THE INTERNATIONAL JOURNAL OF
SCIENCE & TECHNOLEDGE

Extracting and Integrating Author Profiles from Research Publications

Tasleem Arif
Assistant Professor, Department of Information Technology, Baba Ghulam Shah Badshah University,
Rajouri, Jammu & Kashmir, India

Abstract:
In this paper we propose a mechanism for extraction and integration of research and work profiles of academic researchers based on a mix of a token-based and a character-based string similarity measure. The proposed technique uses Web mining techniques for extraction of publications metadata from digital libraries which is augmented from various online resources for the purposes of profile creation and integration. In particular we mine publications data from DBLP which is then populated with additional publications features obtained from the Web in a resource bound manner. On the basis of this populated publications data author profiles are extracted and integrated and presented them in terms of their important research attributes. These profiles present a clear picture of research output, research collaborations, research partners, frequency of publications with co-authors and shifting pattern of author affiliations, if any.

Keywords: *Web Mining, Profile Extraction and Integration, Virtual Environment, Digital Libraries, DBLP.*

1. Introduction

We live in an information age where everything has to be well defined so that it can be treated as a piece of information. In this contemporary world everyone wants to look different from the crowd and researchers are no exception. In order to define or describe a researcher in an easy and convenient manner, it is important to profile her on the basis of various research attributes. These attributes include frequency of publications, co-authorships and research collaborations, potential partners, frequency of publications with these co-authors, shifting of affiliations, etc. In a virtual environment where it is hard to differentiate between similar entities, profiling of authors is not a straightforward task and one needs to employ some disambiguation techniques to differentiate similar entities first. The widespread use of digital libraries and digital literature management has provided much needed basic resources for profile creation and integration. This data in itself may not be sufficient for profile creation and additional data has to be obtained from the Web.

Profiles help define persons, places, organizations, etc. In the era of Web 2.0 profiles have become a much more common term and one often finds profiles of Facebook users defining them in terms of their important attributes like name, place of living, education and work details, etc. The transition from casual facebook profiles to formal profiles in LinkedIn, Research Gate, etc. has been an instant phenomenon. We can find profiles of professionals and researchers on these websites defining their important features like name, work place and designation, education, publications, etc. It may seem unnecessary to have a new solution if we already have such services which provide detailed information about a researcher, but this is not a straightforward task. These websites require a user to register for their services, manually update his education and work information, and update his research credentials from time to time. Although this may seem to be a trivial task but in this busy world it may be unfair to expect every researcher to update his credentials from time to time in a disciplined fashion. Therefore, it becomes imperative to have an automated system that shall be capable of extracting and integrating researcher profiles from a small set of data sources.

It is important for organizations and interested parties, like funding agencies, to have almost an accurate idea of the research output of a person. Funding agencies need this information to evaluate the research potential of a prospective researcher, finding appropriate people/research groups to start a new research project, etc. In such cases, the profile of an author/researcher provides a crisp overview of the research credentials of that person. Of all the attributes, the work published by a researcher is the most important one. The importance of publications, either individually or jointly can be gauged from the fact that a number of studies, including [Luukkonen et al., 1992], [Georghiou, 1998], [Gla¨nzel, 2001] [Gla¨nzel, 2002], [Gla¨nzel and Schubert, 2004], [Zitt, and Bassecoulard, 2004], [Wagner and Leydesdorff, 2005a], [Wagner and Leydesdorff, 2005b], [Lorigo and Pellacini, 2007], [Wagner and Leydesdorff, 2008], [Vidican et al., 2009], [Chang and Huang, 2014], have used publications data of authors for understanding a number of research related facts and hence answering related questions.

The rest of this paper is organized as follows: Section 2 presents a brief idea about various elements that define a researcher; Section 3 presents the proposed profile extraction and algorithm, whereas Section 4 presents the experimental results. In section 5 we conclude the work and give some future directions.

