XML based on Clustering Method for personalized Product Category in E-Commerce

KwonSoo Lee, MoonHyun Kim

noxangel@ece.skku.ac.kr , mhkim@simsan.skku.ac.kr

School of Information and Communication Engineering, SungKyunKwan University
300 Chunchun Jangahn Suwon, 440-746, KOREA

Abstract. *In data mining, having access to large amount of data sets for the purpose of predictive data does not guarantee good method, even where the size of Real data is Mobile commerce unlimited. In addition to searching expected Goods objects for Users, it becomes necessary to develop a recommendation service based on XML. In this paper, we design the optimized XML Recommender product data. Efficient XML data preprocessing is required, include of formatting, structural, and attribute representation with dependent on User Profile Information. Our goal is to find a relationship among user interested products from E-Commerce and M-Commerce to XDB. Firstly, analyzing user profiles information. In the result creating clusters with analyzed user profile such as with set of sex, age, job. Secondly, it is clustering XML data which are associative products classify from user profile in shopping mall. Thirdly, after composing categories and goods data in which associative objects exist from the first clustering, it represent categories and goods in shopping mall and optimized clustering XML data which are personalized products. The proposed personalized user profile clustering method has been designed and simulated to demonstrate it's efficient.*

1. Introduction

Now days, concerning a large amount of various data available on the Internet and E-Commerce, there are existing websites and shopping mall which provide recommended service for users to search interest data. For text and image data in web pages, the websites providing user log files, URL-based navigation, keyword-based searching and e-mail service or recommendation are developed, such as the recommendation system of Amazon, content-based filtering system of Web Watcher, personalization recommendation system ACRNews and search engine of Yahoo! (Yahoo shopping mall). In Recommendations system, however, the shopping malls providing such kinds of services are still limited. To solve the problem, i) Develop on structural shopping mall, that is, most of them are on the XML built for E-Commerce and M-Commerce. ii) Creating a XFCM Algorithm based on optimized recommendation data. iii) Having User Profile Information create cluster group, create user profile data based on created cluster group K-means Algorithm by Aprioi. Clustering group compose re-group with XFCM (XML based on Fuzzy Clustering and Merging) algorithm in filtering. As Re-grouped structural XML data of cluster, recommend goods data for users. In this paper, we propose about clustering method to be personalized to category and product data based on structural pattern.

The paper is organized as follows: Section 2 described clustering algorithm. Section 3 described cluster creation and composition based on FCM algorithm as proposal system. Section 4 presented the results on implementation and simulation using the XFCM applied user profile of patterns. Section 5 we proposed conclusion.

2. Related Work

Until present, most recommended systems that operate in web-based divide by three classifications; Manual Decision Rule System, Collaborative Filtering System, and Content-based Filtering Agent System are come hereupon. Manual Decision Rule System describe rule base that web site operator are collecting static profile, user session history through user's registration instruction. Representative example of this system is Broad vision. Broad vision provides an array of business solutions that is Content Management Solutions, Personalization Solutions, Commerce Solutions, and Enterprise Portal Solutions and so on, to meet your greatest challenges, from the power of personalization to robust content management to leading edge applications and enterprise portals.[1][7] In this way, Rules are influenced in contents that are offered to particular users. Collaborative Filtering System provide predicted information's that suitable with user's preference degree through correlation engine based on clarified information that equipped user's estimation or type of preference degree. Representative example is Net Perceptions. Content-based Filtering such as Web Watchers puts weight from user by express or in content similarity of web document about getting individual Profile allusively. [3][6] we are shown as follow Fig. 1., Fig. 2.

