PMKT
BRAZILIAN JOURNAL OF MARKETING, OPINION, AND MEDIA RESEARCH

Chief Editor: Fauze Najib Mattar | Deputy Editor: Claudio Alberto de Moraes
Evaluation System: Triple Blind Review | Posted: Brazilian Association of Research Companies (ABEP) | Languages: Portuguese and English

AMiner-Metadata of Academic Research through Artificial Intelligence

AMiner - Metadados de Pesquisas Acadêmicas por meio da Inteligência Artificial

Norberto de Almeida Andrade[1]Giuliano Carlo Rainatto[2], Genésio Renovato da Silva Neto[3], Jucilene Moreira de Barros Faria[4]

Submission: September 14, 2019
Approval: September 18, 2019

Summary

In this article, we present a new online academic system of research and mining, AMiner. ArnetMiner the second generation system. A free search engine and based on Artificial Intelligence (AI) that aims to overcome Google Scholar, is expanding its corpus of articles to cover about 10 million research articles in administration, computer science and neuroscience, among other areas of equal importance. Since its launch in 2008, they joined several other mechanisms of academic pursuits IA classify promising academic work using a more sophisticated understanding of its content and context. The algorithms and data from academic research AMiner are available to researchers through an application programming interface (API). This paper is organized as follows, we first present the overall architecture of the system. Section 2 discusses related work and Section 3 presents our proposed approaches in the system. Section 4 shows some applications AMiner. Section 5 lists the data sets we built. Finally, section 6 makes a conclusion of the article.

Keywords: Bibliometrics, AMiner, Metadata, Artificial Intelligence.

[1] Master in Business Administration the United Metropolitan Colleges with Specialization in Research Marketing and Consumer Behavior. Professor at the University of São Caetano do Sul. Marketing consultant digital. Address: Rua Coronel Oscar Porto, 70, Paradise, 04003-000, São Paulo, SP, Brazil. Email:norbertofatecsp@hotmail.com

[2] Master Administration the United Metropolitan Colleges with Specialization in Research in Innovation and Innovative Organizations. Professor at the University Anhanguera.Email: giulianorainatto@yahoo.com.br

[3] Master Administration the United Metropolitan Colleges with Specializing in Research Marketing and Consumer Behavior. Business consultant. Email: genesiorenovato@yahoo.com.br

[4] Master Administration the United Metropolitan Colleges with Specializing in Research Marketing and Consumer Behavior. Email: jucil.faria@gmail.com

Resumo

Neste artigo, apresentamos um novo sistema acadêmico on-line de pesquisa e mineração, o AMiner. É a segunda geração do sistema ArnetMiner. Um mecanismo de busca livre e baseado em Inteligência Artificial (IA) que visa superar o Google Scholar, está expandindo seu corpus de artigos para cobrir cerca de 10 milhões de artigos de pesquisa em administração, ciência da computação e neurociência, entre outras áreas de igual importância. Desde o seu lançamento em 2008, juntaram-se vários outros mecanismos de buscas acadêmicas baseadas em IA prometendo classificar trabalhos acadêmicos usando uma compreensão mais sofisticada de seu conteúdo e contexto. Os algoritmos e dados de pesquisa acadêmica do AMiner estão disponíveis para pesquisadores por meio de uma interface de programação de aplicativos (API). Este trabalho está organizado da seguinte forma, primeiro apresentamos a arquitetura geral do sistema. A seção 2 discute os trabalhos relacionados e a seção 3 apresenta nossas abordagens propostas no sistema. A seção 4 mostra algumas aplicações do AMiner. A seção 5 lista os conjuntos de dados que construímos. Finalmente, a seção 6 faz uma conclusão do artigo.

Palavras-chave: Bibliometria, AMiner, Metadados, Inteligência Artificial.

