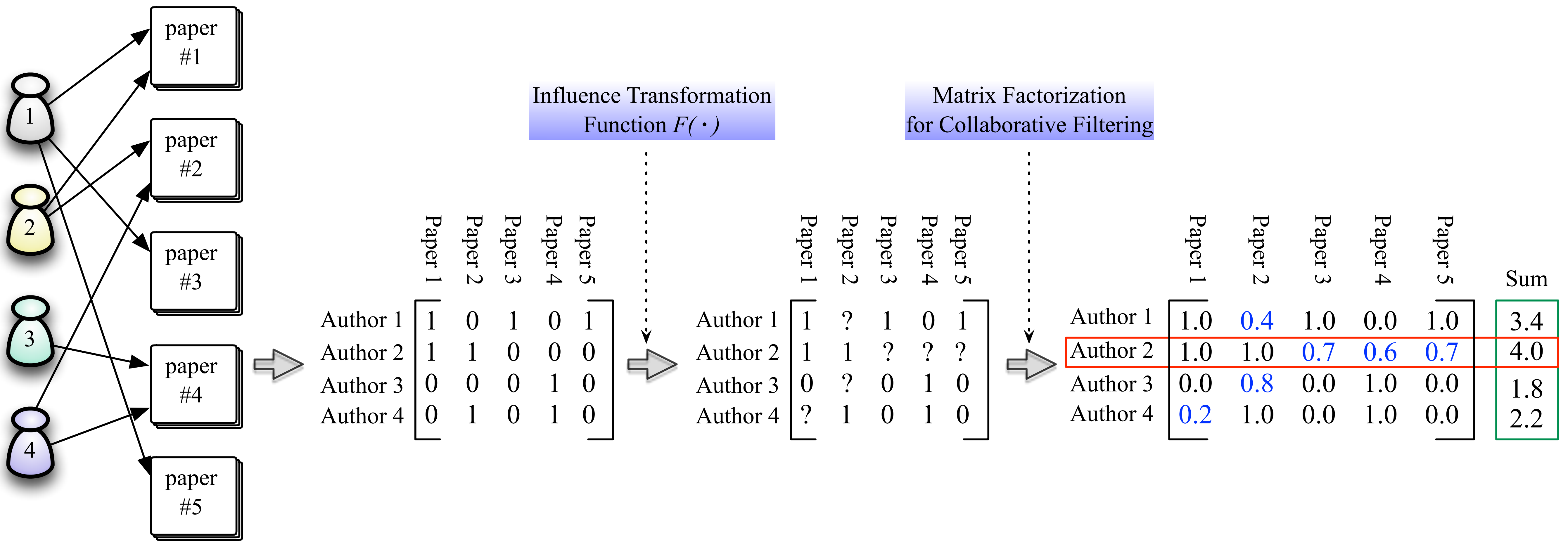
Social Influencer Analysis with Factorization Machines

Ming-Feng Tsai and Zhe-Li Lin
Department of Computer Science, National Chengchi University, Taiwan

Chuan-Ju Wang
Department of Computer Science, University of Taipei, Taiwan


paper #1
paper #2
paper #3
paper #4
paper #5
1
2
3
4
Influence Transformation Function $F(\cdot)$
Matrix Factorization for Collaborative Filtering

(a)

	Paper 1	Paper 2	Paper 3	Paper 4	Paper 5
Author 1	1	0	1	0	1
Author 2	1	1	0	0	0
Author 3	0	0	0	1	0
Author 4	0	1	0	1	0

(b)

	Paper 1	Paper 2	Paper 3	Paper 4	Paper 5
Author 1	1	?	1	0	1
Author 2	1	1	?	?	?
Author 3	0	?	0	1	0
Author 4	?	1	0	1	0

(c)

	Paper 1	Paper 2	Paper 3	Paper 4	Paper 5	Sum
Author 1	1.0	0.4	1.0	0.0	1.0	3.4
Author 2	1.0	1.0	0.7	0.6	0.7	4.0
Author 3	0.0	0.8	0.0	1.0	0.0	1.8
Author 4	0.2	1.0	0.0	1.0	0.0	2.2

(d)

Abstract

This work attempts to observe the collaborative events occurring at individuals involved in a social network to obtain such crucial knowledge. We propose a Factorization Machine approach to find out the latent social influence among the individuals based on their collaborations. Experiments conducted on a real-world DBLP dataset verify that the proposed approach can discover the latent social influence among individuals and provide a better predictive model than several baselines.

