Personalized Celebrity Video Search Based on Cross-Space Mining

Zhengyu Deng, Jitao Sang, and Changsheng Xu

Institute of Automation, Chinese Academy of Sciences,
100190 Beijing, China
{zydeng,jtsang,csxu}@nlpr.ia.ac.cn

Abstract. Online videos are becoming popular these days. Personalized search has been recognized as effective solution for user accessing desired information when facing a daunting volume of videos. Personalized query understanding serves as one of the most challenges in personalized search, which indicates that unique query has distributed meanings and produce different semantics for different users. Take query of celebrity as example, many celebrities are engaged in multiple fields and certain user may be just interested in the field of videos related to his/her own preference. In this paper, we address the challenge of personalized query understanding by focusing on the problem of personalized celebrity video search. An interest-popularity cross-space mining based method is proposed for solution. Specifically, celebrity popularity and user interest distributions are first learned by topic modeling from heterogeneous data of expert knowledge and user online activities, respectively. We then exploit topic-word distribution refinement to correlate the two heterogeneous topic spaces. Finally the candidate videos are re-ranked based on the derived interest-popularity correlations. Carefully designed experiments have demonstrated the effectiveness of the proposed method. The obtained ranking list is highly consistent with the test users' preferences.

1 Introduction

Nowadays, online video propagation has surged up to an unparalleled level. Within the vast video pool, the videos about celebrities appear highly frequently and are closely followed by the users because of the "Celebrity Effect". It's common that celebrities are engaged in multiple domains. Take David Beckham for example, he is famous for his specialty in soccer sports; but he is also active in the field of entertainment as a fashion expert. From the perspective of users, some Real Madrid fans may take interest in his soccer videos; whereas some like young ladies may adore his fashion style and search for entertainment videos; even some idolaters may prefer the videos that are related to his daily life. Therefore, when a user searches "Beckham", a real personalized search scheme should consider this phenomenon and rank the videos based on the user's specific interested area, i.e., Beckham's soccer games, fashion style or daily life.

The challenges of realizing such a personalized celebrity video search scheme lie in three aspects. (1) The various fields that certain celebrity gets involved in

W. Lin et al. (Eds.): PCM 2012, LNCS 7674, pp. 455–463, 2012.

is not always clear and needs to be explored. (2) Users seldom explicitly provide their interest profiles and the interest-oriented preferences are not available in topic level. (3) How to connect user interest with celebrity popularity is not trivial. Generally, user interest and celebrity popularity are extracted in different spaces from heterogeneous data sources. Therefore, how to explore the latent association of the two spaces is the key factor to solve the problem.

In this paper, we propose an Interest-Popularity Cross-space Mining based method to address the abovementioned challenges. For the celebrity side, celebrity popularity is explored by leveraging expert information, e.g., the corresponding wikipedia homepages. Standard topic modeling method of Latent Dirichlet Allocation (LDA) is adopted to extract the celebrity popularity distribution in abstract topic level. For the user side, since off-the-shelf user profile is unavailable or hardly informative, we exploit user interest based on his/her online activities, e.g., video sharing, social tagging. LDA is again utilized for user interest topic extraction. Given the derived heterogeneous popularity and interest spaces, we introduce a cross-space correlation method. Semantic and context intra-word relations are refined by random walk to bridge the interest and popularity spaces. The framework of our proposed approach is shown in Fig. 1. The inputs include the celebrities' Wikipedia profile and the users' uploaded and favorite videos with associated tags. The output is the generated video ranking list. The framework contains three components, namely interest and popularity space construction, cross-space correlation and video re-ranking. Video re-ranking is based on joint probability distribution of user, celebrity and videos in interest space. See section 3 for detailed elaboration. To summarize, the main contributions of this paper are as follows.

(1) We introduce the novel problem of personalized celebrity video search, by exploiting the user interest and celebrity popularity in topic level.

(2) We propose a cross-space correlation method to connect heterogeneous spaces, which serves as a feasible solution to other cross-domain problems.

(3) With celebrity as a special case of distributed query, we provide one of the first attempts to address the query understanding challenge in personalized search problem.

