ISSN: 2277-9655
Impact Factor: 1.852

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

A Survey of Expert Finding in Academic Social Network

S. Krishnaveni[*1], K. Sathiyakumari[2]
[*1]Research Scholar, PSGR Krishnammal College for Women, India
[2]Assistant Professor, PSGR Krishnammal College for Women, India
skvmphil12@gmail.com

Abstract

Social networks place an important role in sharing knowledge, retrieving information from various websites.Recent studies suggest that an increasing participation of people in online activities like content publishing, different kinds of relationships and interactions among people in online social network web sites. Web Data Extraction is an important problem that has been studied by means of different scientific tools and in a broad range of application domains. This survey aims at providing a comprehensive overview of the research efforts made in the field of Profile Extraction from the Academic Social Network. In this paper tried to review some of the accomplished research of expert finding and profile extraction. The contribution of this paper is based on the extraction of social networks and a research framework for analyzing the experts in specified topics and co-author relationships in researcher network using various algorithms and tools.

Keywords: Social network, ASN, ArnetMiner, CRF method, ACT model, TPFG model.

Introduction

The Online Social Network is an ongoing trend, where the people increasingly reveal their personal information. The social relationships between people can be identified by recent initiatives such as Facebook's connect MySpace's data availability and Google's Friend Connect by making their social network data available to anyone. Extraction and mining of academic social network aims to provide comprehensive services in the scientific research field.

The extraction of academic network is used for research trend detection/ tends prediction. Trend detection can help a researcher to analyze the thrust area of particular field, and also used to analyze what other researchers are doing in that or related field. Trend prediction can help a research community to predict an idea of the potential research topics/areas in a particular field.

In an academic social network, the people are not only interested in searching for different types of information (such as authors, conferences, and papers), but are also interested in finding semantics-based information (such as structured researcher profiles). This survey includes: 1) extraction of researcher profiles from the Web, 2) integrate the researcher profiles and publications, 3) simultaneously find expertise objects (of different types) on a topic, and 4 find associations between researchers.

The paper is organized as follows: In Section 2, we give a survey of academic social network. In Section 3, we review the related work. We conclude the paper in Section 4.

A Survey of Academic Social Network

A. *Introduction of Social Network*

Data mining is a process of extracting information from the large database. The social network mining is one of the ongoing research trends in web mining.

Web mining is the Data mining technique that automatically discovers or extracts the information from web documents. It is used to extract an interesting and potentially useful patterns and hidden information from activity related to the World.

A social network is a social structure made up of individuals (or organizations) called "nodes", which are tied (connected) by one or more specific types of interdependency, such as friendship, kinship, common interest, financial exchange, dislike, sexual relationships, or relationships of beliefs, knowledge or prestige.

The social network analysis [14] refers the mapping and measuring of relationships and flows of information between people, organization, computers or other information or knowledge processing entities.

ISSN: 2277-9655
Impact Factor: 1.852

B. Methods for Expert Finding

Data mining is a wide spread process that happens in various aspects of life. It is not the only way to analyze extracted data. The extracted data can be turned into a graph, which represents the structural meaning of the data through the use of vertices, edges and weights. Online social networks provide great graph representation as well as data mining opportunities for a variety of people in different fields.

Challenges arise when parsing the webpage data. Some of the challenges are listed below:

a) OSNs contain data with a wide variety of formats e.g. contact lists, photos, videos, etc. When coding for extraction, the formats have to be taken into account. Different formats have different properties and behave in different ways.

b) In OSN user can customize their profile and this can cause problems because customization means adding various effects to an already dynamic web page.

This survey includes the accomplished methods for expert finding, researcher profile extraction and co-author relationships. The methods are as follows:

1) Conditional Random Fields (CRF)
2) Propagation Based Approach
3) ArnetMiner
4) Time-constrained Probabilistic Factor Graph model (TPFG)

Related Work

B. ***Conditional Random Fields (CRF)***

Conditional random fields (CRFs)[3] are a probabilistic framework for labelling and segmenting structured data, such as sequences, trees and lattice. The primary idea is that of defining a conditional probability distribution over label sequences given a particular observation sequence, more a joint distribution over both label and observation sequence.

