International Journal of Scientific Research in Science, Engineering and Technology (www.ijsrset.com)
© 2020 IJSRSET | Volume 7 | Issue 3 | Print ISSN: 2395-1990 | Online ISSN : 2394-4099
DOI : https://doi.org/10.32628/IJSRSET207330

Analysis on Research Paper Publication Recommendation System with Composition of Papers and Conferences Matrices

Htay Htay Win[1], Aye Thida Myint[2], Mi Cho Cho[3]

[1]Faculty of Information Science/ University of Computer Studies (Taungoo) / Taungoo City, Bago Region, Myanmar

[2]Engineering Support Department, Technological University (Hpa-An) / Hpa-An City, Kayin State, Myanmar

[3]Faculty of Computing, University of Computer Studies (Sittway) / Rakhine State , Myanmar

ABSTRACT

For years, achievements and discoveries made by researcher are made aware through research papers published in appropriate journals or conferences. Many a time, established s researcher and mainly new user are caught up in the predicament of choosing an appropriate conference to get their work all the time. Every scientific conference and journal is inclined towards a particular field of research and there is a extensive group of them for any particular field. Choosing an appropriate venue is needed as it helps in reaching out to the right listener and also to further one's chance of getting their paper published. In this work, we address the problem of recommending appropriate conferences to the authors to increase their chances of receipt. We present three different approaches for the same involving the use of social network of the authors and the content of the paper in the settings of dimensionality reduction and topic modelling. In all these approaches, we apply Correspondence Analysis (CA) to obtain appropriate relationships between the entities in question, such as conferences and papers. Our models show hopeful results when compared with existing methods such as content-based filtering, collaborative filtering and hybrid filtering.

Keywords : Recommender Systems, Machine Learning, Dimensionality Reduction, Correspondence Analysis, Topic Modeling, Author Social Network, Linear Transformation.

I. INTRODUCTION

With the advent of the Internet and the growing amount of information available therein, people are increasingly resorting to finding information online. This in rotate has resulted in a handful of challenges, one of the principal a single for users being finding perfectly what they are looking for or for researchers to keep up to date on information of whose existent they may be unconscious and other.

In order to address this problem, we aim to build a recommender system that recommends the most appropriate publication venues for an author. This system is exceptionally useful to budding researchers who have very little knowledge about the research world and also to experienced researchers by saving a lot of their time and effort.

In this work, we aim to approach this problem in the settings of dimensionality reduction and topic modeling. We propose three different methods to recommend conferences for researchers to submit their paper based on the content of the paper and the social network of the authors: two of them involving content-analysis and the third one involving social network of the authors. Our approach is evaluated speculatively using the dataset of recent ACM conference publications and, to compare with existing methods such as content-based filtering, collaborative filtering and hybrid filtering with promising results.

However, there are several obstacles that need to be addressed in advanced. We list out the challenges along with the different claims from our work.

Challenges: We face several challenges when working in this domain, as illustrated.

(a) In all the previous work done related to our problem, only a model using the social network of the authors has been employed. Content analysis of the papers in consideration, to the best of the authors' knowledge, has never been attempted. Just using the network of authors, without even looking at the paper, is not sufficient to decide where the paper should go to. We do incorporate content into our work.

(b) Suggesting conferences to new authors is a very tricky business. If the author has not published any paper before, he does not have a social network. Hence, the current systems would yield a poor recommendation. We are considering content of the paper lead to better results.

(c) For the second method in our work, we construct a Paper × Words matrix and a Words × Conference matrix, where the (i,j)th entry of each of the matrices indicate the frequency of occurrence of wordj in paperi and wordi in the papers published in conferencej respectively. For the process of recommendation, we compose the two matrices to obtain a Paper × Conference on which we apply CA to proceed. But it is not guaranteed that the entries in the matrix obtained are 2**Main Claims:** The abstracts of the papers will be in consideration for content analysis. The challenges raised above are systematically addressed as follows:

(a) As suggested, just relying on the network of the authors is not sufficient to obtain a good recommendation of a conference. We bring in the content of the paper into our work to build a better model, which to the best of the authors' knowledge has not been explored before in the literature. Since the essence of the entire paper is contained within it's abstract, we build the content matrices using just the abstracts of the various papers. We employ term frequency-inverse document frequency (tf-idf) to generate the matrices of important keywords from the abstracts. In two of the methods, we construct Paper × Words and Words × Conference matrices using the above mentioned technique.