2 Related Work

In academic communities, some researchers attempt to employ clustering algorithms to assist video retrieval [10] [9]. For instance, Shepitsen et al. [10] clustered the social tags into several concepts and thereafter connected the user and item through those concepts. Some other researchers [6] [2] [5] [12] adopted a hierarchically-arranged collection of concepts or ontologies, in which each node of the ontologies represents a certain interest. For example, Leung and Lee [6] proposed a concept-based profiling strategy to represent users' preferences using weighted concept vectors. Evans et al. [2] defined user interest as a distribution over the category nodes of an ontology. Furthermore, some papers have reported to build up one united space for users and items [10] [3] [8] [9] [11]. Among them,

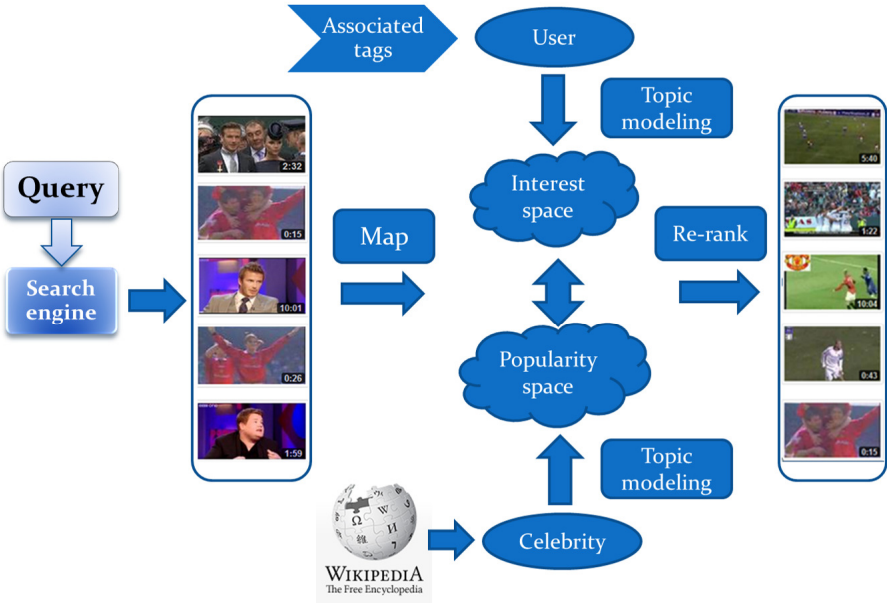

Fig. 1. The framework of our proposed approach

Xu et al. [11] proposed a topic-based personalized search scheme, which maps the sources (user profiles and web pages) onto a unified topic space and decide the ranking score both by the term similarity matching and the topic similarity matching in the unified space. However, these work considers users and items in the same topic space, which is not applicable in our scenario. Besides these works predetermine the subject of interests; hence they fail to represent user interest flexibly. Our work is the first to build up the interest and popularity spaces for user and celebrity separately and re-rank the videos based on the cross-space mining on the topic level.

3 Approach

3.1 Interest and Popularity Topic Spaces Construction

1) **Interest Space**. Generally speaking, users' registration information is useful to analyze their preferences. But it's not easy to aquire because of privacy problem. In comparison, users' active actions (like "upload" or "favor") on a video strongly indicate their attentions and preferences. And since these videos are easy to retrieve through video sharing websites, the users' profiles could be built up by extracting the tags and categories associated with those videos. However, the tags annotated by web users contain plenty of noises such as meaningless words or typos. To tackle this issue, we utilize WordNet to filter out the noises

and only keep noun tags which are the least noisy representations for users' interests. After building up users' profiles, LDA [1] is adopted to learn the latent topics for interest space.

LDA extracts topics ($z \in Z$) from the set of user profiles ($u \in U$) and generates two distributions: user-topic distribution $p(z|u)$ and topic-word distribution $p(w|z)$. The vocabulary consists of N words ($w \in W$). Denote α and β as the hyperparameters respectively. The joint distribution of the topic mixture θ, a set of K topics Z, and N words W, is given by [1]

$$p(\theta, Z, W|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} \sum_{i=1}^{K} p(z_i|\theta) p(w_n|z_i, \beta). \tag{1}$$

2) **Popularity Space**. Generally speaking, celebrities often have Wikipedia entries introducing their careers, achievements, life and many other aspects, which reflect the popularity distribution of the celebrity. So we use the entry information to represent the celebrity and build up the popularity space accordingly. The formulation is the same as Eq. 1.

After the probability distribution of a celebrity over the latent topics is obtained, the distribution could determine whether the celebrity is dominant in one or multiple domains. Apparently, if the probability distribution score is high in several latent topics, the celebrity may be engaged in multi-domains, vice versa. Therefore, we utilize entropy of information to differentiate celebrities. Given celebrity c, the entropy could be explicitly written as

$$H(c) = \sum_{i=1}^{L} p_c(x_i) \log \frac{1}{p_c(x_i)} = -\sum_{i=1}^{L} p_c(x_i) \log p_c(x_i) \tag{2}$$

where L is the number of the latent topics extracted by LDA and x_i ($i \in \{1, ..., L\}$) is the ith hidden topic. $p_c(x_i)$ is the probability of celebrity c on topic x_i. The bigger the value is, the more even the distribution will be, i.e., the more domains the celebrity is engaged in.

