Tag Propagation Based Recommendation across Diverse Social Media

Deqing Yang[§][*], Yanghua Xiao[§], Yangqiu Song[‡], Junjun Zhang[§], Kezun Zhang[§], Wei Wang[§]
[§]{yangdeqing, shawyh, jjzhang, kzzhang, weiwang1}@fudan.edu.cn [‡]yqsong@gmail.com
[§]School of Computer Science, Shanghai Key Laboratory of Data Science, Fudan University, Shanghai, China
[‡]Department of Computer Science, UIUC, IL, USA

ABSTRACT

Many real applications demand accurate cross-domain recommendation, e.g., recommending a Weibo (the largest Chinese Twitter) user with the products in an e-commerce Web site. Since many social media have rich tags on both items or users, *tag-based profiling* became popular for recommendation. However, most previous recommendation approaches have low effectiveness in handling sparse data or matching tags from different social media. Addressing these problems, we first propose an optimized *local tag propagation algorithm* to generate tags for profiling Weibo users and then use a *Chinese knowledge graph* accompanied by an improved ESA (explicit semantic analysis) for semantic matching of cross-domain tags. Empirical comparisons to the state-of-the-art approaches justify the efficiency and effectiveness of our approaches.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: On-line Information Services-Commercial services; H.2 [**Database Management**]: Database applications-Data mining

Keywords

tag propagation; cross-domain; Chinese knowledge graph; Weibo

1. INTRODUCTION

Recently, *tag-based profiling* has became popular since tags are widely available in many social media, e.g., Flickr, YouTube, Delicous and etc. In these media, users can freely use various tags to label themselves or items (images, videos and URLs). These tags are good candidate profiles of users and items, thus navigate recommender systems to decide which items should be recommended to the users. In this paper, we focus on *tag-based user profiling for cross-domain recommendation*. Specifically, our solution overcomes two obvious shortcomings of previous tag-based recommender systems.

1. Data sparsity. Most prevalent recommendation approaches, such as collaborative filtering (CF) will fail when users have no enough tagging activities due to cold start. By our statistics, nearly 45% of Weibo users have no tags resulting in the loss of tag-based profile and the invalidation of CF. To solve such data sparsity, we propose a *local tag propagation algorithm* (LTPA in short) to label users without tags.

2. Isolated Cross-domain Recommendation. Most existing systems only recommend a user with the items from the same domain [5, 2]. And some works on cross-domain recommendation [4] supposed that user and item are fully involved in each domain. Different to them, we focus a more challenging problem of cross-domain recommendation, i.e, recommending the items in domain B to the users in domain A. A and B isolatedly have only user and item information, respectively. For such *isolated cross-domain recommendation*, the tags of different domains are often hard to be matched. To solve this problem, we utilize a *Chinese knowledge graph* (CKG in short) consisting of millions of linked entities/concepts and further propose an improved *explicit semantic analysis* (ESA in short) [1] algorithm to semantically correlate two syntactically different tags.

2. TAG PROPAGATION ALGORITHM

According to Weibo settings, each user can only tag him/erself instead of others. Although nearly half of Weibo users have no tags, we have verified that homophily takes effects among those users with tags through empirical studies, i.e., a Weibo user tends to share more similar tags with his/er friends than others. To employ our algorithm, we first construct a Weibo graph, namely G, of which each vertex is a user and each directed edge points to a user from his/er followee. Moreover, we assign an *influence weight* to each edge to quantify the extent that a followee can propagate his/er preference on some tags to his/er followers. In our experiments, retweet frequency is set as the influence weight. The algorithm's basic idea is, *a user's preference on a certain tag is propagated from others who can influence him/er and all users' tag preferences are updated iteratively during the propagation.*

Given a target user u, our algorithm tries to iteratively evaluate u's tag score vector $\vec{\mathcal{R}}$ specifying u's preference on all candidate tags. Suppose G has overall M users of whom there are L users using N tags, we define a $M \times N$ matrix R of which each row is a user's $\vec{\mathcal{R}}$. In addition, we import a $M \times M$ influence weight matrix F of which the entry F_{uv} is the influence weight from v to u, F_{uv}=0 if v is not u's followee. Then, the R in round t is computed as $R^t = FR^{t-1} = F \cdot FR^{t-2} = ... = F^tR^0 = F^tT = \begin{bmatrix} F^t_{11} & \cdots & F^t_{1M} \\ \cdots & \cdots & \cdots \\ F^t_{M1} & \cdots & F^t_{MM} \end{bmatrix} \times \begin{bmatrix} \vec{\mathcal{T}}_1 \\ \vdots \\ \vec{\mathcal{T}}_L \\ \mathbf{O} \end{bmatrix}$ where $\mathbf{O}$ is a $(M-L) \times N$ zero matrix and $F^t_{ij}(1 \le i, j \le M)$ represents an entry of F^t. $\vec{\mathcal{T}}_i(1 \le i \le L)$ is the tag distribution of user i. Thus, the tag score vector of u after t rounds equals to

