http://ansinet.com/itj

ITJ

ISSN 1812-5638

# INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 10 (8): 1587-1593, 2011
ISSN 1812-5638 / DOI: 10.3923/itj.2011.1587.1593


# Clique Discovery Based on User Similarity for Online Shopping Recommendation

[1]Qing Yang, [1]Ping Zhou, [2]Huibing Zhang and [2]Jingwei Zhang
[1]Electronic Engineering and Automation Institute,
Guilin University of Electronic Technology, Guilin, 541004, China
[2]Computer and Control Institute, Guilin University of Electronic Technology, Guilin, 541004, China

**Abstract:** Identifying cliques with the same interests is valuable for online shopping which can make the recommendation and advertisements to target different users more accurately and maximize the benefits of advertisers, publishers and users. This study, has proposed an effective and efficient method to discover cliques for online shopping which firstly identifies clique leaders and clusters the most similar users, then computes clique cores among existing clique members and finally generates the complete cliques. A marked improvement is that two key factors, users' behavioral characteristics and regular purchase information, are unified to discover cliques. This method can also remove effectively most of fake purchases through computing the operation similarity among different goods categories.

**Key words:** Clique discovery, clique leader, clique core, online recommendation

## INTRODUCTION

Networks provide us with more shopping space, which also causes users spending more time on considering both the price and quality of their favorite goods. At the same time, many online sellers are distressed because few people care their goods due to the limitation of Web pages. These situations raise some preliminary research, such as SMS advertising (Menon, 2010) and the relationships among advertisements, users and brands (Haque *et al*., 2006; Jiang and Tao, 2011; Khatibi *et al*., 2006). In fact, the benefits of online transactions are depending on recommendation mechanism and efficient advertisements more and more (Rodgers and Thorson, 2000). In other words, it is very important to let appropriate users see your goods timely in Web sites. Recommendation and advertisements are effective communication mechanism between online sellers and users (McCoy *et al*., 2007; Hsu, 2007) but the key problem is how to match users and goods. At present, recommendation decision often depends on individual historical information but ignores the mutual information among a group of users with the same interesting. It should be a more significant strategy for sellers to recommend their goods for different users according to the information of a clique for the growing number of users.

According to CNNIC (2009), as June 2009, there are 87,880,000 online shoppers in China, only 26% of the total Internet users. Among those online shoppers, 48.7% of them know shopping sites through friends' recommendation, and 43.3% show special concerns on buyers' comments for shopping decisions. Obviously, a lot of sellers are preferring selling their goods online, more and more users are being attracted to participate in online shopping for superior quality and competitive price. Especially, those mature users have a huge impact on potential users. The large user base and the mutual influence among users assure the effectiveness of recommendation among a clique.

In advertising areas of Web pages, there are two strategies to place goods information. One is to broadcast sales promotion in which all users see the same goods information without any consideration of users' actual requirements, the other is to customize the information for different users, which often depends on the users' historical click and purchase information. For personalized recommendation, there are two primary categories, one focuses on user similarity (Bhuiyan *et al*., 2010; Cheng *et al*., 2010), the other more concerns goods correlation (Linden *et al*., 2003; Yang *et al*., 2010). Schafer *et al*. (1999) also gave a primary taxonomy for recommendation technologies, non-personalized, attribute based, item-to-item correlations and people-to-people correlations. The last two technologies are more popular in today's Web sites. Linden *et al*. (2003) started from the similarity of goods according to every user's purchase information and used item-to-item collaborative filtering to find every goods' nearest neighbors, which are then combined into a recommendation list. Yang *et al*. (2010)

**Corresponding Author:** Qing Yang, Electronic Engineering and Automation Institute, Guilin University of Electronic Technology, No. 1, Jinji Road, Qixing District, Guilin, 541004, China Tel: 86-773-2291522

introduced frequent item set mining to find the most welcome goods which is a coarse-grained recommendation and does not consider any user characteristics. Bhuiyan *et al.* (2010) made use of users' interest similarity to model trust networks, then used trust networks to find the neighbors for recommendation making. Cheng *et al.* (2010) actively predicted user intent only by user browsing behaviors, whose research shows that about 19.3% of browsing session can cause the users' further action, such as search. Information representation is also a key problem for recommendation technologies, Huang *et al.* (2004) modeled users, goods and their links into a graph model which is enough flexible to support different recommendation approaches. Takacs *et al.* (2009) modeled user and goods into different feature matrix and then combined matrix factorization with a neighbor correction to rank items for users. Archak *et al.* (2010) analyzed user-level advertising data, modeled individual user history into a graph structure to mine the long-term behavioral patterns. Kabutoya *et al.* (2010) extended Probabilistic Latent Semantic Analysis into a probabilistic topic model which is used for recommendation decision in the situations that multiple individuals share one account.