3.2 Cross-Space Correlation

After the latent topics of each space have been extracted via LDA, the two spaces could be inter-correlated by these topics. We have known that each topic contains a number of semantic words and there is a probabilistic distribution over them. Thus, we can compute the similarity between these topics by KL-divergence on these words. However, since the user and the celebrity are from heterogeneous spaces, the vocabulary of the interest space is not identical with that of the popularity space. Besides, when we extract latent topics using LDA, we have not considered the correlation between the semantic words, which are very important for information propagation, especially when the training data is very sparse. Therefore, we merge the two vocabularies together and set up an transition matrix by WordNet. After that, the random walk which has been

widely applied in machine learning and information retrieval fields [4] [7], is utilized to update the probability distribution of topic-word.

We use s_{ij} to denote the similarity of word i and j, which is obtained from WordNet. Given a word graph composed of N words and each word is regarded as a node, the transition matrix is denoted by $\mathbf{P}_{N\times N}$. Its element p_{ij} indicates the probability of the transition from the node i to node j and is computed as $p_{ij} = s_{ij}/\sum_k s_{ik}$. Denote $r_k(i)$ as the relevance score of the node i at iteration k, the relevance scores of all the nodes in the graph at iteration k form a column vector $r_k = [r_k(i)]_{N\times 1}$. So the random walk process is formulated as

$$\mathbf{r}_k = \lambda \sum_i \mathbf{r}_{k-1}\mathbf{P} + (1-\lambda)\mathbf{y} \tag{3}$$

where $\mathbf{y}$ is the initial probability distribution vector of the topic-word, and $\lambda \in (0, 1)$ is a weight parameter. The above process will make the similar words have the similar scores and strengthen the words that have many close neighbors. The iteration of Eq. 3 converges to a fixed point $\mathbf{r}_\pi = (1-\lambda)(\mathbf{I} - \lambda\mathbf{P})^{-1}\mathbf{y}$ [7].

Random walk makes each topic has a probabilistic distribution over the whole vocabulary of words. Afterwards, KL-Divergence is utilized to connect user and celebrity at the topic level. Since KL-divergence is direction-related, the average value of the two directions is used here. Assume that z and x are topics from interest and popularity space respectively, the KL-Divergence between them is defined as

$$D_{KL}(z \parallel x) = \frac{1}{2}(\sum_i z(i) ln \frac{z(i)}{x(i)} + \sum_i x(i) ln \frac{x(i)}{z(i)}) \tag{4}$$

where $z(i)$ and $x(i)$ denote the distribution scores of topic z and x on word i. The similarity s_{zx} of topic z and x is defined as the inverse of KL-Divergence. $s_{zx} = 1/D_{KL}(z \parallel x)$

3.3 Video Re-ranking

In this section we will elaborate how to re-rank the video list based on the user-celebrity correlation. Given a user u, when u query certain celebrity c, we search c in video retrieval engine. Afterwards we re-rank the top-n videos by the correlation of interest-popularity space. Specifically, we first project the celebrity videos on the interest space (Notably, for the sake that we want to provide potential interesting videos to user, it's more reasonable to focus on interest space of users). Assume Φ is a $K \times M$ (K is the topic number of interest space. M is the dimension of the vocabulary of the semantic words) Markov matrix, each row of which denotes the probability distribution vector of a topic over each word in the vocabulary. For any video vector $v_{M\times 1}$, we project it to interest space as $v'_{K\times 1} = \Phi v$

For any celebrity video v, the relevance score is computed as

$$\begin{aligned} & p(score \mid v, u, c) \\ &= \sum_{i=1}^{K} p(z_i \mid v) p(z_i \mid u) p(z_i \mid c) \\ &= \sum_{i=1}^{K} p(z_i \mid v) p(z_i \mid u) \sum_{j=1}^{L} p(x_j \mid c) p(z_i \mid x_j) \end{aligned}$$

where L is the topic number of popularity space, z_i (x_j) is the ith (jth) topic of interest (popularity) space, $p(z_i \mid x_j)$ is approximated by KL-Divergence (see Eq. 4). After getting the score for each of celebrity videos, we re-rank the videos according to the scores and return a ranking list to the target user.

