Venue Recommendation: Submitting your Paper with Style

Zaihan Yang Brian D. Davison
Department of Computer Science and Engineering, Lehigh University
Bethlehem, PA 18015
Email: {zay206,davison}@cse.lehigh.edu

***Abstract*—One of the principal goals for most research scientists is to publish. There are many thousands of publications: journals, conferences, workshops, and more, covering different topics and requiring different writing formats. However, when a researcher that is new to a certain research domain finishes the work, it is sometimes difficult to find a proper place to submit the paper. To solve this problem, we provide a collaborative-filtering-based recommendation system that can provide venue recommendations to researchers. In particular, we consider both topic and writing-style information, and differentiate the contributions of different kinds of neighboring papers to make such recommendations. Experiments based on real-world data from ACM and CiteSeer digital libraries demonstrate that our approach can provide effective recommendations.**

I. Introduction

Have you ever had a difficult time to decide where to submit your paper? As a research scientist, you may have such an experience occasionally. It is well understood that one of the fundamental tasks for most research scientists is to publish their work. Even though some experienced researchers may have a target venue in mind before they finish their work, many others, especially new researchers in a domain, prefer to finish their papers first, and then to decide where to submit. Similarly, if the paper is completed after the deadline or not accepted at the target venue, another choice may be needed.

It is not a trivial task to make such a choice, however, due to the rapid growth in both the quantity and variety of publication venues in recent decades. We now have many different kinds of venues, with different topics and genres and requiring different writing formats.

For different research domains, we have different venues; for example, the 'SIGIR' conference for Information Retrieval (IR) research, and the 'VLDB'conference for database research. Moreover, even within one research domain, we also have multiple venues. To take the 'IR' research domain as an example, we have journal publications such as *Information Retrieval* and *ACM Transactions on Information Systems*, as well as conferences,such as SIGIR, ECIR, WWW, CIKM, WSDM, and JCDL. We also have posters, workshops, technical reports, patents, etc. The creation of an automatic mechanism to help researchers deal with this problem is thus valuable, and recommender systems offer such an opportunity.

Recommender systems have emerged as a good solution for helping people deal with the rapid growth and complexity of information. The technique was first introduced to generate suggestions (e.g., for movies and merchandise) to users, and then introduced in social network analysis and has been widely used in many applications, including tag recommendation, link recommendation, and citation recommendation. However, little effort has been employed to tackle the problem of venue recommendation, where given a paper, with its authors, content, and references provided, a list of venues are recommended for submission of this paper.

A number of challenges arise in this task. First of all, the recommended venue should have a good match with the topics discussed in the paper. Venues have their own topic focus, as we have mentioned before, like information retrieval for SIGIR and databases for SIGMOD. Secondly, venues often have their specific writing format requirement. An interesting question may arise as whether papers with similar writing styles can more easily get accepted in similar venues. Finally, a good venue recommendation should match with the research profiles (e.g., historical venues) of the authors of the paper. We are interested in examining how the previous publication history of an author, along with the relationship between the target paper and other papers will be useful to affect the recommendation results.

Collaborative Filtering (CF for short) is the predominant approach in current recommender systems. It can be further divided into memory-based CF and model-based CF. Memory-based CF is widely used due to its simplicity and efficiency, and it provides a good framework for venue recommendation as both papers' inter-similarity and inter-relationships can be incorporated for recommendation. In this work, we introduce the memory-based CF into venue recommendation, and particularly provide two extensions to the basic algorithm. For the first extension, we consider papers' similarity in terms of their writing style. Content-free stylometric features are identified and extracted for similarity computation. For the second extension, we divide the neighboring papers of the target paper into several groups, each of which represents a certain scientific relationship with the target paper. Contributions from each sub-group of neighbors can be differentiated and optimized.

In summary, our paper makes the following contributions: 1) we provide a novel collaborative filtering based mechanism for automatic venue recommendation; 2) we extract stylometric features to measure the similarity between papers in terms of their writing styles; 3) we identify different groups of neighboring papers and differentiate and optimize their contributions for venue recommendation; and, 4) we demonstrate the effectiveness of the proposed approach on real-world data.

II. Related Work

A recommender system aims to automatically generate suggestions (of movies, merchandise) for users from a collection of items based on users' past behaviors. This technology has also been introduced into social network analysis, and used in many applications, such as tag recommendation [1], [2], friend recommendation [3], and citation recommendation [4], [5].