(b) It would be problematic on suggesting recommending conferences to authors with no prior social network. However, this problem might not arise during content-analysis as the author's social network is not in consideration. Just relying on the content of the abstract, we recommend suitable conferences. In our experiments, to suggest conferences to new authors, we observed that this method far supersedes the one relying on only his/her social network.

(c) In order to avoid such a confusion, our third method does not compose the two matrices. Instead a linear transformation is defined between the two spaces after reduction of dimension. In essence, after constructing the Paper × Words and Words × Conference matrices, we apply CA to each of them to reduce their dimension and then define a linear transformation from one subspace to the other for the process of recommendation.

3. Key tasks of the methods: The key tasks of each of the method proposed are as follows:

Method : Considering the content of the paper and composition of matrices.

• We construct a Paper × Words matrix and a Words × Conference matrix, where the (i,j)th entry of each of the matrices indicate the frequency of occurrence of word_j in paper_i and word i in the papers published in conference_j respectively.

• Then, we compose these two matrices and apply CA to obtain the principal column co-ordinates corresponding to the conferences.

- We obtain the principal row co-ordinates of the paper in need of a recommendation by computing it's tf-idf vector, composing with the Words × Conference training matrix and subsequent CA.
- The conference nearest to the paper in the bi-plot is recommended as the most suitable one.

II. Technical Approach

Different approaches can be taken to solve the considered problem of attempting to recommend conferences to authors. Outline of ideas are provided and their pitfalls, if any, are mentioned. This recommender system unlike most commercial ones like recommending books, movies, etc…, involves people in some sense. Thus, there is an emotional connection involved. What this means is, if a conference suggested by our system gets a paper rejected, it is highly unlikely that he will use this system again. This is not that case with books or movie recommenders. So, there is no room for errors and less accuracies. Some previous work on this has been done by H. Luong et al. [6] who have recommended conferences to authors using the social network i.e. the co-author network with the same dataset. Exploring the possibility of using CA has not been attempted before.

We have implemented method for this application and have done a comprehensive evaluation of the results. Three of the methods use Correspondence Analysis and three of them don't. The first method uses the Author-conference relation without taking into account the content of the paper. The next two methods use the content along with an application of CA to arrive at the results. The abstracts of the paper are used for content-analysis. This makes sense because the essence of the entire paper is contained in the abstract.

Content is obtained in two ways: term frequency-inverse document frequency (tf-idf) and topics. LDA has been used for the latter. For each content-method, number of topics used: 100, 200, 400, 600, 800 and 1000. However, only results for 400 topics are displayed in the evaluation, due to there being a very vast multitude of results and it would be too cumbersome to list all of them. Number of words used in tf-idf: 14082. For computing the resultant conferences, three methods of similarity have been used: euclidean distance, cosine similarity and pearsons correlation.

In method, 2008–2009 set of papers have been used for training and 2010 papers have been used for testing. There are a total of 5447 papers for the years 2008–2010, 3572 for 2008–2009 and 1875 for 2010. There are a total of 16 conferences.

The various similarity metrics used in the experiments are given below:

- Euclidean distance:

$$d(x,y) = \sqrt{\sum_{k=1}^{n}(x_k - y_k)^2}$$

where n is the number of attributes and x_k and y_k are the k^{th} attributes of the data points x and y, document respectively.

- Cosine Similarity: In this similarity measure, items are considered as n-dimensional vectors and their similarity is measured as the cosine of the angle that they form between them. Thus, if the cosine measure is close to 1, i.e. the angle between the two
- vectors is close to 0, the items are considered to be very similar.

$$\cos(x,y) = \frac{(\mathrm{x}.y)}{\mathrm{II}x\mathrm{II}\ \mathrm{II}y\mathrm{II}}$$

Where, indicates vector dot product and IIxII is the norm of vector x. This similarity is also known as the L_2 Norm

- Pearson Correlation: Correlation between items can also measure their similarity, linear relationship in this case. Although several correlation coefficients can be used, the most commonly used one is the Pearson Correlation. Given the covariance of data points x and y, Σ, and their standard deviation σ, we compute the Pearson correlation using:

$$Pearson(x,y) = \frac{\sum(x,y)}{\sigma_x \times \sigma_y}$$

A. Composition of Papers-Words/Topics and Words/Topics-Conferences Matrices

a. Data Construction

A way to remedy the defect in the previous method is to look at the content of the papers, abstracts in particular as they capture the entire essence of the paper. From the data collected, we can construct an *paper × words/topics* matrix and *words/topics × conferences* matrix as shown in Figure 1. We construct three matrices in total: two for training, and one for testing. We construct two training matrices, textitpaper × words/topics and *words × topics-conferences* from the 2008–2009 papers, say A_{train} (3572×14082) and Ctrain (14082×16). We also construct a test matrix A_{test} (1875×14082), *paper × words/topics,* from the 2010 papers, which contain all the papers which need recommendation. We write "word/topic" because the content is represented in both ways.