The Conditional Random Fields (CRF) (Jie Tang, Duo Zhang, and Limin Yao, 2005) is used to extract the academic researcher information from the social network. The contributions in this method include: (1) formalization of the problem of researcher network extraction, (2) proposal of a unified tagging approach to researcher profiling, (3) and proposal of a constraint based probabilistic model to name disambiguation.

There are three steps in this approach: relevant page identification, researcher profiling, and publication integration.

In relevant page identification, while giving a researcher name, we first get a list of web pages by a search engine (Google API) and then identify the homepage/introducing page using a classifier.

In researcher profile extraction, the proposed method called unified approach. This approach can incorporate dependencies between different types of profile properties to do better extraction. In publication integration, the proposed method called constraint based probabilistic model to name disambiguation.

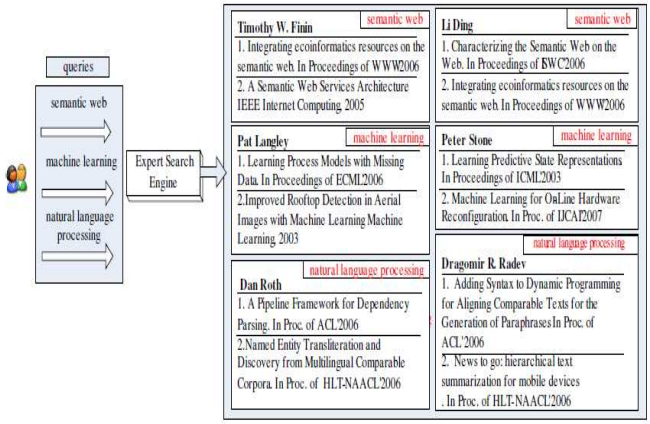

Fig 1. An example of Expert finding

1) A unified approach to profiling

The approach consists of two steps: preprocessing and tagging. In preprocessing, (A) separate the text into tokens and (B) assign possible tags to each token. In this tokens form the basic units and pages form the sequences of units in the tagging problem. In the tagging model, they make use of Conditional Random Fields (CRFs).

(A). Identify tokens in the Web page by using heuristics. There are five types of tokens: 'standard word', 'special word', '<image>' token, term, and punctuation mark. Standard words are unigram words in natural language. Special words \ include email, URL, date, number, percentage, words containing special symbols (e.g. 'Ph.D.' and '. NET'), unnecessary tokens (e.g. '===' and '###'), etc. Identify special words by using regular expressions. '<image>' tokens are '<image>' tags in the HTML file. We identify it by parsing the HTML file. Terms are base noun phrases extracted from the Web pages.

(B). Assign possible tags to each token based on the token type. For special word, we assign tags: Position, Affiliation, Email, Address, Phone, Fax, and Bsdate, Msdate, and Phddate. For '<image>' token, there are two tags: Photo and Email. In this way, each token can be assigned several possible tags. Using the tags, we can perform most of the profiling processing.

ISSN: 2277-9655
Impact Factor: 1.852

2) A constraint-based probabilistic model to name disambiguation

The method is based on a probabilistic model using Hidden Markov Random Fields (HMRF) [8]. This model [14] incorporates constraints and a parameterized-distance measure. The disambiguation problem is direct as assigning a tag to each paper with each tag representing an actual researcher. Specifically, a posteriori probability aims to optimizing the objective function. They incorporate six types of constraints into the objective function. If one paper's label assignment violates a constraint, it will be penalized in turn affects the disambiguation result.

All these constraints are defined between two papers. The first constraint *CoOrg* means the principal authors of two papers are from the same organization. Constraint *CoAuthor* means two publications have a secondary author with the same name, and the constraint *Citation* means whether a paper cites another paper. Constraint *CoEmail* means whether principal authors of the two publications have the same email address (this is a stronger constraint than the others) Constraint *Feedback* denotes user interaction and final constraint τ-*CoAuthor* one common author in τ extension.