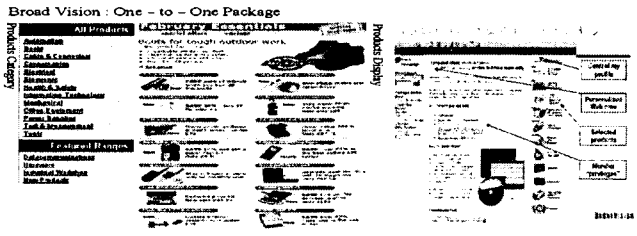

Fig. 1. Broad Vision: One - to - One Package

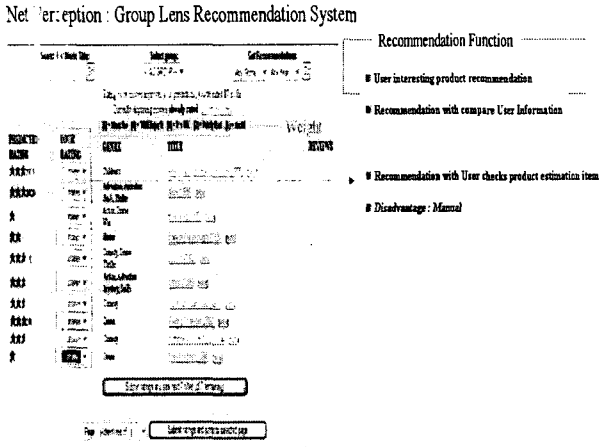

Fig. 2. Net Perception: Group Lens Recommendation System

Recommendation systems of such three types of preference degree are based on recommendation algorithm. Recommendation algorithm start by finding a set of users whose purchased and rated product overlaps the user's purchased and rated product. Recommendation algorithm aggregates product from these similar users, eliminates product the user has already purchased or rated, and recommend the remaining product to the user.[4][5] Two popular versions of these algorithms are collaborative filtering and cluster models. Other algorithms - including search-based methods and our own item-to-item collaborative filtering - focus on finding similar items, not similar users. For each of the user's purchased and rated items, the algorithm attempts to find similar items. It then aggregates the similar items and recommends them.

2.2 Clustering Algorithms

2.2.1 K-means Algorithm

The ***k***-means algorithm (MacQueen 1967, Anderberg 1973) is built upon four basic operations: i) selection of the initial ***k*** means for ***k*** clusters, ii) calculation of the dissimilarity between an goods object and the mean of a cluster, iii) allocation of an object to the cluster whose mean is nearest to the goods object, iv) Re-calculation of the mean of a cluster from the objects allocated to it so that the operating, the other three operations are repeatedly performed in the algorithm until the algorithm converges.[8] The essence of the algorithm is to minimize the cost function.

The k-means algorithm has the following important properties.

1. It is efficient in processing large data sets. The computational complexity of the algorithm is $O(tkmn)$, where m is the number of attributes, n is the number of objects, k is the number of clusters, and t is the number of iterations over the whole data set. Usually, $k, m, t << n$. In clustering large data sets the k-means algorithm is much faster than the hierarchical clustering algorithms whose general computational complexity is $O(n2)$ (Murtagh 1992).
2. It often terminates at a local optimum (MacQueen 1967, Selim and Ismail 1984). To find out the global optimum, techniques such as deterministic annealing (Kirkpatrick et al. 1983, Rose et al. 1990) and genetic algorithms (Goldberg 1989, Murthy and Chowdhury 1996) can be incorporated with the k-means algorithm.
3. It works only on numeric values because it minimizes a cost function by calculating the means of clusters.
4. The clusters have convex shapes (Anderberg 1973). Therefore. it is difficult to use the k-means algorithm to discover clusters with non-convex shapes.

One difficulty in using the k-means algorithm is to specify the number of clusters. Some variants like ISODATA include a procedure to search for the best k at the cost of some performance. But The k-means algorithm is best suited for data mining because of its efficiency in processing large data sets.[10]

2.2.2 ISODATA

This algorithm is based on the k-means algorithm, and employs processes of eliminating, splitting, and clustering. The algorithm is described as following.

1. Start with K_{init} (initial number of clusters) which is user-given. Assign the first K_{init} samples as cluster centers.
2. Assign all samples to the clusters by minimum distance principle.
3. Eliminate clusters that contain less than n_{min} feature vectors and reassign those vectors to other clusters to yield K clusters.
4. Compute a new cluster center as the average of all feature vectors in each cluster.
5. For each kth cluster, compute the mean-squared error $\sigma_n^2(k)$ of each nth component x_n over that cluster and find the maximum $\sigma_{n^*}^2(k)$ component mean-squared error over within cluster k for over n = 1, ..., N, where the index n* is for the maximum component.

6. If there are not enough clusters ($K_{init} < K/2$) and this is not the last iteration, then if $\sigma_{max}(k) > \sigma_{split}$ for any cluster k, split that cluster into two.
7. If this is an even iteration and $K_{init} > 2K$, then compute all distances between cluster centers. Merge the clusters that are close than a given value.

