International Journal of Management Science
Vol 16, No 2, Special Issue, September 2010

User-Created Content Recommendation Using Tag Information and Content Metadata[*]

Byung Woon Rhie
School of Business, Hanyang University, Seoul, Korea

Jong Woo Kim[**]
School of Business, Hanyang University, Seoul, Korea,

Hong Joo Lee
Department of Business Administration, The Catholic University of Korea, Korea

(Received: March 15, 2008 / Revised: December 10, 2009 / Accepted: December 20, 2009)

ABSTRACT

As the Internet is more embedded in people's lives, Internet users draw on new Internet applications to express themselves through "user-created content (UCC)." In addition, there is a noticeable shift from text-centered contents mainly posted on bulletin boards to multimedia contents such as images and videos on UCC web sites. The changes require different way of recommendations comparing to traditional products or contents recommendation on the Internet. This paper aims to design UCC recommendation methods with user behavior data and contents metadata such as tags and titles, and compare performances of the suggested methods. Real web logs data of a major Korean video UCC site was used to empirical experiments. The results of the experiments show that collaborative filtering technique based on similarity of UCC customers' preferences performs better than other content-based recommendation methods based on tag information and content metadata.

Keywords: Contents Recommendation, Collaborative Filtering, User Created Contents, Metadata

1. Introduction

The advance of information technology infrastructure reduces the cost of producing

[*] This work was supported by the Korea Research Foundation Grant funded by the Korean Government (MOEHRD, Basic Research Promotion Fund) (KRF-2007-327-B00212).

[**] Corresponding author, E- mail: kjw@hanyang.ac.kr

and distributing digital contents. Also, Web 2.0 which can be characterized by "share," "involvement," and "openness" brings the paradigm shift on the way to use the Internet. User-created content (UCC) is a good example to represent features of Web 2.0. UCC makes it possible to produce and deliver digital contents with low cost and highly efficient way on the Internet [18]. Currently, the portion of multimedia contents such as images and video clips increases in UCC sites due to the emergence of prosumers and advances of digital technology.

Huge participations for contents creation on UCC sites increases the number of contents exponentially, and causes information overload problem. It brings the need for recommendation functionality on UCC sites for information filtering purpose. However, the recommendation of multimedia UCC requires different considerations and techniques comparing to traditional products and contents recommendation on the Internet. Traditionally, recommendation techniques on the Internet sites use collaborative filtering or content-based approaches [2, 5, 9, 11]. The collaborative filtering approaches use customers' behavior data or explicit rating data, and the content-based approaches that are mainly applied to text-based contents use keyword frequencies of the product descriptions or contents. So, it is interesting whether the collaborative filtering approaches still work for UCCs, and how the contents-based approaches can be applied to UCCs.

This paper aims to design recommendation techniques for UCCs and compare performances of these techniques. Five recommendation methods are designed based on collaborative filtering and content-based approaches. The first method is normal collaborative filtering approach, and from the second to fifth ones are content-based approaches using metadata of the UCCs. The second and third ones use tag metadata and title keywords, respectively. The fourth one uses tag and title metadata together. The fifth one uses representative tags in order to reduce sparseness of tag dimension. To compare performances of the proposed recommendation methods, panel data that include visiting logs on a major Korean video UCC site (auhu.hanafos.com) is used.

The rest of the paper is organized as follows: In section 2, relative works are reviewed which include UCCs, recommendation techniques, collaborative filtering, and content-based filtering. In section 3, five proposed recommendation methods for UCCs are described. Section 4 presents experimental design, and section 5 describes the results and analysis of the experiments. Section 6 finalizes the paper with brief conclusion remarks.

2. Related Work

2.1 User Created Contents

User Created Content (UCC), also known as User Generated Content (UGC) and Consumer Generated Media (CGM) refers to various kinds of media content, publicly available, that are produced by end-users[1]. UCC can be defined as: i) content made publicly available over the Internet, ii) which reflects a "certain amount of creative effort," and iii) which is "created outside of professional routines and practices" [18]. UCC is representative of Web 2.0 era because UCC is mainly developed by users on the Internet. Internet portal service providers attach importance to UCC because UCC provides new way to provide and distribute digital contents with low price and it also generates huge network traffics on their portal sites. Reflecting the interests on UCC, Times selected YouTube[2] as the best invention in 2006, and Economist prospected that UCC will be megatrend [6]. Also, currently many academic studies have been started on UCC, which mainly focus on UCC trends and prospects, copyright issues on UCC, and UCC business models [4, 10, 12, 17].