Figure 1 is an example of expert finding. The left part of the figure describes three queries: semantic web, machine learning, and natural language processing and the right part of the figure shows experts for each query.

C. ***Propagation Based Approach***

The Propagation based approach (Jing Zhang, Jie Tang, and Juanzi Li, 2007) [15] is used to find the person local information and relationships between persons in a unified approach. And also used for finding expert in a social network. The approach consists of two steps. In the first step, person local information is used to estimate an initial expert score for each person and select the top ranked persons as candidates. The selected persons are used to construct a sub-graph.

In the second step, the propagation-based approach, this propagates one's expert score to the persons with whom he/she has relationships.

In Initialization, the person local information to calculate an initial expert scores for each person. The basic idea in this stage is that if a person has authored many documents on a topic or if the person's name co-occurs in many times with the topic, then it is likely that he/she is a candidate expert on the topic. The method calculates the initial expert scores is based on the probabilistic information retrieval model. For a person, first create a 'document' by combining all his/her person local information. Then estimate a probabilistic model for each 'document' and use the model to calculate the relevance score of the 'document' to a topic. The score is then viewed as the initial expert score of the person.

In Propagation, make use of relationships between persons to improve the accuracy of expert finding**.** The vital idea here is that if a person knows many experts on a topic or if the person's name co-occurs in many times with another expert, then it is likely that he/she is an expert on the topic.

Figure 2 shows a snippet of the academic researcher network. In the network, each person has several types of local information, for example, personal profile, contact information, and publications. Two persons can have relationships with each other. The relationship can be directional or bi-directional.

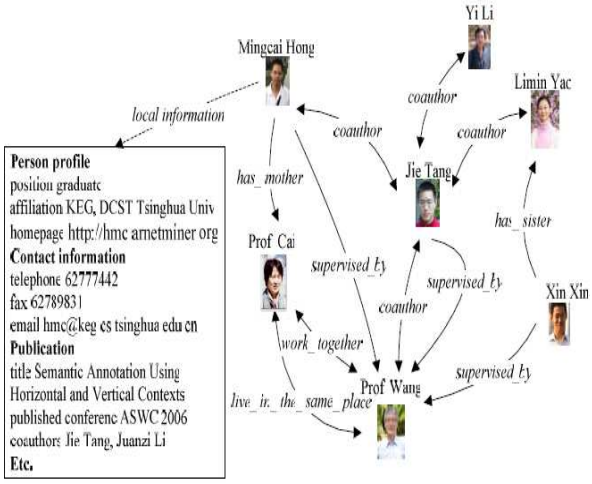

Fig2. An example of academic researcher network

D. ***ArnetMiner***

AretMiner (Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, 2008) [14] is a tool for expert finding. It consists of five main components:

1. *Extraction*: it focuses on automatically extracting the researcher profile from the Web.
2. *Integration*: it integrates the extracted researcher profiles and crawled publications.
3. *Storage and Access*: it provides storage and indexing for the extracted/integrated data in the RNKB.
4. *Search*: it provides three types of searches: person, publication, and category based searches.
5. *Mining*: it provides mining services, e.g., expertise search on a given topic and people association finding.

1. Researcher Profiling

The researcher profile [15] schema is extended by the FOAF ontology. In the profile, 24 properties and two relations are defined. It is non-trivial to perform the profile extraction, as the layout and content of the

ISSN: 2277-9655
Impact Factor: 1.852

researcher homepages/ introducing pages may vary largely depending on the authors. Several research efforts have been made for extracting person profiles.

For evaluating the unified profiling method is used. Randomly chose 1,000 researcher names from ArnetMiner and conducted human annotation. Experimental results show that the proposed approach can achieve a performance of 83.37% on average in terms of F1- measure, against Support Vector Machine based method (73.57%) and Amilcare (53.44%).