How to cite (APA):
Andrade, N. de A., Rainatto, G. C., Silva, G. R. da, Neto, & Faria, J. M. de B. (2019). AMiner - Metadados de pesquisas acadêmicas por meio da inteligência artificial. *PMKT – Revista Brasileira de Pesquisas de Marketing, Opinião e Mídia (on-line), 12*(2), 69-82. Recuperado de www.revistapmkt.com.br

How to cite (ABNT NBR 6023/2018):
ANDRADE, N. de A.; RAINATTO, G. C.; SILVA, G. R. da, Neto; & FARIA, J. M. de B. AMiner - Metadados de pesquisas acadêmicas por meio da inteligência artificial. PMKT – Revista Brasileira de Pesquisas de Marketing, Opinião e Mídia (on-line), São Paulo, Vol. 12, N. 2, 69-82, 2019. Disponível em: www.revistapmkt.com.br. Acesso em:

1 Introduction

Bibliometry is the use of statistical methods to analyze the data of the citation index. They can be analyzed to determine the popularity and impact of articles, authors and specific publications. Citation analysis is a bibliometric commonly used method which is based on building quotes graph, a network representation or chart of citations between documents. Many research fields use bibliometric methods to explore the impact of his field, the impact of a number of researchers, the impact of a particular article or to identify particularly impactful documents within a specific field research

The ArnetMiner (aka AMiner) is a free online service used to index, search and extract large scientific data. The system is designed to search and perform data mining in academic publications in the Internet operations, using social network analysis to identify connections between researchers, conferences and publications. This allows providing services such as discovery of experts, geographical research, trend analysis, recommendation of reviewers, association research, ongoing research, academic performance evaluation and modeling topics. The ArnetMiner was created as a research project on the analysis of social influence ranking social networking and extraction of social networks. The ArnetMiner is commonly used in academia to search for scholarly, educational literature, newspaper universities and miscellaneous items and draw statistical correlations on research and researchers. It attracted more than 10 million independent IP access 220 countries and regions. The product was used in the SciVerse platform Elsevier, and academic conferences as SIGKDD, ICDM, PKDD, WSDM.

The ArnetMiner automatically extracts the Web crawler profile. It collects and identifies relevant pages and uses a unified approach to extract data of the identified documents. It also extracts digital publications online libraries using heuristic rules. Integrates the profiles of researchers extracted and the extracted publications. Employs the researcher's name as the identifier. A probabilistic framework was proposed to deal with the problem of ambiguity name in integration. Integrated data is stored on a network knowledge base of researchers (RNKB). The other main products of the area are Google Scholar, the Scirus of Elsevier and the open source project CiteSeer. The ArnetMiner published several sets of data for academic research purposes, including Academic Open Graph, DBLP (a set of data increasing citations in DBLP data dblp) Disambiguation Names and analysis of social ties. A variety of academic social networking sites, including Google Scholar, Microsoft Academic, Semantic Scholar ResearchGate Academia.edu and gained great popularity over the last decade. The common goal of these academic social network systems is to provide researchers with an integrated platform to query information and academic resources, share their own achievements and connect with other researchers. ResearchGate Academia.edu and gained great popularity over the last decade. The common goal of these academic social network systems is to provide researchers with an integrated platform to query information and academic resources, share their own achievements and connect with other researchers. ResearchGate Academia.edu and gained great popularity over the last decade. The common goal of these academic social network systems is to provide researchers with an integrated platform to query information and academic resources, share their own achievements and connect with other researchers.

Several issues within the academic social networks were investigated in these systems. However, most problems are investigated separately by independent processes. As such, there is no consistent process or a series of methods for mining different academic social networks. The lack of such methods can be attributed to two reasons:

1) Lack of semantic-based information. The user profile information obtained only from user entered your information or extracted by heuristics are sometimes

incomplete or inconsistent. Users do not meet personal information just because they are not willing to do so;

2) Lack of a unified modeling approach to effective mining of the social network. Traditionally, different types of information sources in academic social network were modeled individually and therefore the dependencies between them can not be captured. However, there may be dependencies between social data. high quality search services need to consider the intrinsic dependencies between the different sources of heterogeneous information.

In AMiner, our goal is to answer four questions:

1) Automatically extract the existing Web researcher profile?
2) How to integrate the extracted information (ie, profiles and publications of the researchers) from different sources?
3) How to model the different types of information sources into a unified model?
4) How to provide advanced search services built on a network?

To answer the above questions, a number of new approaches are implemented within the AMiner system. The overall system architecture is shown in Figure 1.