4 Experiments

4.1 Experimental Settings

We conduct our experiments on a dataset collected from YouTube. Firstly we pick up 330 most popular or powerful celebrities from Forbes[1]. Afterwards we shortlist 106 multi-domains engaged celebrities from them via information entropy (see Eq. 2). Then we extract top-200 related videos on the average for each celebrity from YouTube. At the same time, we collect 143 users from YouTube. The average number of videos uploaded or favored by these users is 205. For each user (denoted by u), some of the videos he/she uploaded or favored are related with certain celebrity (denoted by c) among the abovementioned 106 celebrities. Therefore, in our experiments, we assume that u will issue a query c and obtain a ranking list. Afterwards we count how many videos (not in the training dataset) in the target user's video list are recalled in the ranking list. In order to evaluate the performance of our approach, we compare with 1) the method that just learn a united topic space for users and celebrities and 2) non-personalized search. The performance assessment measure is F-score.

$$F - score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{5}$$

We empirically fix the hyperparameters according to the prior expectation about the data. The hyperparameter β controls the smoothing and sparsity of topic-word distribution. Small β encourages more words to have high probability in each topic. Enlightened by this, we empirically choose a relatively small value of $\beta = 0.1$. Similarly, α is fixed as 0.25.

[1] `http://www.forbes.com/wealth/celebrities`

4.2 Experiment Results

Parts of the discovered latent topics of interest and popularity spaces are displayed in Table 1, which confirms our hypothesis that the latent topics of user and celebrity could be well extracted via LDA. Besides, through the monitoring of the user and celebrity data, it is found that the learnt interest and popularity distributions for user and celebrity are well tallied with their profiles. Take Beckham as an example, his probability distributions on Topic 1, 4 and 7 are 0.49, 0.25 and 0.10 respectively, which correspond to the fields of sports, entertainment and daily-life, and is consistent with the actual situation. The key words of each topic are shown in Table 1.

Table 1. Parts of the Latent Topics

No	Interest Space	No	Popularity Space
2	movie disney film story	1	season game team sports
3	education book central	4	interview perform romance
5	music cover rock	7	family life dating ceremony
6	interview season party	9	film role series character
8	comedy funny humour	11	album released song awards

The comparison of average F-score at different depths is illustrated in Fig. 2 (the number of latent topics and the weight of random walk is tuned to its optimal value.). We can see that our approach outperforms other methods consistently. Additionally the influence of random walk is shown in Fig. 3. In this figure, the number of latent topics in interest space and popularity is fixed as 30, and the weight of random walk λ is within the range from 0 to 1 with the interval

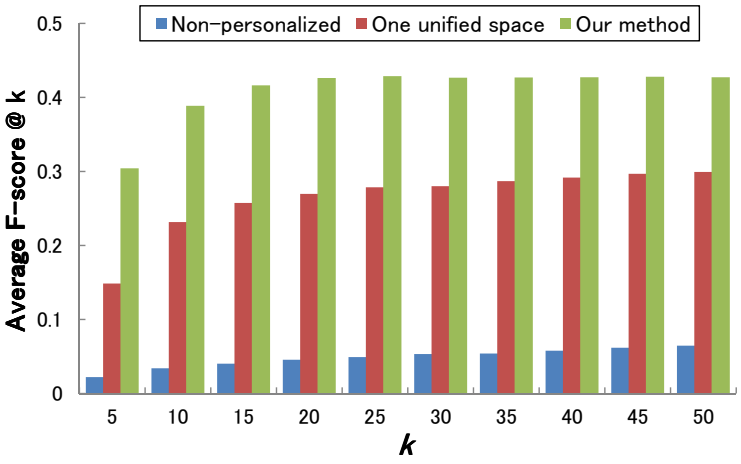

Fig. 2. Different approaches comparing with F-score.

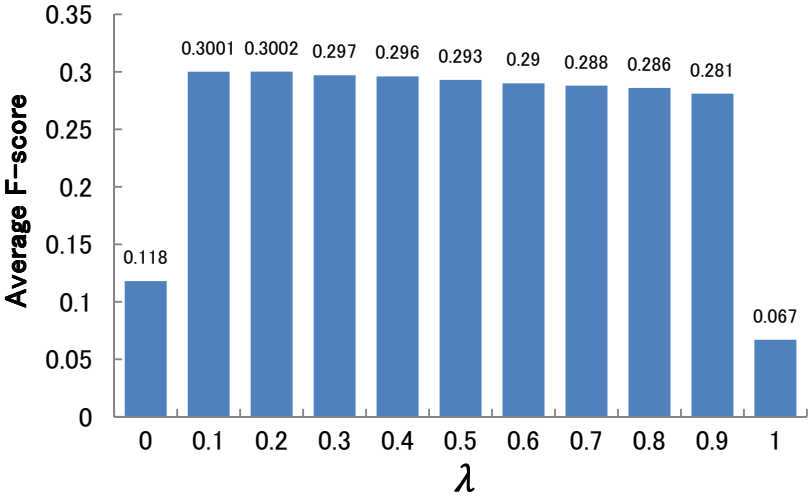

Fig. 3. The influence of random walk. The average F-score of top-50.