2. Expertise Search

The goal of expertise search is aimed at answering: "Who are experts or which are expertise of conferences/papers on topic X?". Here the problem viewed as a ranking problem using either language model to directly calculate the relevance random walk model to estimate importance of each object. They Latent Dirichlet Allocation-style model [13], called Author-Conference-Topic (ACT) model to model the dependencies between different types of objects in the researcher network.

In the ACT model, for each paper, an author is first drawn from a uniform distribution; a topic *z* is then drawn from a mixture weight of the chosen author and a distribution from a symmetric Dirichlet prior; next a word is generated from the topic *z* and a conference stamp is generated from the topic *z*. In this way, the dependencies between different types of objects are modeled using the topic.

Another advantage of the model is that we can use this model to capture the 'semantic'/hidden relevance between the query and the target objects. After applying the ACT model to the research network, again employ a random walk model on the heterogeneous network and finally output a combined score for each object to the query.

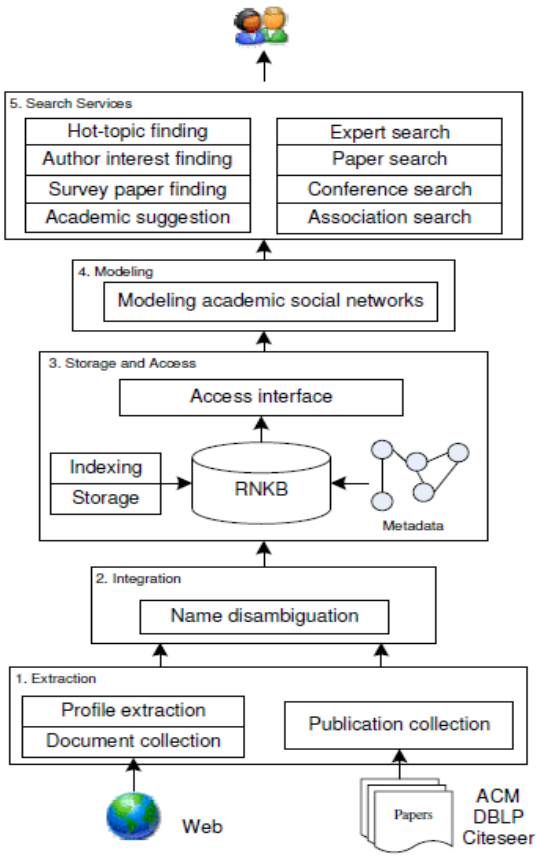

Fig 3. Architecture of ArnetMiner

Conduct experiments on Arnetminer with seven queries and compared the results with two baselines of using language model and Page Rank method, as well as results of two existing systems (Libra and Rexa). Experimental results show that the proposed method outperforms them from 4.26% to 29.2% in terms of MAP.

3. Association Search

Finally, the problem of association search: finding connections between researchers. The formalized association search as that of near-shortest paths and use a two stage approach to deal with it. First, employed a shortest path search to find shortest path from all persons in the network to the target person and then we use a depth-first search method to find top *K* ranked results. This method can find the top *K* results in 2-5 seconds for a general query on the social network with about half million researchers and 1 million publications.

ISSN: 2277-9655
Impact Factor: 1.852

E. ***Time-constrained Probabilistic Factor Graph model***

The Time-Constrained Probabilistic Factor Graph model (TPFG) (Chi Wang, Jiawei Han, Duo Zhang, 2012) is used to model the Dynamic collaboration network. Specifically, the advisor of each author and the advising period are modeled together as a joint probability of as many hidden variables as authors. First make basic assumptions as the prerequisite of this approach, then propose a two-stage framework and present the approach for each stage.

The main idea is to leverage a time-constrained probabilistic factor graph model to decompose the joint probability of the unknown advisor of every author. The Time-related information associated to the hidden social role is captured via factor functions, which form the basic components of the factor graph model.

