CARP: Correlation-based Approach for Researcher Profiling

Hassan Noureddine
EDST
Lebanese University
Beirut, Lebanon
HES-SO/FR
Fribourg, Switzerland

Iman Jarkass
IUT
Lebanese University
Saida, Lebanon

Hussein Hazimeh*
HES-SO/FR
Fribourg, Switzerland

Omar Abou Khaled
HES-SO/FR
Fribourg -Switzerland

Elena Mugellini
HES-SO/FR
Fribourg –Switzerland

***Abstract*—The accelerating progress in science with the active role of the communication media – mainly the web – make person in front of a difficult task, in finding appropriate information during a brief time. In a narrower context, many researches were created in the expertise retrieval domain, as an interesting and complicated task for the scientific community, in face of this huge amount of data scattered across the web. Benefiting from the semantic web technologies and the efforts of data structuring, in this paper we propose a novel approach of correlation based profile building, by exploiting heterogynous web sources. The aim is to generate comprehensive and validated profiles about researchers and experts in the computer science domain.**

Keywords- Expertise Retrieval; Profile Matching; Profiling; quality of data; Semantic Web

I. INTRODUCTION

Since the web has established, it is still growing in a rapid manner. Where, every second millions of bytes are added around the world. Inconsistent with this growth, many web technologies have been emerged and participate in enhancing the efficiency of the web, like semantic web technologies. Therefore this massive growth forms the main motive to use web as a rich source of information and interactions. But at the same time, create more complex problems in the information retrieval domain. For instance, extracting specific and accurate web information must take into consideration the problems of conflicted, repeated and outdated data. In this context, the essential role that played by the web in the scientific progress, make the scientific community interested to solve this problem, especially in the profiling and expertise retrieval domains [1,8].

In this paper we proposed a Correlation based Approach for Researcher Profiling: CARP. The profiling task is going worsen with this massive and scattered amount of increased information across the world of web. As we proposed that we are going to cover the part of the problem relating to researcher profiling. The problem can be briefed as follows: if someone wants to search for a profile related to a specific researcher X, this will be a time-consuming process, especially there are no such standard sources that contain confirmed-content researchers' profiles. Even if we can find many systems as in [8,2,12] and others that provide scientific information related to researchers in several domains. However these data are still lacking to the quality in several cases. Cases lack such researchers' information, and others contain conflicted or outdated ones. Therefore we propose a new profiling approach based on correlating information from heterogeneous web sources, which contain confirmed data about researchers. The objective is to overcome the quality of data issue, and provide comprehensive and validated information about researchers, passing through a matching procedure.

In the rest of paper, the proposed approach is described as follow: The section 2 reviews and discusses the related work. The next section gives an overview on the proposed approach and describes the system architecture. The section 3 presents the obtained results, which are evaluated in the section 4. Finally, the paper is concluded in the last section.

II. RELATED WORKS

This section is composed of two parts. The first one mentions and explains the recent approaches in expert finding, and the other lists the latest approaches related to profile matching among multiple web resources.

A. Expert Finding Systems

The approaches submitted in this area are dealing with finding experts, where the most critical issue is what sources they are going to choose to find experts and create their profiles. The most popular system is Arnetminer, this system is based on finding and creating experts profiles in computer science domain and represents them semantically [8]. Microsoft Academic Search, another expert finding system, offers a diversity of functions for searching experts in several domains of sciences [2]. Other systems like INDURE, are limited to a set of organizations or universities, it provide functions for exploring profiles across these organizations in multiple disciplines [1,3]. The majority of the mentioned systems operate by extracting information from a single source, and even if some use multiple sources, they focus on a single source as the principal one compared to other sources. For example, Arnetminer is based on the home pages as source to extract the basic profile attributes, and then complete the profiles with the information extracted from DBLP [14]. While, our approach is to apply the concept of correlation between multiple web sources, leading to merge the discarded information in a unified profiles. Therefore, we consider that each source has his separate profiles, and all profiles from different sources must pass through a profile matching stage.