Internet sales market is depending on the active goods pushing more (Pricewaterhouse Coopers, 2007; Hsu, 2007), such as recommendation and advertising, which will have essential difference from supermarkets. The rise of social networks and the wide spread of mobile terminals have also a great demand for personalized recommendation (Clemons *et al.*, 2007). All Internet sellers must try their best to present their goods to the users actively, not wait for users' picking out and buying, just like in supermarkets. The basic difficulty for goods pushing is to locate accurately the target users. This study has advised an approach to cluster users into cliques according to their behaviors and historical purchase information, the mutual information in a clique will be used to match items with corresponding users.

This study puts together users' behavioral characteristics, such as operation frequency and users' historical purchase information to establish cliques, the members in a same clique can be applied the same recommendation.

## PROBLEM SETTING

For E-commerce Web sites, the sites should try their best to match goods with interesting users since every user can only browse a limited number of goods and the advertisement area is also limited. Different groups or users should enjoy personalized goods presentation in advertising and recommendation areas. In order to push the interesting goods for different users, this study identifies those different user groups and uses the group information for goods recommendation.

A Web site is modeled as a union of a user set U, a goods set G and a purchase set P, denoted as $WS = U \cup G \cup P$. A purchase process is modeled as two periods, click and buy. The purchase process can be considered as a series of click operations and ended by a buy operation. A click or buy operation is represented as a tetrad, (oper, t, uid, gid). oper is one of click and buy, t is the accurate time of the corresponding action, uid the user identifier and gid the goods identifier. For a specific user and a concrete goods, a purchase process can be considered as a list as following:

$$\begin{aligned} P = \{ & < oper_1, time_1, uid_1, gid_1 >, \\ & < oper_2, time_2, uid_2, gid_2 >, \cdots, \\ & < oper_N, time_N, uid_N, gid_N > \} \end{aligned}$$

There are two kinds of operation list, user operation list UOL and goods operation list GOL. For a specific user, a set of triples can be formed as following:

$$\begin{aligned} UOL = \{ & < oper_1, time_1, gid_1 >, \\ & < oper_2, time_2, gid_2 >, \cdots, \\ & < oper_M, time_M, gid_M > \} \end{aligned}$$

For a concrete goods, the similar set is provided as:

$$\begin{aligned} GOL = \{ & < oper_1, time_1, uid_1 >, \\ & < oper_2, time_2, uid_2 >, \cdots, \\ & < oper_K, time_K, uid_K > \} \end{aligned}$$

**Problem definition:** Given a user set U and a goods set G, every user can be associated with a user operation list, and every goods can be related with a goods operation list. Some user cliques should be discovered according to these operation lists so that they can provide an effective recommendation on these cliques. All cliques, $C_i$, should satisfy the following constraints, $C_i \subseteq U \wedge \cup C_i = U$.

Here, a clique means a group of users who often show similar interests in some specific goods (Wood, 1997). When discovered a clique, the same recommendation strategy will be applied to all users of the clique. The basic philosophy of discovering cliques for online shopping is if a group of users are in one clique, one user bought something, the other users will also buy this goods with high probabilities because of their similarity. In fact, the similarity of both users' behavioral characteristics and purchase information are used to decide the members of a clique and then use the mutual

purchases information for recommendation among members of a clique.

**Example:** In Fig. 1, Every goods belongs to some category, every user carried out some operations in corresponding goods, here solid line is buy and dotted line is click. $U_1$'s operation list is {<buy, $G_1$>,<buy, $G_5$>}, and $U_3$'s is {<buy, $G_3$>,<buy, $G_5$>}. Time information is omitted. For $G_1$ and $G_3$ belong to the same category, $U_1$ and $U_3$ will be in the same clique.

## CLIQUE IDENTIFICATION

In this section, the purchase process is modeled to a hierarchical model (Fig. 1) to discover the cliques. Some distance metrics are defined to measure the similarity of different users and different groups of goods. Two key concepts, clique core and clique leader, are introduced.