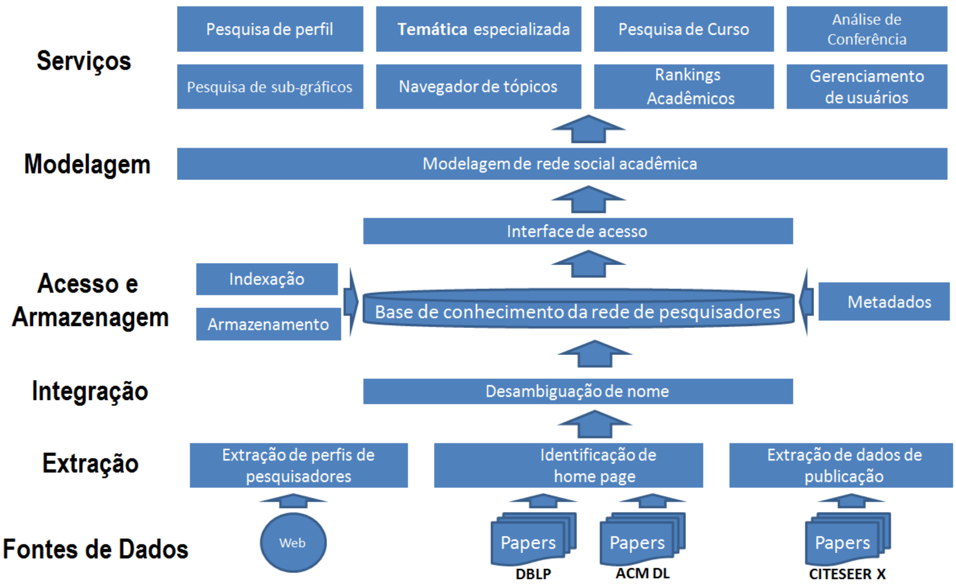


Figure 1- The overall architecture of the system AMiner.

The system mainly consists of five components:

1) **Services** - Provides various services based on the results of modeling: profile search, discovery experts, conference analysis, survey course, subgraph research, topics browser, academic ratings and user management;
2) **Modeling -**Uses a probabilistic generative model to model both the different types of information sources. The system estimates a mixture distribution of topics associated with the different sources of information;
3) **Access and Storage -**It provides storage and indexing for the data extracted and integrated into the network knowledge base of researchers. Specifically, storage, employs Jena, a tool to store and retrieve data ontological; Index employs the indexing method of inverted files, an established method to facilitate the retrieval of information;
4) **Integration -**Joins and integrates the profiles of researchers extracted and the extracted publications. The application employs the researcher's name as the identifier. A probabilistic model and a comprehensive framework were developed to deal with the

problem of ambiguity name in integration. The integrated data is then stored, classified and indexed in a database of research knowledge network.

5) **Extraction -**The focus is on automatic extraction of Web searchers profiles. First, the collection service and identifies the relevant pages (eg home pages or introductory pages) Web and uses a unified approach to extract data of the identified documents. It also extracts digital publications online libraries using heuristic rules. In addition, a simple approach is adopted, but very effective, for the Web user profile, harnessing the power of big data.

For various system resources, for example, extraction profiles, names disambiguation, modeling academic topics, skill and mining research of academic social networks, we propose some new approaches to overcome the disadvantages that exist in conventional methods.

The remainder of this paper is organized as follows. Section 2 discusses related work and Section 3 presents our proposed approaches in the system. Section 4 shows some applications AMiner. Section 5 lists the data sets we built. Finally, Section 6 is a conclusion.

2 Related Work

Previously a number of issues in academic social networks have been investigated and some systems were developed as discussed below.

Google Scholar: Provides a search engine to identify hyperlinks to publications that are publicly available or can be obtained through institutional libraries. Google Scholar is not a social networking site in the general sense, but still became an important platform for academic research resources, track the latest research, promote own achievements and track the academic impact.

***Microsoft Academic*:** employs machine learning technologies, semantic analysis and data mining to help users explore academic information more powerfully.

***Semantic Scholar*:** It is designed to be a search engine "smart" to help researchers find better academic publications faster. Compared to Google Scholar and Academic Microsoft, the Semantic Scholar can quickly highlight the most important items and identify the connections between them.

***ResearchGate*:** It aims to connect geographically distant researchers and allow them to talk continuously. Registered users of the site have a user profile and can share their research output, including articles, data, book chapters, patents, research proposals, algorithms, presentations and open-source software. Users can also follow the activities of others and participate in discussions with them.