Metadata is defined as 'data about data,' and describes the data. Metadata includes data about who, when, why, and how the data is made. Metadata on UCC includes various items such as title of the content, keywords (is also called *tags*) entered by a creator when she/he registered the content, the number of views, the number of scraps which means how many it is stored in other blogs and sent through emails. The metadata on UCC help to understand the content and to access the content easily.

Previously, the major portion of UCC was in text format such as mini homepage, blog, and Bulletin Board System (BBS), and it rapidly shifts to multimedia contents such as video clips and images. It has more powerful communication power, because it uses visual languages rather than textual languages. Since the number of multimedia UCC increases exponentially, it is not easy to find UCC items which meet users' need or preference. It makes a sort of information overload problem. In addition, multimedia UCC has different content structure comparing to text-based UCC, we need to different information filtering techniques for multimedia UCC.

[1] From Wikipedia, http://en.wikipedia.org/wiki/User-generated_content.

[2] www.youtube.com.

2.2 Recommendation Techniques

Recommendation techniques are information filtering methods to provide personalized contents based on analysis of customer profiles, transaction data, and web page visiting logs [1, 9, 11, 13, 15, 16]. Recommendation techniques can be classified to collaborative filtering and content-based approach. Collaborative filtering approach uses customers' ratings on contents or purchase and visiting history to find other customers who have similar preference to target customer [9, 11]. Content-based techniques usually extract keywords from the contents which the customer show her interest on, and make keyword frequency vectors from the extract keywords and used to select appropriate contents [2, 5].

Collaborative filtering techniques are representative recommendation techniques for Internet storefronts [3, 7, 8, 9, 11, 13, 14]. Generally, collaborative filtering methods consist of three phases. First, collaborative filtering methods construct customer-product matrix U based on customer preference data sets. The customer preference data set can be customers' rating on products or purchase and visiting history data. Second, similarities between customers are calculated from customer-product matrix U. Three formulae, Pearson correlation coefficient, cosine vector, and Jaccard similarity coefficient are used to calculate the similarities between customers [11]. The third phase is to predict a customer's preference score to a product using similarities between customers which are calculated in the second phase and other customers' interest or rating information on the product.

Collaborative filtering methods select products to recommend based on preferences of customers who have similar preferences. However, content-based methods select products to recommend based on the similarities between product descriptions or contents about the products. The vector space model that is mainly used for content-based methods generates a keyword vector for a customer that consists of weights of keywords to represent her preference. Also, a keyword vector for a product is constructed to represent features of the product. The similarity between user's keyword vector and product's keyword vector is used to determine customer's preference of the product.

3. Recommendation methods for UCC

In the paper, five methods for UCC recommendation based on collaborative filtering and content-based approaches are designed and compared their performance. The first one is based on collaborative filtering, and second to fifth methods are based on content-based approach.

Based on users' visiting history to UCC items, we can make a user-item matrix *U* that consists of 0 or 1 elements. The element u_{ij} is 1 when customer *i* visited UCC item *j*, otherwise 0. User similarity matrix *S* is calculated using *U* and Jaccard similarity coefficient.

In the paper, content-based methods for UCC use metadata of UCC such as titles, tags, and representative tags. First, user-item matrix *U* is generated as the same way as the above collaborative filtering method. UCC item-metadata matrix *C* is constructed as a binary matrix, element c_{jm} is 1 if *j* item has *m* metadata, otherwise 0. User-metadata matrix *E* is generated using *U* and *C* as follows.

$$e_{im} = \begin{cases} 1 \text{ if } \sum_j u_{ij} c_{jm} > 0 \\ 0 \text{ otherwise} \end{cases} \tag{1}$$

The matrix *E* describes preference of users in terms of metadata. The similarity of user *i* and item *j* is calculated by Jaccard similarity coefficient as shown in Equation 2.

$$q_{ij} = \frac{|i \cap j|}{|i \cup j|} \tag{2}$$

In Equation (2), $|i \cap j|$ is the number of metadata which customer *i* has interest and item *j* has the metadata, and is calculated by $\sum_{for\ all\ m} \mathrm{e}_{\mathrm{im}} \times c_{jm}$, Also, $|i \cup j|$ is the number of metadata which customer *i* has interest or item *j* has, and is calculated by $\sum_{for\ all\ m} \mathrm{e}_{\mathrm{im}} + \sum_{for\ all\ m} \mathrm{c}_{\mathrm{jm}} - \sum_{for\ all\ m} \mathrm{e}_{\mathrm{im}} \times c_{jm}$.

