A Product Recommendation System using Vector Space Model and Association Rule

Debajyoti Mukhopadhyay[1,4], Ruma Dutta[2,4], Anirban Kundu[2,4] and Rana Dattagupta[3]
[1]Calcutta Business School, D.H. Road, Bishnupur 743503, India
debajyoti.mukhopadhyay@gmail.com
[2]Netaji Subhash Engineering College,Garia, Kolkata 700152, India
{rumadutta2006, anik76in}@gmail.com
[3]Jadavpur University, Kolkata 700032, India
rdattagupta@cse.jdvu.ac.in
[4] WIDiCoReL, Green Tower C- 9/1, Golf Green, Kolkata 700095, India

Keywords: e-commerce, vector space model, association rule, Cellular Automata (CA), Single Cycle Multiple Attractor Cellular Automata (SMACA).

Abstract: This paper presents an alternative product recommendation system for Business-to-customer e-commerce purposes. The system recommends the products to a new user. It depends on the purchase pattern of previous users whose purchase pattern are close to that of new user. The system is based on vector space model to find out the closest user profile among the profiles of all users in database. It also implements Association rule mining based recommendation system, taking into consideration the order of purchase, in recommending more than one product. To make the association rule memory-efficient, cellular automata is used.

1 Introduction

World Wide Web makes the life of customers simpler by introduction of e-commerce where commercial activities can be done from own location. Most customers like to have a recommender system by which customers can see the feedback from other users who already purchased the products. This need gives rise to the demand of a product recommendation system. E-commerce makes use of recommender systems in one step ahead which not only shows the feedback from other users but also suggests interesting and useful products to customers. They provide consumers with information that is intended to support their recommendation activities. Recommender systems research is mainly motivated by the need to cope with information overload, lack of user knowledge in a particular domain. There are two types of widely used product recommendation techniques. They are Automated Collaborative Filtering (ACF) and Case-Based Reasoning [2].

The first one, ACF [1] is widely used as the technique for product recommendation in an online store. This approach is based on the feedback given by the previous customers and their feedback is used to recommend the product to new customer. For example, let us suppose there are three customers 1, 2 and 3. They all have purchased A, B, C products. Additionally, customer 1 and 2 purchased the products D and E. As customer 3 has common interest with customers 1 & 2, the products D and E are recommended for customer 3. Some ACF system can provide the reason and data behind recommendation. Most ACF recommendation system use the formula given in equation (i) which also known as **mean squared difference formula.**

$$\delta_{UJ} = \frac{1}{|InCommon|_{fe}} \sum_{InCommon} (U_f - J_f)^2 \quad \text{(i)}$$

This formula is used to calculate the difference between two persons U and J, in terms of their interests on a product. U_f and J_f are the ratings of U and J on the feature f of the product. This paper utilizes the vector space model rather than mean squared difference formula.

ACF approaches can be classified as non-invasive and invasive approaches, based on how the user's preferences are recorded in an ACF system [1]. In invasive approach, user's ratings are floating numbers between 0 and 1. Non-invasive approach, the preferences are Boolean values i.e. 0 & 1. For an example, there are four products P1, P2, P3 and P4. User 1 has used the products P1, P2, P4. In non-invasive approach, the ratings would be 1, 1, 0, 1. In invasive approach it may be 0.2, 0.5, 0 and 0.6. The values 0 indicate that the User 1 has not rated the product P3. Obvious problem with non-invasive approach is, the product P1 is rated low by User1 but

978-0-7695-3513-5/08 $25.00 © 2008 IEEE
DOI 10.1109/ICIT.2008.48

user's rating will be 1 for P1. For this reason, the non-invasive approaches require feedback from more users to recommend any product.