Introduction

How will the reputations of individuals in a social network be influenced by their neighbors? Such important knowledge is unfortunately not obtainable nowadays, and we attempt to observe its manifestation in the form of collaborative events occurring at the individuals involved in a social network to understand the problem. Therefore, the aim of this paper is to infer the latent social influence among individuals based on their patterns of collaborations. In order to tackle the problem, we use the idea of Collaborative Filtering (CF) to discover the latent social influence among individuals. Considering that Factorization Machines (FM), which is one of the state-of-the-art CF techniques, provide some advantages over other existing CF approaches, we apply the FM method to model the latent social influence. Specifically, we first present an influence transformation function to build up the influence matrix of individuals based on their patterns of collaborations. Then, the social influence of each individual is obtained via FM with the influence matrix; furthermore, some auxiliary information is utilized to help the latent influence discovery. Experiments, which are conducted on a real-world DBLP dataset including 3,662 authors and 5,122 papers, attest the proposed method can discover the latent social influence among individuals and produce a better predictive model than several baselines.

Methodology

Fig. 1 gives an illustrative example to introduce the core idea of the modeling process for the latent social influence. Fig. 1(a) depicts the relationships between the authors and their papers. These relationships can be transformed to the matrix representation in Fig. 1(b), in which each element x_{a_i,p_j} equals to 1 if a_i is the author paper p_j and otherwise that equals to 0. We then define an influence transformation function $F(\cdot)$ to build up the influence matrix (see Fig. 1(c)); this is a key step to transform the relationships in Fig. 1(a) to the input of a standard CF algorithm. The transformation function $F(\cdot)$ can be designed variously; in this paper, $F(\cdot)$ is defined as

$$F(x_{a_i,p_j}) = \begin{cases} 1, & \text{if } a_i \text{ is the author of } p_j,\\ ?, & \text{if } \exists a_k \in C_{a_i} \text{ and } a_k \text{ is the author of } p_j,\\ 0, & \text{otherwise,} \end{cases}$$

where C_{a_i} is the set of the authors who have coauthored with ai. After the transformation, we can obtain the resulting matrix in Fig. 1(d) via any CF algorithms. In Fig. 1(d), each number in blue color can be explained as the estimated latent social influence; the numbers in the green box are the sum of the influence scores of each author on all papers. As shown in the figure, we can observe that although author 2 has only written 2 papers, his/her social influence score (i.e., 4) is larger than that of author 1 (i.e., 3.4), who has written the most papers among the 4 authors. Even though author 2 is not the author of papers 3, 4, and 5, we consider that author 2 should still have (latent) influence on these three papers and the influence can be modeled with the patterns of collaborations among the authors. FM provides an advantage over other existing CF approaches, which make it possible to incorporate with any auxiliary information that can be encoded as a real-valued feature vector. Thus, via using FM, this paper integrates with text information to model latent social influence.

Experiments and Discussions

The experimental dataset is built for a certain research community from the DBLP data set, which contains the information of papers and the coauthors of each paper. We first collect the top 20 authors in the field of data mining from Microsoft Academic Search. Then, these 20 authors, their coauthors, and all of the papers of the 20 authors, are constructed as our experimental dataset, which consists of 3,662 authors and 5,122 papers. Note that the ranking of these 20 authors from Microsoft Academic Research is also considered as the ground truth of the experiments. Two rank correlation metrics are used to evaluate the performance in our experiments: Spearman's Rho and Kendall's Tau. The FM library, libFM [3], is adopted to conduct the experiments.

Table 1 tabulates the preliminary experimental results, in which we compare the results of three baselines and those of the proposed FM approach. The three baselines are the ranking via the numbers of coauthors, papers, and citations per author from Microsoft Academic Search. In addition, the fourth (fifth) row denotes the results of our proposed FM method without (with) the textual information, which is described by a bag-of-words model with term frequency. Note that only title words of the papers are used as the text information; the resulting vocabulary size is 4,057. Due to the randomization of the algorithms implemented in libFM, the values in the fourth and fifth rows are the averages of the 20 times experiments. As shown in Table 1, the performance of both FMs without and with texts reach significantly better results than the baseline methods. In addition, we can observe that incorporating the supplementary textual information did greatly improve the performance, which confirms that the textual information is beneficial to the latent social influence discovery.

	Spearman's Rho	Kendall's Tau
#coauthor (baseline)	0.233	0.179
#paper (baseline)	0.388	0.284
#citation (baseline)	0.469	0.347
FM without texts	$0.478^{*\dagger\ddagger}$	$0.349^{*\dagger}$
FM with texts	$0.556^{*\dagger\ddagger}$	$0.409^{*\dagger\ddagger}$

Table 1: The Experimental Results. The notations $*$, $\dagger$, and $\ddagger$ in Table 1 denote the result is significant better than the three baselines #coauthor, #paper, and #citation, respectively, with $p < 0.05$.

Conclusions and Future Work

This study attempts to model the latent social influence among individuals based on their patterns of collaborations in a social network via the FM approach. Preliminary experimental results on the small DBLP dataset for the data mining community show that the proposed approach provides a better predictive model than several baselines. In future work, we will conduct experiments on larger data sets with various fields of research communities. In addition, other auxiliary information, such as the temporal information of the publications [5], will be included and analyzed in our further experiments.

WebSci 2015