IPM-G: Enabling Collaborative Filtering Using Multi-Application Interest Models

Zhurong Zhou[#1], Sampath Jayarathna[#2], Atish Patra[#3],Frank Shipman[#4]

[#]*Computer and Information Science, Southwest University*
Beibei, Chongqing 400715, China

[#]*Computer Science & Engineering, Texas A&M University*

College Station, TX 77840 USA

[#]*Computer Science & Engineering, Texas A&M University*

College Station, TX 77840 USA

[#]*Computer Science & Engineering, Texas A&M University*

College Station, TX 77840 USA

[1]zhouzrcq@gmail.com

[2,3,4]{sampath, apatra, shipman}@cse.tamu.edu

***Abstract*— The Interest Profile Manager (IPM) plays the central role in inferring user interest during document triage. The IPM collects information about interest-related activity from the potentially many triage applications. In this paper, we extend the IPM framework to enable community-based navigation using inferred user interests from information gathering tasks involving the use of multiple applications. We call IPM running on server, Global IPM (IPM-G), and IPM-G can generate similarity assessments, and thus recommendations, based on three different levels: tasks, documents, and annotations. As a result, CF methods can be applied to each level to get results at these three levels of granularity. By representing inferred interests based on the features of their tasks, documents, and annotations, we make possible six potential collaborative filtering (CF) modes in the IPM-G. This paper describes why collaborative filtering based on multi-application interest models is important, abstractly describes the representation of the interest models, and presents details of one of these filtering modes.**

Keywords-Collaborative filtering; user interest modeling; relevance feedback

I. INTRODUCTION

A search engine presents lists of potentially relevant documents to the user. Users have to skim documents to get a sense of their content, evaluate documents to assess their worth in the context of the current activity, and organize documents to prepare for their subsequent use and more in-depth reading. This type of sensemaking task is called Document Triage [4]. A search interface is most often a necessary but not sufficient environment to enable document triage and, in practice, is used in combination with other applications.

The set of applications involved in triage include tools to locating potential documents (e.g. through search or browsing interfaces), tools to examining, skimming, or reading the documents, and tools for recording initial reactions to the documents. For example, the Visual Knowledge Builder (VKB) [8] is a software tool that supports document triage by providing a visual environment for rapidly expressing initial assessments of and relationships among documents.

The IPM plays the central role in inferring user interest during document triage. The IPM collects information about interest-related activity from the potentially many triage applications. This information is aggregated and saved in the user's interest profile. A user interest profile is composed of a set of independent interests. Each interest is represented as a set of document and content features with associated data for computing the overall strength of the interest. Based on the inferred interest profile, the IPM then (1) broadcasts user interests to the participating applications that include application-specific algorithms for computing likely interest of new documents, and (2) acts as an interest assessor for applications that do not include their own interest assessment techniques [3].

This paper presents an architecture and approach to modeling individual interest profiles to provide multiple forms of CF-support via the IPM. The rest of this paper is organized as follows. In section 2, we briefly discuss related work on collaborative filtering and document triage. Section 3 presents the client – server architecture for CF-supported IPM and extends the traditional user-based and item-based CF distinction to account for annotation-level, document-level and task-level assessments. Section 3 also discusses the many features of users' activities collected by IPM and the application of multi-criteria rating [1] to take full advantage of these features by IPM. Finally, we present conclusions and point towards future work in section 4.

II. RELATED WORK

In traditional User-based CF, the fundamental assumption is that if users X and Y rate N items similarly, then they will

rate or act on other items similarly [9]. CF has been identified as a way to provide personalized recommendations to active users of websites where different elements (music, films, products, etc.) can be rated. As a result, when selecting films to watch or choosing products to buy, it is common to take into account the tastes, opinions and experiences of others.

Item-based CF creates a model of the features of items to compute the similarity among items to use alongside the ratings of items to predict the value of a yet unrated item. Thus, item-based CF is another form of content-based (CB) recommendation system [9]. Hybrid approaches combine user-similarity assessments and document-similarity assessments to generate predicted ratings.