Goods are divided into some categories, every goods belongs to a specific category. Every user has an operation list. According to different user operation lists, the users' similarity will be computed. In a given time bucket, one user is always associated with a bag of goods which is the set of all goods that the user has bought or clicked. Every goods is also related with a group of users which includes two parts, click subset and buy subset. Our basic strategy is to find enough leaders according to the operation lists and cluster other users with corresponding leaders to establish cliques.

**Leader identification:** Firstly, weight is introduced to represent the popularity of goods which is helpful to give a priority to those leaders who are active for popular goods. For any goods type, BN is used to denote its total number of buy and CN the total number of click. The weight for one goods type can be defined as $BN^* W_1 + CN^* W_2$, $w_1$ and $w_2$ are weight for buy and click operations which are here 1 and 0.1. The weights are normalized as following:

$$\frac{BN_j * W_1 + CN_j * W_2}{\sum_i (BN_i * W_1 + CN_i * W_2)} \quad (1)$$

For a clique, there are two primary elements, clique leader and clique core. A clique leader is a user who is very active in online shopping, has often a steady operation frequency and provides big support for online shopping in corresponding clique. A clique core is composed of a set of goods which are often bought by most of members in this clique. The set of goods of a clique core can help to extend clique and are also valuable for recommendation to other members.

In real online purchase, there are often some fake purchases, for example, some users are employed to do

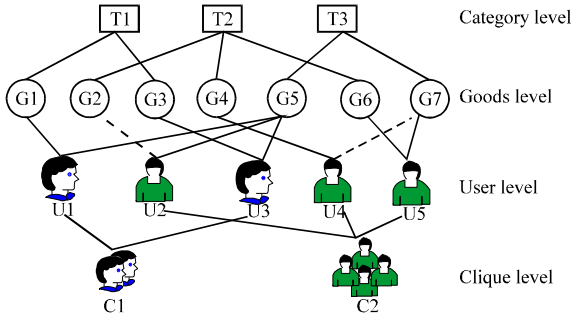


Fig. 1: A hierarchical model of purchases

fake purchases so that some goods can have a good rank. For a group of rigged users often serve for a range of goods, goods operation similarity of different categories are used to remove those fake purchases.

**Definition 1:** Goods Operation Similarity (GOS) . Given two kinds of goods, $G_i$ and $G_j$, every goods has a related click set CS and buy set BS. CS is composed of those users who only click this goods but do not buy. BS consists of those users who click and buy the goods. Their operation similarity is defined as the arithmetic mean of weighted Jaccard similarity,

$$\frac{W_1 * \frac{|BS_i \cap BS_j|}{|BS_i \cup BS_j|} + W_2 * \frac{|CS_i \cap CS_j|}{|CS_i \cup CS_j|}}{W_1 + W_2} \quad (2)$$

The above formula is abbreviated as GOS ($G_i$, $G_j$) which can be easily extended for a group of goods.

All users are ranked according to their support which is used to measure the users' contributions for online shopping and defined as:

$$GTW \sum_i BN_{user,i} * W1 + CN_{user,i} * W2 \quad (3)$$

$BN_{user,i}$ is the total number of buys that user has done to the $i_{th}$ goods. $CN_{user,i}$ is click number that user done to the $i_{th}$ goods. GTW denotes the goods type weight. The algorithm for identifying leaders is shown in Algorithm 1.

Algorithm 1: Leader identification

**Input**: a user set U,
a goods set G,
a set of online operation information <oper, t, uid, gid >
**Output**: a set of leaders' uid

1: label every goods by corresponding category
2: compute the operation similarity of different categories pairwise according to GOS
3: divide categories into different regions on operation similarity
4: compute users' support in different regions
5: rank users on support
6: return the top-k users

**Clique discovery:** In the following, the whole cliques will be generated based on clique leaders. Here, an iterative process is used to find the complete cliques. Firstly, leaders are used to find their follows, when enough users are added into the cliques, the clique core will be formed, and then use these cores to find more members until the process ends.

In order to get a good performance, the users' behaviors and purchase information are considered to model a user. Users' behaviors are measured by two new distance metrics, click to buy distance and user tag similarity. Purchase information are measured by user operation similarity.

Click to buy distance and user tag similarity are both used to measure the users' shopping habits which are independent with the concrete goods.