Traditional approaches for recommendation can be divided into two categories. One is the content-based filtering, which analyzes the content information associated with items and

users, such as user profiles and item descriptions, and matches that information for recommendation. Another typical approach, the collaborative filtering (CF) approach [6], however, predicts users' preferences over items by learning past user-item relationships. CF-based method can be further divided into two groups: the memory-based approach and the model-based approach.

Memory-based CF [7] predicts the rating of target user over an item as a weighted average of similar users or items. Depending on how past ratings are used, memory-based CF can be further classified into user-based, in which predictions are made by first finding similar users to the target user and averaging their ratings over a given item; or item-based, in which ratings are averaged over items similar to the target item that have been rated by the target user are found. For both user-based and item-based approaches, similar users (items), which are referred to as neighbors, are often found by applying the K-nearest neighbor approach.

Memory-based CF methods are simple and efficient, but sometimes they are suboptimal. Model-based approaches can overcome some of the limitations. The model-based approach trains a compact model via learning the observed user-item rating patterns, and directly predicts ratings of unknown user-item pairs instead of searching ratings of neighbors. It is, however, more time-consuming than the memory-based approach, and is complicated to implement. Typical model-based CF includes clustering [8] and the aspect model [9]. More recently, the matrix factorization [10] approach has been used as a kind of model-based CF for recommendation.

Two previous works consider the problem of venue recommendation. Lau and Cohen [11] develop a combined path-constraint random walk-based approach, not only for venue recommendation, but also for citation recommendation, gene recommendation and expert finding. In their work, they would present each term in the paper title as a node, combined with other entities, like author names and venue names to construct a big graph. Complex optimization approaches are carried out to learn the weights on each edge of the graph. Pham et al. [12], [13] define the task of venue recommendation as predicting the participating venues of users, and therefore their input is users instead of papers, which is different from our work. They use a clustering-based approach to group users that share similar patterns in choosing publishing venues.

III. Problem Identification

Let p be any given paper, and v be any candidate venue in the data corpus. The venue recommendation task can be defined as follows:

Given a paper p, what is the probability of it being published in venue v ?

It is essentially a ranking problem. Namely, we need to determine $p(v|p)$, rank the candidate venues according to this probability, and return the ranked list of candidate venues as recommendations of venues to which this paper could be submitted.

In order to compute this probability, we adopt the basic idea of collaborative-filtering, and utilize other papers with known venues to predict or recommend venues for the target paper. Moreover, we make two extensions to the original traditional collaborative-filtering based approach: one is to incorporate stylometric features to better measure the similarity between papers; the other is to differentiate the importance of those papers that share some similarity with the target paper, to further improve recommendation performance.

IV. Approach

A. CF-based method

In a traditional user-item recommendation system, when the memory-based collaborative-filtering approach is used to predict the ratings of users over items, the user-item relationship is often represented as a two-dimensional matrix. Similarly, we can represent the relationship between papers and venues in a two-dimensional matrix, where the rows represent papers, and the columns represent venues. For each particular paper-venue pair (p, v), the corresponding entry on matrix represented as $I(p, v)$ indicates whether paper p is published in venue v.

We can apply the memory-based CF into our paper-venue matrix, with the underlying assumption that it would have a higher probability for a paper to get published in venues in which other similar papers have been published. However, the paper-venue matrix is different from the user-item matrix in that one paper can only be published in one venue, and thus it is unsuitable to use the item-based method, where the similarity between items (venues) rated (published) by the target user (paper) is going to be compared. We therefore choose to apply the user-based CF.

Formally, the process of applying user-based CF to the venue recommendation task can be described as follows.

- Given a target paper p_i, we first compute its similarity with all other papers in the data set, and collect the K most similar papers to target paper p_i. The collection of these Top K papers is indicated as $S(p_i)$. K is a system parameter, and can be tuned via experiments.
- We collect all publishing venues of the papers in $S(p_i)$, and denote the collection as $V(p_i)$. For each venue v_j in collection $V(p_i)$, we predict the probability of having p_i published in v_j by computing $P(v_j|p_i)$ by

$$P(v_j|p_i) = \frac{\sum_{p_k \subset S(p_i)} s(p_i, p_k) I(p_k, v_j)}{\sum_{p_k \subset S(p_i)} I(p_k, v_j)} \quad (1)$$

  where $s(p_i, p_k)$ is the similarity score between paper p_i and p_k, and $I(p_k, v_j)$ is an indicator function. We have: $I(p_k, v_j) = 1$, if p_k is published in v_j; otherwise, $I(p_k, v_j) = 0$.
- Rank all candidate venues in $V(p_i)$ by $P(v|p)$.