2. What Defines a Researcher?

Number of publications produced during a particular period is one of the mostly commonly used attribute of researcher which can be obtained directly from publication records. But there are a host of other attributes which are present in publications and if extracted intelligently they can be used to build a profile which defines a researcher. A researcher goes through different phases of his life as any other normal individual does. These phases help define important attributes like education, employment, research, etc. However attributes that define a researcher include research publications, collaborations, employment details, patents filed and received, thesis supervision, etc. In this work we are interested in defining the profile of an author on the basis of some of the features extracted from the disambiguated publications data. These features include the total number of publications (*#TotalPublications*), affiliations, (*Works/WorkedFor*), co-authors (*Co-Authors*), publication frequency with each of the co-authors (*#PapersWith*) and recently published in (*RecentVenue*). A brief description of these attributes is as follows:

- Publications: The number of publications authored by a researcher that are indexed by the target source for a particular author.
- Affiliations: It is possible that a researcher may be associated with more than one organization and institutions. The prestige of an author can also be inferred from the prestige of institutions she has been associated with.
- Co-authors: The co-authors may be fellow colleagues, team members, mentors, students, etc. This attributes defines the co-authorship network of a researcher.
- Publication Frequency: There may be a number of co-authors of a researcher. The number of publications co-authored with each of them may differ. It is possible that the strength of relationship with one co-author or a group of co-authors may be more than that of others. This attribute quantifies the number of publications co-authored with each of the co-authors.
- Recent Venue: An author may publish in a various journals and conferences. This is called as publishing venue. The choice of publishing venue depends upon the quality of work carried out by the researcher. This venue is an indirect indication of the current level or quality of research work carried out by the researcher.

3. Profile Extraction and Integration

Majority of the sources which provide publications data contain raw information. This information has to be pre-processed before it can be supplied to a profile extraction and integration mechanism. This pre-processing is called name disambiguation whereby confusing publications are grouped into unambiguous publication clusters or groups. The proposed technique needs to make use of the disambiguated publications data of authors to integrate their profiles. For the purposes of creating unambiguous clusters the proposed technique uses a modified version of name disambiguation technique proposed in [Arif et al., 2014] for disambiguation of publications. The explanation of name disambiguation mechanism is out of scope of this paper.

Since publications data has been considered as a rich source of implicit information about a researcher, we extract some of the critical information like number and pattern of research publications and research collaboration, affiliations and shifting pattern, if any, etc. from them. This publication and profile information can also be used for extraction and analysis of academic social networks derived from this disambiguated publications data [Arif et al. 2012].

Figure-1 shows the architecture of the proposed system for extraction of previously defined elements of a researcher profile.

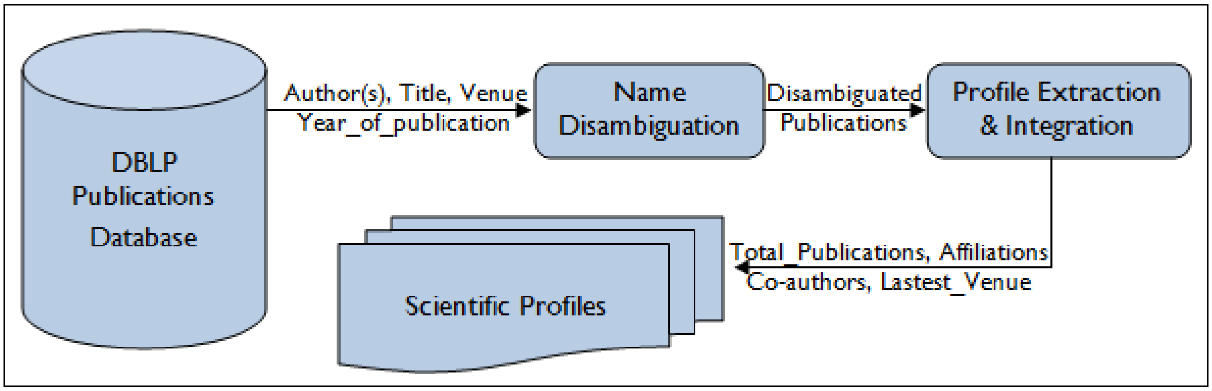

Figure 1: Architecture of the Proposed Profile Extraction & Integration System.

The algorithm used for profile extraction and integration is shown in Table 1. This algorithm extracts the previously defined attributes from disambiguated publications data and then integrates them in the form of a profile which can be presented in a tabular form.