The advantages of the ISODATA are its self-organizing capability, its flexibility in eliminating clusters that are too small, its ability to divide clusters that are too dissimilar, and its ability to merge clusters that are sufficiently similar. But Some disadvantages are i) multiple parameters must be given by the user, although they are not known a priori, ii) a considerable amount of experimentation may be required to get reasonable values, iii) the clusters are ball shaped as determined by the distance function, iv) the value determined for K depends on the parameters given by the user and is not necessarily the best value, and v) a cluster average is often not the best prototype for a cluster.[9]

2.2.3 Fuzzy Clustering Algorithms

Fuzzy clustering plays an important role in solving problems in the areas of pattern recognition and fuzzy model identification. A variety of fuzzy clustering methods have been proposed and most of them are based upon distance criteria. One widely used algorithm is the fuzzy c-means (FCM) algorithm.[11],[12] It uses reciprocal distance to compute fuzzy weights. A more efficient algorithm is the new XFCM. It computes the cluster center using Gaussian weights, uses large initial prototypes, and adds processes of eliminating, clustering and merging after filtering in user clusters.[2] In the following sections we propose and simulation XML based on Fuzzy Clustering and Merging(XFCM).

3. Proposed System

3.1 System Architecture

The proposed system is intended to support all the purchasing activities, range from product search to payment, for E-Commerce and M-Commerce on the World Wide Web. The system consists of six main components as follows: i) ***Shopping Module*** gathers and filters product information from shops on the web. It translates product information into XML form automatically that is then transmitted to the mobile device of client. ii) ***Personalization Module*** provides a user data on preference and transaction. iii) ***Ordering Module*** gathers information related with order form of each shop and order product to the corresponding shop by filling in the order form. iv) ***M.P. Browser (Mobile Product Browser)*** is middle application which supports various types of product search on the mobile device. v) ***DATABASE*** keeps information on product that is gathered by the shopping Module including user's personal information and transaction information. vi) ***Ontology Server*** keeps a standard category information on product categories that are used in a number of representative online shopping malls.[13]

- ***Shopping Module*** : Search Agent, Filtering Agent and XML Translator
- ***Ordering Module*** : Order Agent and Retrieval Agent
- ***Personalization Module(Recommender System)*** : User Data Management Module and User Adaptation Module

User Data Management Module. This module manage user's personal data such as User ID, Name, Age, Sex, etc., which are necessary for settlement and login to each shop, and users historical purchase data about ordering and recommendation of goods.

User Adaptation Module. This module have recommend product with user's personal data and purchase data. It consists of two functional components, Clustering Module and Mining Module.

Mining Module. This Module selects the recommendation product by using XML based on K-Means algorithm by Apriori. It uses data about purchase information of customers who are in the same cluster. This data is clustered by Clustering Module.

Clustering Module. Clusters group based on the similarity of personal data of users. XFCM (XML based on fuzzy Clustering and Merging) algorithm is used for clustering users by FCM algorithm. The XFCM algorithm has efficient recommender product with structural cluster grouping that can solve the multiple parameters must be given by the user, although they are not known K-means, a considerable amount of experimentation may be required to get reasonable values, the clusters are ball shaped as determined by the distance function, the value determined for K depends on the parameters given by the user and is not necessarily the best value, and a cluster average is often not the best prototype for a cluster. Therefore, it can organize by itself, and control the intensity of classification through the vigilance parameter.

In this paper, we propose Clustering Module: Entire System Module is shown in Fig. 3. and FCM Algorithm pattern Model is shown in Fig. 4.

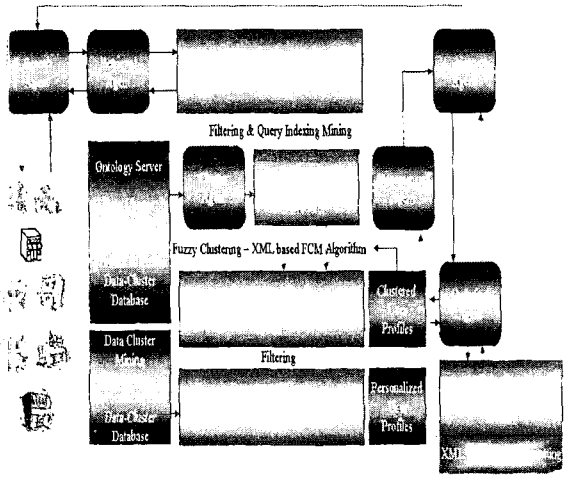

Fig. 3. Architecture of System

2 FCM Algorithm pattern

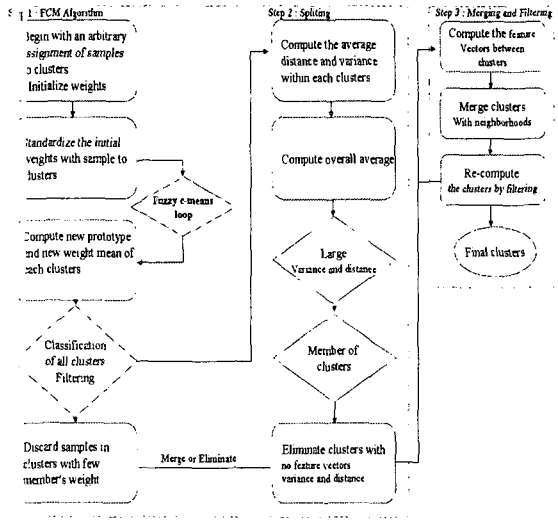

Fig. 4. FCM Algorithm Pattern Model

s shown in Fig 4, 5):