B. Extension 1: Stylometric Features

As indicated in the above formula, one crutial component in this CF-based method for venue recommendation is the paper-paper similarity measurement. Dominant similarity measures in the traditional CF method include the Pearson Correlation Coefficient and Vector Space Cosine Similarity measurement. We make use of the latter method.

Papers differ in their content and topics. Moreover, papers as well as venues are also distinguishable by their writing styles [14]. To better measure papers' similarity, we need to consider both the content and stylometric features. To represent papers' content, we take use of Mallet [15], which is open source software implementing LDA [16], to retrieve the papers' content distribution over 100 topics; To capture the writing styles of papers, we identified 25 types and over 300 distinct features that are non-topic related and context-free. These features can be grouped into three categories, which measure a paper's writing style from lexical, syntactic and structural aspects.

TABLE I
FEATURES

Type	Features	Description
Lexical	TokenNum	Total number of words
	TypeNum	Total number of distinct words
	CharNum	Total number of characters
	SentenceNum	Total number of sentences
	AvgSenLen	Average sentence length
	AvgWordLen	Average word length
	ShortWordNum	Total number of short words (less than 3 characters)
	HapaxVSToken	Frequency of one-occuring words
	HapaxVSType	Frequency of one-occuring words
	ValidCharNum	Total number or characters excluding the non-digital, non-alphabetical and non-white-space characters
	AlphaCharNum	Total number of alphabetic characters
	DigitalCharNum	Total number of digital characters
	UpperCaseNum	Total number of characters in upper-case
	WhiteSpaceNum	Total number of white-space characters normalized by CharNum
	SpaceNum	Total number of space characters
	TabSpaceNum	Total number tab spaces
	Vocabulary Richness	A vocabulary richness measure defined by Zipf
Syntactic	FuncWordNum	Total number of function words
	PunctuationNum	Total number of punctuations (‘.’, ‘?’, ‘!’, ‘,’, ‘:’, ‘;’, ‘”, ‘ / ’)
	FuncWordFreq	Frequency of function words (298 feature)
Structural	SectionNum	Total number of sections
	FigureNum	Total number of figures
	EquationNum	Total number of equations
	TableNum	Total number of tables
	ReferenceNum	Total number of references

Lexical features [17] reflect a paper's preference for particular character or word usage. Typical features within this category include number of distinct terms, number of alphabetic and digital characters, average sentence length, etc. Syntactic features [18], however, focus on extracting the different formats and patterns in which sentences of a paper are organized. The most representative syntactic features include function words, punctuation and part-of-speech tags. In our work, we make use of the first two syntactic features. Structural features [19] represent the layout of a piece of writing. We adopt in our work five structural features specifically for scientific papers, including the number of sections, figures, tables, equations and references. The entire feature sets is presented in Table I.

In summary, we represent each paper by a feature vector, which is a combination of the paper's content topic distribution and stylometric features.

C. Extension 2: Neighbor Differentiation

Another crucial component in the memory-based CF model is to retrieve proper neighbors that share similarity with the target paper. Normally, this is done by finding the top K neighboring papers in terms of their cosine similarity score with the target paper. However, papers do not only differ in the value of the similarity scores, but also in their different relationships with the target paper. For example, given a paper, we can find other papers that are written by the same authors (authorship), papers that are cited by the target paper, and papers that share the same citations with the target paper (bibliographic coupling). All of these kinds of papers should play different roles in their influence on the target paper in selecting future venues in which to publish.

We divide the Top K similar papers into four categories. The first category is called ‘author-neighbors’, which are papers written by at least one author in common with the target paper. The second category is referred to as ‘reference neighbors’, which are the papers that have been cited by the target paper. The third category is named as ‘sibling neighbors’, which are papers that have at least one common reference paper with the target paper. All other papers that share similarity with the target paper, yet do not fall into any of the three categories mentioned above are referred to as ‘other neighbors’. Since we rely on the historical data for prediction or recommendation, for any given paper p which is finished in year y_1, and is to be predicted, we would only consider neighboring papers that have been published before y_1.

To differentiate their influence on the target paper, we introduce four parameters, each of which indicates the importance of neighbor papers of one category. To compute $P(v_j|p_i)$, the updated CF model can then be indicated as:

$$P(v_j|p_i) = \sum_{c:1\rightarrow 4} \alpha_c \frac{\sum_{p_k \subset N_c(p_i)} s(p_i,p_k) I(p_k,v_j)}{\sum_{p_k \subset N_c(p_i) I(p_k,v_j)}} \quad (2)$$

where $N_c(p_i)$ $(1 \leq c \leq 4)$ indicates the four categories of neighbor papers of the target paper p_i. $\alpha_c \in [0,1]$ is the parameter that needs to be tuned to reflect the influence of neighbor papers of category c.