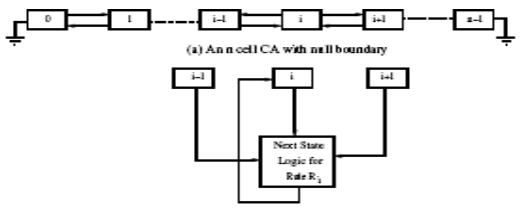
The other approach of product recommendation is Case-Based Reasoning. A widely used formula for CBR in identifying and recommending similar products is nearest neighbor retrieval, which is based on weighted Euclidian distance [3]. Learning the similarity and the utility of the retrieved cases is another important problem addressed by some CBR approaches. These approaches give importance on:

- Learning the similarity measures between cases without a need for a pre-classification among them [3]
- Acquiring the preferences of users from the return sets of products

This paper proposes a hybrid recommender system based on ACF and Association Rule mining, taking into consideration the sequence of purchase. Using association rule derived from access log of the users, this system recommends more than one product. This recommendation can result into some products which are not available in the recommendation system database. This paper also outlines the implementation of association rule through cellular automata to minimize the memory.

2 Cellular Automata Preliminaries

An $n\psi$ cell Cellular Automata (CA) consists of $n\psi$ cells (Figure 1(a)) with local interactions. It evolves in discrete time and space. The next state function of three neighborhood CA cell (Figure 1(b)) can be represented as a rule as defined in Table 1 [7]. First row of Table 1 represents $2^3 = 8$ possible present states of 3 neighbors of $i^{th}\psi$cell - (i-1), i, (i+1) cells. Each of the 8 entries (3 bit binary string) represents a minterm of a 3 variable boolean function for a 3-neighbourhood CA cell.



(b) The i^{th} cell configured with rule R_i

Figure 1. Local Interaction between Cellular Automata Cells

In subsequent discussions, each of the 8 entries in Table 1 is referred to as a Rule Min Term (RMT). Each of the next five rows of Table 1 shows the next state (0 or 1) of $i^{th}\psi$ cell. Hence there can be $2^8 = 256$ possible bit strings. The decimal counterpart of such an 8 bit combination is referred to as a CA rule [7].

The decimal equivalent of 8 minterms are 0, 1, 2, 3, 4, 5, 6, 7 noted within ()-below the three bit string.

In this paper, Single Cycle Multiple Attractor Cellular Automata (SMACA) has been used. Typically, a non-linear SMACA consists of 2^n number of states where $n\psi$ is the size of SMACA. The structure of a non-linear SMACA has attractors (self-loop or single length cycle), non-reachable states, and transient states. The attractors form unique classes (basins). All other states reach the attractor basins after certain time steps. Typical SMACA is shown in Figure 3.

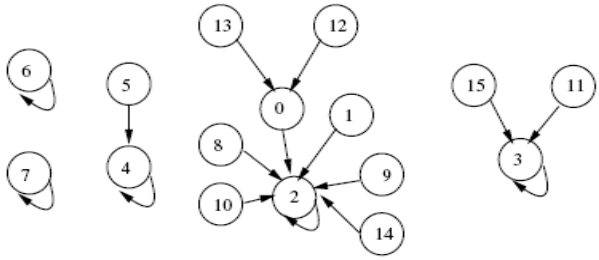


Figure 2: State Transition Behavior of Non-linear SMACA with Rule Vector RV <192 12 207 200>

Definition 1: Reachable State- A state having 1 or more predecessors is a reachable state.
Definition 2: Non-reachable state- A state having no predecessors is termed as non-reachable state.
Definition 3: Transient state- A non-cyclic state of a nongroup CA is referred to as transient state.
Definition 4: Cyclic state- A state in a cycle of the state transition behavior of a CA.
Definition 5: Attractor Cycle- The set of states in a cycle is referred to as an attractor cycle.

3. Our Approach

Our recommendation system emphasizes three key areas. These are

- Introducing Vector space model in ACF which is dealt in section 3.1
- Space optimization in FP-growth algorithm dealt in 3.2
- Recommending more than one product dealt in section 3.3

3.1 Vector Space Model in ACF

Vector Space model is mainly used in information retrieval [5]. Every document is looked upon as vector. The frequencies of terms are the components of vectors. Then Cosine Similarity (equation (ii)) is the process by which the similarity between two document vectors is measured [6]. The Cosine Similarity of two vectors (d1 and d2) is defined as

$$\text{Cos(d1,d2)} = \frac{dot(d1, d2)}{\| d1 \| \| d2 \|} \qquad \text{(ii)}$$

Table 1.Truth Table of sample rules of a CA cell showing the next state logic for the Minterms of a 3 variable boolean function-The 8 minterms having decimal values $0 \leftarrow \psi 1 \leftarrow \psi 2 \leftarrow \psi 3 \leftarrow \psi 4 \leftarrow \psi 5 \leftarrow \psi 6 \leftarrow \psi 7$-are referred to as Rule Minterms (RMTs)

