International Journal of Scientific Research in Science, Engineering and Technology (www.ijsrset.com)
© 2020 IJSRSET | Volume 7 | Issue 2 | Print ISSN: 2395-1990 | Online ISSN : 2394-4099
DOI : https://doi.org/10.32628/IJSRSET207254

A Study on Publication Recommender System with Content Modelling

Htay Htay Win
Faculty of Information Science/University of Computer Studies (Taungoo)/ Taungoo City, Bago Region, Myanmar
htayhtaywin243@gmail.com

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 Modelling, Linear Transformation, Author Social Network, Content Modelling

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 [11].

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 based on the massive literature survey.

1. 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.

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) 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.

(b) Maximum essence of the relationship between the attributes in a table is obtained only in lower dimensional subspaces. Thus, when reducing the dimension of the matrices using CA, we essentially throw out the redundant information while maintaining the crucial and important part of them that are responsible for the relationships. As an added bonus, the reduced dimension increases the efficiency of the methods.

(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:

(a) Method 1: 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.

(b) Method 2: Considering the content of the paper and a linear transformation.

- We construct the Paper × Words and Words × Conference matrices as before, but instead of composing them, we reduce them to lower dimensional subspaces individually using CA.

- Then, the linear transformation from the reduced paper space to the reduced conference space will be defined.
- This prior step enables us to take a paper, in need of recommendation, to the space of conferences and suggest a conference closest to it.

II. DATA SET AND TOOLS USED

A. Data Used

Techniques based on the network analysis of authors and content analysis of the publications, have been explored for the purposes of recommendation. Each of the following subsections describes the data collected and techniques/tools applied on the data. For uniformity, we have used the publications in ACM conferences over the years 2008 to 2010. The selected conferences include

1. SIGBED - Special Interest Group on Embedded System
   - CASES - Compilers, Architecture, and Synthesis for Embedded Systems
   - CODES + ISSS - International Conference on Hardware/Software Codesign and Systems Synthesis
   - EMSOFT - International Conference on Embedded Software
   - SENSYS - Conference On Embedded Networked Sensor Systems
2. SIGDA - Special Interest Group on Design Automation
   - DAC - Design Automation Conference
   - DATE - Design, Automation, and Test in Europe
   - ICCAD - International Conference on Computer Aided Design
   - SBCCI - Annual Symposium On Integrated Circuits And System Design
3. SIGIR - Special Interest Group on Information Retrieval
   - CIKM - International Conference on Information and Knowledge Management
   - JCDL - ACM/IEEE Joint Conference on Digital Libraries
   - SIGIR - Research and Development in Information Retrieval
   - WWW - World Wide Web Conference Series
4. SIGPLAN - Special Interest Group on Programming Languages
   - GPCE - Generative Programming and Component Engineering
   - ICFP - International Conference on Functional Programming
   - OOPSLA - Conference on Object-Oriented Programming Systems, Languages, and Applications
   - PLDI - Programming Language Design and Implementation

All together there are 16 conferences, which are from the 4 special interest groups. SIGBED is special interest group on embedded systems and accepts contributions related to embedded computer systems including software and hardware. SIGDA is special interest group on design automation. It accepts contributions on design and automation of complex systems on chip. SIGIR accepts contributions related to any aspect of Information Retrieval (IR) theory and foundation, techniques and applications. SIGPLAN is special interest group on programming languages and accepts contributions related to design, implementation and principles of programming languages and others in [1], [3], [7].

B. Co-Author Network

We have downloaded the DBLP database, which contains the conference proceedings. This database contains the XML records of all the publications. Each record contains its publication information such as: author names, publication venue, title, year, and the DOI (Digital Object Identifier) of the publication. These attributes and generated a co-author network are extracted whereas each node in the co-author network represents an author and each edge represents the co-authorship between the author nodes and other in [2].