By maximizing the joint probability of the factor graph we can infer the relationship and compute ranking score for each relation edge on the candidate graph. One can apply general algorithms for inference on factor graph, e.g., sum-product and Junction Tree. However, these algorithms undergo the problem of low efficiency. So the new message-passing algorithm on the candidate graph is designed that approximates the computation and greatly improves the efficiency.

Data Sets. This model uses the DBLP Computer Science Bibliography Database maintained by Michael Ley as the dynamic collaboration data set G to infer the advisor-advisee. It consists of 654,628 authors and 1,076,946 publications with time provided (from 1970 to 2008). To test the accuracy of the discovered advisor-advisee relationships, this approach adopts three data sets: One is manually labeled by looking into the home page of the advisors, and the other two are crawled from the Mathematics Genealogy and AI Genealogy.

We refer to them as MAN, MathGP and AIGP respectively. They only poetically cover the authors in DBLP. Further separate MAN into three sub data sets: Teacher, PhD and Colleague. Teacher contains all kinds of advisor-advisee pairs, while PhD only contains graduated PhDs pairing with their advisors. Colleague contains colleague pairs, which are negative samples for advisor-advisee relationship. And we use these data to generate random data sets for test.

Method. We compare the proposed TPFG with the following baseline methods:

- Sum-Product+Junction Tree (JuncT). It computes the exact joint probability as the ranking score.
- Loopy Belief Propagation (LBP). It employs an approximate algorithm for inference.
- Independent Maxima (IndMAX). It computes the maximal local likelihood for each variable independently.
- SVM. It is a supervised approach and requires labeled pairs, both positive and negative, as training data.
- RULE. For each author, from all the collaborators that satisfyAssumption 2, choose the one with most coauthored papers.

Figure 4 gives an example of advisor -advisee relationship analysis on a research publication network. The left figure shows the input: a temporal collaboration network, which consists of authors, papers, and paper-author relationships. The middle figure shows the output of our analysis: an author network with solid arrow indicating the advising relationship, and dotted arrow suggesting potential but less probable relationship.

For example, the arrow from Bob to Ada indicates that Ada is identified as the advisor of Bob. The triple on the edge, i.e., (0.8, [1999, 2000]), represents Ada has the probability of 80% to be the advisor of Bob from 1999 to 2000. Such results can benefit many potential applications such as research community detection and evolution analysis. The right figure gives an example of visualized chronological hierarchies. The parent-child relation in the tree corresponds to the advisor-advisee relationship. We can see the advising path from root to leaf.

ISSN: 2277-9655
Impact Factor: 1.852

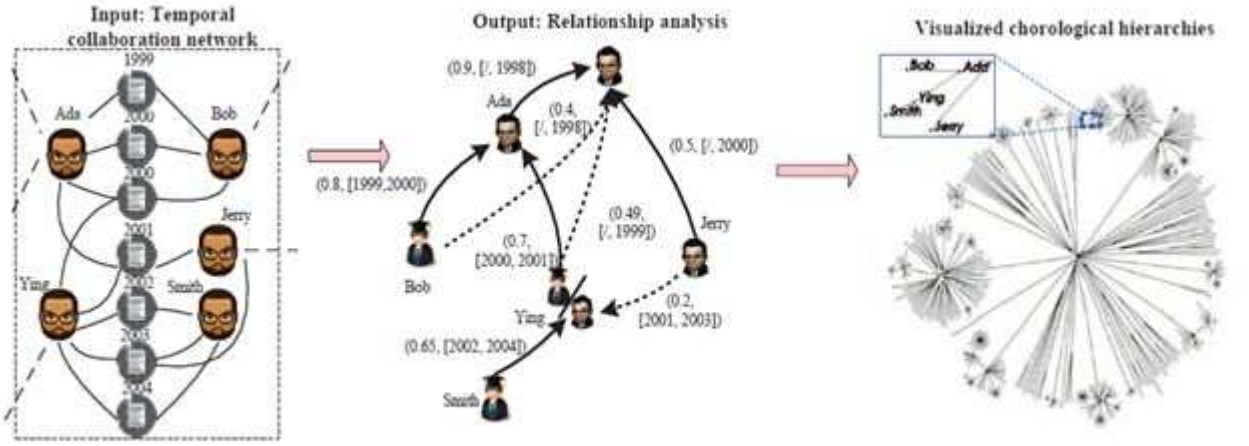

Fig 4. Example of advising relationship analysis on the co-author network.