Note : Set of Minterms $T\psi = \neg 7 \leftarrow \psi 6 \leftarrow \psi 5 \leftarrow \psi 4 \leftarrow \psi 3 \leftarrow \psi 2 \leftarrow \psi 1 \leftarrow \psi 0 \diamondsuit \leftarrow$represented as ¶T(7), T(6), T(5), T(4), T(3), T(2), T(1), T(0)$\diamondsuit \leftarrow$(T(m)=m, m=$0 to 7$) in the text, are noted simply as q.

Present states of 3-neighbours $(i\psi \bullet \leftarrow 1) \leftarrow \psi i \leftarrow \psi$and $(i\psi + 1)$-of $ith\psi$cells (Minterms of a 3 variable boolean function)	111 (7) T(7)	110 (6) T(6)	101 (5) T(5)	100 (4) T(4)	011 (3) T(3)	010 (2) T(2)	001 (1) T(1)	000 (0) T(0)	Rule Number
Next state of $ith\psi$cell	0 1 1	1 0 1	0 0 0	1 1 1	1 0 0	0 1 0	1 1 1	0 0 0	90 150 210

where,
dot(d1,d2) = d1[0]*d2[0]+d1[1]*d2[1]+……
||d1|| = sqrt (d1[0]^2+d2[0]^2……..)
Same Cosine Similarity can be used in ACF. The Product rating of each user can be viewed as a vector and for a user the rating of each product is the component of the vector. Let us suppose, there are two users U1 and U2. There are five products P1, P2, P3, P4, P5. The rating of User1 is 0.5, 0.6, 0, 0.7, 0.8. The vector of User1 is represented as (5, 6,0,7, 8). Similarly for User2 whose rating is 0.5, 0.6, 0.6, 0.2, 0.9 will be represented by vector (5, 6, 6, 2, 9). The User3 had purchased first three products P1, P2, P3 and ratings for those products are .4,.5,.6. The User3 vector will be represented as (4,5,6). Now, product has to be recommended for User3. As User3 purchased only P1, P2, P3, the vector considered for User1 will be (5, 6, 0) and User2 will be (5, 6, 6). Now the Cosine Similarity value for User 1 and User 3 and for User2 and User3 is .99. So User2 is more similar to User3 than User1 and the product P5 is recommended. As User 2 has rated low for product P4, this product is not recommended for User3.

3.2 Association Rule

The association rule [5] can be found out by popularly known FP-Growth Algorithm. The Apriori algorithm is avoided purposefully to avoid the problem of candidate generation. The main disadvantage of FP-growth Algorithm is , it takes huge memory while storing in tree structure.
The main disadvantage of FP-growth Algorithm is, it takes huge memory while storing in tree structure. This problem can be overcome by using SMACA. In this approach, the root of FP-tree is the attractor in SMACA. Each non-reachable and transient state is the product. Each node of FP-tree is represented a state in SMACA structure, randomly generated. There is a mapping of each state to product and there is tag of count with each state. In Implementing SMACA the actual nodes need not be stored in memory, only the rules satisfying the structure of nodes need to be stored. For example if the state transition behavior is like Figure 2, then only the rule vector <192 12 207 200> needs to be stored. That means 16 nodes can be realized through 4 bytes. For large item set, substantial amount of memory can be saved in this approach. The theory behind the synthesis of SMACA rule is given in [6]. Let us suppose there are 5 transactions and 5 items. The transactional data is shown in table 2. The order of purchase time is considered here.

TID	List of Item_ID
T100	P2,P1, P5
T200	P2, P4
T300	P2, P3
T400	P2, P1, P4
T500	P1, P3

Table 2: Transaction database example

The first scan of the database derives the set of frequent item sets (1-itemset) which counts the occurrence of each item. The 1-itemset of the example database is {P1:3, P2: 4, P3:2, P4:2, P5:1}. Let us consider the minimum support count is 2 and items are sorted in decreasing order of frequencies.

