Document Recommendation in Organizations with Personal Folders

Chiu-Wen Huang, Chin-Hui Lai, Duen-Ren Liu
Institute of Information Management
National Chiao Tung University
Hsinchu, Taiwan

Abstract: In organizations, knowledge workers usually have their own personal folders that store and organize needed codified knowledge (textual documents) in taxonomy. In such personal folder environments, providing knowledge workers needed knowledge from other workers' folders is important to facilitate knowledge sharing. This work adopts recommendation techniques to provide knowledge workers needed textual documents from other workers folders. Experiments are conducted to verify the performance of various methods using data collected from a research institute laboratory. The result shows that the CBF approach outperforms other methods.

Keywords: Knowledge Management, Document Recommendation, Collaborative Filtering, Content-based Filtering.

I. Introduction

Sharing sustainable and valuable knowledge among knowledge workers is a prominent activity of knowledge management. Organizational knowledge and expertise are usually codified into textual documents, including forms, letters, papers, manuals and reports, to facilitate knowledge capture, search and sharing [7].

Knowledge workers tend to keep their codified knowledge in their own personal folders. Textual documents stored in each worker's personal folder are usually organized into categories in taxonomy. In such personal folder environments, providing knowledge workers needed knowledge from other workers' folders is important to facilitate knowledge sharing. Conventional knowledge management systems (KMSs) have provided search function to help knowledge workers find needed knowledge. However, very few KMSs have considered the issue of proactively providing knowledge workers needed knowledge in personal folder environments.

Recommender systems [2] seem to be an effective solution for proactively providing knowledge workers needed knowledge. Conventional application domains of recommender systems are "Music", "Movie" or "Product" recommendations. Various recommendation methods have been proposed for recommender systems. Collaborative Filtering (CF) assumes that items (e.g. documents) from like-minded users are often relevant. Collaborative filtering utilizes preference ratings given by various users to determine recommendations to a target user based on the opinions of other similar users. Content-based Filtering (CBF) utilizes profile matching to determine recommendation to target users. In application to recommend documents, Content-based filtering provides recommendations by matching user profiles (e.g., interests) with content features (e.g., feature vectors of documents). Each user profile is derived by analyzing the content features of documents accessed by the user.

LIBRA system [6] is an example of content-based filtering, which recommends books based on book information extracted from Web pages. Siteseer [8] used collaborative filtering to provide Web page recommendations based on the bookmarks of the user's virtual neighbors without considering the categorization of bookmarks. Knowledge Pump [4] used CF techniques to recommend documents based on personal profiles of interest. In these systems, each user stores his/her documents in a commonly agreed classification scheme rather than a personalized one.

RAAP [3] is an example of hybrid system developed to classify and recommend bookmarks retrieved from the Web. The InLinx system [1] also supports the classification and recommendation of bookmarks retrieved from the Web based on content analysis and virtual clusters. Middleton et al. [5] presented an ontological user profiling approach to recommend academic papers. Recommended papers are those match the user's profile and have also been read by similar users.

This work investigates recommendations of textual documents in personal folder environments. Each knowledge worker has his/her own folder that store documents into personalized categories, namely categories defined by himself/herself. We adopt recommendation techniques to provide knowledge workers needed textual documents from other workers folders. Conventional document recommender systems assume a common category schema without considering personalized categories. Our proposed approaches combine filtering and text categorization to recommend documents to target worker's personalized categories. The recommendation proactively notifies knowledge workers regarding peer-reviewed documents, and therefore knowledge diffusion is evolved from 「Pull」 to 「Push」. By means of knowledge diffusion, knowledge workers can learn from each other and eventually elevate work productivity and efficiency. Experiments are conducted to verify the performance of various methods using data collected from a research institute laboratory.

Proceedings of the Fifth International Conference on Electronic Business, Hong Kong, December 5-9, 2005, pp. 528 - 532.