Algorithm: Author Profile Integration			
Input:		Target-Author Name and Disambiguated Publications Data	
Output:		Profile of Given Target-Author Name	
	1.	[Extract ***#TotalPublications***] Extract total number of publications from the cluster for the current author.	
	2.	[Extract ***Works/WorkedFor***]: Extract and assign to a list all affiliations of Target Author from Disambiguated Publications Clusters.	
	3.		[Remove Duplicates from ***Works/WorkedFor***] Find and remove matching co-authors from the list generated in Step-2 using fuzzy string matching techniques (***Cosine Similarity*** in 1st phase and ***Jaro-Winkler Similarity*** in 2nd phase, if CosineSimilarity > threshold).
	4.	[Extract ***Co-Authors***] Extract all the Co-authors of Target Author from Disambiguated Publications Clusters.	
	5.		[Remove Duplicates from ***Co-Authors***] Find and remove matching co-authors from the co-author list generated in Step-4 using fuzzy string matching techniques (***Cosine Similarity*** in 1st phase and ***Jaro-Winkler*** Similarity in 2nd phase, if CosineSimilarity > threshold).
	6.	[Extract ***#PapersWith***] Extract the number of publications of the Target Author with each of the Co-Authors generated in Step-5.	
	7.	[Extract ***RecentVenue***] Extract the venue of the conference or name of the journal in which the Target Author has published most recently.	

Table 1: Algorithm for Profile Integration

4. Experimental Results

After disambiguation the web mined publications are provided to the proposed profile extraction and integration mechanism. The proposed technique extracts the intended attributes from the explicit and implicit information contained in the extracted publications. Some of these values are converted from qualitative to quantitative terms. Table-2 presents the extracted attributes for each of the authors with a given name. Each of these profiles contains some interesting facts about the profiles generated for each author using the proposed author profile extraction and integration approach.

Rashid Ali	2	Department of Biochemistry, Faculty of Medicine, Jawaharlal Nehru Medical College, Aligarh Muslim University, Aligarh.	Fozia Khan	1	Biochemistry (Moscow), 2006
			Zafar Rasheed	1	
			Mohd. Wajid Ali Khan	2	
Rashid Al-Ali	3	Division of Biomedical Informatics, Sidra Medical and Research Centre, Qatar	Shane R. Reti	3	MedInfo 2013
			H. Al-Jalahma	2	
			Henry J. Feldman	2	
			Charles Safran	1	
			Steven Bedrick	1	
Rashid Ali	3	Vishveshwarya Institute of Engineering and Technology, Dadri, G. B. Nagar, U.P. India	Anuj Bhardwaj	3	Journal of Wavelet Thoery & Applications, 2010
			Bani Singh	2	
Rashid J. Al-Ali	9	Department of Computer Science Cardiff University, UK	Sanjay Jha	1	Journal of Systems Architecture 52(2), 2006
			David W. Walker	5	
			Omer F. Rana	7	
			Abdelhakim Hafid	2	
			Gregor Von Laszewski	4	
			Kaizar Amin	4	
			Mihael Hategan	2	

Author	*#Total Publications*	*Works/ WorkedFor*	*Co-Authors*	*#PapersWith (Co-Authors)*	*RecentVenue (Conference/ Journal)*
Rashid Ali	23	1: College of Computers & IT, Taif University, Saudi Arabia 2: Department of Computer Engineering, A. M. U., Aligarh, 202002, India	Jamshed Siddiqui	3	International Conference on Contemporary Computing-IC3, 2014
			Shahab Saquib Sohail	3	
			Mirza Mohd. Sufyan Beg	9	
			Tanushi Vashishtha	1	
			M. Asger	4	
			Tasleem Arif	3	
			Supriya Kamthania	1	
			Om Prakash	1	
			Nesar Ahmad	2	
			S. M. Zakariya	3	
			Manzoor Ahmad Lone	1	
			Majid Bashir Malik	1	

Table 2: Profiles & Statistics for Author-Name 'Rashid Ali'

5. Conclusions & Future Directions

Profiles are one-stop shop for finding information about an entity. It may an organization, person, o a place. Profiles help answer number of questions related to credentials or attributes of an entity. In case of researchers profiles help in various different ways to funding agencies, peer group, prospective guides or students, prospective employers, etc. Research publications and research collaborations are an important indicator for defining the profile of a researcher. In this paper we proposed a method for profile integration and extraction from DBLP publications data. Various important attributes like co-authors, total_publications, works/worked_for etc. have been extracted from publications data using web mining techniques. These attributes are used for profile extraction and integration. These profile help answer various important questions about the researcher in question.

