Ontology Based System for Expert Searching in Academia using SWRL and SPARQL

Furqan Hussain Essani[1] Quratulain Rajput[2]

Abstract

Searching an expert with relevant experience and expertise is an important and a challenging task in academics. A lot of work has been carried out in this regard, however the semantic web technologies for modelling the information to search an expert is not being explored extensively. This paper proposed an ontology based system to search an academic expert of a particular field of study. The system comprised of the ontology which consists of an academic contribution of an individual. Additionally, SWRL (Semantic Web Rule Language) rules were created based on the academic contribution as publications made by an individual to infer their field of expertise. Finally, the SPARQL queries were performed to search an expert. This research developed a tool to experiment the proposed system and used IEEE explore Digital Library to retrieve academic contribution of an individual. The ontology based system of an expert searching is foundan efficient in terms of reducing the data modelling cost and making the system easily extendable and reusable for other applications.

Keywords: Expert searching, Ontology, SPARQL, SWRL.

1 Introduction

Expert finding is a challenging problem that has practical applications in many fields. Academia is one such domain where the question of finding an expert on a particular topic frequently arises. The expertise of an individual in academia is primarily judged by the individual's contribution in the form of his/her published work. Due the explosive growth of information over web made it extremely difficult to search an expert from unstructured and heterogeneous web sources. Therefore, several platforms have already been introduced to provide experts information such platforms are DBLP[3], CiteSeer[4], ACM digital Library[5], IEEE explore[6], Google Scholar[7], and so on. In contrast to traditional keyword based searching from unstructured sources these platforms provide structured representation of information of academic contribution of an individual. However, still, none of these platforms used semantic technologies to represent semantic information for searching experts.

In the recent years, the semantic web technologies have emerged as a much needed platform that provide structured information with semantically rich data model. Moreover, semantic technologies overcome the problem of data integration as well as provide easily scalable and reusable data modelling technique[1]. Ontologies as data model is one of the key

[1 2] Institute of Business Administration. Karachi, Pakistan

[1]fessani@iba.edu.pk | [2]qrajput@iba.edu.pk

[3]www.dblp.com

[4]www.citeseer.com

[5]www.acm.org

[6]www.ieee.org

[7]Scholar.google.com

components of semantic technologies. RDF, RDFS and OWL are Ontology languages from low high semantic expressivity respectively. In ontology, information is represented as statement that consists of a subject, object and predicate. The collection of these statements makes data model that has potential to directly apply logical inference in the data model to infer new knowledge without any pre-processing task.

The task of expert finding gained significant attention and a lot of work has been suggested in this regard after the inclusion of expert search task in the TREC Enterprise track[2]. The objective of expert finding task is, given a query, to find out a ranked list of experts. Some of the closely related works include fine-grained expert search model by Bao[3], topic-based and language hybrid model by Deng[4] and ranking workgroup members using citation analysis by Bogers[5]. Similarly social networks have also been used for expert searching for which notable work includes ArnetMiner[6].

All the above mentioned work is focused on the use of probabilistic approach and revolves around the traditional text mining techniques. However, not much attention is given to the semantics of the information. This paper proposed ontology based system for searching experts that can be integrated and reused across different applications built using semantic technologies.

The rest of the paper is organized as follows. Section II provides an overview of the related work while Section III describes the proposed methodology of ontology based system for searching experts and section IV describes the application to understand the potential of the presented work. Finally Section V, concludes the paper and provides future research directions.

2 Related Work

The problem of expert searching in academics has been addressed with different perspectives. Most of the work carried out in this regard use traditional probabilistic based approach. Deng[4] in his work suggested three different models for achieving the task. One of these models assigns a prior probability to a document to signify its importance and impact. The document prior probability is used along with the query-topic relevancy to rank a particular expert. In his other model he considered the fact that an expert may have expertise in various fields and hence ranking of expert is based on the aggregate expertise on various topics. His last model uses a hybrid approach by combining the two models.

Another probabilistic based work, carried out by Bao[3] proposes a more specific evidence oriented model. The model extracts evidence when a topic and a person with a specific relation are found in a given document. The evidence is then evaluated on different measures and each collected measure is then used to score and rank an expert.

The work carried out by Mangaravite in [7] to find expert in academia suggests modelling the document-person association as an estimate rather than a Boolean variable. The suggested method introduces probability of a document being informative of the expertise of the author.