B. Profile Matching

Many approaches have proposed in this context and each one address this issue from his perspective, in this section we

DOI reference number: 10.18293/SEKE2015-145
*The first and third authors contributed equally to this work.

will focus on those who concentrate on web and social networks as a main source of information. In some approaches as in [4], they address this problem at the level of only two social networks, also they suppose that we have only one person profile among each social network, this approach and others use machine learning algorithms to resolve their decisions regarding the matching process. In [5], they proposed an expert finder system based on semantic matching between user profiles, they use the process of spreading to include additional related terms to a user profile by referring to an ontology (Wordnet or Wikipedia) [5]. Jain, Kumaraguru and Joshi [6] proposed an approach that matches profiles across Facebook and Twitter, by exploiting syntactic and image matching methods to discover the similarity between user profiles. In [7], they propose a vector based comparison algorithm that computes the similarity between two profiles according to their vector of attributes, and then classify whether they are the same or not based on a specific threshold. The mentioned approaches solve the problem partially. On the one hand they always apply the correspondence between social networks that are similar and almost have the same profile attributes. On the other hand they ignore the problem of name disambiguation by assuming that there is a unique profile for each person in different social networks. In contrast, we are working on matching profiles between multiple sources with different types, and we consider also the problem of name disambiguation by investing the detected similarity between profiles, as described in the next section.

III. PROPOSED APPROACH

Our proposed approach CARP is aiming to find a solution that addresses the problem of researchers' profiling, by benefiting from the heterogeneity of structured and unstructured data distributed across the web, this will carried out through a complete architecture composed of six main components as illustrated in figure 1.

Problem Definition and formulation: the main goal of CARP approach is to produce researchers' profiles by correlating information coming from several web resources. Let R_i be a specific web source (DBLP, MAS, LinkedIn), contains a set of profiles that belongs to a specific author name: $R_i = \{P_1, P_2, \ldots, P_n\}$, and each profile P_j contains a set of attributes $P_j = \{A_1, A_2, \ldots, A_n\}$, where $R_i.P_j.A_k$ is a specific attribute for a profile that belongs to a specific web resource. The aim is to find similar profiles among these sources by matching information extracted from their attributes, and then merge this information to produce complete profiles.

A. Ontologies

The initial stage in our architecture is to construct the system ontology. It covers all classes and properties describing the researchers' profiles, their relationships and their scientific products. It support and facilitates the information extraction and storage processes. Our ontology is based on the SWRC ontology (Semantic Web for Research Communities) [13], and it is composed of four major classes: the class person, document, education, position and organization, where each person (researcher) has a set of object and data properties. For instance a researcher has an education (PhD, Master, Bachelor), or he is an author for a document.

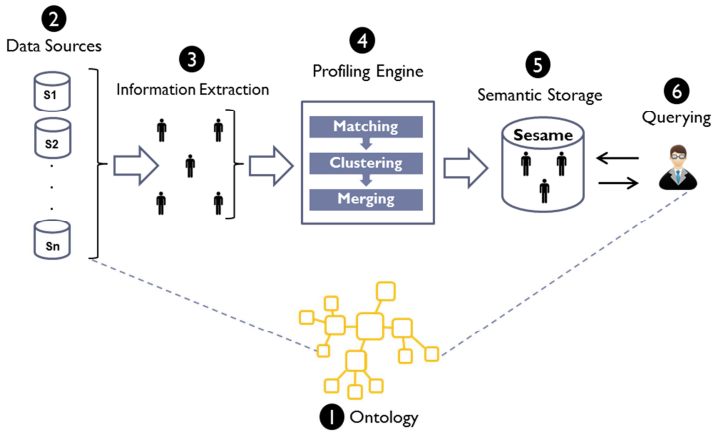

Figure 1. Architecture Components

B. Data Sources

Since our proposal is to apply the concept of correlation between multiple web sources, we have analyzed the data granted from different web sources. Then, we decided to use two types of sources: The bibliographic sources and the social networks. On the one hand, bibliographic sources provide essential information about researchers, their scientific activities and publications. On the other hand, there is a big trend to use social networks and especially professional networks. In this context we have chosen MAS, DBLP [2,14] as bibliographic sources and LinkedIn [9] as a social network.