hase 1 : ***FCM Algorithm***

Step 1 . Begin with an arbitrary assignment of samples to clusters
: Initialize weights
Step 2. Standardize the initial weights with sample to clusters
Step 3. Fuzzy c-means loop
Step 4. Compute new prototype and new weight mean of each clusters
Step 5. Classification of all clusters Filtering
Step 6. Discard samples in clusters with few member's weight : merge or eliminate

Phase 2 : ***Splitting (Re-Filtering)***

Step 1 . Compute the average distance and variance within each cluster
Step 2 . Compute overall average
Step 3 . Large Variance and distance
Step 4 . Member of clusters
Step 5 . Eliminate clusters with no feature vectors variance and distance

Phase 3 : ***Merging and Filtering***

Step 1 . Compute the feature Vectors between clusters
Step 2 . Merge clusters with neighborhoods
Step 3 . Re-compute the clusters by filtering
Step 4 . Final clusters

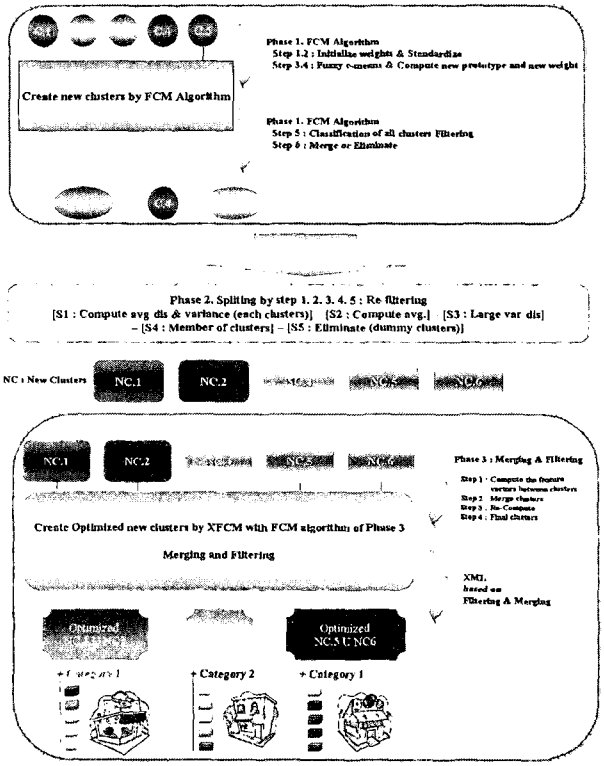

Fig. 5. Advanced FCM Algorithm Pattern Model

3.3 XFCM Algorithm applied FCM

XFCM Clustering is new cluster merging or eliminating from small cluster filtering part and cluster's central distance and similarity. This algorithm measure user propensity from user profile information, action, etc., and provide personalized recommends product rely on this measurement. Small cluster filtering part measure that rating of entire data and clustered data, similarity of other cluster. Measured small cluster except from merging step (Phase3-step2), improve efficiency of merging step.[12] Merging Part based on the i) most closer central distance (similarity

of cluster), ii) after selected 1 part, iii) performed using clustering validity between clusters. This clustering solved the Xie-Beni Validity. The Xie-Beni Validity result on compactness and separatio measurement by prototype. XFCM Algorithm use user's cluster group structural method by XML. XML is Data Independence, Improved domain knowledge, Improved data searches, Extension data.

This algorithm relies on XML Pattern Matching to analyze structured user profile pattern. The concept of this algorithm is as follow: i) Match the root element information (match= "Age, Sex, Job") that is selected via match = "/, *, |, @" pattern of four cases attribute to xsl : template by user profile. ii) The select= attribute of the xsl : value - of element works almost identically.

3.4 Proposal Module

XML Database Module: This Module compose Product Data, Action Data, Product and Action Data, Recommendation Data (New product, Event product, and so on). DataBase keeps information on product that is gathered by the Shopping Mall Data (Product Data, Action Data) including user's personal information, transaction information, basket placement (the placement of the product in the shopping basket).