A seminal work that uses semantic approach is presented by Nazimuddin[8]. In his work he has suggested indexing of academic information from ontological perspective. The approach combines academic ontology with the academic social network and finds the score of an expert based on his/her academic contribution and relationships with other experts.

However this model, like others, uses feature vector representation of an academic topic and cosine similarity measure to calculate similarity between the given query and the topic. In this suggested work, this task is simplified by representing the topic with a set of keywords and writing Semantic Web Rule Language (SWRL) rules to identify the similarity [6]. This suggested model is applied to find the experts in different fields of Computer Science.

The work presented in this paper is similar to the Nazim Uddin's work [5] in the use of ontology, however, in contrast to calculate similarity between topics, the novelty of our approach lies in the publications retrieval mechanisms from IEEE explore. Further the set of keywords in each publication were used to create SWRL rules to infer the experts of the field.

3 Ontology Based System For Experts Searching

In this section we have discussed the research that provides ontology based system for expert's searching. The system utilized ontology that incorporated academic information as well as SWRL rules to infer experts from academic information. The main components of the presented system as shown in Fig. 1 comprised of the following three major steps:

A. Ontology construction for academic information
B. Ontology reasoning
C. Ontology Querying

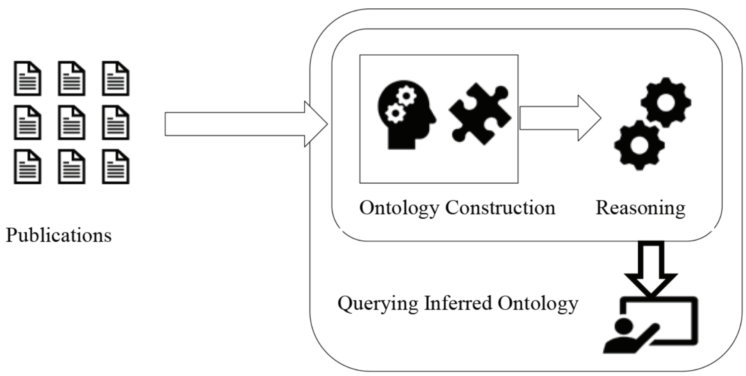

Figure 1: Ontology based System for Expert Searching

A Ontology Construction for academic information

In this component we semantically stored all the academic information of an individual. The research publications of an individual could be utilized to categorize the person's expertise into different fields. For this purpose, we suggested to retrieve the set of publications given fields as keyword on the available online digital libraries. Then each of the retrieved publication associated with that field in the ontology and the author of the publication will be considered an expert of the field. This ontology model consists of several classes, object properties, data type properties, and domain range restrictions as discussed below:

Classes

In order to semantically store the academic information from publication's details, there was a need to specify several classes in ontology those are commonly used in the publication.

Publication: This class includes the set of published articles in the different subject areas of computer science.

Journal: A journal is a type of publication. Therefore, it is a subclass of Publication.

Conference Proceeding: A conference proceeding is a type of a publication. It is a subclass of Publication.

Book: A book is a type of publication. It is a subclass of Publication.

Researcher: This class includes the set of people who have made contribution in the subject areas through their published work.

Field: This class includes set of different subject areas in computer science domain.

Computer Vision: Computer Vision is one of the fields of Computer Science.

NLP: Natural Language Processing is one of the fields of Computer Science.

Image Processing: Image Processing is one of the fields of Computer Science.

Similarly, other subject areas with their sub-files can be added to extend the ontology knowledge base.

Object properties

The relations (object properties) in the ontology describe how the classes and their individual members are related to each other. In this case we have defined five different relations as explained below and also shown in Table 1.

hasFirstAuthor: A publication (p) has a first author whose contribution is more significant than the other contributing authors

isFirstAuthorOf: A researcher who has contributed as a first author to a publication (p). It is an inverse relation of hasFirstAuthor

isCoAuthoredBy: A publication (p) may have contributing authors besides the first author

hasCoAuthored: A researcher who has contributed as a co-author to a publication (p). It is an inverse relation of is Co Authored By

Include: This relation is used to identify all the publications relevant to a particular subject area.