C. Information extraction

The system starts operations with the structured information extraction, where the provided information are granted by the API of each source. However due to the limitation of the provided structured information, our system also extracts information from unstructured text, from home pages, publications and biographies. Two methods are used for this task. The first one is GATE (General Architecture for Text Engineering) as rule based method. It is used to extract the existing contact information (affiliation, email and location) from the publications headers. Thus GATE is suggested because it shows an average precision and recall of 90-95% on extracting contact information [10]. The second method is CRF (Conditional Random Fields), this method is employed to extract other attributes (education and the list of historical positions) from biographies existed in publications, homepages and LinkedIn profiles, by tagging them based on a built training set. We decide to use CRF, because it has lowest error rates for POS tagging compared to other methods [11]. Based on the chosen methods, the extraction process produces a set of preliminary profiles, presenting the attributes available in each source.

D. Profiling Engine

The main goal of this engine is to generate unified and confirmed profiles, passing through a correlation between the preliminary profiles. The correlation process is composed on three steps: matching, clustering and merging, as shown in the figure 2. The profile engine starts operating firstly with the matcher M1 that aimed at finding the similarity between

profiles from DBLP and MAS. We decide to use these two sources according to the permanent availability of two common profile attributes (affiliation and publication title), and to achieve this we have employ two string matching algorithms. We chose Jaro-Winkler (1) to calculate the similarity between affiliations, because Jaro-Winkler metric seem to be intended primarily for short strings (e.g., personal names), and Jaccard index (2) to calculate the similarity between publication titles.

$$Sim_{Jaccard}(R_1.P_2.A_2, R_2.P_3.A_2) = \frac{R_1.P_2.A_2 \bigcap R_2.P_3.A_2}{R_1.P_2.A_2 \bigcup R_2.P_3.A_2} \quad (1)$$

$$Sim_{JW}(R_1.P_2.A_2, R_2.P_3.A_2) = Sim_{Jaro}(R_1.P_2.A_2, R_2.P_3.A_2) + \frac{p'}{10}.(1 - Sim_{Jaro}(R_1.P_2.A_2, R_2.P_3.A_2)) \quad (2)$$

Let S_a be the similarity result of matching between two affiliations where $S_a = Sim_{JW}(R_1.P_i.affiliation, R_2.P_i.affiliation)$. S_p is the similarity result of matching between two publications titles where $S_p = Sim_{Jaccard}(R_1.P_i.Publication, R_2.P_i.Publication)$, and S_c is the similarity result of matching between two coauthors titles where $S_c = Sim_{JW}(R_1.P_i.coauthor, R_2.P_i.coauthor)$. Additionally, t_a , t_p and t_c are the threshold numbers, which represent the percent of matching between affiliations, publications and coauthors respectively, where $S_a >= t_a$, $S_p >= t_p$ and $S_c >= t_c$. The matching process between each profile from DBLP with each profile from MAS starts by comparing the list of publications, if the number of matched publications $>= t_p$ we decide that the two profiles belong to the same entity, else we continue the matching process by comparing the rest of publications using affiliation extracted from each publication, if the matching remains null we resolve the similarity based on the coauthors attribute. For each set of matched profiles we create a cluster C_i, and populate each matched profile to its parent cluster. After obtaining set of clusters, each cluster must undergoes to a merging operation, this step aims at unifying the set of profiles in each cluster into one profile and validate its attributes by applying several merging rules for each attribute.

After finishing the first correlation by matching (M1), clustering and merging between DBLP and MAS, we obtain a set of unified profiles. These profiles will act as an input for the matcher M2 that aim to complete the profiling operation by complementing the rest of profile attributes from LinkedIn. Each unified profile will be matched by a set of LinkedIn profiles using three attributes: affiliation, publication and education. The matching process starts by comparing publication titles, if there is common publication between two profiles we decide that the two profiles belong to the same person, else we compare the affiliations if there are the same we decide that the two profiles are for the same person, else we compare the list of education organizations to detect the similarity between two profiles. In this case, deciding whether two profiles belong to the same person will be easier, because the LinkedIn data are typed by the users themselves, and consequently there is an absence of the name disambiguation problem. This is the reason to not repeat the clustering method, and merging directly the matched profiles into the final unified profiles (the output) as shown in the figure 2.

