Microblogging Hyperlink Recommendation with Tensor-based Clustering

Lei Wang[1], Dehong Gao[2], Chunlei Chen[1]

[1]5147 Dongfeng Street, Weifang University, Weifang, China, wanglpqpq@gmail.com
[2]B2B-1688, Alibaba Group, Hangzhou, China, Gaodehong_polyu@163.com

Abstract — **Micro-blogging service, one of the most popular Web2.0 services, collects millions of user status every day. Due to the length limitation, users usually need to explore other ways to enrich the content of their micro-blogs. Some studies have provided findings to suggest that users can benefit from added hyperlinks in their micro-blogs. In this paper, we focus on the hyperlinks in micro-blogs and propose a new application, called hyperlink recommendation. We expect that the recommended hyperlinks can be used to enrich the information of user micro-blog. A three-order tensor is used to model the user-'micro-blog'-hyperlink relations. Two tensor-based clustering approaches, tensor decomposition-based clustering (TDC) and tensor approximation-based clustering (TAC) are applied to group the users, micro-blog and hyperlinks with similar interests, or similar use contexts. Recommendation is then made based on the reconstructed tensor using cluster information. The evaluation results in terms of Mean Absolute Error (MAE) shows the advantages of both the TDC and TAC approaches over a baseline recommendation approach, i.e., memory-based collaborative filtering. Comparatively, TAC approach achieves better performance than TDC approach.**

Keywords - Micro-blogging, Hyperlink recommendation, Tensor-based clustering

I. INTRODUCTION

The growth of the Web 2.0 technologies has led an explosion of social network media sites in industry, among which micro-blogging services are undoubtedly one of the most popular ones now. For example, Twitter (*www.twitter.com*) and Weibo (*www.weibo.com*) collect billions of short messages every day. In academia, micro-blogging related researches already become a new research hotspot, and the research topics range from the micro-blogs' content to the micro-blogging social networks. However, one difficulty in these researches is that the tiny micro-blog is too short to delivery rich information. For instance, the length of a tweet is limited to 140 characters only.

The significances of hyperlink are miscellaneous. For the micro-blogging service itself, the hyperlink can enrich the information of the micro-blog content to attract more users. Reading linked Web pages will also prolong the amount of time that a user remains in the micro-blogging websites. The stickiness is important from the Search Engine Optimization (SEO) perspective. The longer a visitor stays, the more valuable the micro-blog websites will be. In contrast, for every single micro-blogging user, according Microsoft' work [22], the tweets with hyperlinks can improve the tweets' reliability. And it gives the tweets more chances to be retweeted. To a certain degree, it can be viewed as one method for a user to achieve recognition and accumulate reputation. To comprehend the **quantity** and the **purpose** of the hyperlinked contents, a set of the hyperlinks (included in the downloaded tweets and weibos) are manually annotated. The statistics shows that 84 percent of the hyperlinks bring relevant and helpful information to the topics and 67 percent of the relevant hyperlinks provide the exact information source. Therefore, it is reasonable to conclude that there is considerable amount of high quality hyperlinks in the micro-blogging network and these high quality hyperlinks play an important role.

Taking into consideration the above hyperlinks' values, we design a new micro-blogging scheme, called *hyperlink recommendation* [8], and this scheme can recommend hyperlinks to users with the purpose of helping them to enrich the content of their micro-blogs. Existing recommendation approaches fall into two primary categories, i.e., content filtering and collaborative filtering. Content filtering manages to identify whether the content of an item (i.e., what to be recommended) matches the profile of a user (i.e., whom to recommend to). Unlike content filtering, collaborative filtering predicates the item that would be interesting to a user by discovering the commonality among the users based on their previous correlated behaviors. Considering the hyperlinked Web pages are full of non-text images, videos and audios, we choose the Collaborative Filtering (CF) framework to approach the problem of hyperlink recommendation. With this framework, the most straightforward would examine the user-hyperlink relations and define it as a correlated behavior when two users cite the same hyperlink. However, it will lose very important information, i.e., micro-blog, the context where a hyperlink is embedded in.

Micro-blog can more accurately bridge user and hyperlink. For example, a user may post two distinct tweets but cite the same hyperlink. While other users indirectly cite (reply or retweet in Twitter) the hyperlink through different tweets. Without tweets we cannot see the difference. Moreover, the relation between hyperlinks can be estimated by micro-blogs. For example, in certain situations, one tweet may cite two hyperlinks, which implies that the two hyperlinks are highly correlated. Thus in hyperlink recommendation, we consider three different types of items, user, micro-blog and hyperlink. To make recommendation with this three-way relation, one idea is to separate the relation to three sub-relations, like user-micro-blog, user-hyperlink and micro-blog-hyperlink and then heuristically combine the results on each pairwise relation. However this will change the original meaning of relations. To explicitly model these inherently inter-related relations, we leverage the three-order collaborative *Tensor*. Whereas the sparsity of collaborative tensor puts forwards a challenging issue to three-way hyperlink recommendation. Considering the disadvantages of memory-based CF methods in terms of resolving sparsity problems, we investigate on the model-based CF approaches in this work and further develop existing matrix-based clustering to tensor-based clustering. Two algorirthms, i.e., *tensor decomposition-based clustering* (TDC) and *tensor approximation-based clustering* (TAC) are adopted to handle this problem. TDC applies spectrum clustering to tensor Tucker decomposed matrixes, and TAC is dedicated to approximate the tensor using cluster information with best efforts. The tensors are reconstructed based on the generated clusters, and recommendation is then made relying on these tensors. The proposed approaches are evaluated with regards to Mean Absolute Error (MAE). Compared to memory-based CF approaches, the experimental results indicate the advantages of both TAC and TDC. And the former slightly outperforms the latter.