Conclusion

Web Mining is powerful technique used to extract the Information from past behavior of users. Various algorithms are used to mining or extracting the data from a web page. The main focus is to extracting the user/research profile from a social network web sites.

This survey was designed to provide researchers with a snapshot of the current state of Academic Social Network. The manual entering process is very obviously tedious and time consuming for extraction of the researcher profile information. Recent work has shown the feasibility and promise of information extraction technologies for extracting the structured data from the Web, and it is possible to use the methods to extract the profile of a researcher.

In this paper was explained extraction of researcher's profiles, expert finding and co-author relationship in the academic social network. It would be interesting to further investigate new extraction models and algorithm for improving the accuracy of profile extraction.

References

[1] R. Bekkerman and A. McCallum. Disambiguating web appearances of people in a social network. In Proc. of WWW'05, pages 463–470, 2005.

[2] D. Blei, A. Ng, and M. Jordan. Latent Dirichlet Allocation. Journal of Machine Learning Research, 2003, 3, 993-1022.

[3] F. Ciravegna. An adaptive algorithm for information extraction from web-related texts. In Proc. of IJCAI'01 Workshop, August 2001.

[4] C. P. Diehl, G. Namata, and L. Getoor. Relationship identification for social network discovery. In AAAI'07, pages 546–552. AAAI Press, 2007.

[5] Herring, S.C., Kouper, I., Paolillo, J.C., Scheidt, L.A., Tyworth, M., Welsch, P., Wright, E., and Yu, N. "Conversations in the Blogosphere: An Analysis "From the Bottom Up"," 38. Hawaii International Conference on System Sciences (HICSS-38), IEEE Press., Hawaii, 2005.

[6] T. Kristjansson, A. Culotta, P. Viola, and A. McCallum. Interactive information extraction with constrained conditional random fields. In Proc. of AAAI'04, 2004.

[7] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: probabilistic models for segmenting and labeling sequence data. In Proc. of ICML'2001, pp.282-289.

[8] Li-Chun, Y., Kretschmer, H., Hanneman, R.A., and Ze-Yuan, L. "The evolution of a citation network topology: The development of the journal Scientometrics," International Workshop on Webometrics, Informetrics and Scientometrics & Seventh OLLNET Meeting, Nancy, France, 2006.

[9] Lorrain, F., & White, H. C. (1971). The structural equivalence of individuals in social

ISSN: 2277-9655
Impact Factor: 1.852

networks. Journal of Mathematical Sociology, 1, 49–80.

[10] Y.F. Tan, M. Kan, and D. Lee. Search engine driven author disambiguation. Proc. of JCDL'2006. pp. 314-315.

[11] J. Tang, M. Hong, J. Zhang, B. Liang, and J. Li. A New Approach to Personal Network Search based on Information Extraction. Demo paper. In Proc. of ASWC'2006.

[12] J. Tang, D. Zhang, and L. Yao. Social network extraction of academic researchers. In Proc. of ICDM'07, pages 292–301, 2007.

[13] Wasserman, S., & Faust, K. (1994). Social network analysis: Methods and applications. New York: Cambridge University Press.

[14] D. Zhang, J. Tang, J. Li, and K. Wang. A constraint-based probabilistic framework for name disambiguation. Proc. Of CIKM'2007. pp. 1019-1022.

[15] Zhang, J., Tang, J., Li, J.: Expert Finding in a Social Network. In: Kotagiri, R., Radha Krishna, P., Mohania, M., Nantajeewarawat, E. (eds.) DASFAA 2007. LNCS, vol. 4443, pp. 1066–1069. Springer, Heidelberg (2007)