Recommending Publication Venue in Context Using Abstract

Tian-mu MA* and Wei FANG
School of Computer Science, BUPT, Beijing, China, 100876
*Corresponding author

Keywords: Topic modeling, Information retrieval, Recommending system, Topic coherence.

Abstract. One of the key steps of composing a scientific work is selecting the most appropriate publication venue. All a researcher can rely on are experiences from previous study or retrieving similar topics provided by publication venues. A reliable submission recommender is thus in great demand. Unlike the existing works with the same endeavor, we propose a topic modeling based recommending system leveraging briefly abstracts which avoids the somehow unappeasable demand of full texts, references, etc. In this paper, we set up a model with carefully estimated parameters for fine-grained representation of the corpus. Combining with several schemes of generating candidate venues, a context filter based on publication dates is included to optimize the results. Our proposal was evaluated and compared with several baselines on a dataset of 23,766 papers of 18,862 authors from 373 venues. Extensive experiments illustrate that our recommending system can product more satisfactory accuracy among other methods while requiring far more less information.

Introduction

When it comes to defining a preferred venue to submit paper, a couple of factors are commonly considered. Venues’ topic, of course, is a cardinal principle of filtering unrelated venues. Although the organizers have released some subjects corresponding to their venue, researchers have always been expecting to understand more accurately, which means a more fine-grained topic-matching method is required. Some other practical factors include venue reputation and time arrangement including deadlines and reviewing time. The agreement on metrics for evaluating reputation of scientific papers, as discussed in [1], has long been unreachable. Reputation metrics for venues, which is based on integrating its papers’ reputation, is thus an ambiguous region to be explored.

In this paper, we propose a recommending system based on Topic Modeling for recommending publication venues to researchers with hesitation in selecting an appropriate venue to submit their scientific work. Our fundamental strategy is using abstracts as main data sources according to the fact that consideration of future decision usually begins at the early stage of paper compiling. In order to better capture the latent relations on topics, we introduce the metrics of topic coherence to determine an approximately optimal granularity of our topic model. Cosine similarities are calculated among topic distribution representations inferred by our model, upon which several schemes with different emphasis are built to generate candidate venues. Also, a context filter is attached to the schemes with the idea that papers with closer publication date contribute more to the accuracy of prediction. After all these preliminaries, a series of targeted experiments are conducted on a dataset of 23,766 papers from 373 venues to determine the parameters involved in our model and evaluate the performance of various schemes. Elaborate experiment results analysis along with other approaches by previous works and conventional baselines illustrates that our system is capable of recommending publication venue with even more higher accuracy while requiring far more less information from users.

Related Work

Although recommending systems in various scenes have been widely studied, those specifically designed for recommending academic venues are rarely seen. Papers work on academic venue recommendations are focused in mainly two aspects: participation recommendation and publication
recommendation. Participation recommendation aims at assisting researchers to participate in talks, discussions and so on. Representative papers like [2], [3] and [4] relies mainly on social network built on co-authorships and citation networks to characterize researchers. Publication recommendation focuses more specifically on the activity of contributing research papers.

Social network of authors is introduced, such as in [5], [6], to help locate other authors who share the similar preference in paper submission, hence a list of venues they submitted to is accessible as recommendation candidates. Such social network is built by linking authors between whom co-authorship exists. As described in [5], they collected co-authorship recursively starting from the core author, until a network of three-level depth was generated. This process may be well working for a senior researcher, but still likely to become instable considering our original purpose of assisting those fresh to a brand new discipline and are not necessarily involved in a well-connected network.

Other previous works employed methods on mining content. Among them, a representative example is [7]. In order to capture all aspects of content, topics and writing styles, a set of features is extracted from each paper’s full text, including content and stylometric features. An LDA (Latent Dirichlet Allocation) [8] approach is used to retrieve paper’s content distribution over 100 topics, and over 300 other features are grouped into three categories which are lexical, syntactic and structural features. Based on the above, a user-based CF(Collaborative Filtering) is used to generate candidate venues. An extension of improvement in similarity measurement leverages co-authorship and reference information to differentiate neighbor weights into four discrete values.