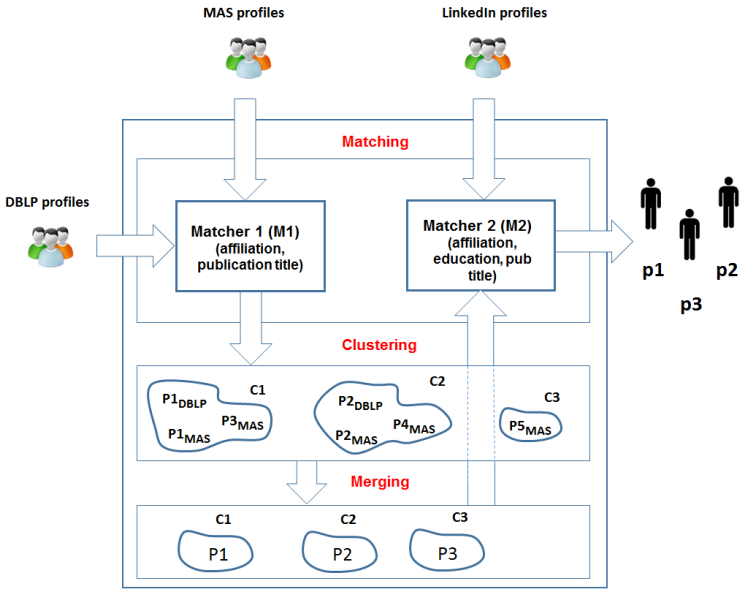

Figure 2. Profiling engine steps

E. *Semantic storage*

The Semantic Web technologies enhance the ability to discover relations between properties more than current traditional databases, thus we propose to store the extracted profiles in a semantic database in form of RDF triples according the system ontology.

F. *Quering*

The final step in our architecture is to retrieve information about researchers, where the query will be a researcher name. The query language used for this task is SPARQL.

IV. PERFORMANCE EVALUATION AND RESULTS

Referring to the architecture described in figure 1, we have implemented the various system elements, and thus provided web interface for receiving user requests and respond with relevant results. The prototype of our architecture is implemented using JavaEE, where all the tests are performed on Intel 2.93 core i7, 8GB of RAM PC.

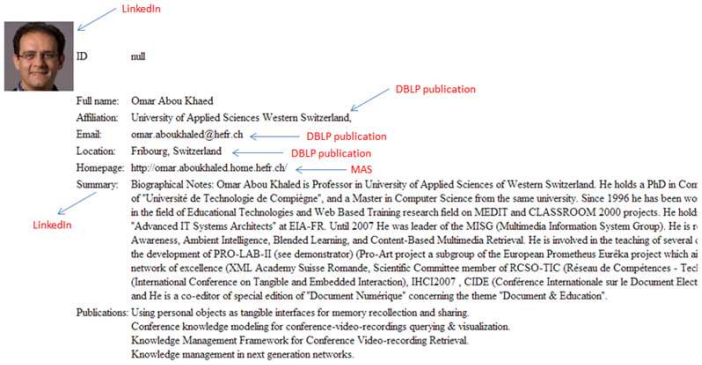

Figure 3. An example of researcher profile

The figures 3 present an example of profile generated by the system. We can see the benefit of the correlation, mainly by

obtaining comprehensive and confirmed profile, with attributes retrieved from various sources. The obtained results show that the same attribute is not always recovered from the same source, so that the missed attribute from some source can be provided in the other. This increases the possibility of retrieving information. For instance, the attribute "summary" are extracted from LinkedIn, and in case of absence, it can be extracted from the biography inside publications.