The contributions of this work are threefold:

1. There have been some studies on micro-blogging hyperlinks. But as far as we know, there is not existing work that concentrates exclusively on hyperlinks. We do statistics on millions of tweets and weibos, and manually annotate certain amount of hyperlinks. The statistical analysis results drive us to further explore the significance and value of hyperlinks.
2. We propose a novel micro-blog application, i.e. hyperlink recommendation. And this application aims to help users to enrich the content of their micro-blogs by including the recommended hyperlinks. We insight into this problem and design a three-order tensor to model the relations among users, micro-blog and hyperlinks.
3. We apply and experiment with two solutions, *tensor decomposition-based clustering* (TDC) and *tensor approximation-based clustering* (TAC), to the three-order tensor-based recommendation problems and evaluate the effectiveness and efficiency of them.

The rest of this paper is organized as follows. Section 2 introduces the previous work on collaborative filtering and recommendation in micro-blog. Section 3 presents our preliminary statistical analysis that motivates the hyperlink recommendation application. Section 4 describes the proposed approaches including collaborative tensor construction, three-order tensor clustering approaches and tensor-based recommendation strategy. Section 5 compares different approaches and discusses the experimental results. Finally, Section 6 concludes the paper.

II. RELATED WORK

A. Collaborative filtering

Collaborative filtering is one of the most successful methods to build recommendation systems. The fundamental assumption is if two persons have rated n items similarly, or have similar taste like forwarding, endorsement, tagging, therefore they will rate or act on the other similarly [31], as 'like minds behavior alike'.

GroupLens [28] is one of the standard collaborative filtering systems in the early age, which aims to enable people to identify their interesting articles from the flood of articles. The rating servers forecast scores according to users' previous rating information. Similar works can be found in [9, 20, 21]. These methods are usually called memory-based collaborative filtering algorithms since they use the entire or a sample of user-item rating information to generate the prediction [31]. The prevalent procedures of memory-based CF algorithm are: 1) calculating the similarity based on the rating correlation between two users or items; 2) producing a prediction for the active user by taking the average of all the ratings of users or items [29]. It has been adopted by some remarkable commercial recommendation systems like Amazon[1], due to its easy implementation and some other advantages [16]. However several shortcomings limit the memory-based CF methods. For example, 1) usually the correlated data are highly sparse, which means the common items seldom appear. This makes the

[1]http://www.amazon.com

DOI 10.5013/IJSSST.a.17.28.20

ISSN: 1473-804x online, 1473-8031 print

computation of similarity unreliable [31]. 2) With the coming of big data, the increase of data size usually makes memory-based CF method not scalable for massive data. To overcome these shortcomings, model-based CF methods are investigated. Ungar and Forster [35] adopts the clustering method to solve movies recommendation problems. Based on rating matrix, users and movies are clustered separately, and prediction is made within the user and movie clusters. O'Conner and Herlocker [24] cluster the rating data and leverage memory-based CF method to perform predictions within every cluster. Chee et al. [5] propose the *RecTree* method to improve scalability using a divide-and-conquer approach. Their method executes an efficient K-means-like clustering iteratively. The prediction is made within an advisory cluster that the active user is affiliated to.

All above works just consider the two-dimensional recommendation, which means that they focus on the user-item rating. However, with the requirements of personalized recommendation, usually the recommendation system needs to consider the three-way nature of information. Thus tensor-based recommendation systems appear in these years. Symeonidis et al. [32] apply three-order tensor to the social tagging recommendation problem. Tensor is constructed based on user, item and tag relation, and then the HOSVD technique is used to reduce the dimensions. Recommendation is produced by re-construction of the tensor. Tensor is used by Zhang et al. in Flickr group recommendation [37]. CANDECOMP/PARAFAC (CP) tensor decomposition is employed to capture the underlying pattern in the tensor. Recommendation is made by user and group associated score matrix. Zheng et al. [38] make use of a tensor representation to model user-location-activity. They put forward a regularized tensor and matrix decomposition method to solve the sparse data problem. The real-world test shows it can outperform several state-of-the-art solutions.

B. Recommendation in Micro-blogging

As an information platform, the micro-blogging service not only wins a broad massive of users' favor, but draws a considerable interest in academia. These researches involve information delivery [14], real-time event detection [19] and micro-blog related summarization [23, 30]. At present the related recommendation in micro-blog can be categorized into two types: real-time news recommendation and user recommendation.

For news recommendation, researchers usually use the real-time property of micro-blog because it can collect the freshest micro-blog about news topics. For example, Phelan et al. [25, 26] describe a new method to recommend news with the help of Twitter micro-blogging. Later this work is extended. They implement one system, called Buzzer [26], to rank personal RSS subscriptions, and to mine trending terms from public Twitter timelines and user's own friend subscriptions. Gao et al. [7] carry out their research on the interplay between personal preferences and public tendency. They design a framework to enrich the semantics of tweets and trending modeling. They find that by combining user and trend profiles they can further improve recommendation quality. For user recommendation, research mainly focuses on the followee recommendation, which means for an active user to recommend other users to follow. To this end, Chen et al. [6] evaluate four algorithms to expand the user friends list. The social-network- information-based algorithm can produce better-received recommendations. Hannon et al. [9] exploit recommendation source from online Twitter's information . They explore several techniques, including content-based recommendation, which rely on the content of tweets, collaborative filtering style approaches (based on the follower and followee relation), and a number of hybrid strategies. Koga and Taniguchi [10] hypothesize that users had latent connections if the KL divergence distance is short. They develop a user recommendation engine which extracts latent topics of users based on followings, lists, mentions and RTs. Finally a questionnaire survey is done to evaluate the effectiveness of algorithm.

III. HYPERLINK RECOMMENDATION

Ahead of designing the hyperlink recommendation approaches, the very beginning question we ask ourselves is whether enough high-quality hyperlinks are available in micro-blogs. Analysis is carried on Twitter and Weibo, two most popular micro-blogging services in America and in China. We show some statistics below to interpret what motivates us to investigate on hyperlink recommendation.