As the same pursue of ours, [9] proposed a system using only title and abstracts. In their system, topic distribution alone is considered in the analysis of abstracts. Three topic modeling methods are introduced to profile each paper with topics making use of abstracts. Two of them are derived from LDA, perform poorly due to the somehow unsuccessful attempts in perfectly profiling publication venues. While, an N-gram based text categorization method from [10] outperforms the others, which will be discussed in our work and chosen as a baseline method. Experiments carried out in [9] showed that they had achieved aligned performance with previous works while requiring less information.

**Notation and Terminology**

A publication venue for scientific research papers can be in various categories such as conferences, workshops, journals and so on. Actually our dataset is indeed composed of papers from different forms of publication venues. For simplicity, we use “venue” in this paper to represent these publication organizations. Similarly, when dealing with text processing, we use “paper” for intuition where actually only abstracts are involved and analyzed.

Notations used in our paper are displayed (see Table 1) for reference.

| Symbol | Description |
|--------|-------------|
| $p_t$  | the target paper to which we recommend venues. |
| $p_a$  | an arbitrary paper. |
| $pf_t$ | the profile of the target paper. |
| $pf_a$ | the profile of an article(paper). |
| $pf_v$ | the profile of a venue. |
Our Approach

N-Gram-Based Text Categorization

Since venues can be viewed as a group of scientific publications with the same kind, publication venue recommendation can be described as deciding a list of candidate venues of the most similar kind to \( p \), namely a text categorization problem. Hence we first take N-Gram-based Text Categorization (NGTC), which is introduced in [10] and brought in as a baseline of our approach. An N-gram here is a contiguous N-character slice of a longer string, overlapping each other in some degree. N-grams of different lengths are used simultaneously from one to \( N \). In order to embody the head and tail of each word, a blank is appended to both sides. (The underscore character ‘\_’ is used to represent blanks.)

In the profiling phase, a profile \( pf_a \) is generated for each paper by calculating the frequency of each slice’s occurrence after a process of N-grams in which \( N = 5 \) and sorting the list in descending order. The detailed frequency number will be ignored and the most frequently appeared 300 slices (as used in [10]) will be regarded as \( pf_a \). Similarly, the profile \( pf_v \) is generated for each venue by collecting all papers’ N-grams that are published in it. The list of slice-frequency pairs of descending order will be truncated to 300 most frequent slices as well.

In the recommending phase, we firstly use a statistical measure of distance based on the method called “out-of-place” in with some adjustment. The distance of \( p_t \) to another paper \( p_a \) is calculated as:

\[
\text{Dist}_{p_t \rightarrow p_a} = \sum_{s \in pf_t} |\text{pos}(s, pf_t) - \text{pos}(s, pf_a)|
\]

where \( \text{pos}(s, pf) \) is a function similar to indexing, which returns the position of the slice \( s \) in profile \( pf \). The adjustment here is that we define the returned value as the maximum, i.e., 300 as in the previous description, if \( s \) is not found in \( pf \). The original method simply omitted the subtraction here and augment the distance by the maximum value, however we believe this will leave such slices in \( pf \) with different positions acting exactly the same. \( p_a \) can be replaced with a venue and with \( pf_v \), the distance from \( p_t \) to a certain venue is obtained.

Similarity Calculation with Topic Modeling

Two basic schemes of generating candidate venues can be concisely summarized as:

1. Compare the target paper to each venue, and collect a subset as candidate venues.
2. Compare the target paper to each paper in the corpus, collect a subset and take their corresponding venues as candidate venues.

While the former exhibits an intuitive and easy-understanding idea, the problem of how to generate a profile for each venue in the comparison becomes a principal obstacle due to the uncertainty of the weight distribution over different affiliated papers. Here, we bypass this intractable issue and concentrate on the latter scheme, where a more specific metric of estimating paper similarity is required. Inspired by the fundamental requirement of coincidence of topics in publication venue selection, we resolve to settle this pair-wise relevance judgement problem between papers, namely similarity calculation, by leveraging some prevalent topic modeling algorithms.

From the most popular schemes such as tf-idf [11], pLSI [12], etc., we take LDA on behalf for its remarkable appearance in text modeling, text classification and collaborative filtering, and for intuitional comparison against previous works that had also employed LDA.