II. Methodology

This section presents the proposed document recommendation methodology that aims to fulfill the goal of push-mode knowledge diffusion. In the organization, documents, manuals, reports, know-how and the like from people in the same project team or from people with similar working experience are of great help. Fig. 1 shows knowledge sharing in personal folder environments. Knowledge workers used to manage his owned document repository by storing documents in different categories. Each knowledge worker shares their own documents to others.

The proposed methods recommend documents stored in other knowledge workers' folders to the target worker's right category. One of the ways to reuse the knowledge in the enterprise is to sharing the knowledge by interflow of knowledge documents. However, the received documents are another burden because knowledge workers have to spend time on managing them. Classification is the basic to manage documents for users to quickly access and store them. However, different people have different criteria to classify documents.

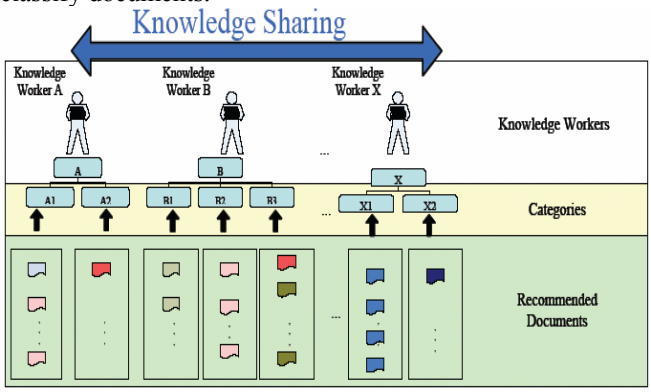

FIG. 1. Recommendation for knowledge sharing

Our approach tries to find the recommendation candidates by examining the document and category profiles to predict if a document is suitable to be recommended to the target category. We proposed two kinds of recommendation methods: content-based filtering and collaborative filtering in order to recommend documents in the personal folder environment. As a result, explicit knowledge embedded in knowledge workers' personal folders is circulated from peer to peer to facilitate knowledge sharing.

II. 1 Methods Based on Content-Based Filtering

Based on the concept of content-based filtering recommendation, we divide this methodology into three phases. Phase 1 is profile generation; phase 2 is document filtering; and the last phase is recommendation list generation.

In phase 1, the method generates three kinds of profiles, including Document Profile (DP), Category Profile (CP) and User Profile (UP).

Document Profile (DP)

Let d_j be a document, and let $DPj = <dt_{1,j}: dw_{1,j}, \ dt_{2,j}: dw_{2,j}, \ ..., \ dt_{n,j}: dw_{n,j}>$ be the feature vector (document profile) of d_j where $dw_{i,j}$ is the weight of a term i that occurs in d_j. Notably, the weight of a term represents its degree of importance to represent the document (codified knowledge). The well-known *tf-idf* approach is often adopted for term weighting [9][10]. Let the term frequency $df_{i,j}$ be the occurrence frequency of term i in d_j, and let the document frequency df_i represent the number of documents that contain term i. The importance of term i to a document d_j is proportional to the term frequency and inversely proportional to the document frequency, which is expressed as Eq. (1).

$$dw_{i,j} = \frac{1}{\sqrt{\sum_i \left(dt_{i,j} \times (\log \frac{N}{df_i} + 1) \right)^2}} dt_{i,j} \times (\log \frac{N}{df_i} + 1) \quad (1)$$

where N is the total the number of documents and the denominator in the right side of Eq. (1) is a normalization factor to normalize the weight of term.