Task-based recommender systems [10] rely on a long-term history of user activity from which to mine patterns, e.g. query log data collected by search engines. The query logs combined with click-through data provide a starting point to build user behavior models, but these may lack task-specific semantic labels to extract tasks from the logs. Instead, similarity of tasks is assessed through recognizing similar behavior and often requires considerable log data to generate.

Adomavicius et al. [1] discussed two approaches – the similarity-based approach and the aggregation function-based approach – to incorporating and leveraging multi-criteria rating information in recommender systems.

In [5], Ganta presents a methodology to compute an aggregated interest model accumulated from partial models across multiple triage-related applications. This work shows that IPM models can be merged from different applications, and from different users, to combine with the CF techniques.

III. CLIENT-SERVER ARCHITECTURE

Up until now, the IPM has been a service that resided on the local computer for a user. As shown in Fig.1, we now extend the IPM to a client-server architecture. This makes IPM interest profiles not only sharable by various local client applications, but also sharable by a community of users.

We call IPM running on server, Global IPM (IPM-G), and IPM running on client Local IPM (IPM-L). In this new architecture, IPM-L is still responsible for collecting activity data, generating Interest Profiles (IP), and sharing IPs among the user's various client applications such as web browser, word processing application, PDF readers etc. The IPM-G communicates with IPM-L to mine users' IPs and aggregate them into Task Interest Profiles (TIP).

An individual user's activity is clustered to generate multiple interests that are modeled independently [5]. Each of these interests are assumed to be related to a particular task or set of tasks. It is these task-based interest models for particular users that are shared at the IPM-G. Tasks in the IPM-L are currently described by the set of applications, documents, and user activity observed by the system. The IPM-G enables, but does not require, additional features to be recorded for tasks. This automatically-collected observations and additional features are used to compare and cluster tasks across members of the community. We next describe more formally this framework.

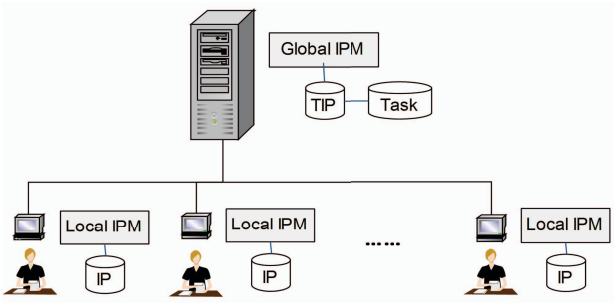

Fig. 1 Client-Server Architecture of IPM

A. IPM-based Collaborative Filtering

An Interest Profile is a list of documents and document segments, user activity [2] associated with each document or segment, a term vector that characterizes each document or segment, and a set of visual and metadata features for each document or segment. User interests are computed as needed based on this data. Because the interests identified in the IPM-L are aggregated into a TIP after a document triage task is complete, the Task is added to IP to ensure the associations among user, IP and task. We put all of this together into the formal definition of IP.

$$IP = (Task, User, DC, TV, VF, UA, R) \quad (1)$$

$User = \{u_1, u_2, \cdots, u_n\}$ is the set of n users engaged in a shared task.

$DC = \{d_1, d_2, \cdots, d_m\}$ is the set of m documents associated with a task.

$TV = \{tv_1, tv_2, \cdots, tv_m\}$ is the set of m term vectors, tv_i characterizes d_i,

$VF = \{vf_1, vf_2, \cdots, vf_m\}$ is a set of visual features, vf_i describe d_i,

$UA = \{a_1, a_2, \cdots, a_l\}$ is the set of l user activities.

$$R = \{(u_i, a_j, d_k) \mid u_i \in User, a_j \in UA, d_k \in DC\} \quad (2)$$

(u_i, a_j, d_k) is a 3-tuple which represents one activity associated with a document by a user.