**Definition 2:** Click to Buy Distance (CBD). Given a user operation list, CBD is defined as the average of the click number that all goods have been bought. If BS is the set of goods bought by the given user, $CN_i$ denotes the click number of the $i_{th}$ goods in BS, CBD can be computed as:

$$CBD = \frac{\sum_i CN_i}{|BS|} \quad (4)$$

If two users have close CBD values, the two users' shopping behaviors are often similar. In Fig. 2, CBD ($user_1$) = 3, CBD ($user_2$) = (1 + 2 + 1)/3 = 1.33. Obviously, $user_1$ often decides to buy some goods after multiple clicks, but $user_2$ more tends to buy directly. From Fig. 2, we can also see that only the goods set that have been bought are considered which is accurate only in the situations that the majority of operations are buy. CBD can't give a good description for those users who click a lot of goods but buy little. User tags are introduced to solve this problem which combines basic statistics and CBD to label every user by a tag. For a user, there are four primary behavioral patterns in online shopping, which are listed in Table 1.

Corresponding to the above behaviors, users are divided into four types, elementary, impulsive, chary and faithful users. The four types are considered as user tags, every user will be labeled with one of them. When two users are labeled with the same tag, they are considered to be very similar in shopping behavioral habits. The degree of tag similarity is gradual from elementary to faithful, for example TS (faithful, faithful) > TS (faithful, chary)>TS (faithful, impulsive), here TS is a computing function for tag similarity. In Fig. 2, $user_1$ can be labeled with chary, $user_2$ with faithful. Every user is given a tag, for those users with same tag, CBD can be computed to rank them.

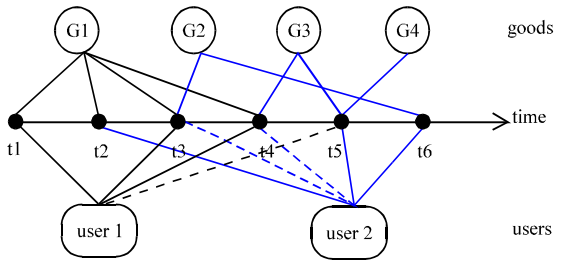


Fig. 2: An Illustration of user shopping

Table 1: User Shopping Behavioral Patterns

| Patterns | Details |
|---|---|
| Only click | The user only click some goodsbut do not buy them |
| Click and buy | The user click some goods and often buy them at once |
| Click to buy | The user click some goods once or many times, and then buy it |
| Buy and buy | The user buy some goods many times |

The purchase information similarity is related with the concreted goods that a user has clicked and bought, which is similar to goods operation similarity and can be computed as:

$$\frac{W_1 * \frac{|BG_i \cap BG_j|}{|BG_i \cup BG_j|} + W_2 * \frac{|CG_i \cap CG_j|}{|CG_i \cup CG_j|}}{W_1 + W_2} \quad (5)$$

The above formula is abbreviated as PIS ($U_i$, $U_j$). Here, BG is a set of goods that a user has bought and CG is a set of goods that the user has clicked. Based on both behavioral similarity and purchase information similarity, every user can be unified with corresponding leaders to generate some basic cliques. The core of these basic cliques will be used to expand them. The process is repeated until all users find their cliques. The detailed algorithm for clique discovery is presented in Algorithm 2.

Algorithm 2: Clique discovery

**Input:** a user set *U*,
a goods set *G*,
a set of online operation information<operation, t, uid, gid>
**Output:** cliques
1: find all leaders according to *algorithm* 1, get the set of leaders *LS*
2: compute distance between user *u* and leader *l*
$Dist_{ul}$=CBD (u, *l*)+TS (u, *l*)+ PIS (u, *l*)
3: organize the most similar users with their leaders into corresponding cliques and remove them from U
4: while U is not null
5: find the clique core
6:1add new users into cliques according to clique core
7: remove those users from U
8: end while