$$\begin{array}{c|cccc|} & \omega_1 & \omega_2 \dots & & \omega_L \\ a_1 & g_{11} & g_{12} \cdots & & g_{1L} \\ a_2 & g_{21} & g_{22} \dots & & f_{2L} \\ \vdots & \vdots & \vdots & & \vdots \\ a_N & g_{N1} & g_{N2 \dots} & & g_{NL} \end{array} \times \begin{array}{c|cccc|} & c_1 & c_2 \dots & & c_M \\ \omega_1 & h_{11} & h_{12} \dots & & h_{1M} \\ \omega_2 & h_{21} & h_{22} \dots & & h_{2M} \\ \vdots & \vdots & \vdots & & \vdots \\ \omega_N & h_{L1} & h_{L2 \dots} & & h_{LM} \end{array}$$

Figure 1: The Paper-Word and Word-Conference Matrices

Here, g_{ij} is the number of times author a_i has used the word w_j in all of his considered publications. h_{ij} is the number of times word w_i has been used in the conference c_j in total, i.e. considering all the papers that have been accepted in conference c_j, all of them combined use the word w_j, h_{ij} number of times. We generate the conference matrix by computing the centroid from those entries of the paper matrix which corresponds to this particular conference in [1], [2], [3], [4], [5].

b. Applied Method

The algorithm followed is given in the following steps:

- We multiply the training matrices, A_{train} and C_{train}, to obtain M_{train}. The result M_{train} is a paper × conference matrix.
- We compute the standardized residual matrix S_{train} from M_{train} as mention.
- We then obtain the coordinate matrices (both standard and principal for rows and columns), after decomposing Strain using SVD (Singular value decomposition).
- After this, we multiply the test matrix A_{test} with the training matrix $C_{train}$to obtain M_{test}.
- Using the matrix M_{test} as a supplementary row matrix, we compute its principal coordinates using the standard column coordinates of M_{train}.
- Then, for each paper in M_{test}, we compute its similarity with each of the conferences and sort the result.
- Then, for each paper in M_{test}, we compute its similarity with each of the conferences and sort the result.
- We then get a ranked list of recommendations for each paper.

In this method, we multiply the author-words and words-conference matrices and apply CA after that, to recommend a conference to an author. But, this may not capture the relations between the authors and conferences well. An alternative would be to reduce the author-words matrix and the words-conference matrix individually first. Then, defining a transformation from the first subspace to the other might help capture the relations better, which is the next method.

Instead of words, a paper can also be represented in terms of topics. This is more meaningful because if a paper is about information retrieval but does not have much of the IR jargon, then the chances of recommending an IR conference for this paper is less. But, if we capture the topics, then this solves that problem.

B. EVALUTION AND RESULTS

In this section, we detail the evaluation procedures and discuss the results obtained. We have used a metrics to evaluate the performance of the algorithms described above. They were applied on the ranked list of recommendations generated by the above methods: In this section, we detail the evaluation procedures and discuss the results obtained. We have used a total of 7 metrics to evaluate the performance of the algorithms described above. They were applied on the ranked list of recommendations generated by the above methods:

- Mean Precision at *K* (MP@*K*): The mean Precision at K for a set of queries is defined as the mean of the Precision at K values for each of those queries. Precision at *K*, *P(K)*, is defined as:

$$P(K)=\frac{\text{No.of relevant documents retrieved in the top K results}}{K}$$

- Mean Recall at *K* (MR@*K*): The mean Recall at *K* for a set of queries is defined as the mean of the Recall at *K* values for each of those queries. Recall at K, *R (K),* is defined as:

$$\text{R(K)}=\frac{\text{No.of relevant documents retrieved in the top K results}}{\text{Total number of relevant documents}}$$

- Mean Average Precision at *K* (MAP@*K*): Mean average precision at *K* for a set of queries is the mean of the average precision at *K* values for each of those queries.

$$\text{MAP} = \frac{\sum_{q=1}^{Q} \text{AveP}(q)}{Q}$$

where Q is the number of queries. Here *AveP*(*q*) is the average precision for the q^{th} query. Average precision is defined as:

$$\text{AveP} = \frac{\sum_{k=1}^{n} \text{AveP}(P(k) \times rel(k)\ )}{\text{no. of relevant documents}}$$

where rel*(k)* is an indicator function equaling 1 if the item at rank k is a relevant document, zero otherwise. *P(k)* is the precision at *k*.