V. EVALUATION

A. Experimental Setup

We introduce in this section the experiments we carried out for the task of venue recommendation. In particular, we wish to explore the following questions:

- What would venue recommendation results be if we utilize stylometric features alone to measure paper similarity?
- Can we achieve improved performance if we combine both the content and stylometric features for paper similarity measurement?
- Which category of paper neighbors would play the most important role in helping to predict publishing venues?
- Under what combination of the four categories can the best recommendation performance be achieved?

We carried out experiments on two real world data sets. The first data set is a subset of the **ACM Digital Library**, from which we crawled one descriptive web page for 172,890 distinct papers having both title and abstract information.

For each published paper, we extracted the information about its publishing venue and references. Due to possible venue name ambiguity, we first converted all upper-case characters into lower-case, and removed all non-alphabetic symbols. We further removed all digits as well as the ordinal numbers, such as the 1st, the 2nd, and applied the Jaccard Similarity match to merge duplicate venue names. We finally obtained 2,197 distinct venues. To remove author names' ambiguity, we represent each candidate author name by a concatenation of the first name and last name, while removing all the middle names. We then use exact match to merge candidate author names. Finally, we obtain 170,897 distinct authors.

The second data set we utilized is the **CiteSeer Digital Library** scientific literature distributed by the 2011 HCIR challenge workshop[1]. The data corpus is divided into two parts. Meta-data about a paper, such as its title, publishing venue, publishing year, abstract, information about citation references are kept in XML format; the full content of that

[1]http://hcir.info/hcir-2011

paper is in pure text format. We applied the same name disambiguation process as we did for the ACM data set, and obtained 119,927 papers that have abstract, full content and venue information, resulting in 478,805 authors and 48,797 venues. We further select 35,020 papers published across 739 venues, each of which has at least 20 papers published in it, to serve . as the experimental papers for the CiteSeer data set. We randomly choose 10,000 papers from ACM and CiteSeer data sets respectively as our target papers whose venues are to be predicted.

As introduced previously, we have identified three categories, and 25 different types of features. For papers in the CiteSeer data set, where the full content of papers is available in pure text format, we can simply count the number of times the word 'figure' or 'Figure' appears in the paper to obtain the number of figures. We did the same for number of sections, number of tables and number of equations. Finally, we extracted 371 stylometric features for papers in the CiteSeer data set, and 367 features for papers in the ACM data set.

To test venue recommendation performance, we match the predicted venues with the real publication venues of the target papers. Two standard metrics: **Accuracy@N** (N varies among 5, 10, and 20) and **MRR** are adopted for evaluation.

B. Results Analysis

1) **Stylometric Features:** We first examine whether paper similarity based on stylometric features can lead to good recommendation performance. By doing this, we represent each paper by a vector composed of only the stylometric features of that paper, and compute papers' similarity based on those paper vectors.

For comparison, we construct paper vectors by only making use of their paper content information, that is, the paper's content distribution over 100 topics learned from LDA. We also combine both content and stylometric features to get merged features for paper similarity measurement. In all the experiments, we set the parameters $\alpha_c (1 \leq c \leq 4)$ to be 0.25.

We collect the Top K most similar papers with known venues the possible publishing venue of the target paper. K is a system parameter, whose value might affect the prediction performance. To examine its effect, and varied the value of K among 500, 1000, 2000, 5000, 10000. We also experimented with using all neighboring papers of the target paper. Experimental results for ACM and CiteSeer data sets are described in Table II.

Several observations can be found from the results on the ACM data set. First of all, there is a significant improvement as we combine both stylometric and content-based features as compared to working on either stylometric or content-based features separately, whose performance is competitive with each other. The improvement is nearly or more than 50% when a subset of paper neighbors are considered, and is 10.92% working on all paper neighbors in terms of Accuracy@5. Secondly, there is no obvious increase in terms of Accuracy@5, Accuracy@10 and Accuracy@20 as the value K (the Top K most similar papers to the target paper) increases from 500 to 10000 working on either stylometric or content features seperately. However, we achieved consistent improvement on the average MRR value. When working on combined features, performance in terms of all metrics also obtained constant improvement. We achieve significant improvement when we collect all paper neighbors for consideration. The best performance is achieved when working on all neighbors with combined features. Over 55.72%, 69.81% and 78.32% papers can have their publishing venues be correctly predicted within Top 5, Top 10 and Top 20 results respectively.