Category Classifier (CC)

A classifier for a category is constructed through *tf-idf* approach which is employed to extract the discriminating terms and their weights among categories of a knowledge worker. Let $CC_r = <cct_{1,r}: ccw_{1,r}, \ cct_{2,r}: ccw_{2,r}, \ ..., \ cct_{n,r}: ccw_{n,r}>$ be the category classifier of c_r where $ccw_{i,r}$ is the weight of a term i that occurs in CC_r. Let the term frequency $ctf_{i,r}$ be the occurrence frequency of term i in c_r, and let the category frequency cf_i represent the number of categories in target user u that contain term i. The weight of term i in a category c_r is proportional to the term frequency and inversely proportional to the category frequency, which is expressed as Eq. (2).

$$cw_{i,r} = \frac{1}{\sqrt{\sum_i \left(ctf_{i,r} \times (\log \frac{L_u}{cf_i} + 1) \right)^2}} ctf_{i,r} \times (\log \frac{L_u}{cf_i} + 1) \quad (2)$$

where L_u is the total the number of categories in user u. Notably, the denominator in the right side of Eq. (2) is a normalization factor to normalize the weight of term.

User Profile (UP)

The profile of a user u_x is represented as a feature vector of weighted terms derived by analyzing documentation set owned by u_x. After the documents are pre-processed and represented in the form of term vectors, UP_x is derived by averaging the feature vectors (i.e. centroid approach) of documents in u_x. Let D_x denote the set of documents in u_x. Furthermore, the user profile (feature vector) UP_x of user u_x

is defined as the centroid vector obtained by averaging the feature vectors of documents in D_x. Let $uw_{i,x}$ denote the weight of a term i in UP_x. $uw_{i,x}$ is derived as Eq. (3).

$$uw_{i,x} = \frac{1}{|D_x|} \sum_{d_j \in D_x} dw_{i,j} \quad (3)$$

Phase 2 applies recommendation scheme to filtering documents of low similarity. Content-based recommendation mainly computes the similarity between category classifier and document profile.

The cosine formula is a widely adopted scheme to measure the similarity degree between two items x and y. The cosine of the angle between their corresponding feature vectors Q and R is computed as given by Eq. (4). The degree of similarity is higher if the cosine similarity is close to 1.0.

$$sim(x, y) = \text{cosine}(Q, R) = \frac{Q \bullet R}{|Q||R|} \quad (4)$$

The method considers both the similarity of document profile to the category classifier and user profile. The predicted rating $\hat{p}_{a,j}$ of recommending document d_j to the category c_a of the target user u_x is expressed in Eq. (5):

$$\hat{p}_{a,j} = (1 - \alpha_{CBF}) Sim(CC_a, DP_j) + \alpha_{CBF} Sim(UP_x, DP_j) \quad (5)$$

where $sim(CC_a, DP_j)$ is the similarity of CC_a and DP_j and $sim(UP_x, DP_j)$ is the similarity of UP_x and DP_j (u_x is the owner of C_a). α_{CBF} ranges from 0 to 1 and will be decided by the analytical experiments.

The last phase is to generate a recommendation list of document-category pairs for allocating documents to destination categories. The document-category pairs are sorted according to their predicted ratings. The pairs with top-N highest raking are selected for recommendation. Notably, those documents that the target user already has are not included in the recommendation list.

II. 2 Methods Based on Collaborative Filtering

This method uses the opinions of other knowledge workers with similar profiles to make recommendations. Two approaches are developed including Collaborative Filtering (CF) and Collaborative Filtering based on Joint coefficient (CF-J).

II .2. 1 Collaborative Filtering (CF)

There are also three phases in the proposed CF approach. Phase 1 generates the needed profiles. The category classifier is mainly used to determine which category a document should be allocated to, and thus is suitable for classification purpose. However category classifier is not suitable to derive similar neighbors, since the discriminating terms may distort the similarity of categories in different users. Thus, category profile is defined to compute the similarity of categories, and is further used to find category neighbors. Similar to the generation of user profiles described in section 2.1, the centroid approach is used to derive the category profile.