$$\vec{\mathcal{R}}^t_u = \sum_{i=1}^{L} F^t_{ui}\vec{\mathcal{T}}_i \qquad (1)$$

[*]Correspondence author. This work was supported by National NSFC(No.61003001,61170006,61171132,61033010); Shanghai STCF No.13511505302; NSF of Jiangsu Prov. No. BK2010280.

WWW'14 Companion, April 7–11, 2014, Seoul, Korea.
ACM 978-1-4503-2745-9/14/04.
http://dx.doi.org/10.1145/2567948.2577285.

Obviously, the computation cost of $\vec{\mathcal{R}}$ is up to F^t whose cost is $O(LtM^3)$ which is unaffordable on large influence graphs. Next, we introduce how to reduce the computation cost.

In order to reduce computation cost, we propose the definition of *social influence* about tag propagation. Given a path p=$u_0, u_1, ..., u_t$ in G, we define the social influence along p from u_0 to u_t as $si(p)$=$\prod_{i=0}^{t-1} \frac{w_{u_i u_{i+1}}}{\sum_{u:u \to u_{i+1}} w_{u u_{i+1}}}$ where $w_{u_i u_{i+1}}$ is the influence weight of edge $u_i \to u_{i+1}$. Let $P_t(v,u)$ be the set of all paths of length t from v to u, thus the social influence of v on u at radius t is $si_t(v,u)$= $\sum_{p \in P_t(v,u)} si(p)$. Furthermore, we define $si_0(u,u)$=1 and $si_0(v,u)$=0 if $v \neq u$. Then, the total social influence of v on u is $si(v,u)$=$\sum_{t=0}^{\infty} si_t(v,u)$. Accordingly, we recognize that $F_{u,v}$=$si_1(v,u)$ if v is u's in-neighbor. Moreover, we can prove that $si_t(v,u)$=$F^t_{u,v}$ by induction on t. Inspired by the findings that more than 95% of information diffusion in Twitter network is less than the scope of 2 hops from the origin [3], we can restrict t $\leq$2 when computing Eq. 1. It means that our algorithm seeks u's profile tags within a local scope which is named as *local tag propagation algorithm* (LTPA in short).

3. SEMANTIC MATCHING

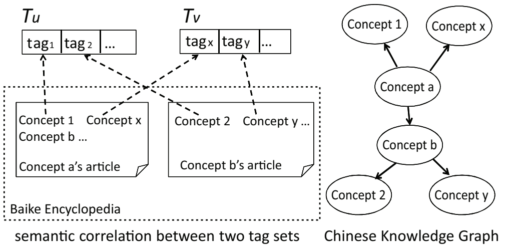
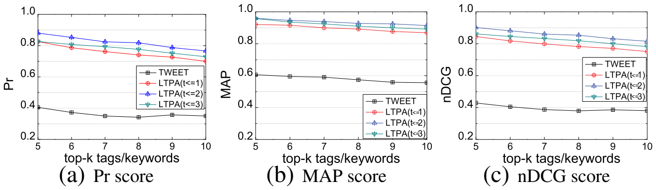


Figure 1: Tags correspond to CKG Concepts by dashed arrows and solid arrows are hyperlinks between concepts.