B. PM-G Recommendation Methods

As previously described, CF recommendation techniques fall into two main categories: user-based and item-based. Each of these techniques requires similarity functions with which to determine similar and dissimilar users and/or items. IPM-G can generate similarity assessments, and thus recommendations, based on three different components in the above model: tasks, documents, and annotations (captured as visual features in the above model). As a result, CF methods can be applied to each level to get results at these three levels of granularity.

There are six possible combinations of CF methods with IPM-G similarity assessments, shown in Fig 2. Users can be identified as similar based on their tasks, based on their

inferred document assessments provided by IPM-L, and based on their annotations or reading/organizing behavior patterns. Similarly, items (documents or document components) can be identified as similar based on whether they are used in similar tasks, have similar inferred assessments, and have similar annotations.

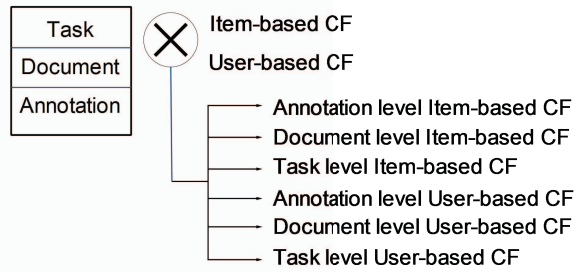

Fig. 2 CF-based IPM

Consider the case of annotation level item-based CF shown in Fig. 3. In this approach, IPM-G recommends document segments that are similar to the annotations in the documents he already annotated.

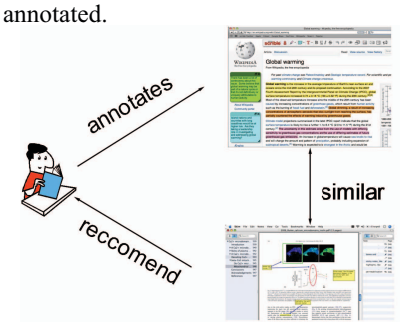

Fig. 3 Annotation level Item-based CF

The next section describes the process of generating suggestions in one of the six approaches (document level user-based CF) using multiple criteria (i.e. multiple features of the users).

C. Document Level User-based CF

The document level user-based CF discussed in this section is a multi-rating recommender method, which recommends documents to a user based on what similar users have consumed. The major problem with single-rating recommender systems is they tend to hide the underlying heterogeneity of the user's preference during a document triage activity. Multi-criteria ratings can help to understand the individuality of each user and their interaction pattern with each application separately. In this section, we will discuss extending IPM to incorporate Multi-Criteria Ratings in a neighborhood-based collaborative filtering recommendation approach [6] .

IPM collects three types of data about user activity and the documents with which they interact: document attributes, document reading activity, and the document organizing activity. This data tends to inherently noisy both for traditional user modeling (due to users having idiosyncratic tasks and document sets) and for collaborative filtering (due to users having idiosyncratic work practices). IPM ' s collection of data from multiple applications broadens the set of information on which to find patterns within this data, resulting in better user models [8]. Here we adopt a multi-rating recommender method for the same reason – variance due to different work practices can be captured by the different ratings.

Based on the interest model, and the analysis in Sections 2 and 3, we established a data model of the Document Level User-based CF (Fig. 4).

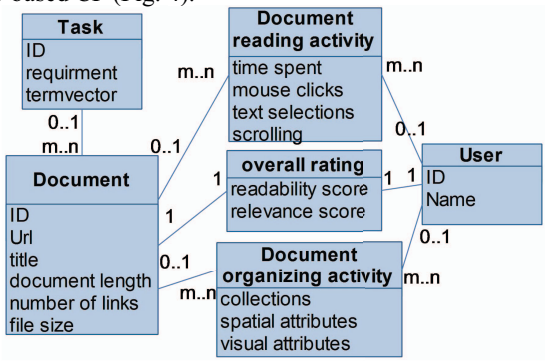

Fig. 4 Data Model of IPM-based Multi-rating Recommender

Formally, the general form of a rating function in the Document Level User-based CF is,

$$R = Users \times Documents \rightarrow R_0 \times R_1 \times R_2 \cdots R_l \quad (3)$$