The profile of a category c_a is derived by analyzing the set of documents in c_a. Each document d_j is pre-processed and represented as a feature vector DP_j. Let D_a denote the set of documents in c_a. Furthermore, the category profile CP_a of category c_a is defined as the centroid vector obtained by averaging the feature vectors of documents in c_a. Let $cw_{i,a}$ denote the weight of a term i in CP_a. $cw_{i,a}$ is derived as Eq. (6). Notably, category profile does not consider the effect of terms in discriminating the category of a user.

$$cw_{i,a} = \frac{1}{|D_a|} \sum_{d_j \in D_a} dw_{i,j} \quad (6)$$

Phase 2 identifies the neighbors of the target category. The similarity between CP_s is derived to decide neighbors. For recommending a document d_j to the target category c_a, the neighboring categories (neighbors) of c_a is selected from categories that contain d_j. The cosine formula is used to determine the similarity of CP_s. We use the k-NN based method for choosing neighbors.

Phase 3 derives the predicted rating of document-category allocation. In addition to the profiles, the CDR (Category-Document-rating) / UDR (User-Document-rating) matrix is needed to record the rating of categories/users on documents.

There are two approaches to derive the ratings, binary approach and profiling approach. The binary approach derives the ratings based on the criteria whether the category/user contain the document. If a category c_a contains a document d_j, the rating value of c_a on d_j, $CDR_{a,j}$, is 1; otherwise, the value is 0. If the category c_a is owned by the user u_x, i.e., u_x has document d_j, the rating value of u_x on d_j, $UDR_{x,j}$, is 1; otherwise, the value is 0. The profiling approach uses the similarity of category/user profile and document profile to derive the rating. The rating value of c_a on d_j, $CDR_{a,j}$, equals $sim(CP_a, DP_j)$, i.e., the similarity of the category profile of c_a and the document profile of d_j. The rating value of u_x on d_j, $UDR_{u,j}$, is set to $sim(UP_x, DP_j)$, i.e., the similarity of the user profile of u_x and the document profile of d_j.

The CDR/UDR generated by the binary approach is called binary CDR/UDR, while the CDR/UDR generated by the profiling approach is called non-binary CDR/UDR. Eq. (7) computes the predicted rating of recommending document d_j to the category c_a of the target user u_x.

$$\hat{p}_{a,j} = \frac{\sum_{c_b \in c_a\text{'s neighbor}} (1 - \alpha_{CF}) Sim(CP_a, CP_b) \times CDR_{b,j} + \alpha_{CF} Sim(UP_x, UP_y) \times UDR_{y,j}}{Number\ of\ c_a\text{'s neighbors}} \quad (7)$$

where $sim(UP_x, UP_y)$ is the similarity between UP_x and

UP_y, $sim(CP_a, CP_b)$ is the similarity between CP_a and CP_b, c_b belongs to c_a's neighbors, u_y is the owner of c_b. $CDR_{b,j}$ is defined in phase 3. α_{CF} is a parameter to adjust the relative importance of category similarity and user similarity.

Finally, the scheme generates a list of candidate document-category allocation. The procedure is the same as the CBF described in phase 3 of section 2.1.

II. 2. 2 Collaborative Filtering Based on Joint coefficient (CF-J)

CF-J is similar to CF. The difference between CF and CF-J is the similarity computation. CF calculates the similarity by weighted term profiles. The joint coefficient approach (CF-J) calculates the similarity based on the joint coefficient, which represents the relationship between two categories/users decided by the number of the documents they have in common. The more they have, the more similar they are. Equation (8) is the formula to compute the joint coefficient *(Jcof*) in CF-J.

$$Jcof(c_a, c_b) = \frac{2 \times N_{a \cap b}}{N_a + N_b} \tag{8}$$

where N_a and N_b is the number of documents in c_a and c_b respectively, and $N_{a \cap b}$ represents the intersection of documents that both c_a and c_b have. The binary CDR is used to derive N_a, N_b and $N_{a \cap b}$. Similarly, joint coefficient between two users u_x and u_y can be defined as $Jcof(u_x, u_y)$.