C. Content-Analysis

The ACM site provides abstracts for all the publications on its website. In order to perform content analysis, we crawled the ACM site and extracted the abstracts over the years 2008 to 2010 from the above-mentioned conferences. We extracted a total of about 5447 abstracts published in these conferences and used them for content-based recommendations and other in [5].

D. Tools Used

a. Neo4j Graph Database

For constructing the co-author network, Neo4j graph database has been used. It is an open-source project for graph databases. The python bindings were used to interact with the database.

MySQL Database : We relied on MYSQL to store the information on publications like year, DOI and venue [9], [5]. Programming Languages

Latent Dirichlet Allocation (LDA) was written in C++. All the other applied methods were written in Python and R [8].

III. CONCLUSION

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]. Marko Balabanovi'c and Yoav Shoham. Fab: content-based, collaborative recommendation. Communications of the ACM, 40(3):66–72, 1997.

[2]. Michael Pazzani and Daniel Billsus. Learning and revising user profiles: The identification of interesting web sites. Machine learning, 27(3):313–331, 1997.

[3]. 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.

[4]. Robert M Bell, Yehuda Koren, and Chris Volinsky. The bellkor solution to the netflix prize, 2007.

[5]. G'abor Tak'acs, Istv'an Pil'aszy, Botty'an N'emeth, 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.

[6]. Arkadiusz Paterek. Improving regularized singular value decomposition for collaborative filtering. In Proceedings of KDD cup and workshop, volume 2007, pages 5–8, 2007.

[7]. 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.

[8]. Ken Goldberg, Theresa Roeder, Dhruv Gupta, and Chris Perkins. Eigentaste: A constant time collaborative filtering algorithm. Information Retrieval, 4(2):133–151, 2001.

[9]. Manolis G Vozalis and Konstantinos G Margaritis. Using svd and demographic data for the enhancement of generalized collaborative filtering. Information Sciences, 177 (15):3017–3037, 2007.

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

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

[12]. Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. Incremental singular value decomposition algorithms for highly scalable recommender systems. In Fifth International Conference on Computer and Information Science, pages 27–28. Citeseer, 2002.

[13]. Panagiotis Symeonidis. Content-based dimensionality reduction for recommender systems. In Data Analysis, Machine Learning and Applications, pages 619–626. Springer, 2008.

[14]. Markus Zanker, Matthias Fuchs, Wolfram H¨opken, Mario Tuta, and Nina Mu¨ller. Evaluating recommender systems in tourism—a case study from austria. Information and communication technologies in tourism 2008, pages 24–34, 2008.

[15]. Steven Bethard and Dan Jurafsky. Who should i cite: learning literature search models from citation behavior. In Proceedings of the 19th ACM international conference on Information and knowledge management, pages 609–618. ACM, 2010.

[16]. C Lee Giles, Kurt D Bollacker, and Steve Lawrence. Citeseer: An automatic citation indexing system. In Proceedings of the third ACM conference on Digital libraries, pages 89–98. ACM, 1998.

[17]. Kannan Chandrasekaran, Susan Gauch, Praveen Lakkaraju, and Hiep Phuc Luong. Concept-based document recommendations for citeseer authors. In Adaptive Hypermedia and Adaptive Web-Based Systems, pages 83–92. Springer, 2008.

[18]. Xiang Chen, Cheng-Zen Yang, Ting-Kun Lu, and Hojun Jaygarl. Implicit social network model for predicting and tracking the location of faults. In Computer Software and Applications, 2008. COMPSAC'08. 32nd Annual IEEE International, pages 136–143. IEEE, 2008.

Cite this article as :

Htay Htay Win, "A Study on Publication Recommender System with Content Modelling", International Journal of Scientific Research in Science, Engineering and Technology (IJSRSET), Online ISSN : 2394-4099, Print ISSN : 2395-1990, Volume 7 Issue 2, pp. 399-403, March-April 2020. Available at doi : https://doi.org/10.32628/IJSRSET207254
Journal URL : http://ijsrset.com/IJSRSET207254