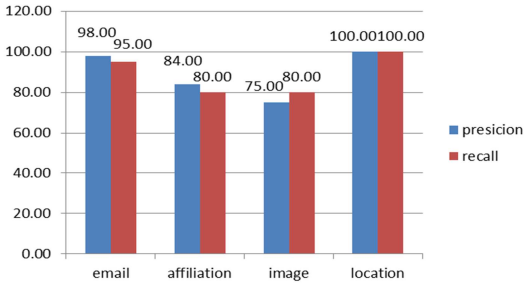

Figure 4. Precision and recall measures for each attribute

The figure 4 presents the precision and recall for four different attributes among 25 tested profiles, which they form the principal profile attributes. Based on these measures we are analyzed the results, thus the attributes "email" and "affiliation" are extracted from publication using GATE. Hence the strength of the precision and recall depends partially on the accuracy of GATE, and as we observe that we are obtaining 98 percent and 84 percent for email and affiliation respectively. The affiliation is sometimes failed to be extracted by GATE because of some problems. For instance, the language in which the affiliation is written, however in our case we are considering only the English language. The attribute "image" is extracted from three different resources: biographies, LinkedIn and MAS, resulting a precision of 75 percent. However we still need a strong face recognition method to validate the correctness of this attribute. Finally, the attribute "location" is extracted from two different sources publications and LinkedIn, this attributes has 100 percent of both precision and recall because location names are easy to be validated due to their limitation unlike affiliation and other attributes.

Additionally, the study of the availability of each attribute before and after the correlation has proved the efficiency in increasing it, especially for the attribute not strongly available. For instance the "image" attribute as shown in the Table I.

TABLE I. AVAILABILITY OF IMAGE ATTRIBUTE BEFOR AND AFTER CORRELATION

	Image from biography(publication)	Image from LinkedIn	Image from MAS
Before	53%	38%	57%
After	89%		

On another side, our approach was able to address the issue of name disambiguation in a low proportion, by benefiting from the partitioning of profiles among resources, which allows us to detect the diversity between profiles. Table II shows four profiles with name disambiguation tested between DBLP and MAS. This issue is directly affected by the resolution rate of this problem by each source. In LinkedIn, it does not exist because users enter information by themselves. In MAS the problem is opposed, where we can find several profiles for the same researcher. Therefore the problem must be resolved in DBLP, where the disambiguation exists in various cases.

TABLE II. NAME DISAMBIGUATION RESULTS TESTED ON FOUR DIFFERENT AUTHOR NAMES

Author name	Num. of MAS profiles	Num. of DBLP profiles	Actual Num. of profiles	Num. of profiles after merging
Kai Eckert	4	2	2	2
Hong Shen	11	1	4	3
Michael Wagner	1	3	12	4
Feng Liu	1	1	4	2

V. CONCLUSION AND FUTURE WORK

In this paper we present CARP approach, which based on the concept of correlating information from several web resources, to satisfy the production of qualified profile information, our investigated approach shown promised results. Moreover, this approach has overcome the problem of name disambiguation in some cases by benefiting from the variety of profiles among the different sources. However we still need a strong approach addressing this problem, and as a future work, we propose to add a name disambiguation block aiming to split the target profiles before the correlation.

REFERENCES

[1] K. Balog, Y. Fang, M. de Rijke, P. Serdyukov and L. Si. Expertise Retrieval.(2012)

[2] (2014,July). Microsoft Academic Search [Online]. Available: http://academic.research.microsoft.com/.

[3] E. A. Jansen. A Semantic Web based approach to expertise finding at KPMG.(2010).

[4] Raad, E., Chbeir, R., Dipanda, A. User Profile Matching in Social Networks. In NBiS, 2010.

[5] Thiagarajan, R., Manjunath, G. and Stumptner, M. Finding experts by semantic matching of user profiles. Technical Report, HP Laboratories. (October 2008).

[6] P. Jain, P. Kumaraguru, and A. Joshi. @ i seek 'fb.me': identifying users across multiple online social networks. IW3C2, 2013.

[7] Vosecky, J.; Hong, D.; and Shen, V. Y. User identification across multiple social networks. In Int. Conference on Networked Digital Technologies (2009).

[8] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su. ArnetMiner: Extraction and Mining of Academic Social Networks. In SIGKDD 2008.

[9] (2014,July). LinkedIn [Online]. Available: https://www.linkedin.com/.

[10] H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. GATE: An architecture for development of robust HLT applications. In ACL, 2002.

[11] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Eighteenth international Conference on Machine Learning, 2001.

[12] (2014,July).Google scholar. Available: https://scholar.google.com

[13] (2013,July) The Semantic Web for Research Communities Ontology (SWRC). [Online]. Available: http://ontoware.org/swrc/

[14] (2014,July). DBLP XML dataset [Online]. Available: http://dblp.uni-trier.de/xml/