A. Statistical Analysis of Micro-blog Hyperlinks

Using the public Twitter APIs[2], we tracked all the Twitter official trending topics within the time span from March, 2nd to June, 4th in 2011. And we also downloaded the related tweets in real-time. Totally, 13,211,258 valid tweets for 2,113 trending topics were obtained. We divided the topics into categories according to *whatthetrend*[3]. Similarly, we use the public Weibo APIs[4] to track all the weibos trending topics in the period from

[2] dev.twitter.com
[3] http://www.whatthetrend.com/
[4] http://open.t.sina.com.cn/

Sept, 1st to Nov, 1st in 2014 and the related weibos to the trending topics. Totally, 3,323,221 weibos for 1,212 trending topics were downloaded. These trending topics are already categories according to Weibo official site.

With these categorized tweets and weibos at hand, we first check the hyperlinks' **quantity** . As demonstrated in Fig. (1), the overall ratios of hyperlinks in Twitter and Weibo are 26.2% and 22.4%, respectively. We also noticed that 58.4% of hyperlinks are present at the end of tweets, which seems to serve as additional remarks or explanations. These findings tell that there are numerous hyperlinks for our research. More than that, users sometimes have to add a hyperlink to a micro-blog to make their messages more complete in some topic categories.

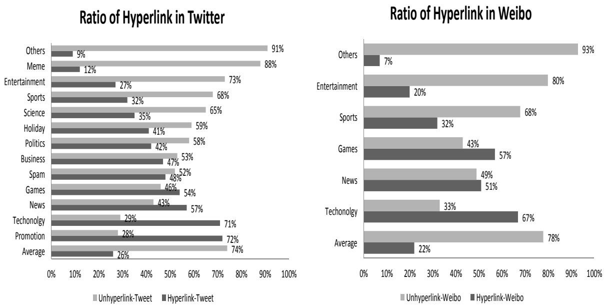

Fig.1 Ratios of Hyperlinks in Twitter and Weibo

Besides we investigate what information can be brought by relevant hyperlinked webpages. The information brought in by hyperlinks is categorized as *information source* (*Src Info*), *background information* (*Bkg Info*), *comment*, *support information* (*Spt Info*) and others. These categories can direct researchers to explore different applications with hyperlinks, e.g. information source web pages can help in trending topic summarization while comment webpages tend to be important in opinion mining. 67% and 72% of related hyperlinks convey information sources about the trending topics in Twitter and Weibo.

B. Hyperlink Recommendation Modeling

Considering aforementioned factors and hyperlinks' potential values , we design a novel application, i.e. hyperlink recommendation. And this application is dedicated to recommending hyperlinks to users such that the content of their micro-blogs is enriched. In addition, we expect that adding proper hyperlink can improve the credibility of micro-blogs, and makes the micro-blogs more likely to be forwarded. This will further more improve the popularity of the user when more and more other users read his/her micro-blog [22]. In hyperlink recommendation, besides the user and hyperlink, another important object is the user's micro-blogs. Micro-blogs can be regarded as a bridge from user to hyperlink. Through micro-blogs, user and hyperlink can be connected accurately and understandably. To model this user-'micro-blog'-hyperlink relation, a three-order collaborative tensor is necessary. Each cell in tensor represents a citation from a user to a hyperlink in his/her micro-blog. We designate two kinds of citation: direct citations and indirect citations. The direct citation refers to the case where a user first introduces a hyperlink in his/her micro-blog. The indirect citation refers to the case where a user forwards/replies/mentions a micro-blog that contains a hyperlink. These behaviors are indicators of endorsement to the micro-blog and we assume they also want to show endorsement to the hyperlink. So we regard both the users who first introduce the hyperlinks in their micro-blogs and the users who indirectly cite the hyperlinks as citing hyperlinks.

A similar and even worse problem with a three-order tensor is that it is much sparser than a two-way matrix. The memory-based CF approach is weak in handling this problem because it estimates similarity according to the correlated items. In our sparse tensor, the co-cited hyperlinks between two users are really few, which make the similarity of users unreliable [31]. Hence, we are favor of using the model-based CF approaches in order to alleviate this problem. It has been claimed that the free-following strategy and the phenomenon of homophily cause users prone to form groups easily in the micro-blog network [36]. This motivates us that to incorporate clustering techniques in model-based CF approaches. This allows us to identify the groups of users and hyperlinks and make recommendation according to the group information. The advantage of this model-based CF method is it can handle the sparsity better than memory based CF methods. This also helps to improve scalability with large data sets. The clusters of users can help to identify the taste of the group users. The clusters of hyperlinks can help to improve the diversity of recommendation, which can further improve the user experience.

Hyperlink recommendation is executed in three pipelined processes: (1) user-'micro-blog'-hyperlink tensor is constructed based on citation and forwarding behaviors, and the tensor acts as the collaborative rating; (2) users, micro-blogs and hyperlinks are clustered according to the tensor, and this process is the essential step to accomplish recommendation; (3) the tensor is re-construct based on the clustering results and the hyperlink is recommended to a potential user who might be of interest.

IV. TENSOR CLUSTER CF RECOMMENDATION

In this section we will fully describe our tensor-based CF recommendation approach, which includes construction of the collaborative tensor, clustering based on tensor and generation of recommendation. Two approaches, *tensor decomposition-based clustering* (TDC) and *tensor approximation-based clustering* (TAC), are adopted in tensor clustering. The notations used in following paper is listed as follows: tensors are denoted by Lucida Calligraphic uppercase letters ; random variables and matrices in uppercase bold letters *U*, *M*, *H*; the element (*i*, *j*) of matrix and element (*i*, *j*, *k*) tensor by lowercase letters u_{ij}, x_{ijk}; The sets such as are enumerated as and the index *i* over the set {1,…, *n*} is denoted by .