Where R_0 and R_1 are the set of possible overall rating values, and represents the all possible implicit rating values for each individual criterion k (k = 2, ⋯, l). In our data model, the overall document relevance score while infers the readability of the document. Data gathered from document reading activities are chosen as the multi-criteria items, because this information represents the users ' reading/organizing pattern. Document reading activity includes different user actions such as reading time in the reading application, total time spent in viewing a document, number of mouse clicks, characteristics of the user ' s scrolling behavior, and frequency of document access. These implicit ratings need to be normalized to a specific range [0, 15] during pre-processing as they initially belong to significantly different ranges.

We construct the IPM-based multi-rating recommender in the following steps.

Step 1: Aggregate data from IPM

When a document triage task is completed by a user, the local IPM uploads the collected data to the IPM-G as an XML file. IPM-G parses the data and stores it as a Task Interest Profile (TIP) in relationship tables.

A relation schema for the TIP can be expressed using the following representation,

$$TIP = \left(d_j, u_i, OR1, OR2, click, scroll, TSDC, TRT\right) \quad (4)$$

When user u_i is reading document d_j

OR1 is the overall rating to the relevance of to the task.

OR2 is the overall readability rating

Click is the frequency of all click events
Scroll is the frequency of scrolling events
TSDC is the times of the change of scroll direction.
TRT is the total reading time.

	Document d_1	Document d_2	Document d_3	Document d_4
User u_1	(5,10,5,2,5)	(2,5,5,2,5)	(4,8,5,2,5)	(5,15,5,2,5)
User u_2	(4,8,5,2,5)	(5,11,5,2,5)	(5,15,5,2,5)	(5,10,5,2,5)
User u_x	(5,8,5,2,5)	(4,8,5,2,5)	(4,8,5,2,5)	?
User u_3	(5,12,5,2,5)	(1,2,5,2,5)	(5,10,5,2,5)	(5,8,5,2,5)
User u_4	(5,14,5,2,5)	(1,2,5,2,5)	(3,7,5,2,5)	(4,10,5,2,5)

user to be predicted | neighbors of user u_x | Overall Rating to be predicted

Fig. 5 Collaborative Filtering in a Multi-Criteria Setting

Step 2: Construct multi-rating matrix

For m users and n documents, the user × document rating matrix R is represented as a m × n matrix, As shown in Fig. 5. Each rating that usrer ui gives to document j consists of an "overall" rating OR1, OR2, and l -1 (l=5) multi-criteria ratings click, scroll, TSDC, TRT:

$$r_{ij} = (\text{or1}, \text{or2}, click, scroll, TSDC, TRT) \quad (5)$$

Step 3: Calculate similarity estimations between two users

The $l+1$ different similarity estimations can be obtained by measuring the similarity between users u_x and u_y,

$$S = \{sim_k(u_x, u_y) | 0 \le k \le l\} \quad (6)$$

This is a set of l+1 individual similarities between users ux and uy. We use Person Correlation Coefficient (PCC) [7] as our metric to measure similarity between users u_x and u_y :

$$sim_k(u_x, u_y) = \frac{\sum_{di \in Dxyk} \left(r_{xik} - \hat{r}_{x \uparrow k}\right)\left(r_{yik} - \hat{r}_{y \uparrow k}\right)}{\sqrt{\sum_{di \in Dxyk} \left(r_{xik} - \hat{r}_{x \uparrow k}\right)^2 \sum_{di \in Dxyk} \left(r_{yik} - \hat{r}_{y \uparrow k}\right)^2}}$$

where $D_{xyk} = \{d_i \in D | r_{xik} \neq \theta, r_{yik} \neq \theta\}$ denotes the set of documents co-rated by both u_x and u_x on on criterion . $\hat{r}_{x \uparrow k}$ and $\hat{r}_{y \uparrow k}$ are the average rating by user u_x and u_x on criterion k, respectively.

