not just from the main data but also from the meta data and the discovering data from association rule.
2. The idea of using association rules in collaborative filtering or collaborative methods not only using the meta data of the content-based materials, but add what in the users baskets from the associate items "the meta data of these items", from the double items, triple items, and so on.
3. Hyam Hirsh, Chumki Basu, Brian D, Davison put question in "Learning to personalize". How to predict user interests? They answer that the

collaborative methods represent one method of collect information about user interests, with in the principle of "word of mouth", means others opinion of items, and say content-based methods exploit one kind of information, and collaborative methods exploit a second kind of information."[8]. And here in this article tried to put a third method by using association rules to users baskets to collect the most association items in baskets to recommend the same others interests to the customers.

4. Working with data mining should be done carefully from the beginning because the prepared query will relate a lot on different items in the huge data base, so to give a suitable and right results it must be done carefully.
5. Data Mining and Personalization are process methods based on statistical approaches, which depend on the related items (or different kind of items), this leads it's depend of correctness of the information gathered from customers or automatically, this correctness of information will led to correct the results of extract information from the data mining phase and then led to stable personalization system.

References

[1] Barry, snyth and Paul, cotter, "A personalized Television listings services", communication of the ACM, August 2000.

[2] Belkin, Nicholas J. "Helping People Find What They Don't Know", Communication of the ACM, august 2000, vol 43, No.8.

[3] Bonett Monica, "An Architecture for Personalization services within subject gateways", 2002, http://www.imesh.org/toolkit/work/components/personalization/ architecture.

[4] Bonett. Monica, ”Personalization of Web Services: opportunities and challenges", 2001, Vol 43, No.8. www.ariadne.ac.uk/issue28/personalization/intro.html .

[5] Edith Schonberg, Thomas Cofino, "Measuring Success", communication of the ACM, August 2000, Vol 43, No.8.

[6] Han, Jiawei. Kamber, Micheline. ,"data mining: concepts and techniques",Morgan Kaufmann Publisher, 2001.

[7] Riecken, doug, "Personalized Views of Personalization", Communication of the ACM, august 2000, vol 43, No.8.

[8] Robert Seidman (Seidman's On Line Insider), August 1998, Vol. 5, Issue 24. http://www.goodreports.com.

[9] S.Saxe, Richard,"website personalization using data mining and active database techniques".

[10] Siegel, Carolyn, "internet marketing: foundations and applications", Houghton Mifflin Company, 2004.

[11] Spiliopoulou, Myra. "Web Usage Minning for Web Site Evaluation", Communication of the ACM, august 2000, vol 43, No.8.

[12] The Personalization Consortium , www.personalization.org/personalization.html.

[13] Volokh, Eugene, "Privacy in Personalized website", Communication of the ACM, august 2000, vol 43, No.8.

خلاصة:

في هذا البحث تم دراسة مسألة وضع تقنيات تنقيب البيانات (data mining) في خدمات الويب المخصصة (personalization) كأداة مساعدة لتقنية التنقية التشاركية (collaborative filtering) حيث تقوم بجمع المعلومات عن الزبون من خلال رأي المشاركين الأخر في مجال معين ، باستخدام الأداة وحدها يتم مطابقة ما يفضله الزبون مع ما يفضله المشاركين في نفس المجال بمطابقة البيانات ومتعلقاتها، في هذا البحث تم توظيف خوارزمية (A-prior) من طريقة قانون التشارك (association rule) في تنقيب البيانات كأداة مساعدة للـ (collaborative filtering) لفعاليات الزبون لإيجاد الأمور المشتركة التي يفضلها باقي الزبائن مع البيانات المشتركة في جولاتهم بالموقع.