So the list L after first scan is {P2:4, P1:3, P3:2, P4:2, P5:1}. The FP-tree is constructed as shown in Figure 3 .This tree can be realized as shown in Figure 4.

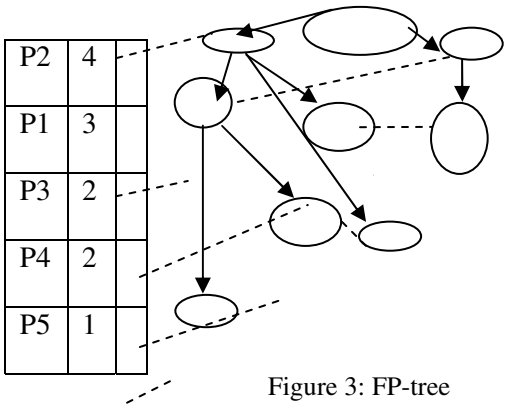


Figure 3: FP-tree

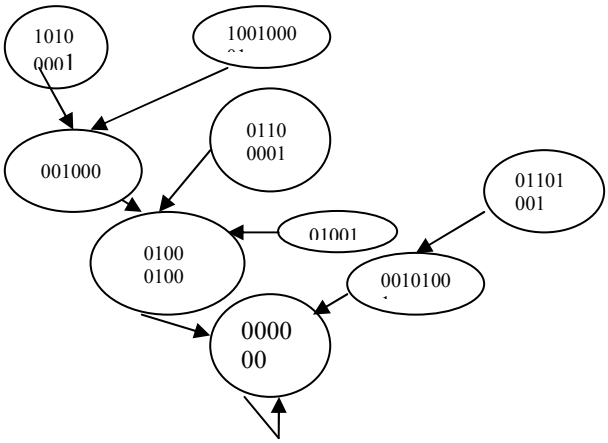


Figure 4: SMACA structure of FP-tree

3.3 Use of Association rule in Product Recommendation

For recommending more than one product, we can make use of association rule. Suppose there are more products in addition to above products.

Those products are P5, P6, P7 and these products are bought after P4. Suppose we find the association rule P4 => P6. Then in addition to the product P4, the product P6 will be recommended also. The Algorithm 1 contains the steps necessary to recommend the products

Algorithm 1: Recommendation of product

Input: D = {Pl1, Pl2,….Plk} (Database of Plans), Plr (The sorted Plan for User for whom the recommendation is being carried out), s (Support)

Output: Recommended products

Step 1: Use Vector Space model on D and Plr to find the recommended product

Step 2: Incorporate the recommended product in Plr

Step3: Find all Association rule using FP-growth algorithm

Step 4: If the Plr conforms the Association rules
Then recommend the product
Recommend the products which satisfy the association rules and sequence of D
Else
Recommend the product which is next rated satisfying association rule and sequence of D
Repeat Step3 thru Step4.

Step5:End

4. Conclusion

This paper introduced vector space model in e-commerce area. Recommendation more than one product has been proposed in this paper. This recommendation system uses Web log in making recommendation. This paper also used SMACA in implementing association rule to minimize the memory requirement

References

[1] B. Prasad, 2007, "A Knowledge-based product recommendation system for e-commerce," *International Journal of Intelligent Information and Database Systems,* Volume 1, No.1, ISSN:1751-5858. Inderscience Publishers..

[2] A. Stahl, 2001, "Learning feature weights from case order feedback," *Proceedings of the 4th International Conference on Case-Based Reasoning.* Lecture notes in Artificial Intelligence , Springer 2080, pp.78-84.

[3] D.Wettschereck, D.W. Aha, 1995, "Weighting features," *1st International Conference on Case-Based Reasoning.* Springer, New York, USA.

[4] J.Han, M.Kamber, 2005, "*Data Mining Concepts and Techniques*," Elsevier. India.

[5] G. Salton, 1989, "Automatic Text Processing, The Transformation, Analysis, and Retrieval of Information by Computer," Addison-Wesley.

[6] A. Kundu, R.Dutta and D. Mukhopadhyay, "Generation of SMACA and its Application in Web Services," *9th International Conference on Parallel Computing Technologies, Pact 2007 Proceedings*, Russia, September 3-8, 2007.

[7] S. Wolfram, "Theory and Application of Cellular Automata," World Scientific, 1986