- Mean Normalized Discounted Cumulative Gain at P (MNDCG@P): Discounted Cumulative Gain (DCG) at P is defined as:

$$DCG_P = \sum_{i=1}^{p} \frac{2^{rel_i} - 1}{log_2(i\ +\ 1)}$$

where $irel_i$ is the relevance score of result *i*. DCG uses a graded scale of relevance and this allows us to have preferences in the predicted results. Let us assume an ideal sequence of predicted results which would yield the maximum DCG_P. We call this the ideal DCG_P, denoted by $IDCG_P$. The normalized DCG_P, $NDCG_P$, is the ratio of the obtained $\text{DCG}_\text{P}$with that of the ideal $IDCG_P$. This would thus always yield a value between 0 and 1. The mean normalized DCG_P for a set of queries is then the mean of the $NDCG_P$ values for each of those queries.

- Mean Reciprocal Rank (MRR): The reciprocal rank of a query response is the multiplicative inverse of the rank of the first correct answer. The mean reciprocal rank is the average of the reciprocal ranks of results for a sample of queries Q:

$$\text{MRR} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{rank_i}$$

- Mean F-Measure at *K* (MF-M): The mean F-measure at *K* for a set of queries is the mean of the F-measures at *K* for each of those queries. F-measure is defined as the harmonic mean of precision and recall:

$$F=\frac{\text{2.precision.recall}}{\text{(Precision+recall)}}$$

This is the balanced F-score, where the weights of precision and recall in the harmonic mean are equal. We can also have cases of uneven weights.

- Mean *R-Precision* (MR-P): The mean R-Precision for a set of queries is the mean of the R-Precision values for each of those queries. R-Precision is defined as the Precision at *R*, where *R* is the number of relevant documents. At this position, the precision and recall values become equal.

For the experiments, we have chosen the value of *K* and *P* to be 5. This means that the measures are evaluated (which are @*K* and @*P*) considering only the top 5 of the returned results. For the purpose of calculating the metrics, we have defined relevant conferences in two cases:

1. A predicted conference is relevant if it is same as the actual conference the paper was originally published in (we have that information from the 2010 data set). For computing DCG in this case, the relevant conference (which is the original conference) is given a score of 1 and the rest are given scores 0.

2. A predicted conference is relevant if it belongs to the Special Interest Group (SIG) of the actual conference the paper was originally published in. For computing DCG in this scenario, the original conference is given a score of 2, the other conferences in the SIG are given a score of 1 as they are considered to be partially relevant. The rest of the conferences get a score of 0.

For calculating similarity to determine the ranking of the retrieved results, we have used three different metrics as previously mentioned:

- Euclidean Distance
- Cosine Similarity
- Pearson Correlation

TABLE I

Parameter	Parameter
Number of Iterations	1000
Dirichlet Prior α	0.5
Number of Topics	400
Number of Training Papers	3572
Number of Test Papers	1875

Table 1: Experimental Parameters for LDA

Earlier it was explained that the dimension of the lower-dimensional subspace for an I ×J matrix is ≤ min{I −1, J −1}. Since, we have only 16 conferences and more than 1000 papers, the minimum is always 15. Although the experiments were evaluated for more than one subspace, due to lack of space and vast multitude of results, we only show the results for a 10-dimensional subspace. We call this d in [13], [14].

The experimental parameters used for LDA. For tf-idf, 14082 words were used.

For displaying the results of the experiments, the following conventions are used:

- ✓ MAP@5: *Mean Average Precision at 5*
- ✓ MNDCG@5: *Mean Normalized Discounted Cumulative Gain at 5*
- ✓ MRR: *Mean Reciprocal Rank*
- ✓ MR-P: *Mean R-Precision*
- ✓ MF-M: *Mean F-Measure*
- ✓ MP@5: *Mean Precision at 5*
- ✓ MR@5: *Mean Recall at 5*

Method: Composition of Paper-Words/Topics and Words/Topics-Conference Matrices

Here we present the results for the method, which composes two matrices and reduces the dimension. We have two cases: one using tf-idf matrices and one using topic matrices. The results for both are given below:

- ✓ Case 1: Using tf-idf representation (14082 words). d = 10. The results are displayed in Table 2.
- ✓ Case 2: Using topic representation (400 topics). d = 10.