CF-J uses joint coefficient instead of profile similarity to derive the predicted rating as expressed in Eq. (9).

$$\hat{p}_{a,j} = \frac{\sum_{c_b \in c_a\text{'s neighbor}} (1-\alpha_{CF-J}) Jcof(c_a, c_b) \times CDR_{b,j} + \alpha_{CF-J} Jcof(u_x, u_y) \times UDR_{y,j}}{Number\ of\ c_a\text{'s neighbors}} \tag{9}$$

III. Experiemnt and Evaluation

Experiments using a real application domain were carrie d out for recommending research papers in a research i nstitute laboratory.

III. 1 Experimental Setup

Knowledge workers have their own folders storing documents (research papers) that assist them in writing theses or accomplishing research projects. There are 11 users, 35 categories and 1062 documents. The sparsity in the data sets is 99.96%. For each category, there are at least ten documents in order to provide enough information of the codified profiles. We also limit the level of categories to be one. Those categories with level higher than one will be aggregated into their level-1 ancestors. The data set was divided into an 80% training set and a 20% testing set. The training set includes documents stored in workers' personal folders, and was used to generate recommendation list. Testing data was used to verify the recommendation quality of various methods.

Two metrics, precision and recall, are commonly used to measure the quality of recommendation. These two metrics are also extensively used measures in information retrieval [10]. Recall is the fraction of relevant documents that can be located.

$$\text{Recall} = \frac{\text{number of correctly recommended documents}}{\text{number of relevant documents}} \tag{10}$$

Precision is the fraction of recommended documents (predicted to be relevant) that are really relevant to workers.

$$\text{Precision} = \frac{\text{number of correctly recommended documents}}{\text{number of recommended documents}} \tag{11}$$

Documents relevant to a worker *u* are those documents owned by *u* in the test set. Each relevant document is associated with its corresponding category owned by *u*. Such relevant document with associated category is called a relevant document-category pair of *u*. Correctly-recommended documents are those in the recommended document-category pairs that match the relevant document-category pairs of *u*.

F1-metric can be used to balance the trade-off between precision and recall [10]. F1-metric assigns equal weight to precision and recall, and is given by,

$$\text{F1} = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} \tag{12}$$

III. 2 Experimental result

We compare different factor of CF. CF-Binary, CF-Profile and CF-J uses binary ratings, profiling ratings and joint coefficient, respectively, as described in section 2.2. The $\alpha_{CF} / \alpha_{CF-J}$ is used to tune the weight of predicted rating contributed from Category similarity and user similarity. The α values for CF-Binary, CF-Profile and CF-J, are 0.5, 0.0 and 0.2, respectively, which are decided according to the highest average value. Fig. 2 shows the comparison (F1-metric) of CFs under different Top-N. CF-Binary is relatively better than the CF-Profile method. This indicates that the rating part in the CF-Profile approach does not provide useful rating information by using the profiling approach. The failure of CF-Profile might result from the rating part of formula (the similarity of category and document) which could not truly represent user's rating on the documents. Therefore, the CF-Profile method could not effectively recommend the right document to the right category. Consequently, we adopt the CF-Binary method rather than the CF-Profile Method to represent the CF

method in further comparisons.

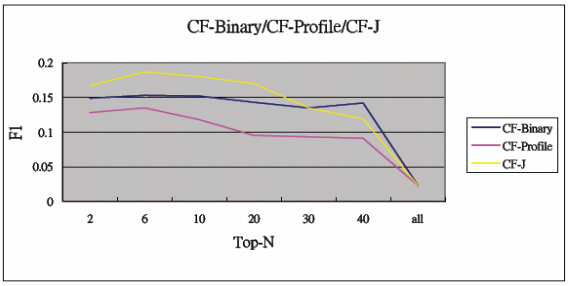

FIG. 2: Comparison of CF and CF-J under different Top-N

Fig. 2 also shows that CF-J achieves better result in smaller Top-N and CF-Binary works better in larger Top-N. The number of overlapped documents among different categories is usually small. Hence, CF-J performs worse when recommending more documents. αCF-J and αCF are set to 0.2 and 0.5, respectively. This indicates that the opinions from the similarity of user profiles provide constructive effect in improving recommendation quality.