TABLE II
VENUE RECOMMENDATION RESULTS ON ACM AND CITESEER DATA

	ACM			CiteSeer		
*T*op K=500						
	Style	Content	S+C	Style	Content	S+C
Accuracy@5	0.084	0.103	**0.150**	0.065	**0.125**	0.108
Accuracy@10	0.150	0.190	**0.291**	0.086	**0.172**	0.148
Accuracy@20	0.265	0.352	**0.526**	0.141	**0.251**	0.231
MRR	0.002	0.003	**0.005**	0.010	0.013	0.013
*T*op K=1000						
	Style	Content	S+C	Style	Content	S+C
Accuracy@5	0.081	0.081	**0.150**	0.086	**0.122**	0.116
Accuracy@10	0.138	0.151	**0.286**	0.105	**0.157**	0.152
Accuracy@20	0.239	0.272	**0.504**	0.137	**0.212**	0.209
MRR	0.003	0.004	**0.009**	0.008	0.009	0.009
*T*op K=2000						
	Style	Content	S+C	Style	Content	S+C
Accuracy@5	0.079	0.071	**0.166**	0.114	0.122	**0.130**
Accuracy@10	0.128	0.124	**0.319**	0.131	0.156	**0.162**
Accuracy@20	0.224	0.221	**0.520**	0.158	0.197	**0.209**
MRR	0.005	0.006	0.013	0.006	0.007	**0.008**
*T*op K=5000						
	Style	Content	S+C	Style	Content	S+C
Accuracy@5	0.080	0.075	**0.214**	0.153	0.117	**0.158**
Accuracy@10	0.128	0.124	**0.375**	0.177	0.148	**0.196**
Accuracy@20	0.220	0.217	**0.559**	0.203	0.197	**0.236**
MRR	0.009	0.008	**0.022**	0.006	0.006	**0.007**
*T*op K=10000						
	Style	Content	S+C	Style	Content	S+C
Accuracy@5	0.086	0.082	**0.249**	0.190	0.118	**0.195**
Accuracy@10	0.134	0.140	**0.422**	0.221	0.161	**0.231**
Accuracy@20	0.230	0.241	**0.604**	0.250	0.227	**0.272**
MRR	0.011	0.009	**0.027**	0.006	0.006	**0.007**
*A*ll Neighbors						
	Style	Content	S+C	Style	Content	S+C
Accuracy@5	0.502	0.367	**0.557**	0.238	0.124	**0.239**
Accuracy@10	0.623	0.492	**0.698**	0.286	0.178	**0.290**
Accuracy@20	0.716	0.600	**0.783**	0.332	0.250	**0.337**
MRR	0.032	0.016	**0.046**	0.006	**0.007**	0.006

We noticed consistent performance when working on the CiteSeer data set, where paper full content is used for generating both content and stylometric features. Content-based features work better than stylometric features when a small set of top-returned paper neighbors are adopted; however, the performance on using stylometric features gradually outperform that of content-based features when more top-returned paper neighbors are considered. When combining both stylometric and content-based features, there is no improvement as compared to using pure content-based features, however, we observe improved performance for such a combination when more than 2000 top neighbors are considered. The best performance is also achieved when all paper neighbors and all features contribute, where 23.87%, 28.99% and 33.74% papers can have their venues correctly predicted within Top 5, Top 10 and Top 20.

2) **Weights among neighbors, Parameter tuning:** We expect that different categories of neighboring papers can have different contributions when making venue recommendations.

We gradually change the weight for each particular type of neighbors from 0 to 1, and let the other three kinds of neighboring papers share the remaining weight. Results are reported in Figure 1.

When the weight for a particular type is set to be 1, it actually indicates the individual contribution of that type of neighbors. As shown in the results, author neighbors contribute the most in both data sets, while the other neighbors are less important. It indicates that when authors finish their work, they often submit the paper to those venues in which they have had