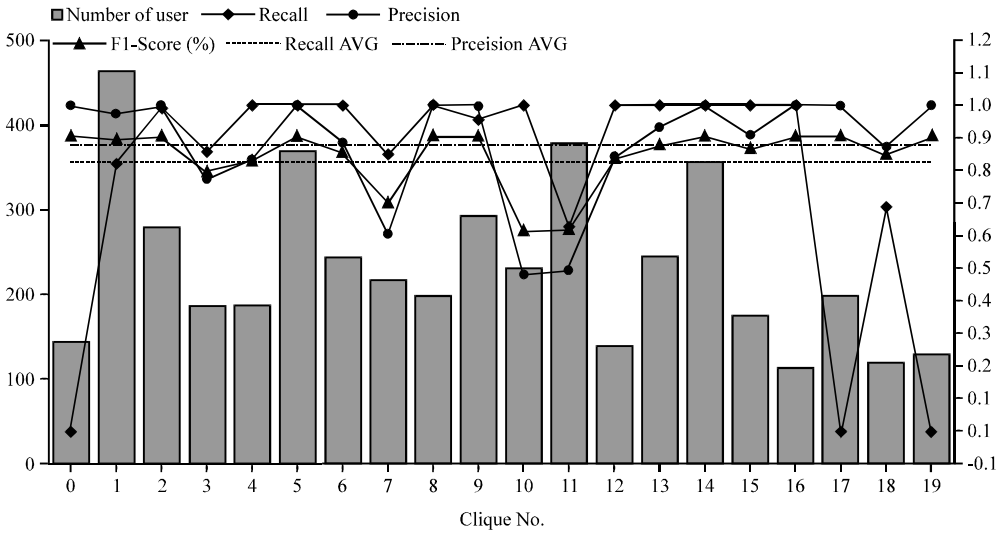


Fig. 3: Experimental Results for Recall, Precision and F1-Score (The abscissa denotes the No. of cliques, from 0 to 19, the left ordinate denotes the number of clique members, and the right corresponds to the recall, precision and F1-score)

## EXPERIMENTS

In this section, a series of experiments are carried out to evaluate the effectiveness and efficiency of our methods.

**Data set:** A basic data set is constructed that consists of a month of purchase information. 5,000 users and 800 goods are involved in this data set. The original format of the data set is just tetrad, (oper, t, uid, gid). In order to assure the data quality, some data cleaning are done before our experiments. Some purchase information related with invalid goods and invalid users are removed from this data set. Every user is labeled with corresponding clique ID in advance so that we can verify the accuracy of our method. The experimental environments are Core 2 Duo CPU 2.2 GHz and 2 GB memory under Windows OS.

**Experimental results:** In order to show the effectiveness of our study, recall, precision and F1-score are used as metrics. There are a total of 20 cliques in our data set. We consider the data set as a whole to discovery cliques. Except the average recall and precision, the respective recall and precision for every clique are also computed. The experimental results are reported in Fig. 3 which contains both the number of clique members and related computation results. From Fig. 3, we can see that our method has a good performance; especially it

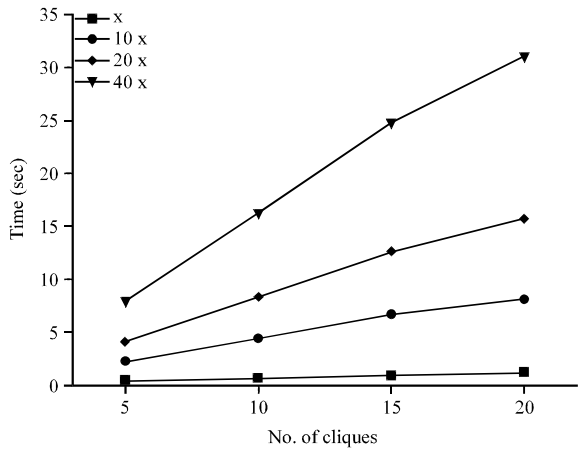


Fig. 4: Time performance comparison

provides a high precision and stable F1-score. Because of considering the similarity of both behavioral characteristics and purchase information, those close users can be clustered together well. Though the recalls of several cliques are low, the average recall is still retained at a high level, more than 80%. The reason for low recall has two factors, one is that the information provided by the basic clique core is not enough, the other is that those users in corresponding cliques also present similar characteristics with some users in other cliques which causes our method gave out a wrong decision.

Considering the efficiency, different combinations of clique numbers and purchase information numbers are used to carry out experiments. In order to show the scalability, we randomly choose 5 cliques of data for the first test and then add 5 cliques of data into current test data per time until all data are tested. For each round of experiments, we also enlarge the basic purchase data to 10, 20 and 40 times. The time performance is reported in Fig. 4. Obviously, present method has a nearly linear time complexity for different combination of data and can response to large-scale data processing.