With a fixed number, LDA generates multiple topics represented by word distribution and the distribution of the obtained topics for \( p_a \) in the corpus. For each paper, a feature vector is created as thus, with the \( i \)th element indicating the weight of the \( i \)th topic. We next occupy cosine similarity as our similarity evaluation metric, where a higher similarity value means a closer relationship in topics.
Granularity of Topics

A preparatory work before topic modeling is to transform texts into Bag of Words (BoW) representation, as applied in [8]. When it is performed in our framework, our corpus composed of abstracts reveals the characteristic of sparsity in vector space to some extent. Thus a fine-grained set of topics is necessary in pursuing a more human-interpretable and discriminative topic model. However, most conventional LDA implementations require a manually assigned parameter of number of topics for training and there are no recognized selection criteria on it. To address this issue of determining a proper number of topics, we introduce the concept of topic coherence.

There are some prevalent measures for topic coherence such as \( C_V, C_{UX}, C_{UAM}, C_{NPMI} \), specifically discussed in [13]. Here we select \( C_V \) for its outstanding performance in the evaluation. \( C_V \) is based on a sliding window, a one-set segmentation of the top words and an indirect confirmation measure that uses normalized pointwise mutual information (NPMI) and the cosine similarity. By conducting a series of trials of training models of different amounts of topics with the same corpus and observing the trend how \( C_V \) value varies, we can find an approximately optimal value of the number of topics.

Recommendation with Context

A universal goal of recommendation is to generate a ranked list of candidates so as to provide the top K most probable reference wherein K can be 3, 5, 10, etc. Based on the thinking of collaborative filtering, here we propose two practical schemes of generating a ranked list of candidate venues for further recommendation.

RANK BY QUANTITY: First of all, a threshold is determined which is a value between 0 and 1 indicating an appropriate similarity level and all papers under this threshold are filtered out. Each one of the remaining papers have its corresponding venue recorded and contribute an increment in quantity to the venue. After such a traversal, a rank of descending order is obtained embodying venues containing large quantities of most similar papers.

RANK BY SIMILARITY: With a relatively straight-forward strategy, only a few explorative steps are required, which is far more effective than the former scheme. While traversing the paper list ranked by similarity, we keep appending the corresponding venue of the current paper to our final ranking before reaching its maximum volume only if the venue has not been included yet.

CONTEXT FILTER: We introduce context factors herein as a filter while selecting candidate papers to vastly enhance accuracy. We believe that papers written within a smaller interval of time contribute more than those written a bit longer ago. Here we take advantage of accessible time-stamps of the papers in the corpus, which are the corresponding publication dates in this case, and the conditional possibility working in the filter is given as follows:

\[
P(I|p_i, p_a) = \exp\left[-T(p_i) - T(p_a)\right].
\]

where, \( T(p) \) is a function by which we correspond the time-stamp to a paper \( p \). Note that in our experiments, interval \( T(p_i) - T(p_a) \) must be a positive value because practically we only compare papers already published before to the target paper. \( I \) is a probability event indicating “ \( p_a \) is chosen to compare with \( p_i \)”. While the interval increases for \( p_a \) from \( p_i \), its \( P(I|p_i, p_a) \) will rapidly decrease, thus less probably “approved” to be included as a candidate paper.

Experiments

Dataset

Our dataset is obtained from the University of Michigan’s CLAIR Group’s ACL Anthology Network interface, which contains information of papers included in the many ACL (the Association of
Computational Linguistics) venues. The dataset covers 23,766 papers of 18,862 authors published in 373 venues.

A fundamental preprocessing method operated by NLTK (Natural Language Toolkit) was exerted to filter and normalize texts. For each abstract, we first tokenized the text and removed all digits, punctuations, stop-words and tokens with length being three or less. Then we converted all tokens to its stems, after which removal in previous step was carried out once more.

**Experimental Procedure and Results**

We cut off 10% newest papers each as validation and test set, and the rest as training set. In a single-pass experiment, 100 samples are randomly chosen from the validation set and recommended separately. A successful recommendation here refers to the appearance of the actual venue in the N recommended venues. Venue-Accuracy@N is selected as our evaluation metric and is calculated by averaging all results after several times of repeated experiments.

**Baseline: NGTC**

We adjusted some details in NGTC and use it here as a baseline because it was listed as a best performing method in [9]. The method is operated to the training set with \( N = 5 \) and slice lists truncated at 300. After all papers of the training set had corresponded profiles, we merged them into venue profiles with respect to the subordinative relationships.