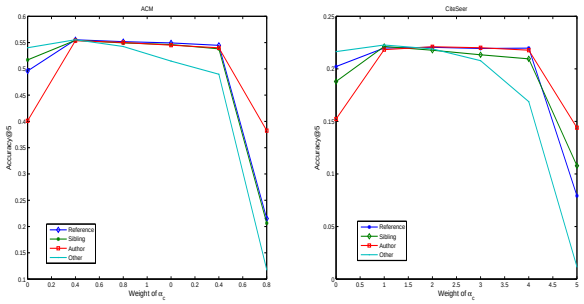

Fig. 1. ACM and CiteSeer: Weight of Neighbors

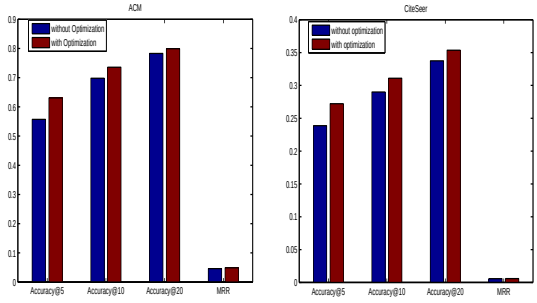

Fig. 2. ACM and CiteSeer: Parameter Optimization

a previous paper successfully published. This is on one hand due to researchers continuing to focus on similar or related topics, at least within similar research domains. On the other hand, authors will gain more reputation and thus confidence in certain venues, so that they are always willing to submit to those venues, and it also has higher probability to have their work accepted. Reference neighbors and Sibling neighbors are competitive with each other, which matches our initial expectation, as reference neighbors and sibling neighbors both are topic-related with the target paper.

We also notice from both results that we need to incorporate all types of neighbors, since we can retrieve better performance when all four categories of neighboring papers contribute rather than giving any of them zero weight. Moreover, even though the author neighbors are the most important source of information, when giving extra weight to them, predictive performance decreased.

3) **Weights among neighbors, Parameter Optimization:** Parameter tuning, as we addressed in the last section, tells us the different importance of different categories of neighboring papers. We are more interested, however, to find parameter settings that can give us the best recommendation performance. To implement that, we can apply parameter optimization approaches.

Given a paper p_i, which is the target paper, and any candidate venue v_j in the data set, we can compute the probability $P(v|p)$ based on formula (4). Suppose A_j, R_j, S_j and O_j represent the normalized accumulated similarity score between the target paper and author neighbors, reference neighbors, sibling neighbors, and other neighbors respectively; the formula can be re-written as: $P(v_j|p_i) = \alpha_1 A_j + \alpha_2 R_j + \alpha_3 S_j + \alpha_4 O_j$

Let us suppose the real publishing venue for the target paper p_i is venue v_j, then in an ideal venue recommendation system, for any other venue candidate v_k rather than v_j, the computed probability score $P(v_k|p_i)$ should be less or at most equal to $P(v_j|p_i)$; that is, we need to have $P(v_j|p_i) - P(v_k|p_i) \geq 0$ for all v_k $(k \neq j)$. Naturally, our goal is to learn the values of the four parameters α_c $(1 \leq c \leq 4)$, such that $\sum_{k:1\rightarrow V}(P(v_j|p_i) - P(v_k|p_i))$ can be maximized, where V is the number of candidate venues. Therefore, we introduce our objective function as: $h = argmax \sum_{k:1\rightarrow V} s(P(v_j|p_i) - P(v_k|p_i))$ where $s(x)$ is the sigmoid function: $s(x) = \frac{1}{1+e^{-x}}$.

To achieve the optimal combination of weights, we use gradient descent in which the four parameters are updated in each iteration until they converge.

As shown in Figure 2, we achieved more than a 13% improvement in Accuracy@5 for both ACM and CiteSeer.

4) **Comparisons with other approaches:** In order to demonstrate the effectiveness of our proposed approach, we compare results across several baseline algorithms:

Simple Counting: For each target paper p_i, we simply count the occurring frequency of venues of three kinds of neighboring papers of paper p_i, i.e., the reference neighboring papers (papers cited by p_i, referred as SimpleCount-Ref), sibling neighboring papers (papers that share at least on citation with p_i, referred as SimpleCount-Sibling) and author neighboring papers (other papers written by authors of p_i, referred as SimpleCount-Author). We also count the frequency of venues of the combination of all three kinds of neighboring papers (referred as SimpleCount-All). We would then rank and return the venues in terms of their frequency.

Content-based LDA: We construct a profile for each venue by concatenating all the papers published in it. We use LDA topic model implemented by Mallet [15] to retrieve the topic distribution for each paper and venue over 100 topics. We then compute and rank venues by their similarities with the target paper.