A. Collaborative Tensor Construction

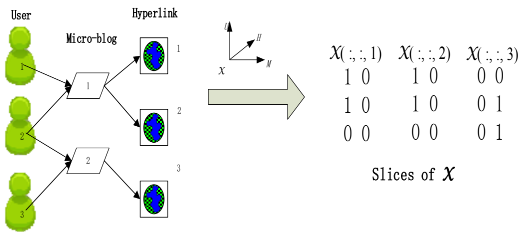

Fig.2 Tensor Construction: User-'Micro-blog'-Hyperlink

Tensor is an *N*-order array. We use a three-order tensor to model the user-'Micro-blog'-hyperlink relation in hyperlink recommendation. We initialize a three-order collaborative tensor, where *u*, *m*, *h* are the numbers of users, tweets and hyperlinks, respectively. Each element in the tensor denotes user *i* citing hyperlink *k* in his/her micro-blog *j*. The direct citation and indirect citation are used to construct the tensor, for example, if user 2 forwards micro-blog 1 of user 1, they both have one link to micro-blog 1. Fig. (2) illustrates one example of tensor construction and $\mathcal{X}(:,:,i)$ denotes the i^{th} slice of the tensor.

Unfolding (or *matricization*) is an important operation of tensor which transforms an *N*-order tensor into a matrix. A tensor can be arranged as a matrix , which is denoted by, as shown in Fig (3). The other background information of tensor is well illustrated in [13].

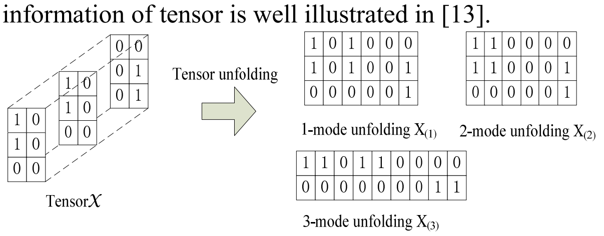

Fig.3 Illustration of three-mode Tensor Unfolding

B. Tensor Decomposition-based Clustering

To cluster this three-order tensor, one idea is to find ways to decompose the tensor into matrixes, and apply matrix cluster algorithms to cluster the decomposed matrixes. Two commonly used decomposition methods are *CANDECOMP/PARAFAC* [13] and *Tucker* decomposition [33]. Due to its easy implementation, Tucker decomposition is usually used in tensor decomposition [32], which can be applied to our tensor clustering as well.

1) Tensor Decomposition and HOSVD

The *SVD technique* is commonly adopted to factorize a matrix to into three matrixes (Equation (1)).

$$L_{I\times J} = U_{I\times I}\Sigma_{I\times J}V_{J\times J} \quad (1)$$

where is the left singular vector of *L* and is the transpose of matrix with the right singular vectors. is the diagonal matrix of singular values of *L*.

Like matrix factorization, in order to extract the inner relation of each mode (user, micro-blog, hyperlink), we attempt to factorize our collaborative tensor into a multiply of three orders, user order, micro-blog order and hyperlink order, as

$$\mathcal{X} \approx g \times_1 U \times_2 M \times_3 H = \sum \sum \sum g_{abc} u_a \otimes \quad (2)$$

are the orthogonal factor matrices and can be viewed as the principal components in each mode and the symbol "" represents the vector outer product. The core tensor shows the interaction of different ways. The *U*, *M*, *H* can be explained as the independent way matrixes, user, micro-blog, and hyperlink. Tucker [33] introduces one method to factorize the tensor x into a core tensor multiplied by a matrix along each way. Today, this method is better known as *High Order Singular Value Decomposition* (*HOSVD*) [13]. From Equation (2), the unfolding forms on each mode are

$$X_{(1)} \approx U g_{(1)} (H \otimes M)^T \quad (3)$$

$$X_{(2)} \approx M g_{(2)} (H \otimes U)^T \quad (4)$$

$$X_{(3)} \approx H g_{(3)} (M \otimes U)^T \quad (5)$$

where denotes the *n*-mode unfolding of Tensor . From Equations (3~5), the user, micro-blog, hyperlink way matrixes can be obtained from the left leading singular vectors of 1/2/3-mode unfolding of tensor . The HOSVD algorithm is described in Table I, and the visualization of HOSVD is illustrated in Fig. (4).

DOI 10.5013/IJSSST.a.17.28.20

ISSN: 1473-804x online, 1473-8031 print

TABLE I. TUCKER DECOMPOSITION OF COLLABORATIVE TENSOR

Algorithm of Tucker Decomposition(HOSVD)
Input: three-order tensor
Output: $\mathcal{g}$, U, M, H
1. Calculate the user-mode unfolding , 'micro-blog'-mode unfolding and hyperlink unfolding;
2. U, M, H are set as the P, Q, R leading left singular vectors of , and , respectively;
3. The core tensor is calculated as .
Return $\mathcal{g}$, U, M, H

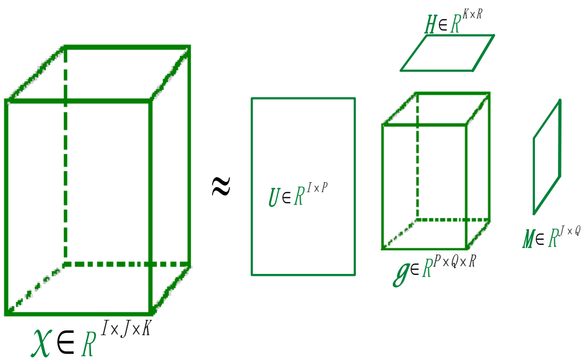

Fig.4 Illustration of Tucker Decomposition (three-order)

2) *Tensor-based Spectrum Clustering*

Spectral Clustering is one popular clustering algorithm, which is used to explore data analysis with many applications ranging from statistics, computer science to psychology. It makes use of the spectrum of data to perform dimensionality reduction for grouping the similar items.