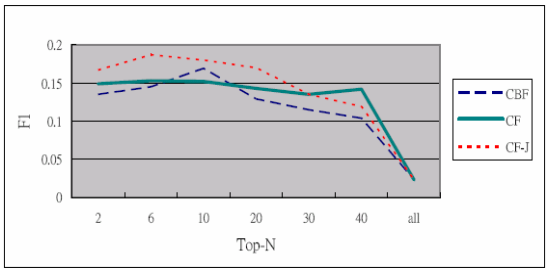

FIG. 3: Comparison of CBF and CF methods

Fig. 3 shows the comparison of CBF, CF (CF-Binary) and CF-J under different Top-N. The CBF uses Category Classifier (CC) to provide content-based filtering. α_{CBF} is set to 0 for CBF. The result shows that the CF and CF-J performs better than CBF. CF-J gain better performance when Top-N is smaller; however, when Top-N is getting larger, CF (CF-Binary) method provides better recommendation quality.

IV. Conclusions

This work investigates the issue of sharing codified knowledge stored in workers' personal folders. Various recommendation approaches are proposed to recommend codified knowledge to the right category of workers' personal folders. The proposed approach provides workers needed relevant documents from other workers' folders to facilitate knowledge sharing. The explicit codified knowledge can circulate around the organizations by sharing codified knowledge from personal folders. The proposed work can reduce the efforts and manpower in document classification and improve knowledge sharing among organizations.

Acknowledgement

This research was supported in part by the National Science Council of the Taiwan (Republic of China) under the grant NSC 95-2416-H-009-002.

References

[1] Bighini, C., Carbonaro, A., Casadei, G., "InLinx for Document Classification, Sharing and Recommendation," *the 3rd IEEE International Conference on Advanced Learning Technologies (ICALT'03)*, 2003, pp. 91.

[2] Burke, R.D., "Hybrid recommender systems: survey and experiments," *User Modeling and User-Adapted Interaction*, 12(4), 2002, pp. 331-370.

[3] Delgado, J., Ishii, N., and Ura T., "Intelligient Collaborative Information Retrieval," *Proceedings of the 16th IBERO-American Conference on AI*, 1998, 148, pp.170-182.

[4] Glance, N., Arregui, D. and Dardenne, M, "Knowledge Pump: community centered collaborative filtering," *In Proceedings of the Fifth DELOS Workshop on Filtering and Collaborative Filtering*, 1998, pp. 83-88.

[5] Middleton, S. E., Shadbolt, N. R. and De Roure, D. C., "Ontological User Profiling in Recommender Systems," *ACM Transactions on Information Systems (TOIS)*, 2004, 22(1), pp. 54-88.

[6] Mooney, R., and Roy L., "Content-Based Book Recommending Using Learning for Text Categorization," *Proceedings of the Fifth ACM Conference on Digital Libraries*, 2000, pp.195–204.

[7] Nonaka, I., (1994), "A Dynamic Theory of Organizational Knowledge Creation," *Organization Science*, 5(1), pp.14–37.

[8] Rucker, J., and Polanco, M.J., "Personalized navigation for the Web," *Communications of the ACM*, 40(3), 1997, pp. 73-75.

[9] Salton, G.., and Buckley, C., "Term Weighting Approaches in Automatic Text Retrieval," *Information Processing & Management*, 1988, 24(5), pp. 513-523.

[10] Salton, G.., and McGill, M., *Introduction to modern information retrieval*, McGraw-Hill, 1983.

[11] van Rijsbergen, C.J. 1979. Information Retrieval. Butterworths, London, 2nd edition.