As a part of future work we intend to include other research and academic attributes for integrating the profiles of a researcher. In fact we intend to integrate their profiles from various online resources using Web mining techniques. As part of our future work we intend to integrate this profile extraction and integration system as part of our proposed academic social network system.

6. References

i. [Arif et al., 2012] Arif, T., Ali, R. and Asger, M. (2012). Scientific co-authorship social networks: A case study of computer science scenario in India.International Journal of Computer Applications, 52(12), pp: 38-45.

ii. [Arif et al., 2014] Arif, T., Ali, R., and Asger, M. (2014). Author name disambiguation using vector space model and hybrid similarity measures. In Proceedings of 7th International Conference on Contemporary Computing-IC3'2014, Noida, India: IEEE. pp: 135-140.

iii. [Chang and Huang, 2014] Chang, H.-Wen and Huang, M.-Hsuan (2014). Cohesive subgroups in the international collaboration network in astronomy and astrophysics. Scientometrics, 101(3), pp. 1587-1607.

iv. [Gla¨nzel, 2001] Gla¨nzel, W. (2001). National characteristics in international scientific co-authorship relations.Scientometrics, 51(1), pp. 69–115.

v. [Gla¨nzel, 2002] Gla¨nzel, W. (2002). Co-authorship patterns and trends in the sciences (1980–1998): A bibliometric study with implications for database indexing and search strategies. Library Trends, 50(3), pp. 461–473.

vi. [Gla¨nzel and Schubert, 2004] Glanzel, W. and Schubert, A. (2004). Analysing scientific networks through co-authorship. Handbook of Quantitative Science and Technology Research, Kluwer Academic Publishers.

vii. [Lorigo and Pellacini, 2007] Lorigo, L. and Pellacini, F. (2007). Frequency and structure of long distance scholarly collaborations in a physics community. Journal of the American Society for Information Science and Technology, 58(10), pp. 1497–1502.

viii. [Luukkonen et al., 1992] Luukkonen, T., Persson, O. and Sivertsen, G. (1992). Understanding patterns of international scientific collaboration. Science, Technology and Human Values, 17(1), pp. 101–126.

ix. [Vidican et al., 2009] Vidican, G., Woon, W. L. and Madnick, S. (2009). Measuring innovation using bibliometric techniques: The case of solar photovoltaic industry. Working Paper CISL# 2009-05, Massachusetts Institute of Technology, Cambridge, MA 02142, 2009.

x. [Wagner and Leydesdorff, 2005a] Wagner, C. S. and Leydesdorff, L. (2005). Mapping the network of global science: Comparing international co-authorships from 1990 to 2000. International Journal of Technology and Globalisation, 1(2), pp. 185–208.

xi. [Wagner and Leydesdorff, 2005b] Wagner, C. S. and Leydesdorff, L. (2005). Network structure, self-organization, and the growth of international collaboration in science. Research Policy, 34(10), pp. 1608–1618.

xii. [Wagner and Leydesdorff, 2008] Wagner, C. S. and Leydesdorff, L. (2008). International collaboration in science and the formation of a core group. Journal of Informetrics, 2(4), pp. 317–325.

xiii. [Zitt, and Bassecoulard, 2004] Zitt, M., & Bassecoulard, E. (2004). Internationalisation in science in the prism of bibliometric indicators. In H. F. Moed, W. Gla¨nzel, & U. Schmoch (Eds.), Handbook of quantitative science and technology research: The use of publication and patent statistics in studies of S&T systems (pp. 407–436), Dordrecht: Kluwer Academic Publishers.