Clustering Module: This module based on the similarity of User Profile Information. User profile compose Age, Sex, Job that is three terms resolve root, element of each node and each data inside database with user profile create clusters. Created clusters merge and filter, Result of optimized clusters reflected Mining Sequential XML Patterns Model. K-means algorithm is used for clustering users by Apriori. The K-means algorithm has the Characteristics that Classification, precedence work for estimate are use in various analysis such as loss value processing work as many as error value in addition to throng analysis.[11] It uses data about purchase information of customers who are in the same cluster. Clustering creation method is as follows :

Product Data

Ps = {Age} , {Job} : Clustering 1
{Age Union Job} , {Age Adjunction Job} : Clustering 2
{Sex} Union {Age Union Job} , {Sex} Adjunction {Age Adjunction Job} : Clusterin 3

Action Data

As = {Age} , {Job} : Clustering 1
{Age Union Job} , {Age Adjunction Job} : Clustering 2
{Sex} Union {Age Union Job} , {Sex} Adjunction {Age Adjunction Job} : Clustering 3

Product and Action Data

PAs = {Age} , {Job} : Clustering 1
{Age Union Job} , {Age Adjunction Job} : Clustering 2
{Sex} Union {Age Union Job} , {Sex} Adjunction {Age Adjunction Job} : Clustering 3

Recommendation Data

Rs = {Age} , {Job} : Clustering 1
{Age Union Job} , {Age Adjunction Job} : Clustering 2
{Sex} Union {Age Union Job} , {Sex} Adjunction {Age Adjunction Job} : Clustering 3

Mining Sequential XML Patterns Module: This Module selects the recommendation products by using XFCM algorithm based on FCM. It uses data about purchase information of users who are in the optimized cluster. This data is clustered by Clustering Module. This Module has three data types. i) Product data composed category goods in shopping mall, ii) Action data composed purchase action, shopping basket, click stream, and so on., Product and Action data composed common and similar goods data. On the basis of above three types of data, this algorithm relies on Structural method based on association rule to analyze the sequential user action patterns.

4. Evaluation

The implementation environment is as follows: we employed IIS Web Server based on Windows 2000 and MS SQL 2000 for server side, XML, Weka Tool as simulator for client side.

4.1 System Implementation

We simulated using three algorithms based on cluster module. With K-means algorithm, utilized each to product data, action data, product and action data using user set = {Sex, Age, Job}. We result that experiment with K-means by Apriori Algorithm, Product data Clustering (PC), Action data Clustering (AC), Product and Action data Clustering (PAC). The result is shown as following on Fig. Also, we experiment all of them with WEKA Tool based on K-means by Apriori algorithm. The result is shown as following on Fig. 6., 7.