of 0.1. $\lambda = 0$ (see Eq. 3) means that there is no random walk applied; whereas $\lambda = 1$ means that the effect of initial value is ignored. It is shown from the figure that basically setting λ in (0.1, 0.9) is better than setting λ to 0 or 1. The former case means that the two spaces are linked directly and the connection would be minimized. The latter case means that the random walk takes fully effect, which makes the word distribution for each topic become identical, so the result is the least satisfactory. The optimal performance is achieved at around $\lambda = 0.2$. This result confirms the significance of random walk in updating the distribution of topic-word. On the other hand, the statistics on the tags of user and celebrity spaces shows an overlapping rate of around 70% (see Table 2), which in turn confirms the necessity to adopt random walk.

Table 2. Statistic of Semantic Words

-	Interest Space	Popularity Space	Total
No. of words	5602	5524	7980
% in total	70.2%	69.2%	-

5 Conclusions

In this paper, we have presented a cross-space mining method to exploit the correlation between user preferences and celebrity queries. Personalized video search is conducted by matching the user's interest distribution with distributed celebrity popularizes. Promising experimental results have demonstrated the effectiveness of our approach. In the future work, we will develop our current studies in the following direction. (1) Utilize visual information to facilitate the

learning from interest space or popularity space. (2) Instead of returning a ranking list, we will try to visualize the search results into semantically consistent groups. (3) We will investigate the problem of personalized query understanding in more general personalized search applications.

Acknowledgment. This work was supported in part by National Program on Key Basic Research Project (973 Program, Project No. 2012CB316304) and the National Natural Science Foundation of China (Grant No.90920303, 61003161).

References

1. Blei, D., Ng, A., Jordan, M.: Latent dirichlet allocation. The Journal of Machine Learning Research 3, 993–1022 (2003)
2. Evans, A., Fernández, M., Vallet, D., Castells, P.: Adaptive multimedia access: from user needs to semantic personalisation. In: Proceedings of the 2006 IEEE International Symposium on Circuits and Systems, ISCAS 2006, 4 p. IEEE (2006)
3. Gauch, S., Chaffee, J., Pretschner, A.: Ontology-based personalized search and browsing. Web Intelligence and Agent Systems 1(3/4), 219–234 (2003)
4. Hsu, W., Kennedy, L., Chang, S.: Video search reranking through random walk over document-level context graph. In: Proceedings of the 15th International Conference on Multimedia, pp. 971–980. ACM (2007)
5. Kim, H., Chan, P.: Learning implicit user interest hierarchy for context in personalization. In: Proceedings of the 8th International Conference on Intelligent User Interfaces, pp. 101–108. ACM (2003)
6. Leung, K., Lee, D.: Deriving concept-based user profiles from search engine logs. IEEE Transactions on Knowledge and Data Engineering 22(7), 969–982 (2010)
7. Liu, D., Hua, X., Yang, L., Wang, M., Zhang, H.: Tag ranking. In: Proceedings of the 18th International Conference on World Wide Web, pp. 351–360. ACM (2009)
8. Liu, F., Yu, C., Meng, W.: Personalized web search for improving retrieval effectiveness. IEEE Transactions on Knowledge and Data Engineering 16(1), 28–40 (2004)
9. Ma, Z., Pant, G., Sheng, O.: Interest-based personalized search. ACM Transactions on Information Systems (TOIS) 25(1), 5 (2007)
10. Shepitsen, A., Gemmell, J., Mobasher, B., Burke, R.: Personalized recommendation in social tagging systems using hierarchical clustering. In: Proceedings of the 2008 ACM Conference on Recommender Systems, pp. 259–266. ACM (2008)
11. Xu, S., Bao, S., Fei, B., Su, Z., Yu, Y.: Exploring folksonomy for personalized search. In: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 155–162. ACM (2008)
12. Zhou, X., Wu, S., Li, Y., Xu, Y., Lau, R., Bruza, P.: Utilizing search intent in topic ontology-based user profile for web mining. In: IEEE/WIC/ACM International Conference on Web Intelligence, WI 2006, pp. 558–564. IEEE (2006)