TABLE II

Metrics	Euclid		Cosine		Pearson	
	Actual	SIG	Actual	SIG	Actual	SIG
MAP@5	0.5800	0.7124	0.5937	0.778	0.5820	0.7616
MNDCG@5	0.6573	0.7213	0.6755	0.7571	0.6648	0.7452
MRR	0.9543	0.8475	0.6041	0.8545	0.5933	0.8507
MR-P	0.3781	0.7205	0.3829	0.7888	0.3696	0.7686
MF-M@5	0.2956	0.6910	0.3061	0.7477	0.3036	0.7356
MP@5	0.1773	0.6219	0.1836	0.6729	0.1821	0.6620
MR@5	0.889	0.777	0.918	0.841	0.910	0.8276

Table 2: Results for Method: Using tf-idf matrices, d = 10

TABLE III

Metrics	Euclid		Cosine		Pearson	
	Actual	SIG	Actual	SIG	Actual	SIG
MAP@5	0.3433	0.4801	0.3880	0.5820	0.3818	0.5715
MNDCG@5	0.4112	0.4861	0.4600	0.5476	0.4531	0.5383
MRR	0.3818	0.6330	0.4191	0.6616	0.4136	0.6584
MR-P	0.1975	0.5068	0.2261	0.5898	0.2218	0.5824
MF-M@ 5	0.2058	0.5025	0.2259	0.5662	0.2229	0.5534
MP@5	0.1235	0.4522	0.1355	0.5096	0.1337	0.4981
MR@5	0.6176	0.5653	0.6778	0.6370	0.6688	0.6226

Table 3: Results for Method: Using topic matrices, d = 10

Here, it is observed that using tf-idf representation for content outperforms its topic counterpart in [7].

III. CONCLUSION

Although each of the above methods has its own merits, from the results obtained we observe the following: Although each of the aforementioned procedures has its own advantages, from the surveys obtained, we observe the following:

The content-based methods proposed easily beat popular methods like collaborative filtering. This shows that for this system, considering content is vital. Computing similarities with content in hybrid filtering also did not prove to be very helpful, as the remainder of the procedure is identical to collaborative filtering.

Content-based filtering is seen to outperform the CA-based methods. This may be attributed to the fact that there is a certain amount of information loss during the dimensionality reduction phase, while content-based filtering utilizes the "pure" raw content.

In the results obtained, using tf-idf for content proved to be better than using topics. This may be due to considering a much larger number of words in tf-idf representation (14082) than it's topic counterpart (400). Also, the method of generating the topic matrices may have influenced the results.

Lastly, we observe that cosine similarity proves to be the best measure to calculate the similarities.

IV. REFERENCES

[1]. Xun Zhou, Jing He, Guangyan Huang, and Yanchun Zhang. A personalized recommendation algorithm based on approximating the singular value decomposition (approsvd). In Proceedings of the The 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology-Volume 02, pages 458–464. IEEE Computer Society, 2012.

[2]. Gabor Takacs, Istvan Pilaszy, Bottyan Nemeth, and Domonkos Tikk. A unified approach of factor models and neighbor based methods for large recommender systems. In Applications of Digital Information and Web Technologies, 2008. ICADIWT 2008. First International Conference on the, pages 186–191. IEEE, 2008.

[3]. Arkadiusz Paterek. Improving regularized singular [6] ON Osmanli and IH Toroslu. Using tag similarity in svd-based recommendation systems. In Application of Information and Communication Technologies (AICT), 2011 5th International Conference on, pages 1–4. IEEE, 2011.

[4]. Manolis G Vozalis and Konstantinos G Margaritis. A recommender system using principal component analysis. Current Trends in Informatics, 1:271–283, 2007.

[5]. Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. Application of dimensionality reduction in recommender system-a case study. Technical report, DTIC Document, 2000.

[6]. Hiep Luong, Tin Huynh, Susan Gauch, Loc Do, and Kiem Hoang. Publication venue recommendation using author network's publication history. In Intelligent Information and Database Systems, pages 426–435. Springer, 2012.

[7]. Eric J Beh. Simple correspondence analysis: a bibliographic review. International Statistical Review, 72(2):257–284, 2004.

Cite this article as :

Htay Htay Win, Aye Thida Myint, Mi Cho Cho, "Analysis on Research Paper Publication Recommendation System with Composition of Papers and Conferences Matrices", International Journal of Scientific Research in Science, Engineering and Technology (IJSRSET), Online ISSN : 2394-4099, Print ISSN : 2395-1990, Volume 7 Issue 3, pp. 120-127, May-June 2020. Available at
doi : https://doi.org/10.32628/IJSRSET207330
Journal URL : http://ijsrset.com/IJSRSET207330