Table 1: Ontology Properties

	Property Name	Domain	Range
Object Properties	hasFirstAuthor	Publication	Researcher
	isFirstAuthorOf	Researcher	Publication
	isCoAuthoredBy	Publication	Researcher
	hasCoAuthored	Researcher	Publication
	include	Field	Publication

Data type properties

Data type properties describe more semantic information as literal that is associated with the particular class in the ontology. Following are some data type properties described in ontology. Table 2 lists the data type properties with domain range restrictions.

hasTitle: Each publication has a title, so it is stored in the ontology. The title is stored as a string.

hasKeyword: Each publication is associated with set of keywords. Each keyword is stored as a string.

hasLastName: Every researcher has a last name. The last name is stored as a string.

hasAffiliation: A researcher has an affiliation with any academic or professional organization. This data property is stored as a string.

hasEmail: A researcher has an email address which is used for correspondence over internet. This data property is stored as a string.

Table 2: Data type Properties

	Property Name	Domain	Range
Data Properties	hasTitle	Publication	String
	hasKeyword	Publication	String
	hasLastName	Researcher	String
	hasAffiliation	Researcher	String
	hasEmail	Researcher	String

In the presented work we considered three fields of computer science that include Computer Vision, Natural Language Processing (NLP) and Image Processing. Instead of manual tagging of publication with the field we suggested retrieving the publications from IEEE explore based on the provided field. For example, we searched IEEE explore for the list of publications by providing NLP keyword. Thus, it is assumed that this list of publications associated with the field NLP.

Further the association of field with publication also introduced the association of set of keywords in publication with the field. This would help to describe the set of keywords related to particular field.

The ontology used to create the data store is shown in Fig. 2. In this figure the circles represent the classes in the ontology, the arcs connecting those circles represent the relations between these classes and the rectangles represent the literal values of the classes.

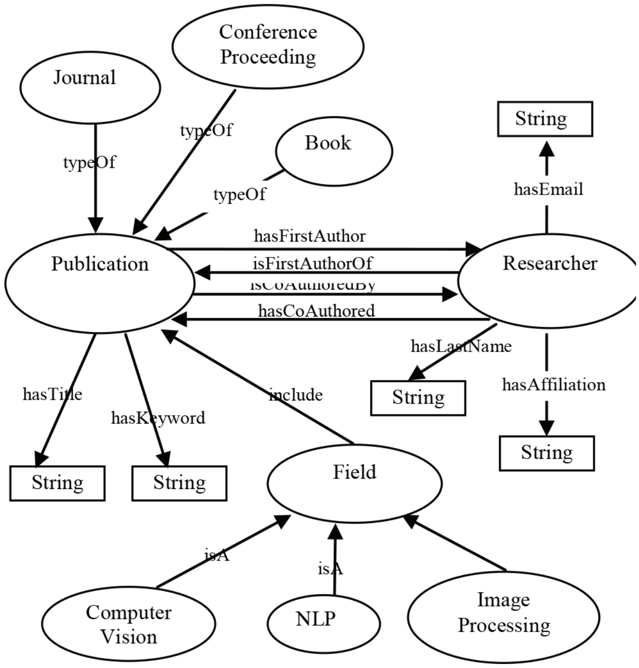

Figure 2: Ontology Diagram

In this approach there is a possibility that one keyword may be represented in more than one field of Computer Science. For example, as shown in Table 2, the keyword, "feature extraction" is being used to represent all the three fields under our consideration.

The set of keywords for different fields will form an overlapping set as shown in Fig. 3.There will be some keywords that will represent only a single field, some keywords that will be common to any two fields, while some will be common to all three fields.

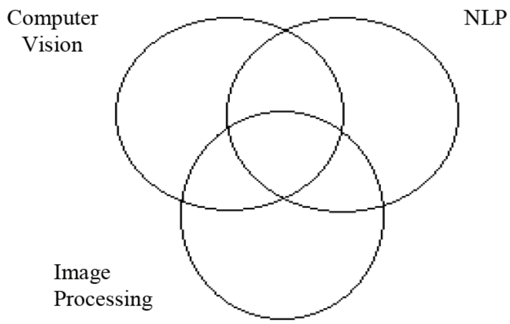

Figure 3: Overlapping set of keywords associated with fields

Table 3: Selected Keywords representing fields in Computer Science

	Field of Computer Vision	Field of NLP	Field of Image Processing
Keywords	Feature extraction	Feature extraction	Feature extraction
	Computer vision	Sentence completion	Accuracy
	Image block representation	Lexical disambiguation	Biomedical imaging
	Pattern recognition	Ngram	Breast cancer
	Vision development	Opinion mining	Image segmentation
	Vision analysis	Recommendation systems	SVM
	HCI	User comments	Fuzzy logic
	Image analysis	Machines learning	Signal processing

B Rules Construction and Ontology Reasoning

Rules help in deducing knowledge that is present in the data set but cannot be expressed through the ontology language. The Semantic Web Rule Language (SWRL) is used to construct the rules which are built on top of the ontology. Using the SWRL rules, a publication is assigned to belong to a particular field of Computer Science based on the keywords. Depending on the scope of the publication, one publication may belong to different fields at the same time.