According to Luxburg [18], the spectrum clustering process can be defined as:

$$\max trace(V^T S_N V)$$
$$s.t.\ V^T V = I$$

The spectral clustering can be re-defined as a Frobenius norm optimization when the normalized similarity matrix is positive semi-definite [17]:

$$\max \|V^T S_N V\|_F^2$$
$$s.t.\ V^T V = I$$

In terms of three-order clustering, in which the clustering involves different items, like user, micro-blog, hyperlink way, Liu [17] optimizes the three-order clustering by adding up individual objective functions as:

$$\max \sum_{i=1}^{n} \|V^T S_N^{(i)} V\|_F^2$$

$$s.t.\ V_i^T V_i = I$$

where **Error! Reference source not found.** is the i^{th}-way matrix. If we regard the decomposed matrixes (U, M and H) as different ways, and apply them to spectrum clustering, the optimization can be written as

$$\max \sum_{i=1}^{n} \|V^T S_N^{(i)} V\|_F^2$$

$$s.t.\ V_i^T V_i = I$$

This algorithm is described in Table II:

TABLE II. TENSOR DECOMPOSITION-BASED CLUSTERING(TDC)

Algorithm of Tensor Decomposition-based Clustering
Input: Decomposed matrix;
Output: clusters of users, micro-blog and hyperlinks.
1. Obtain the user, micro-blog, and hyperlink way matrixes, U, M and H with HOSVD algorithm;
2. Normalized Laplacian and , and degree matrix D is defined as; Similarly, calculate the and ;
3. Cluster the user matrix with the K-means algorithm into clusters ;Similarly cluster and to generate the micro-blog clusters and hyperlink clusters

In TDC, the clustering procedure just operates on the decomposed matrix, like user-mode decomposed matrix or hyperlink-mode decomposed matrix. The reason for only using certain mode decomposition is that the decomposed matrixes stem from the collaborative tensor, and are likely to contain the principal information of the user-'micro-blog'-hyperlink tensor. But we believe that if we can synchronize the clustering of different way items in the collaborative tensor, the clustering in one way can help that of the other ways. Take this into consideration, we utilize the three-order clustering framework by Banerjee et al. [1] in our clustering process.

C. Tensor Approximation-based Clustering

1) Bregman Divergence

Definition 1 Let be a real-valued strictly convex function defined on the convex set , the domain of , such that is differentiable on, the interior of . The *Bregman divergence* is defined as

$$d_\phi(x_1, x_2) = \phi(x_1) - \phi(x_2) - \langle (x_1 - x_2), \nabla\phi(x_2) \rangle$$

where $\nabla\phi$ is the gradient of function ϕ.

Given a Bregman divergence and a random variable, the uncertainty in the random variable can be captured in terms of a useful concept called Bregman Information [3], defined as below:

Definition 2 For any Bregman divergence and any random variable, the *Bregman information* of X is defined as the expected Bregman divergence to the expectation:

$$I_\phi(X) = E_\omega[d_\phi(X, E_\omega[X])]$$

Intuitively, this quantity measures the "information" of the random variable. The objective of the three-order tensor approximation-based clustering (TAC) can be reformulated as seeking an estimation that has the maximal Bregman information with regards to the original tensor. These definitions are used in the rest of this paper.

2) Three-order TAC Formulation

The collaborative tensoris constructed by multiple types of entities, user, micro-blog and hyperlink. We adopt three clustering functions which represent the mapping from the entities to their corresponding clusters, defined as user clustering, micro-blog clustering and hyperlink clustering .

Let and be random variables, and their definition domains are $\{1,\ldots,l\}$, $\{1,\ldots,t\}$and $\{1,\ldots,n\}$,respectively. Be be an estimation of the data tensor**Error! Reference source not found.**. Then the three-order TAC approach's objective function can be measured with regards to the expected distortion between, that is

$$E[d_\Phi(x,\hat{x})] = \sum \sum \sum \omega_{abc} d_\Phi(x_{abc}, \hat{x}_{abc}) = d_\Phi(x,\hat{x}) \qquad (6)$$

Where Φ_ω is a separable convex function induced on the tensor by the convex function . The three-order clustering aims to obtain the best (ρ, γ, δ) that minimizes this expected Bregman distortion.

The accuracy of three-order TAC approach is determined by how close the χ' can approximate the χ. The core processing of three-order TAC approach is the construction of the approximation χ' through generated clusters (ρ, γ, δ) and several summary statistics. The summary statistics store some statistical information, like cluster marginal, which are derived from the original collaborative tensor. Here we investigate the special case where the summary statistics stem from aggregating along the clusters, i.e. the summary statistics preserved are just the cluster means or equivalently the conditional expectation random variable.

With the cluster and summary statistics, Banerjee et al. propose to use a new *minimum Bregman Information* (MBI) *Principle* to find the best approximation [2]. For our problem that uses the cluster means as the summary statistics, MBI principle indicates that the best approximation equals [1].

3) Tensor-based Bregman Clustering

With the best approximation obtained from MBI, we then turn to the original Tensor three-order Clustering problem in Equation (5). According to [1], the objective function can be formulated as the loss in Bregman information between the approximation. and its original collaborative tensor, defined as:

$$E[d_\Phi(x,\hat{x})] = I_\Phi(x) - I_\Phi(\hat{x}) \qquad (7)$$

With Equation (7), the original Bregman clustering problem in Equation (6) can be posed to find the optimal set of clusters defined as

$$(\rho^*, \gamma^*, \delta^*) = \text{agrmin}\, E[d_\Phi(x,\hat{x})] = \text{argmin}[I_\Phi \qquad (8)$$

Given a user, micro-blog and hyperlink clusters, if and based on MBI principle the approximation can be obtained by .. Let be the collaborative tensor means, the optimal approximation corresponding to is given by. The decomposition of the objective function Equation (6) can be defined as

$$E[d_\Phi(x,\hat{x})] = \sum \sum \sum \omega_{abc} d_\Phi(x_{abc}, \mu_{\rho(a)\gamma(b)\delta(c)}) \qquad (9)$$