## CONCLUSIONS AND FUTURE WORK

In this study, both the users' behavioral characteristics and users' purchase information are considered to cluster related users into cliques, the work is valuable for Web recommendation and online advertisements. Based on the user cliques, Web sites can target the users more accurately for different goods and establish an effective information channels between online users and sellers. With the continuous improvement of user participation in online shopping, there are still much work to be improved and solved in Web recommendation. At present, only the users' similarity are cared, in fact, different goods are also correlative, for example, if a user bought a badminton racket, it will be a high-probability event for him to buy a pair of badminton shoes. In the next step, the goods correlation and user similarity will be integrated to discover cliques for more accurate recommendation. Another important problem is the real-time analysis, a timely recommendation and online advertisements for specific users will make online-shopping gain with little efforts.

## ACKNOWLEDGMENTS

We gratefully acknowledge the support of both the National Natural Science Foundation under grants No.60961002 and Education Department of Guangxi under grants No. 201010LX154.

## REFERENCES

Archak, N., V.S. Mirrokni and S. Muthukrishnan, 2010. Mining advertiser-specific user behavior using adfactors. Proceedings of the 19th International Conference on World Wide Web, (WWW'10), ACM New York, USA., pp: 31-40.

Bhuiyan, T., Y. Xu, A. Josang, H. Liang and C. Cox, 2010. Developing trust networks based on user tagging information for recommendation making. Proceedings of the 11th International Conference on Web Information System Engineering, (WISE'10), Hong Kong, China. pp: 357-364.

CNNIC, 2009. 2009 report of China online shopping market research. China Internet Network Information Center, China.

Cheng, Z., B. Gao and T.Y. Liu, 2010. Actively predicting diverse search intent from user browsing behaviors. Proceedings of the 19th International Conference on World Wide Web, (WWW'10), ACM New York, USA., pp: 221-230.

Clemons, E.K., S. Barnett and A. Appadurai, 2007. The future of advertising and the value of social network websites: Some preliminary examinations. Proceedings of the 9th International Conference on Electronic Commerce, Aug. 19-22, Minneapolis, MN., USA., pp: 267-276.

Haque, A., A.K. Tarofder and S. Al-Mahmud, 2006. Internet advertisement: Helps to build brand. Inform. Technol. J., 5: 868-875.

Hsu, H.C., 2007. How advertising affects: A study of the Chinese E-market. Asian J. Manag. Hum. Sci., 1: 539-557.

Huang, Z., W. Chung and H. Chen, 2004. A graph model for E-commerce recommender systems. J. Am. Soc. Inform. Sci. Technol., 55: 259-274.

Jiang, J. and X. Tao, 2011. Responses of Chinese consumers to corporate advertising themes: Cue applicability and contextual priming effects. Asian J. Market., 5: 17-30.

Kabutoya, Y., T. Iwata and K. Fujmura, 2010. Modeling multiple users' purchase over a single account for collaborative filtering. Proceeding of the 11th International Conference on Web Information System Engineering, Dec. 12-14, Hong Kong, China, pp: 328-341.

Khatibi, A., A. Haque and K. Karim, 2006. E-commerce: A study on internet shopping in Malaysia. J. Applied Sci., 6: 696-705.

Linden, G., B. Simith and J. York, 2003. Amazon.com recommendations: Item-to-item collaborative filtering. IEEE Inter. Comput., 7: 76-80.

McCoy, S., A. Everard, P. Polak and D.F. Galletta, 2007. The effects of online advertising. Commun. ACM, 50: 84-88.

Menon, V., 2010. SMS advertisement: Competitve to gulf market. Asian J. Market., 4: 131-143.

Pricewaterhouse Coopers, 2007. IAB internet advertising revenue report. Interactive Advertising Bureau (IAB), pp: 1-20. http://www.iab.net/media/file/IAB_PwC_2007_full_ year.pdf.

Rodgers, S. and E. Thorson, 2000. The interactive advertising model: How users perceive and process online ads. J. Interact. Adv., 1: 42-61.

Schafer, J.B., J. Konstan and J. Riedi, 1999. Recommender systems in e-commerce. Proceedings of the 1st ACM Conference on Electronic Commerce, Nov. 3-5, ACM Press, Denver, Colorado, United States, pp: 158-166.

Takacs, G., I. Pilaszy, B. Nemeth and D. Tikk, 2009. Scalable collaborative filtering approaches for large recommender systems. J. Mach. Learn. Res., 10: 623-656.

Wood, D.R., 1997. An algorithm for finding a maximum clique in a graph. Oper. Res. Lett., 21: 211-217.

Yang, Q., P. Zhou and J. Zhang, 2010. Frequent browsing patterns mining based on dependency for online shopping. Inform. Technol. J., 9: 1246-1250.