For example, a publication with the keyword “Feature extraction” or “HCI” or “Pattern recognition” or “Vision analysis” would belong to the field of Computer Vision in our data store. Any publication with keyword “Ngram” or “Opinion Mining” or “Machine learning” would belong to the field of Natural Language Processing. And a publication with the keyword “Accuracy” or “SVM” or “Fuzzy logic” would belong to the field of Image Processing.

The set of rules for these assignments is given in table 3. The corresponding rules in SWRL rules are mentioned in table 4. Similar rules can be written for other keywords forming the set representing different fields.

Table 4: Rules to assign publication and it’s keyword to a field

Rule No.	Rule Description
1	A publication belongs to the field of Computer Vision if it has a keyword of “Feature extraction”
2	A publication belongs to the field of Computer Vision if it has a keyword of “HCI”
3	A publication belongs to the field of Computer Vision if it has a keyword of “Computer vision”
4	A publication belongs to the field of Computer Vision if it has a keyword of “Vision analysis”
5	A publication belongs to the field of NLP if it has a keyword of “Feature extraction”
6	A publication belongs to the field of NLP if it has a keyword of “Opinion mining”
7	A publication belongs to the field of NLP if it has a keyword of “Machine learning”
8	A publication belongs to the field of Image Processing if it has a keyword of “Accuracy”
9	A publication belongs to the field of Image Processing if it has a keyword of “SVM”
10	A publication belongs to the field of Image Processing if it has a keyword of “Fuzzy logic”

C Ontology Querying

The query component is used to search the expert from the ontology data store. In this component we have used SPARQL(Semantic Protocol and RDF Query Language) queries to retrieve the data from the data store [7]. Discussed in detail in the experiment section.

4 Experiment

A Data Collection and instance population in ontology

There are various academic databases available that store articles which are published in journals and other different repositories. In this paper, we have restricted ourselves to only those academic publications that belong to the area of computer science. Some well-known academic data bases in the area of computer science include the ACM Digital Library from ACM [8], the CiteSeerX provided by the Pennsylvania State University[9], the DBLPfrom University of Trier[10] and the IEEE explore Digital Library by IEEE[11].

For proof of concept, we developed a tool to create the data store. We have used publications retrieved from the IEEE explore Digital Library. IEEE explore Digital library is one of the largest resources in the computer science domain with nearly 3.7 million entries as of now. The information extracted from each publication includes names of authors, affiliation, emails and the set of IEEE keywords associated with the publication.

As seen in Figure 4 we have added instances of different publications few of them listed here, namely p1, p2, and p9. The instance p1 has first author a1 and has two keywords of "computer vision" and "feature extraction"

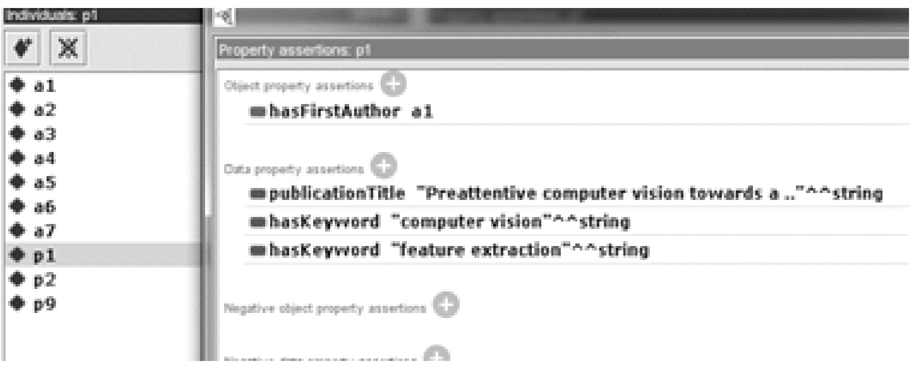

Figure 4: Instances of Publications

B Reasoning with SWRL rules

After populating the instances in the ontology with various publications details, the SWRL rules given in Table 4 were used to infer their membership to different fields. Since the set of

[8] *http://dl.acm.org/*
[9] *http://citeseerx.ist.psu.edu*
[10] *http://dblp.uni-trier.de*
[11] *http://ieeexplore.ieee.org*

keywords describing the fields were overlapping, one publication may belong to more than one field at one time.