Equation (9) is the total of all the three-order, and with the purpose of decreasing the objective function we merely refresh the current selected way each time. For instance, for any user *a*, the contribution to the overall objective function is determined by its current assignment. Assuming we can express the objective function of Equation (9) as the total of the other two-way contributions in the form

$$J_u(\hat{x}) = \sum \sum \sum \sum \omega_{abc} d_\Phi(x_{abc}, \mu_{[ik]}) \qquad (10)$$

DOI 10.5013/IJSSST.a.17.28.20 ISSN: 1473-804x online, 1473-8031 print

The selection of user cluster assignment *i* accurately determines what set of *j*, *k* cluster means occur in Equation（10）. Hence the best possible choice for the new cluster assignment is to pick the value of *i* that has the minimum cost,

$$\rho^{new}(a) = \operatorname{argmin} J_u(i) = \operatorname{argmin} \sum \sum \sum \sum \omega_{abc} d_\phi(x_{abc}, \mu_{ijk}) \tag{11}$$

Since the new assignments effectively change the current approximation tensor to a new tensor which is just a user-permuted version of that achieves a lower cost, that is

$$E[d_\phi(x, \hat{x}_{\rho, \gamma, \delta})] \leq E[d_\phi(x, \hat{x})]$$

The decrease in the objective function value is due to the optimal greedy update in any way cluster assignments. Once all the way assignments have been updated, one needs to re-compute the new minimum Bregman solution corresponding to.

Table III indicates that the TAC algorithm begins with an arbitrary selection of clustering. In a single iteration, the user, micro-blog, and hyperlink clusters are refreshed in order to decrease the objective function Equation (9). After the assignments have been achieved, the cluster means are refreshed, and it further reduces the objective function. This process is iteratively performed until convergence. Since the objective decreases as the iterations proceed, and the objective is lower bounded, the algorithm can definitely converge to the objective's local minimum.

D. Tensor-basedRecommendation

Provided that the user, micro-blog and hyperlink clusters are produced by TDC or TAC in this section we make recommendation based on these clusters. In our three-order tensor clustering, we assume that summary statistics preserve the cluster means. Through the clusters, we can construct the approximation of the original tensor by. Thus, the tensor-based recommendation can be transformed to the problem of link prediction, which predicts the missing citation using the reconstructed tensor. For example, if we want to determine whether hyperlink *h* should be recommended to user *u*, the weight is calculated as:

$$\hat{x}_{uh} = \frac{\sum_{a:\rho(a)=u} \sum_{b} \sum_{c:\delta(c)=h} \omega_{abc} x_{abc}}{\sum \quad \sum \sum}$$

This weight is used to determine if hyperlink *h* should be recommended to user *u*.

From the constructed tensor, we could obtain the weight of micro-blog *m* and hyperlink *h*. This weight can be used to determine if we need to recommend hyperlink *h* to micro-blog *m*. But we think that users may have various interests and the micro-blog can only represent one of the interests. Recommending hyperlink to a single micro-blog can hardly improve the diversity of information that the user receives. In contrast, if we recommend hyperlinks from different clusters, we can make users receive more diversified information. This is much more important in social media network, like Twitter and Weibo [14, 16]. Thus it is another reason that we propose recommending hyperlink to users.

TABLE III. TENSOR APPROXIMATION-BASED CLUSTERING

Algorithm of Tensor Approximation-based Clustering
1. Input: Tensor , probability measure ω, Bregman Divergence d_ϕ , number of clusters *l*, *t*, *n*.
2. Output: three-order clustering that locally optimizes the objective function in Equation (9)
3. Algorithm Begin:
4. Initialize () with an arbitrary clustering
5. Repeat
6. {Step A: update cluster means}
7. For *i* = 1:*l* do
8. For *j* = 1: *t* do
9. For *k* =1:*n* do
10.
11. End for
12. End for
13. End for
14. {Step B: Update User Cluster}
15. For a = 1:u do
16. Update user clusters according to Equation (11)
17. End for
18. {Step C: Similar with step B, update Micro-blog and Hyperlink Cluster}
19. Until Convergence
20. Return ()

DOI 10.5013/IJSSST.a.17.28.20

ISSN: 1473-804x online, 1473-8031 print

V. EXPERIMENT

The dataset introduced in Section 2 is used in our experiments and for our research purpose, we select six trending topics and among these trending topics we select the tweets which have been forwarded more than two times, as well as hyperlinks and users related to the tweets. The reason for this is because in our tensor construction, the tweet is used to bridge the user and the hyperlink. The detailed information is shown in table VI. There are a total of 2377 citation relations in our experiment, among which 4/5 are randomly selected to generate the clusters and the rest are used for prediction.

The technical details of our approaches, TDC and TAC, are presented in Section 4. Here we will describe the detailed parameter settings in these two approaches and baseline approach.

TABLE IV. DETAIL ABOUT EXPERIMENTAL DATASET, WHERE CITATION MEANS THAT USER *I* CITES HYPERLINK *K* IN TWEET *J*

Trending topic	User	Tweet	Hyperlink	citation
Minsk	716	285	184	2094
Angry Birds Game Coming	427	135	63	732
Apple iPhones	456	185	76	798
TEPCO	335	112	101	687
Texas Motor Speedway	516	187	91	877
World Champions	411	279	120	845
Summing up	2270	1183	635	2377

TDC is achieved by applying the HOSVD technique to tensor and then using spectrum clustering to group users and hyperlinks. In equation 2, the *p*, *q* and *r* need to be set before HOSVD. Because the tensor is quit sparse, if we keep too few features, it will make many decomposed matrix with many zero rows. This may harm the tensor-based spectrum clustering later on. Thus in order to keep most information of the original matrix and mine the latent association we set *p*, *q* and *r* as:, and . In tensor-based spectrum clustering, we need to determine the cluster number before performing tensor-based spectrum clustering. The user, tweet and hyperlink cluster numbers are set to , and, where ε varies from 0.1 to 0.7.