Masses of experiments were carried out and the average values are displayed in TABLE 2. The scheme of indirect collecting candidate venues by comparing profiles between papers works barely satisfactorily, yet the other scheme of directly comparing the target paper to the venue profile performs poorly.

**Topic Coherence for Finding Number of Topics**

To simplify the massive calculations of topic coherence, we leveraged a toolkit named genism [14]. In our experiments, we picked some possible numbers as the inputs in different LDA models and trained them with exactly the same corpus. And for each model, the above-mentioned \( \text{C}_v \) was calculated as its topic coherence value. We plotted the results to observe how topic coherence values vary.

According to the definition of \( \text{C}_v \), an outstanding topic model holds a higher score of topic coherence than the unsatisfactory ones. As shown in Figure1, a model holding the amount of 40 topics had received a globally highest score, accompanied by some local maxima in the rear of the curve. Hence, we occupied 40 as the number of topics for the LDA model in the following experiments.

**Schemes of Generating Candidate Venues**

For each \( p \) sampled from the test set, its topic distribution was inferred by the LDA model chosen in previous section. A ranking by similarity value was thereupon obtained, with which we achieved the goal of generating N candidate venues for \( p \).

The method of collecting appropriate candidates differentiates our proposed schemes. In the first scheme of ours, we filtered the whole corpus with a similarity value working as the threshold. Then we checked the corresponding venue of each qualified paper and ranked the venues by frequency of occurrence. Selecting a certain threshold value greatly depends on the corpus itself. As shown in Figure 2, we did a series of experiments to test the performance of different threshold values. The values range from 0.05 to 0.95 with an interval of 0.05. As we can see through the charts, the accuracy fluctuates in all three lines while the threshold value rises. And even a great decline happens when the value rises to 0.95, due to the few candidate papers of an average amount of 7.53. This illustrates us that we must avoid endlessly raising the threshold value in order to pick most similar papers, because there could be necessarily without enough papers or even none fitting an extremely high threshold.
like 0.9 or more. We picked 0.7 for its persuasive high threshold value and requirement of much smaller amount of papers, around 1/3 and 1/10 of that required by 0.55 and 0.7 respectively.

As for the second scheme, the first encountered N venues are selected while traversing the list ranked by similarity in advance. Here we assigned N with 10 for further evaluation. With numerous experiments, we found that the number of papers involved in generating 10 venues is very close to an average of 14.25. Such number is relatively small considering of the weak interrelationship between papers and venues. This led to our thinking of applying additional constraints to weaken the randomness.

**Time Interval in Context Filter**

A crucial part in the definition of our context filter is interval, indicating the difference value of $p_a$ from the target paper $p_t$ that is used to calculate conditional probability $P(f|p_1, p_a)$. To clarify the reason of filtering papers using this context feature, we carried out some more experiments. As our dataset ranges in decades, we picked argument values at year level. In each independent experiment, we used only papers sharing an identical interval to generate candidate venues. Recommendations based on the second scheme mentioned above were performed and accuracies were calculated for assessment.

As we can see in Figure 3, a dramatic slide occurs with interval increases. Accompanied by weakened fluctuations, the curves are convergent to a small scale, demonstrating a trend of decrease in accuracy even though it seems to be not strictly resulted from intervals locally. This obvious deterioration in performance illustrates the requirement of constraint in context according to time interval.
Overall Evaluation

The experimental results of our methods are displayed along with those by some other previous papers in Table 2. We performed two schemes, “rank by quantity” and “rank by similarity”, based on the topic distribution representation inferred by an LDA model. As shown in the table, the first scheme slightly outperforms the second over all three indexes with a gap of around 7% each, smaller at Accuracy@10. While adding the context filter to the pure LDA, all indexes of both schemes welcome a tremendous promotion of 20%, more or less. However, the first scheme is not outperforming over all indexes any more. With a larger increment, the second scheme eventually gets a higher Accuracy@10 than the first one, a satisfying 77.2% over 68.2%.