As seen in Figure 5 and in Figure 6 the publication p1 was assigned to the fields of Computer Vision and NLP based on rules 3 and rule 6 given in table 4.

Table 4: SWRL Rules to assign publication to a field

Rule No.	Rule Description
1	Publication(?p), hasKeyword(?p, “Feature extraction")→ComputerVision(?p)
2	Publication(?p), hasKeyword(?p, “HCI")→ComputerVision(k2?p)
3	Publication(?p), hasKeyword(?p, “Computer vision")→ComputerVision(?p)
4	Publication(?p), hasKeyword(?p, “Vision analysis")→ComputerVision(?p)
5	Publication(?p), hasKeyword(?p, “Feature extraction")→ NLP(?p)
6	Publication(?p), hasKeyword(?p, “Opinion mining")→ NLP(?p)
7	Publication(?p), hasKeyword(?p, “Machine learning")→NLP(?p)
8	Publication(?p), hasKeyword(?p, “Accuracy")→ImageProcessing(?p)
9	Publication(?p), hasKeyword(?p, “SVM")→ImageProcessing(?p)
10	Publication(?p), hasKeyword(?p, “Fuzzy logic")→ImageProcessing(?p)

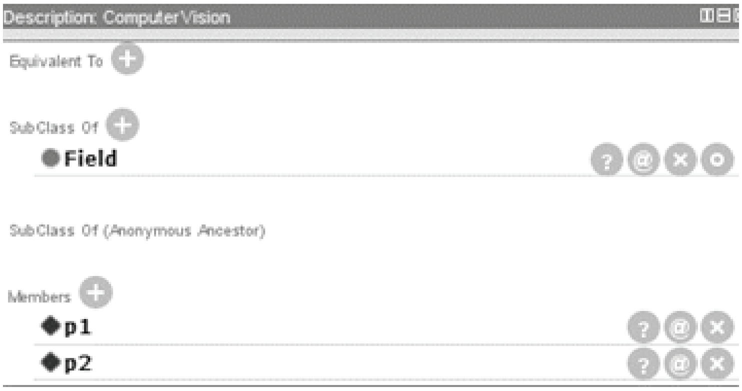

Figure 5: Publication p1 inferred to belong to field of Computer Vision

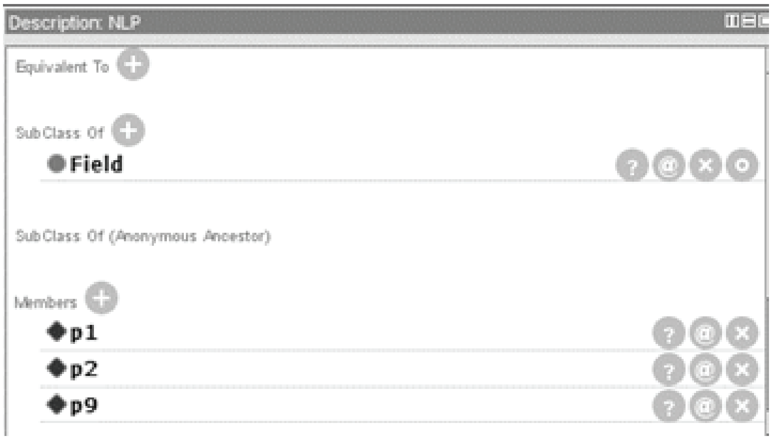

Figure 6: Publication p1 inferred to belong to the field of NLP

C Querying for Expert Searching

The experts were retrieved according to their contributions as a first author of the publication in each field of computer science. For retrieval purpose, SPARQL query was used which first retrieves all the publications depending on the user given query field and consider their first author as experts in that field.

In Figure 7 we have made query to get the list of experts in the field of Computer Vision. The query first evaluates all the publications for the given field and then lists down the name of first author for each publication. Along with the author name, other details like the affiliation and the email address can also be retrieved by extending the same query.