TAC is to find the best approximation of the original tensor. Since the citation is set as binary, the ω is calculated as

$$\omega_{cited} = \frac{(\#citation)}{[illegible]}$$

$$\omega_{uncited} = 1 - \omega_{cited}$$

where the symbol # represents the number, for example, user# means the number of user. Bregman divergence is set square divergence, which is the special case of Bregman divergence. The squared loss is defined as , corresponding to the function. Similar to TDC, the user, tweet and hyperlink cluster numbers are set to, and, where ε varies from 0.1 to 0.7.

We regard the memory-based CF approach as the baseline, which calculates the user similarity matrix *Sim* using the *Pearson Correlation* by Equation (13).

$$Sim_{u,v} = \frac{\sum_{i \in I} (c_{u.i} - \bar{c}_u)(c_{v.i} - \bar{c}_v)}{\sqrt{\sum_{i \in I} (c_{u.i} - \bar{c}_u)^2} \sqrt{\sum_{i \in I} (c_{v.i} - \bar{c}_v)^2}} \quad (13)$$

where $c_{u,i}$ represents the citation of hyperlink *i* and user *u*. $\bar{c}_u$ is the average citation of the co-cited hyperlink of u^{th} user. The is to take into consideration all the hyperlinks that both user *u* and *v* have cited. For the active user, recommendation is performed to take a weighted average of all the citation on that hyperlink according to Equation (14).

$$p_{u,v} = \bar{c}_u + \frac{\sum_{v \in U} (c_{v,j} - \bar{c}_v) \cdot Sim_{u,v}}{\sum_{v \in U} |Sim_{u,v}|} \quad (14)$$

We evaluate the experimental results by *Mean Absolute Error* (MAE), which is the most widely used metric in CF research [12], defined as

$$MAE = \frac{\sum_{(i,j)} |p_{i,j} - r_{i,j}|}{n}$$

The MAE of the memory-based CF approach (BL) and the TDC approach is shown in the table below (smaller is better):

TABLE V. EVALUATION OF BASELINE AND TDC APPROACHES

	BL	TDC (ε)						
		0.1	0.2	0.3	0.4	0.5	0.6	0.7
MAE	0.842	0.814	0.810	0.807	0.811	0.811	0.826	0.819

The results in table V shows that the TDC approach outperforms the baseline approach, which may indicate that in TDC tweets bring in more correlation information the cluster-based approach can discover the users group, and that grouped users' the behaviors are more helpful for recommendation than the pairwise-user Pearson similarity. As ε rises, the cluster number increases in each way and the similarity within clusters will decline. The prediction accuracy may be influenced by weakly connected users. Thus the MAE increases with the increase of ε.

Fig. (5) illustrates the evaluation of TAC. Compared to the baseline evaluation, both the TAC and TDC approaches dramatically surpass the baseline approach. Similar to TDC, the MAE rises with the increment of ε. Besides, the TAC approach is slightly superior to the TDC approach. This may be due to the fact that the TAC seeks the approximation using three way clustering information while TDC just adopts one-way clustering.

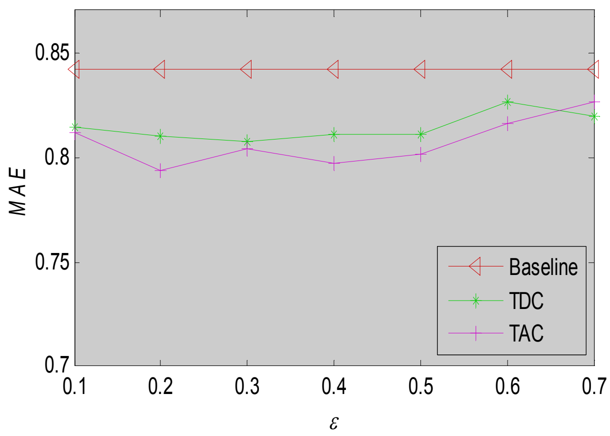

Fig.5 The Evaluation of TAC approach

VI. CONCLUSION AND FUTURE WORK

In this paper, we do statistics on micro-blog hyperlinks out of millions of micro-blogs. And a novel application is designed to enable users to enrich the content of their micro-blogs. Three-order tensor is leveraged to model the user-'micro-blog'-hyperlink relations. TDC and TAC are utilized to cluster the user, micro-blog and hyperlink in three-order tensor. The experimental results indicate TAC exhibits superiority in MAE.

In order to extend our work, we will realize a real interface to describe and evaluate our hyperlink recommendation. Still we try to combine TDC and TAC approaches.

CONFLICT OF INTEREST STATEMENT

We declare that we have no financial and personal relationships with other people or organizations that can inappropriately influence our work, there is no professional or other personal interest of any nature or kind in any product, service and/or company that could be construed as influencing the position presented in, or the review of, the manuscript entitled.

ACKNOWLEDGEMENTS

Our work is supported by Foundation of Science-technology Development Project of Shandong Province of China under Grant No. 2011YD01042, and the Science and Technology Development Program of Weifang (No. 2015GX008).

REFERENCES

[1] Banerjee, A., Basu, S., and Merugu, S. Multi-way Clustering on Relation Graphs. In *Proceedings of SIAM International Conference on Data Mining*, Minneapolis, Minnesota.USA. 2007.

[2] Banerjee, A., Dhillon, I. S., Ghosh, J., Merugu, S., and Modha, D. S. A generalized maximum entropy approach to bregman co-clustering and matrix approximation. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, Seattle, WA, USA. 2004.

[3] Banerjee, A., Merugu, S., Dhillon, I. S., and Ghosh, J. Clustering with Bregman Divergences. In *Proceedings of SIAM International Conference on Data Mining*, Florida, USA. 2004.

[4] Banerjee, A., Merugu, S., Dhillon, I. S., and Ghosh, J. Clustering with Bregman Divergences. *Journal of Machine Learning Research, 6*, 1705. (2005).

[5] Chee, S. H. S., Han, J., and Wang, K. 2001. RecTree: An Efficient Collaborative Filtering Method. In *Proceedings of Data Warehousing and Knowledge Discovery*, Munich, Germany.