Traditional memory-based CF: We use the original traditional memory-based CF approach, in which we do not incorporate stylometric features of papers to compute their similarity, nor do we categorize neighboring papers and differentiate their different contributions. Under this scheme, $P(v_j|p_i)$ can be computed as: $P(v_j|p_i) = \sum_{p_k \subset S(p_i)} s(p_i, p_k) I(p_k, v_j)$, where papers' similarity is determined by their topic distribution obtained from LDA.

Graph-based FolkRank algorithm: We used the FolkRank algorithm [20], which is an adaptation of PageRank, and has been shown empirically to generate high quality recommen-

TABLE III
ACM AND CITESEER: COMPARISON WITH BASELINE ALGORITHMS

ACM Data	Accuracy@5	Accuracy@10	Accuracy@20	MRR
SimpleCount-Ref	0.203	0.212	0.212	0.0006
SimpleCount-Sibling	0.252	0.307	0.344	0.0008
SimpleCount-Author	0.377	0.430	0.446	0.0008
SimpleCount-All	0.470	0.566	0.603	0.0013
contentLDA	0.010	0.018	0.024	0.0008
traditionalCF	0.317	0.467	0.608	0.0283
FolkRank	0.102	0.184	0.252	0.0087
Our method	0.557	0.698	0.783	0.0459
CiteSeer	Accuracy@5	Accuracy@10	Accuracy@20	MRR
SimpleCount-Ref	0.096	0.099	0.099	0.0001
SimpleCount-Sibling	0.112	0.141	0.161	0.0001
SimpleCount-Author	0.129	0.157	0.176	0.0001
SimpleCount-All	0.199	0.239	0.277	0.0002
contentLDA	0.008	0.016	0.022	0.0005
traditionalCF	0.095	0.015	0.224	0.0040
FolkRank	0.037	0.068	0.113	0.0037
Our method	0.239	0.290	0.337	0.0058

TABLE IV
VENUE RECOMMENDATION RESULTS: EXAMPLES

Paper Title	Top 5 Predicted Venues
1. corpus structure language models and ad hoc information retrieval (SIGIR 2004)	annual meeting acl **annual intl acm sigir conf on research and development in information retrieval** journal machine learning research computational linguistics acm ieee cs joint conf on digital libraries
2. induction of integrated view for xml data with heterogeneous dtds (CIKM 2001)	acm sigmod intl conf on management data **intl conf on information and knowledge management** acm symposium on applied computing communications acm vldb journal mdash intl journal on very large data bases
3. multi resolution indexing for shape images (CIKM 1998)	acm intl conf on multimedia intl conf on very large data bases annual acm siam symposium on discrete algorithms conf on visualization annual conf on computer graphics and interactive techniques (rank 8) **intl conf on information and knowledge management**
4. video suggestion and discovery for youtube taking random walks through the view graph (WWW 2008)	intl conf on human computer interaction with mobile devices and services annual sigchi conf on human factors in computing systems acm sigkdd intl conf on knowledge discovery and data mining **intl conf on world wide web** annual meeting on association for computational linguistics

dations in tag recommendation systems. The basic idea of this approach is to run PageRank algorithm twice, giving uniform initial weights to all nodes in the first time, and giving higher weight to targeted nodes in the second time. The difference in terms of the weight of the nodes is then used to generate the final ranking.

We compare the results using our proposed approach with the baseline algorithms, and show the results in Table III. The results we report under our method are the best results we can achieve when both stylometric and content features are combined and all neighboring papers are considered. As indicated from the results, our approach outperforms the baseline algorithms under all evaluation metrics. The content-based approach works the worst. TraditionalCF can work better than the graph-based FolkRank algorithm; moreover, we can achieve better performance when no normalization is introduced. The SimpleCount-based method can provide surprisingly good results, and is the second best algorithm among all compared algorithms. However, our model can improve performance over SimpleCount-All by 18.53% (on ACM) and 19.77% (on CiteSeer) in terms of Accuracy@5.

5) **Case Study Example:** We show in this section several recommendation examples using our proposed approach. We report in Table IV the Top 5 returned venues for three randomly chosen papers in our system. Venue names written in bold indicate the actual publishing venue of that paper. We observed that for each target paper, under most circumstances, the top five returned venues share similarity in topics, and are content-related with the target paper. They are all reasonable candidate venues to which the paper could have been submitted. For papers that concentrate on topics within specific subset of a wide research domain, or discussed topics covering interdisciplinary domains, we can also provide proper recommendation. For example, paper 1 focuses on modeling language, and therefore some computational linguistics related venues are ranked highly, such as ACL. Paper 2 discussed database integrated view design, and therefore venues in the database domain like SIGMOD and VLDB are returned. We also notice that some paper may have other appropriate choices when considering submitting, for example, for paper 3, even though its actual publishing venue is only ranked 8th, several other venues ranked higher than the actual venue are also good places to submit.