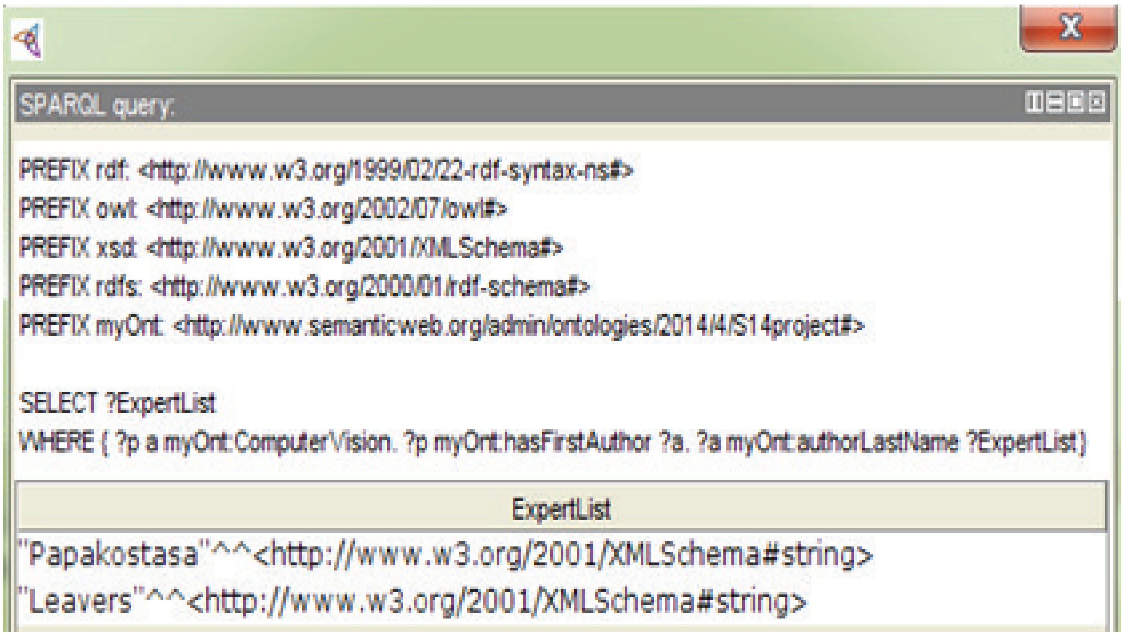

Figure 7: Query to list experts in the field of Computer Vision

5 Conclusion

The expert searching model proposed in this paper used ontology based data model that is well defined, structured and semantically rich academic information. Publications data retrieved from IEEE explore Digital Library for the experiment potential of proposed approach. The different fields in the area of computer science are represented using a set of keywords and the publications were assigned to an individual field and used SWRL using publication to infer expert of the field. The expert in a field is considered to be a researcher who has a contributed work as a first author. Based on the user query a list of experts was generated using the publications that belong to the field that has been queried. This research, further extends in future by ranking the authors in a particular field.

References

[1] N. Shadbolt, T. Berners-Lee, and W. Hall, "The Semantic Web Revisited," IEEE Intell. Syst., vol. 21, no. 3, pp. 96–101, Jan. 2006.

[2] "Text REtrieval Conference (TREC) Home Page." [Online]. Available: http://trec.nist.gov/. [Accessed: 25-Dec-2017].

[3] S. Bao, H. Duan, Q. Zhou, M. Xiong, Y. Cao, and Y. Yu, "A Probabilistic Model for Fine-Grained Expert Search.," in ACL, 2008, pp. 914–922.

[4] H. Deng, I. King, and M. R. Lyu, "Formal models for expert finding on DBLP bibliography data," in Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on, 2008, pp. 163–172.

[5] T. Bogers, K. Kox, and A. van den Bosch, "Using citation analysis for finding experts in workgroups," in Proc. DIR, 2008, pp. 21–28.

[6] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su, "Arnetminer: extraction and mining of academic social networks," in Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, 2008, pp. 990–998.

[7] Vitor Mangaravite and Rodrygo L. T. Santos, "On Information-Theoretic Document-Person Associations for Expert Search in Academia," presented at the SIGIR '16, Pisa, Italy, 2016.

[8] T. H. D. Mohammed Nazim Uddin, "EXPERTS SEARCH AND RANK WITH SOCIAL NETWORK: AN ONTOLOGY-BASED APPROACH," Int. J. Softw. Eng. Knowl. Eng., vol. 23, no. 01, 2013.