We found that the tags generated in different social media are often hard to be matched blocking tag-based recommendation between users and items across different domains. For example, most tags of Weibo users are different to Douban movies' tags. To settle this problem, we resort to a CKG consisting of billions of concepts, to semantically correlate two syntactically different tags. For each concept, there exists an article page in Chinese online encyclopedia to describe it. As well, there are many hyperlinks to other concepts on each article page. In this paper, we use an improved ESA model to evaluate semantic relatedness between two tag sets from different social media. Specifically, two terms are considered semantic related if they co-occur in the same article page. The more such article pages can be found, the more semantically related the two terms are. For example, in Fig. 1 tag set T_u has tag_1 and tag_2, T_v has tag_x and tag_y, and we find an article page of Concept a from CKG where tag_1 and tag_x co-occur in Concept a's article. Such article is an evidence of indirect semantic relatedness between the two tags. So does Concept b. Therefore, although T_u and T_v share no tags, they are still related in the context of semantics. Formally, the semantic interpretation of a tag set T can be represented by a *concept vector* defined as $\vec{C}_T = [c_1, ..., c_C] \in \mathbb{R}^C$. C is the total number of articles/concepts in CKG and each $c_j (1 \leq j \leq C)$ is the accumulative semantic relevance of concept j to all tags in T. For a concept j, its semantic relevance to a tag t is $c_{j,t} = s_j(t) \cdot p_i/|Neg(j)|$. $s_j(t)$ is the TF-IDF score of tag t in concept j's article and $|Neg(j)|$ is the number of distinct reference concepts in concept j's article, and p_i quantifies the extent that tag i profiles the user or item. p_i can equal to the tag score of LTPA. The factor $\frac{1}{|Neg(j)|}$ used to damp $c_{j,t}$ distinguishes our model from generic ESA.

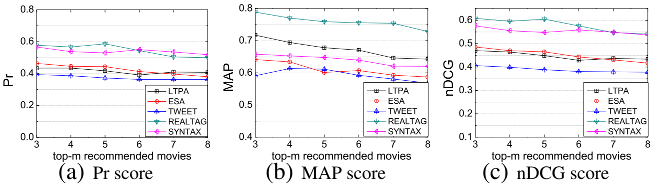
Table 1: Baselines.

Notation	Description
TWEET	profile a user by the keywords of his/er tweets
REALTAG	profile a user with his/er real tags instead of LTPA generating
ESA	use generic ESA to match two tag sets
SYNTAX	match two tag sets based on common tags instead of ESA



(a) Pr score (b) MAP score (c) nDCG score

Figure 2: Volunteer survey results on top-k profile tags/keywords.



(a) Pr score (b) MAP score (c) nDCG score

Figure 3: Volunteer survey results on top-m recommended movies.

4. EVALUATION AND CONCLUSION

In order to evaluate the effects of LTPA and CKG-based semantic matching, we surveyed 41 volunteers' acceptance rates to the top-m Douban movies that were recommended to them by our approaches. In our experiments, each volunteer has Weibo account and was profiled by top-20 tags generated by LTPA and the movies' tags were directly borrowed from Douban. Table 1 lists some compared baselines. We first compared LTPA's performance of tag profiling under different ts (tag propagation radius) and top-k tags/keywords and depicted the results in Fig. 2(a)∼(c) which display that LTPA outperforms TWEET and performs best when $t \leq 2$. Then we also used the competitors to recommend top-m movies from over 4,000 candidates to the volunteers. Fig. 3(a)∼(c) list the volunteers' feedbacks about the recommended movies when setting different ms. REALTAG almost has the best performance under all metrics, but it can not be applied for all Weibo users since nearly 48% of Weibo users have no real tags. Our approaches, namely LTPA, perform very close to ESA in Pr and nDCG but beats ESA in MAP justifying the import of damping factor of $\frac{1}{|Neg(j)|}$. We can not neglect that SYNTAX only worked on 85% of the volunteers even though it wins Pr and nDCG.

In this paper, we propose a novel cross-domain recommendation including LTPA-based tag profiling and CKG-based tag semantic matching. Experiments justify our approaches' effectiveness and the superiority over the state-of-the-art competitors in the context of tag-based recommendation across different social media.

5. REFERENCES

[1] E. Gabrilovich and S. Markovitch. Can movies and books collaborate? cross-domain collaborative filtering for sparsity reduction. In *Proc. of IJCAI*, 2009.

[2] C.-C. Hung, Y.-C. Huang, J. Y. jen Hsu, and D. K.-C. Wu. Tag-based user profiling for social media recommendation. In *Proc. of AAAI*, 2008.

[3] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *Proc. of WWW*, 2010.

[4] B. Li, Q. Yang, and X. Xue. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proc. of IJCAI*, 2007.

[5] S. Sen, J. Vig, and J. Riedl. Tagommenders: Connecting users to items through tags. In *Proc. of WWW*, 2009.