VI. CONCLUSION

We applied the memory-based collaborative filtering approach for venue recommendation, and in particular, we updated the original CF based approach by applying two extensions. The first extension is to incorporate papers' stylometric features to better measure the similarity between papers, the second one is to divide the neighboring papers into four categories. By tuning or optimizing the different contributions of four categories of neighboring papers, we expected to obtain better recommendation performance. Experiments demonstrate our approach to be an effective method for venue recommendation, which outperformed several baseline algorithms. By differentiating the four categories of neighboring papers' contributions, we also find that papers that are published by the same authors are the most reliable source of information for the venue recommendation task.

ACKNOWLEDGMENTS

This work was supported in part by a grant from the National Science Foundation under award IIS-0545875.

REFERENCES

[1] B. Sigurbjornsson and R. Zwol, "Flickr tag recommendation based on collective knowledge," in *WWW'08*, 2008, pp. 327–336.
[2] Y. Song, Z. Zhuang, H. Li, Q. Zhao, J. Li, W. Lee, and C. Giles, "Real-time automatic tag recommendation," in *SIGIR*, 2008, pp. 515–522.
[3] W. Hsu, A. King, M. Paradesi, T. Pydimarri, and T. Weninger, "Collaborative and structual recommendation of friends using weblog-based social network analysis," in *AAAI Spring Symposium '06*, 2006.
[4] Q. He, J. Pei, D. Kifer, P. Mitra, and L. Giles, "Context-aware citation recommendation," in *WWW'10*, 2010, pp. 421–430.
[5] J. Tang and J. Zhang, "A discriminative approach to topic-based citation recommendation," *Advances in Knowledge Discovery and Data Mining, Lecture Notes in Computer Science*, vol. 5476, pp. 572–579, 2009.
[6] G. Adomavicius and A. Tuzhilin, "Toward the next generation of recommendation systems: A survey of the state-of-the-art and possible extensions," *IEEE Trans on Knowledge and Data Engineering*, vol. 17, no. 6, pp. 734–749, Jun. 2005.
[7] J. Breese, D. Heckerman, and C. Kadie, "Empirical analysis of Predictive Algorithms for Collaborative Filtering," in *UAI*, 1998, pp. 43–52.
[8] A. Kohrs and B. Merialdo, "Clustering for Collaborative Filtering Applications," in *Computational Intelligence for Modelling, Control and Automation*. IOS, 1999.
[9] T. Hofmann and J. Puzicha, "Latent Class Models for Collaborative Filtering," in *AI*, 1999, pp. 688–693.
[10] A. Paterek, "Improving regularized singular value decomposition for collaborative filtering," in *KDD Cup and Workshop*, 2007.
[11] N. Lao and W. Cohen, "Relational retrieval using a combination of path-constrained random walks," *Machine Learn*, vol. 81, no. 1, pp. 53–67, 2010.
[12] M. Pham, Y. Cao, and R. Klamma, "Clustering Technique for Collaborative Filtering and the Application to Venue Recommendation," in *Proc. of I-KNOW*, 2010.
[13] M. Pham, Y. Cao, R. Klamma, and M. Jarke, "A clustering approach for collaborative filtering recommendation using social network analysis," *Journal of Universal Computer Science*, vol. 17, no. 4, pp. 583–604, 2011.
[14] Z. Yang and B. D. Davison, "Distinguishing Venues by Writing Styles," in *Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL)*, Jun. 2012, pp. 371–372.
[15] A. McCallum, "MALLET: A Machine Learning for Language Toolkit," in *http://mallet.cs.umass.edu*, 2002.
[16] D. Blei, A. Ng, and M. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.
[17] F. Mosteller and D. Wallace, *Inference and Disputed Authorship: The Federalist*. Addison-Wesley, Reading, Mass., 1964.
[18] E. Stamatatos, N. Fakotakis, and G. Kokkinakis, "Automatic text categorization in terms of genres and author," *Comp. Ling*, vol. 26, no. 4, pp. 471–495, 2001.
[19] O. de Vel, "Mining Email authorship." in *Text Mining Workshop. KDD*, 2000.
[20] A. Hotho, R. Jaschke, C. Schmitz, and G. Stumme, "FolkRank: A Ranking Algorithm for Folksonomies," in *In Proc. of LWA*, 2006, pp. 111–114.