In this study, we compare four variations of content-based method. That is, four variations of UCC item-metadata matrix *C* are considered, C_{tag}, C_{title}, $C_{tag+title}$, and $C_{r\text{-}tag}$. C_{tag} uses user specified tags as metadata, and C_{title} does titles of contents as metadata, $C_{tag+title}$ uses tags and titles together as metadata, and $C_{r\text{-}tag}$ does representative tags of contents only as metadata. $\mathrm{C}_{\text{r-tag}}$ is a set of the most frequently appeared tag among tags of a UCC item.

4. Experimental Design

4.1 Experimental Data Set

To compare the five recommendation methods described in section 3, a panel data set which includes visiting logs of a popular UCC site called andu[3] in Korea during two months from April to May, 2007. The panel data set was provided by Koreanclick[4] which maintains Internet user panel to gather Internet usage statistics. There are several reasons to select andu site. First, andu support user specified tagging feature, and second, the volume of andu visiting data is quite large. Third, using URIs (Uniform Resource Identifiers) in web log data, it is not difficult to access the related web page.

To extract metadata from UCC whose URL is listed in panel log data, a program is developed using Java programming language and Microsoft SQL Server 2005. The original data set from Koreaclick includes visiting time, panel ID, visiting URL, and stay duration. We select 145 panel IDs who visit equal or more than 10 times to andu site from April to May, 2007. 2770 UCCs are selected which has been visited and have at least one tag. There are 418 tags which are appeared equal or more than 2 times. Also there are 425 keywords which are appeared equal or more than 2 times in titles of 2770 UCCs. There are 1696 keywords that is appeared at least two times in tags or titles and 256 representative tags are selected as metadata. Table 1 summaries basic statistics of the experimental data set.

Table 1. Experimental data set

Data Set	Selection criteria	Number of data
Panels	At least ten items visiting during two months	145
UCCs	Visited UCCs with at least one tag	2770
Tags	Tags which is used in at least two UCCs	418
Titles	Title keywords which is used in at least two UCCs	425
Tags+Titles	Keyword which is used at least two times in titles or tags	1696
Representative Tags	The keywords which are the most frequent tag for a UCC	256

[3] http://andu.hanafos.com.

[4] http://koreanclick.com.

Five experiments are performed to compare performances of recommendation methods for UCCs. The first one is based on collaborative filtering and from second to fifth ones are based on content-based approach. The details of five methods were described in Section 3. The data set is divided into training data set and test data set. Randomly selected 70% visiting data of a user are assigned to the training data set, and the remainders are assigned to the test data set. Using five different methods, three most preferable UCCs are selected for each user in the test data set. To compare the five methods, we use *Precision* which is usually used to compare performances of recommendation techniques. Precision in this study means how many UCCs are really visited by the user among recommended UCCs. The formula of Precision is shown as Equation (3). In the formula, N is the number of recommended UCCs, and N_{rs} is the number of UCCs which have been visited by the user among recommended UCCs.

$$Pr = \frac{N_{rs}}{N} \tag{3}$$

The above experiment procedures in Section 4.2 have been performed ten times. That is, the separation of training data set and test data set has been performed ten times, and personalized selection of three most preferable UCC items for each user based on five methods have been performed ten times. Also, Precision is calculated for each time and the average of the precision values is calculated in the final stage.

5. Results and Analysis

Table 2 exhibits the results of the experiments described in Section 4. The numbers in the precision column is the average of ten time experiments. Since we performed ten times of each experiment, the average precision values are not normally distributed. Thus, the differences of five methods are statistically tested based on average precision values with Mann-Whitney test. The test results are displayed in the right side of Table 2 (CF, Tags, Titles, and Tag+Titles columns). In these columns, numbers in a cell are test statistics and numbers in parentheses are p-value.

Table 2. Experimental Result

	Method	Precision	CF	Tags	Titles	Tags+Titles
1	CF	0.1611	-	-	-	-
2	Tags	0.0887	0.000 (0.000)	-	-	-
3	Titles	0.1276	6.000 (0.001)	0.000 (0.000)	-	-
4	Tags+Titles	0.0692	0.000 (0.000)	16.500 (0.011)	0.000 (0.000)	-
5	Representative Tags	0.0729	0.000 (0.000)	19.000 (0.019)	0.000 (0.000)	38.000 (0.362)

The results show that collaborative filtering technique based on similarity of UCC customers' preferences performs better than other content-based recommendation methods based on tag information and content metadata. The performances of the collaborative filtering technique and those of other content-based recommendation methods are different statistically significant. Among content-based recommendation methods, the recommendation method based on title keywords shows the best performance. Also, the performances of the method and those of other content-based recommendation methods using tag information are different statistically significant. The results show that keywords in tags are less informative than those in titles for recommendation purpose. It means that the quality of tags of UCCs is lower than that of titles of UCC. That is partly because UCC producers can assign arbitrary tags on UCCs.