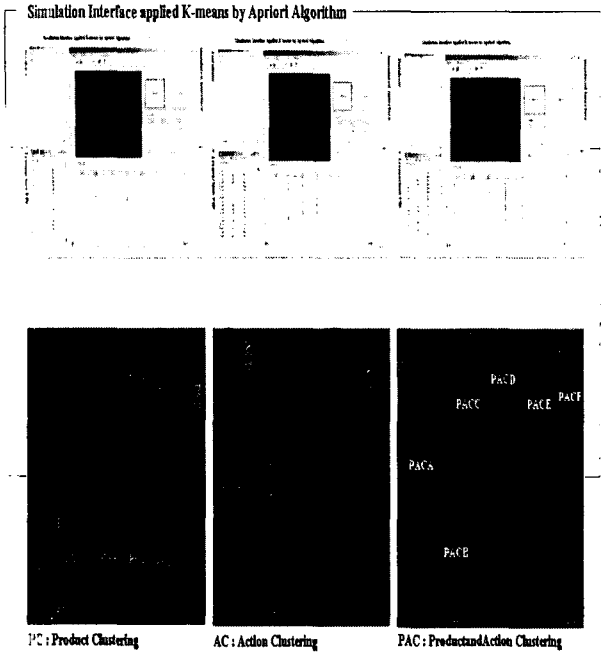

Fig. 6. *Simulation Interface applied K-means by Apriori Algorithm*

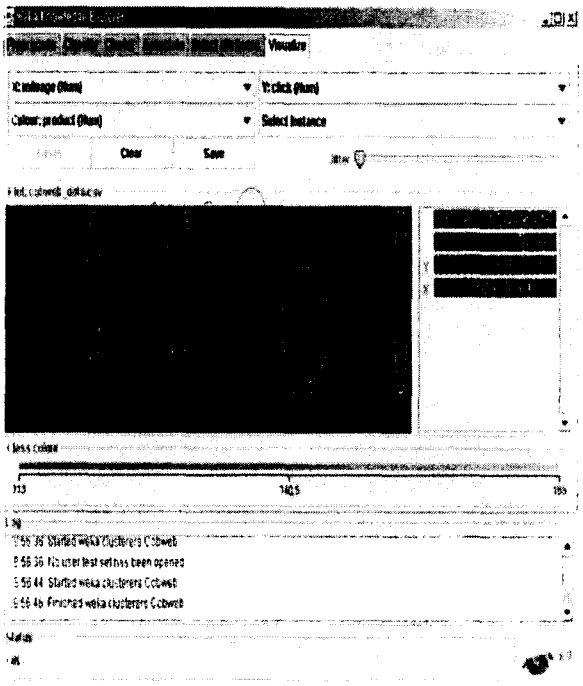

Fig. 7. Result on entire simulation with Weka Tool.

Based on K-means by Apriori algorithm and WEKA Tool, We create clustering groups. This clustering groups recommended product data user's personalization each of clustering groups can be expressed to set and Type of set is as followed.

Product Data

Ps = {Age} , {Job} : Clustering 1
Result : {PCA}, {PCB}, {PCC}, {PCD}, {PCG}, {PCH}, {PSC}

Action Data

As = {Age} , {Job} : Clustering 1
Result : {ACA}, {ACB}, {ACC}, {ACD}, {ACE}, {ACF}, {ACG}

Product and Action Data

PAs = {Age} , {Job} : Clustering 1
Result : {PACA}, {PACB}, {PACD}, {PACE}, {PACF}

Category Structural. Create structural Shopping Mall (large scale portal sites) of data that is Product Data, Action Data, Product and Action Data for E-Commerce and M-Commerce formatted and Category of each data as shown in Fig 8. Then it present category information on products of structural shop, such as Great Classification, Bisection Kind, Subdivision Kind that include IDs(GC, BK, SK), Product Names.

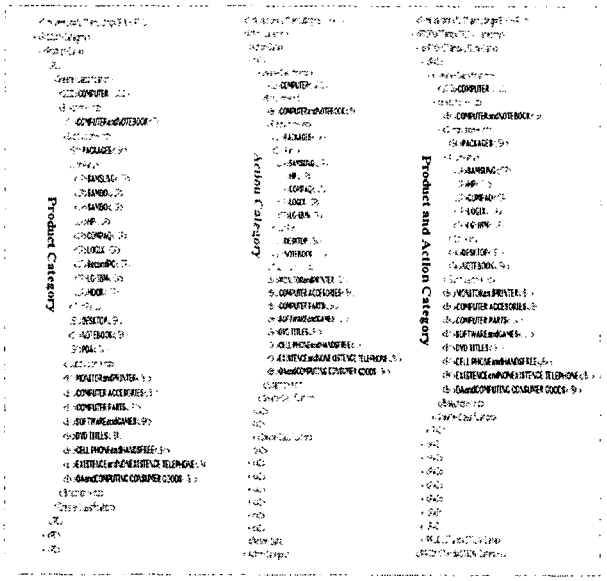

Fig. 8. Category information on products of structural

There were seen some patterns and differences as result that experiment structured category with XML. Firstly, if examine about some patterns, <product-category>, <action-category>, <productandaction-category> with relationship

high position root name and <product-data>, <action-data>, <productandaction-data> with relationship sub element name could be seen in two patterns : i) <template match="category"> that is matches the first category name, ii) <template match="data"> that is matches the first data name. Also each classification is a same structure according to category depth and also appeared sequentially. Secondly, if examine about differences, product categories, action categories, product and action categories. we can see that category depth for the end page that user monitoring is different. The result is as shown in Fig. 9.