Table 2. Experiment results.

| Methods            | N=3  | N=5  | N=10 |
|--------------------|------|------|------|
| venue profile      | 12.5 | 16.6 | 25.3 |
| paper profile      | 24.7 | 33.7 | 45.9 |
| rank by quantity   | 31.8 | 42.4 | 54.4 |
| rank by similarity | 25.0 | 35.6 | 50.7 |
| rank by quantity   | 51.1 | 60.3 | 68.2 |
| rank by similarity | 40.4 | 57.2 | 77.2 |
| Two-step-LDA       | 3.4  | 3.8  | 4.0  |
| LDA+clustering     | 16.1 | 21.7 | 33.2 |
| Cavnar-Trenkle     | 26.8 | 34.0 | 45.6 |
| CF-based           | n.a. | 55.7 | 69.8 |

Our results, with or without context filter, outperform all three methods given by a similar paper to ours which was also pursuing the goal of recommending with abstracts and took advantage of LDA to model the topics of abstracts. The other method of conducted analysis over full texts of papers and made use of citation information, while our method with context filter merely requires abstracts, getting preferable performance with much less information.

Conclusion

In order to assist the puzzled scientific researchers in accomplishing choosing a publication venue where his or her paper is most probable to be published. We proposed schemes for recommending venues based on topic modeling algorithms capturing the elusive topic related connections among papers. Instead of exploring a more efficient method of profiling a venue, we rely on the pair-wise similarities to observe the possibility of co-occurrence in a same venue. The highlight of our work is that our corpus is merely built with abstracts, which is much shorter than full texts in general schemes and such abstracts or perhaps brief introductions are more easily available during composing a paper.
Among the confirmatory experiments, we discovered the influence of publication date and thus a context filter was included to enlarge the contribution of recent papers. Experimental results proved that our best schemes managed to get even better performance than previous works which employed much more information.

References

[1] B. Meyer, C. Choppy, J. Staunstrup and J. van Leeuwen, “Viewpoint: Research Evaluation for Computer Science”. Common. ACM, vol. 52, pp. 31-34, Apr. 2009.

[2] M. Pham, D. Kovachev, Y. Cao, G. Mbogos and R. Klamma, “Enhancing Academic Event Participation with Context-aware and Social Recommendations”. International Conference on Advances in Social Networks Analysis & Mining, 2012, 159(3): 464-471.

[3] F. Xia, N. Asabere, J. Rodrigues, F. Basso and N. Deonauth, “Social-Aware Venue Recommendation for Conference Participants”. Proc.10th IEEE Int. Conf. Ubiquitous Intell. Comput., Dec. 2013: 134-141.

[4] M. Pham, Y. Cao, R. Klamma, “Clustering Technique for Collaborative Filtering and the Application to Venue Recommendation”. Proc. Of I-KNOW, 2010.

[5] H. Luong, T. Huynh, S. Gauch, P. Do and K. Hoang, “Publication Venue Recommendation using Author Network’s Publication History”. Asian Conference on Intelligent Information & Database Systems, 2012, 7198(6): 426-435.

[6] H. Luong, T. Huynh, S. Gauch and K. Hoang, “Exploiting Social Networks for Publication Venue Recommendations”. International Conference on Knowledge Discovery & Information Retrieval, 2012: 239-245.

[7] Z. Yang and B. Davison, “Venue Recommendation: Submitting your Paper with Style”. Machine Learning and Applications (ICMLA), vol. 1, pp. 681-686, IEEE, 2012.

[8] D. Blei, A. Ng and M. Jordan, “Latent Dirichlet Allocation”. Journal of Machine Learning Research, vol. 3, pp. 993-1022, 2003.

[9] E. Medvet, A. Bartoli and G. Piccinin, “Publication Venue Recommendation based on Paper Abstract”. IEEE International Conference on Tools with Artificial Intelligence, 2014: 1004-1010.

[10] W. Cavnar and J. Trenkle, “N-Gram-Based Text Categorization”. Symposium on Sdair, 1994: 161-175.

[11] G. Salton and M. McGill, “Introduction to Modern Information Retrieval”, McGraw-Hill, 1983.

[12] T. Hofmann, “Probabilistic Latent Semantic Indexing”. International Acm Sigir Conference on Research & Development in Information Retrieval, 1999, 42(1): 50-57.

[13] M. Roder, A. Both and A. Hinneburg, “Exploring the Space of Topic Coherence Measures”. Eighth Acm International Conference on Web Search & Data Mining, 2015: 399-408.

[14] Genism: topic modelling for humans, https://radimrehurek.com/gensim/.