6. Conclusion Remarks

In this study, five recommendation methods for UCC videos which are based on collaborative filtering and content-based approach are designed and compared their performances in terms of precision. Four content-based methods use tags, keywords in titles, tags and keywords in titles together, and representative tags, respectively. Experiments using a real UCC web site visiting data show that collaborative filtering provides the best performance in terms of precision.

One of limitation of this study is that experiments had been performed on only one UCC video site data. To generalize the result of this study, the proposed methods will be applied to other UCC web site data. The result of this study is just an initial stage to develop recommendation methods that are specialized to multimedia UCC sites. Currently, we consider the feasibility to apply ontology concept which can relates semantic similarities among keywords in order to improve recommendation performance.

References

[1] Ansari, A., S. Essegaier, and R. Kohli, "Internet recommendation systems," *Journal of Marketing Research* 37, 3 (2004), 363-375.

[2] Balabanovic, M. and Y. Shoham, "Content-based, collaborative recommendation," *Communication of ACM* 40, 3 (1997), 66-72.

[3] Breese, J. S., D. Heckerman, and C. Kadie, "Empirical analysis of predictive algorithms for collaborative filtering," the Fourteenth Conference on Uncertainty in Artificial Intelligence, (1998), 43-52.

[4] Chang, E. A., "Does chatter matter? The impact of user-generated content on music sales," *Journal of Interactive Marketing* 23, 4 (2009), 300-307.

[5] Foltz, P. W. and S. T. Dumains, "Personalized information delivery: An analysis of information filtering methods," *Communication of the ACM* 35, 12 (1992), 51-59.

[6] Grossman, L., "Time, you-yes, you-are TIME's Person of the Year," *TIME* 12 (2006).

[7] Herlocker, J. L., J. A. Konstan, L. G. Terveen, and J. T. Riedl, "Evaluating collaborative filtering recommender systems," *ACM Transactions on Information Systems* 22, 1 (2004), 5-53.

[8] Huang, Z., H. Chen, and D. Zeng, "Applying associative retrieval techniques to alleviate the sparsity problem in collaborative filtering," *ACM Transactions on Information Systems* 22, 1 (2004), 116-142.

[9] Konstan, J. A., B. N. Miller, D. Maltz, J. L. Herlocker, L. R. Gordon, and J. Riedl, "GroupLens: Applying collaborative filtering to usenet news," *Communication*

of the ACM 40, 3 (1997), 77-87.

[10] Lee, H. J., Y. J. Kim, and S. R. Kang, "Understanding personal and cultural factors on the level of UCC participation: Centered on Korea and U.S.A," *The Journal of the Korea Contents Association* 9, 2 (2009), 216-232.

[11] Lee, H. J., J. W. Kim, and S. J. Park, "Parameter selection of collaborative filtering for e-commerce personalized recommendation," *Electronic Commerce Research* 7, 3-4 (2007), 293-314.

[12] Lyou, C. G. and N. Y. Park, "Utilization of UCC in convergence era," *The Journal of the Korea Contents Association* 7, 6 (2007), 89-98.

[13] Linden, G., B. Smith, and J. York, "Amazon.com recommendations: Item-to-item collaborative filtering," *IEEE Internet Computing* 7, 3 (2003), 76-80.

[14] Mild, A. and M. Natter, "Collaborative filtering or regression models for Internet recommendation systems?," *Journal of Targeting, Measurement and Analysis of Marketing* 10, 4 (2002), 304-313.

[15] Mobasher, B., R. Cooley, and J. Srivasta, "Automatic personalization based on web usage mining," *Communication of the ACM* 43, 8 (2000), 142-151.

[16] Mulvenna, M. D., S. S. Anand, and A. G. Buchner, "Personalization on the net using web mining," *Communication of the ACM* 43, 8 (2000), 122-125.

[17] Shim, S. and B. Lee, "Internet portal's strategic utilization of UCC and Web 2.0 ecology," *Decision Support Systems* 47, 4 (2009), 415-423.

[18] Wunsch-Vincent, Sacha and Graham Vickery, Participative web: User-created content, Organisation for Economic Co-operation and Development (OECD), April 2007.