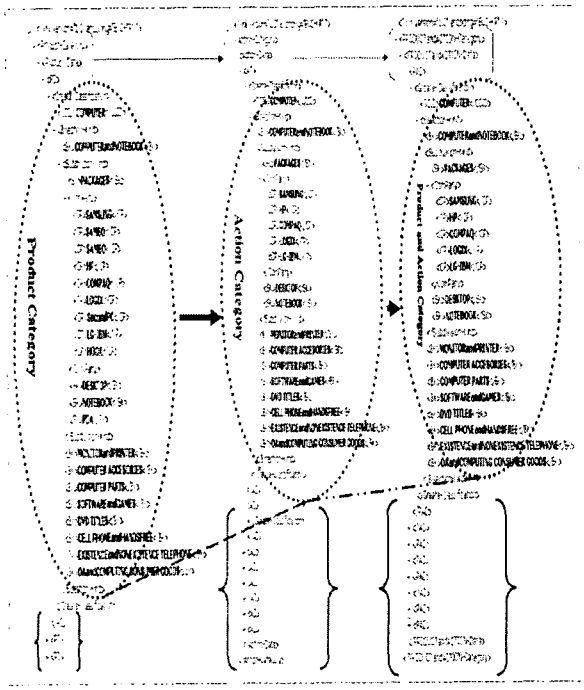

Fig. 9. Each Categories Difference Interface

We examined the first cluster result. We propose examine the second (Refiltering) and third cluster (Optimized XFCM) result as shown in Fig 10. Implementation of XFCM with FCM.

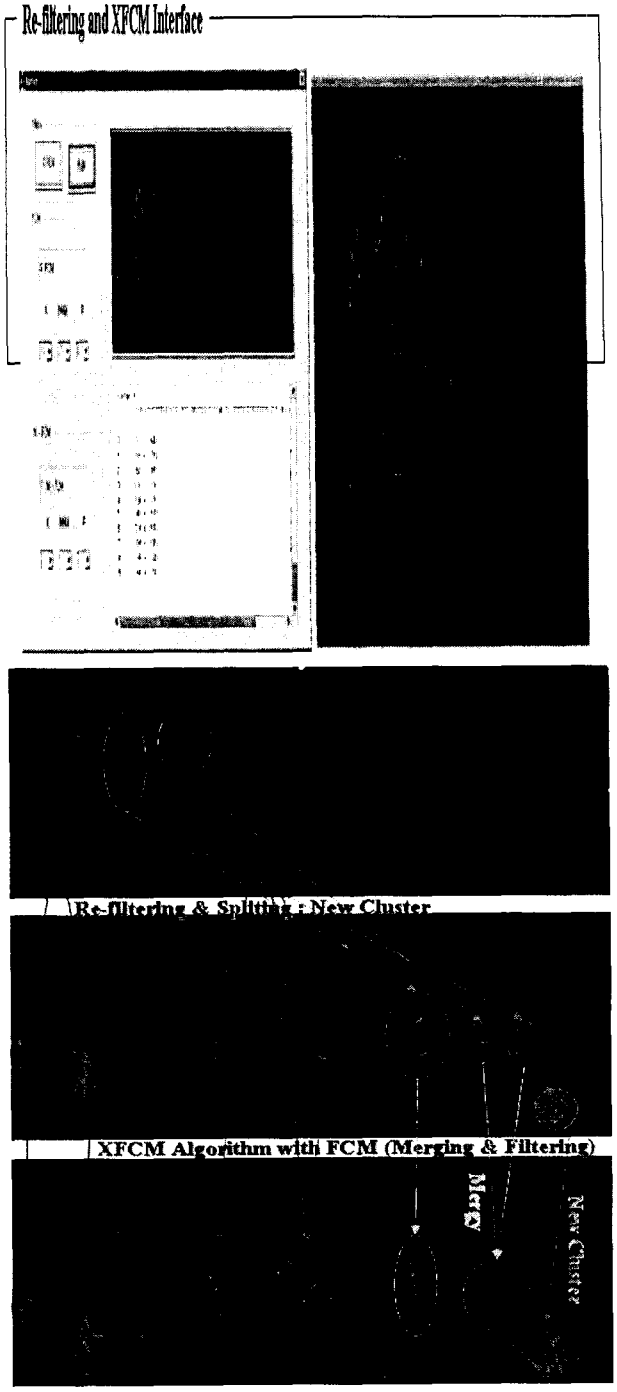

Fig. 10. Re-Filtering and XFCM Interface

XML Transformations. User Profile {Age, Sex, Job} and Data {Ps, As, PAs} show the detailed information of goods and present structural. Finally, The Categories of product reorganized through the standard category by XFCM. As shown in Fig 11.