[6] Chen, J., Geyer, W., Dugan, C., Muller, M. J., and Guy, I. Make new friends, but keep the old: recommending people on social networking sites. In *Proceedings of Computer Human Interaction*, Boston, USA. 2009.

[7] Gao, Q., Abel, F., Houben, G.-J., and Tao, K. Interweaving Trend and User Modeling for Personalized News Recommendation. In *Proceedings of the 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, Lyon, France. 2011.

[8] Gao, D., Zhang, R., Li W. and Hou, Y. Twitter hyperlink recommendation with user-tweet-hyperlink three-way clustering. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, Maui HI, USA. 2012.

[9] Hannon, J., Bennett, M., and Smyth, B. Recommending twitter users to follow using content and collaborative filtering approaches.In *Proceedings of the fourth ACM conference on Recommender systems*. Barcelona, Spain. 2010.

[10] Koga, H., and Taniguchi, T. Developing a User Recommendation Engine on Twitter Using Estimated Latent Topics. In *Proceedings of the 14th international conference on Human-computer interaction: design and development approaches*, Orlando, Florida, USA. 2011.

[11] Julie A. Jacko. Human-Computer Interaction. Design and Development Approaches. In J. Jacko (Ed.), (Vol. 6761, pp. 461-470): Springer Berlin / Heidelberg.

[12] Koji Miyahara, and Michael J. Pazzani. Collaborative altering with the simple bayesian classifier. In *Proceedings of the 6th Pacific Rim international conference on Artificial intelligence*. 679-689. 2000.

[13] Kolda, T. G., and Bader, B. W. Tensor Decompositions and Applications. *Siam Review, 51*(3), 455. 2009.

[14] Kwak, H., Lee, C., Park, H., and Moon, S. What is Twitter, a social Network or a news Media? .In*Proceedings of the 19th International World Wide Web Conference*, Raleigh NC, USA. 2010.

[15] Li, H., Tian, Y., Lee, W.-C., Giles, C.L. and Chen, M.-C. Personalized Feed Recommendation Service for Social Networks. In *Proceedings of The Second IEEE International Conference on Social Computing*, Minneapolis, Minnesota, USA. 2010.

[16] Linden, G., Smith, B., and York, J. Amazon.com Recommendations: Item-to-Item Collaborative Filtering. *IEEE Internet Computing, 7*(1), 76. 2003.

[17] Liu, X. Learning from multi-view data: clustering algorithm and text mining application. Katholieke Universiteit Leuven, Leuven, Belgium. 2011.

[18] Luxburg, U. V. A tutorial on spectral clustering. *Statistics and Computing, 17*(4), 395. (2007)

[19] Mathioudakis, M., and Koudas, N. TwitterMonitor: Trend Detection over the twitter stream.In*Proceedings of the 2010 international conference on Management of data*, New York, NY, USA. 2010.

[20] McLaughlin, M. R., and Herlocker, J. L. A collaborative filtering algorithm and evaluation metric that accurately model the user experience. In *Proceedings of the 27th Annual International ACM SIGIR*, Sheffield, UK. 2004.

[21] Miller, B. N., Albert, I., Lam, S. K., Konstan, J. A., and Riedl, J. MovieLens unplugged: experiences with an occasionally connected recommender system. In *Proceedings of Intelligent User Interfaces*, Miami, FL, USA. 2003.

[22] Morris, M. R., Counts, S., Roseway, A., Hoff, A., and Schwarz, J. Tweeting is believing?: understanding microblog credibility perceptions. In*Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*,Seattle, Washington. 2012.

[23] O'Connor, B., Krieger, M., and Ahn, D. TweetMotif: Exploratory Search and Topic Summarization for Twitter.In*Proceedings of the Fourth International Conference on Weblogs and Social Media*, Atlanta, Georgia. 2010.

[24] O'Connor, M., and Herlocker, J. Clustering Items for Collaborative Filtering. 2001.

[25] Phelan, O., McCarthy, K., Bennett, M., and Smyth, B. On using the real-time web for news recommendation and discovery. In *Proceeding of World Wide Web Conference Series*, Hyderabad, India. 2011a.

[26] Phelan, O., McCarthy, K., Bennett, M., and Smyth, B. Terms of a feather: content-based news recommendation and discovery using twitter. In*Proceedings of the 33rd European conference on Advances in information retrieval*, Dublin, Ireland. 2011b.

[27] Phelan, O., McCarthy, K., and Smyth, B. Using twitter to recommend real-time topical news. In *Proceedings of the third ACM conference on Recommender systems*, New York City, NY, USA. 2009.

[28] Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., and Riedl, J. GroupLens: an open architecture for collaborative filtering of netnews.In *conference of Conference on Computer Supported Cooperative Work*, Chapel Hill, NC, USA. 1994.

[29] Sarwar, B. M., Karypis, G., Konstan, J. A., and Reidl, J. Item-based collaborative filtering recommendation algorithms. In *conference of World Wide Web Conference Series*, Hong Kong, China. 2001.

[30] Sharifi, B., Hutton, M.-A., and Kalita, J. *Summarizing Microblogs Automatically*.In *Proceeding of the Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics.* Los Angeles, California. 2010.

[31] Su, X., and Khoshgoftaar, T. M. A Survey of Collaborative Filtering Techniques. *Advances in Artificial Intelligence*, 1-20. 2009.

[32] Symeonidis, P., Nanopoulos, A., and Manolopoulos, Y. Tag recommendations based on tensor dimensionality reduction. In*Proceedings of the 2008 ACM conference on Recommender systems*. Lausanne, Switzerland. 2008.

[33] TechInfo. Sept. 15th, 2010. Twitter Seeing 90 Million Tweets Per Day, 25 Percent Contain Links. from *http://www.techinfo.com.my/index.php/twitter-seeing-90-million-tweets-per-day-25-percent-contain-links/*

[34] Tucker, L. R. Some mathematical notes on three-mode factor analysis. Psychometrika, 31(3), 279. (1966)