The overall similarity then can be computed by aggregating the individual similarities by averaging all individual similarities.

Step 4: Select neighbors

In this stage, the k-nearest neighbors will be selected, and the neighbors' ratings are treated as samples of the unknown ratings of the active user that need to be predicted.

$N_k(u_x)$ denotes the set of user ' s k nearest neighbors. The neighborhood $N_k(u_x)$ of the active user u_x will be evaluated by the similarity between users. This step helps identify users with similar reading/organizing patterns.

Step 5: Aggregate ratings

The final stage generates a predicted rating $\hat{r}_{xi}$ by user u_x for document d_i by aggregating all the ratings on d_i by users in the neighborhood $N_k(u_x)$:

$$\hat{r}xi = aggr_{ux \in Nk(ux)} r_{xi} = \frac{1}{k} \sum_{ux \in Nk(ux)} r_{xi} \quad (7)$$

Finally, the documents with highest predicted overall relevance ratings are recommended to the user.

IV. CONCLUSIONS

This work extends the IPM architecture to support community-based recommendations. In particular, the framework presented enables combining similarity assessments for the three levels of content modeled by IPM-G (tasks, documents, and annotations) with the two categories of CF (user-based and item-based). As a result, IPM-G can make recommendations based on the similarity of tasks, documents, and annotations. Multi-criteria rating is used to merge the heterogeneous activity data and relevance assessments collected during the performance of tasks and to generate overall assessments based on this heterogeneous data. More research is needed to identify the set of implicit criteria that successfully represent a variety of users' interests efficiently.

ACKNOWLEDGMENT

This work was supported by the Internet of Things application open innovation and application demonstration, which number is MCM20122011.

REFERENCES

[1] Adomavicius, G., Manouselis, N., & Kwon, Y. (2011). Multi-criteria recommender systems Recommender systems handbook (pp. 769-803): Springer.

[2] Badi, R., Bae, S., Moore, J.M., Meintanis, K., Zacchi, A., Hsieh, H., Marshall, C., & Shipman, F. (2006). Recognizing user interest and document value from reading and organizing activities in document triage. Proc. IUI, 218-225.

[3] Bae, S., Hsieh, H., Kim, D., Marshall, C.C., Meintanis, K., Moore, J.M., Zacchi, A., and Shipman, F. (2008). Supporting document triage via annotation based visualizations. Proc. ASIST, 45(1), 1-16.

[4] Bae, S., Marshall, C.C., Meintanis, K., Zacchi, A., Hsieh, H., Moore, J.M., & Shipman, F. (2006). Patterns of reading and organizing information in document triage. Proc. ASIST, 43(1), 1-27.

[5] Ganta, P. (2011). A Comparison of Clustering Methods for Developing Models of User Interest. Master's thesis, Dept. of Comp. Sci. and Eng., Texas A&M University.

[6] Kim, H. R., & Chan, P. K. (2003). Learning implicit user interest hierarchy for context in personalization. Proc. IUI Conf., 101-108.

[7] Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., & Riedl, J. (1994). GroupLens: an open architecture for collaborative filtering of netnews. Proc. CSCW, 175-186.

[8] Shipman, F., Hsieh, H., Maloor, P., & Moore, J.M. (2001). The visual knowledge builder: a second generation spatial hypertext. Proc. Hypertext, 113-122.

[9] Su, X., & Khoshgoftaar, T. (2009). A survey of collaborative filtering techniques. Adv. in Artif. Intell., 2009, 2-2. doi: 10.1155/2009/421425

[10] Tolomei, G., Orlando, S., & Silvestri, F. (2010). Towards a task-based search and recommender systems. Proc. ICDE.