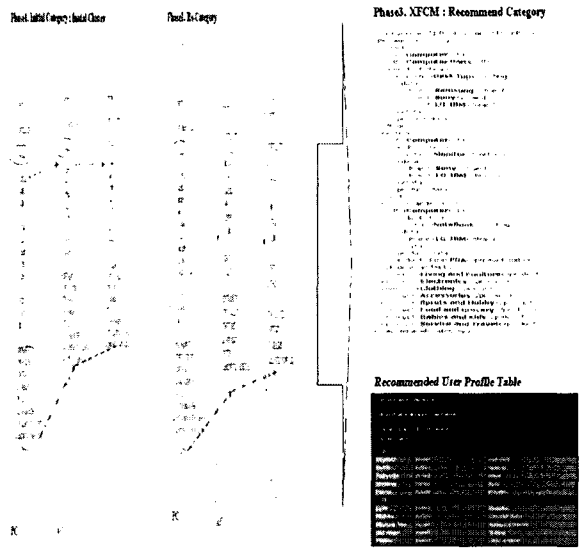

Fig. 11. Recommended structural Categories and goods Interface

4.3 System Evaluation

We create structural user profile , personalization and recommendation data, structured shopping mall as result of recommend algorithm based on user inclination.

We applied on three categories(Product Data, Action Data, Product and Action Data) with Chair Data of Sex based on three algorithms(K-means, FCM, XFCM). The combination of three algorithms and clusters is compose by using nine cells with number of Man, Woman, Neutural cluster. I proposed that i) apply on each of nine cells that selected Chair data , standard category number, Goods include of GC, BK, SK. ii) result in product data cluster with K-means algorithm by Aprioi is same initial declaration. iii) result in data cluster with XFCM algorithm by FCM is well-done personalized recommend product.

5. Conclusion

Recommendation Algorithm provides an effective XML form of targeted E-Commerce and M-Commerce by creating personalized products and categories.

In this paper presents an approach for structural data analysis step with the user profile structure patterns discovery in data.

In this paper, we proposed a recommended system with personalized user profile for E-Commerce and M-Commerce. Advantage of this paper is as following. Firstly, with structural User Profile, Recommend products data efficiently. Secondly, it's easy to apply on M-commerce with XML format. Thirdly, create structured shopping mall directory. Fourthly, efficient products and category recommend creating cluster group which is structural and optimized that XML based. Finally, even if user inclination changes, Cluster group regeneration is possible and apply easily to M-Commerce. We have actually simulated the recommendation algorithm and evaluated effectiveness partially in practice.

In Futher study, More efficient E-Commerce and M-Commerce is expected to be possible by that apply for Agent System in real world.

Reference

[1] http://www.Broadvision.com/OneToOne/SessionMgr/home_page.

[2] J.C.Bezdk, Pattern Recognition with Fuzzy Objective Function Algorithms, Plenun Press, N.Y., 1981.

[3] Resnick, P., et al. Group Lens: An Open Architecture for Collaborative Filtering of Netnews. In Proceedings of ACM CSCW'94 Conference on Computer-Supported Cooperative Work, pages 175-186. 1994.

[4] Paul Resnick and Hal R. Varian, Guest Editor "Recommender Systems," Communication of the ACM, Vol. 40, No. 3, March 1997.

[5] Hirsh, H., Basu, C., and Davison, B. D. (2000) Learning to Personalize. Communications of the ACM, 43, 8, 102-106.

[6] Jermann, P., Soller, A., and Muehlenbrock, M.(2001) From Mirroring to Guiding: A Review of State of the Art Technology for Supporting Collaborative Learning. Proceedings of the Computer Support for Collaborative Learning(CSCL) 2000, 324-331.

[7] Basu, C., H. Haym, and W. W. Cohen, "Recommendation as classification: Using social and content-based information in recommendation," Proceedings of International Conference on User Modeling, June 1999.

[8] Alsabti, K., Ranka, S., & Singh, V. 1998. "An Efficient K-Means Clustering Algorithm. Proc, First Workshop on High-Performance Data Mining".

[9] G. H. Ball and D. J. Hall, "ISODATA: an interactive method of multivariable analysis and pattern classification," Proc. IEEE Int. Communications Conf., 1996.

[10] Yiu-ming Cheeung. "K-means - A Generalized K-means Clustering Algorithm with Unknown Cluster Number," LNCS 2412. Intelligent Data Engineeing and Automated Learning IDEAL 2002.

[11] M.J.Sabin, Convergence and Consistency of Fuzzy c-means / ISODATA Algorithms, IEEE Trans. Pattern Anal. Machine Intell., September 1987.

[12] Liyan Zhang "Camparison of Fuzzy c-means Algorithm and New Fuzzy Clustering and Fuzzy Merging Algorithm," May 2001 Univ. Nevada, Reno.

[13] Eunseok Lee, Sera Jang. "A User Adaptive Mobile Commerce System with a Middlet Application" SungKyunkwan Univ. LNCS 2412. Intelligent Data Engineeing and Automated Learning IDEAL 2002.