Academia.edu: It is an academic social networking site for profit. It allows its users to create a profile, share their work, monitor their academic impact, select areas of interest and follow the research that evolves in specific fields. Although most systems have built up a huge amount of academic resources and provided abundant means of research and consultation social networking functions, they did not conduct systematic analysis of the semantic level or mining. Consequently, our main goal is to provide a unified modeling approach for a greater and deeper understanding of the semantic connection in large heterogeneous academic networks, composed of authors, articles, conferences, journals and organizations. As a result, the system can provide specialized research and research focused on the researcher.

3 Methodology

In this section, we present in detail the challenges of mining of academic social networking data AMiner through the present system and methods and solutions.

3.1 Extraction profile

We define the researcher profile scheme, extending the FOAF ontology, as shown in Figure 2. In the scheme, 24 properties and two relationships are defined.

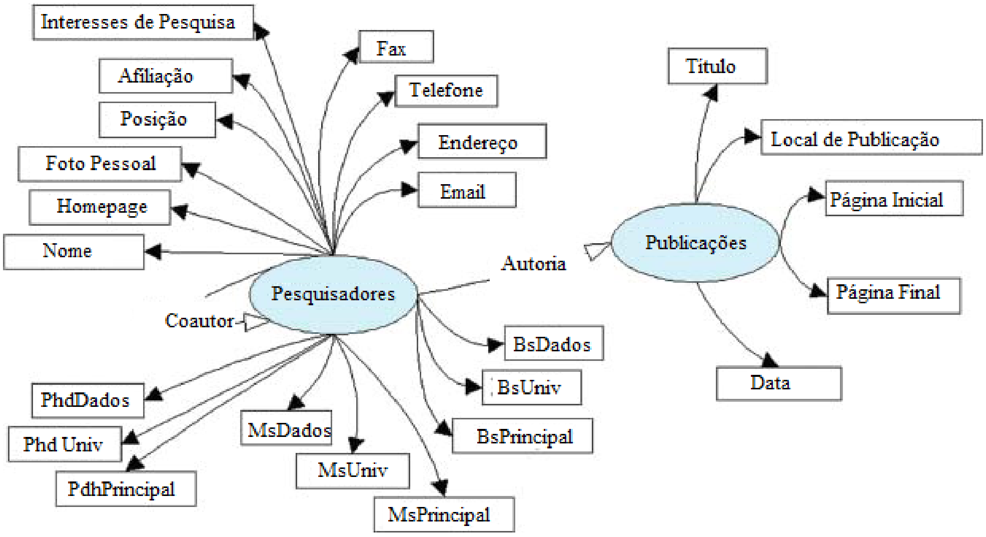


Figure 2 - Researcher profile Scheme extending FOAF ontology.

It is certainly not a trivial task to extract the Web search network. The researchers from different universities, institutes and companies have different page templates, profile, and data feeds. Therefore, an ideal extraction method should consider the processing of all types of models and formats. The approach we propose consists of three steps:

1) **Relevant page ID**- Given the name of a researcher, first we get a list of web pages by a search engine (Google API is used) and then identify the home page or the introduction page using a classifier. We define a set of features, for example, if the page title contains the person's name and the address of the URL (partly) contains the person's name and is used SVM for classification;
2) **Preprocessing**- the text is separated into tokens and attach possible tags to each token. Tokens form the basic units and the pages form the sequences in the following units labeling step;
3) **Marking -**Given a sequence of units, we determine the corresponding tags sequence most likely using a dial model trained. The tag type corresponds to the property set in Figure 2. We define five types of tokens (standard word special word, image token, word and punctuation mark) and uses heuristics to identify tokens on the Web. After that, we assign several possible tags each token based on the token type and then a CRF trained model is used to find the optimal allocation of tag with the highest probability.

Recently, we revisit the Web user profile problem in Big Data and propose a simple approach, but very effective, called MagicFG to create Web user profiles, harnessing the power of Big Data. To prevent the spread of errors, the approach integrates the page ID and profile extraction into a unified structure. To improve the profile of the performance, we present the concept of contextual credibility. The proposed framework also supports the incorporation of human knowledge. Defines human knowledge as logical statements Markov and formalized in a model of factors graphics. The method was implemented in MagicFG AMiner system to create millions of researchers profiles. Figure 3 gives an example of a researcher profile.

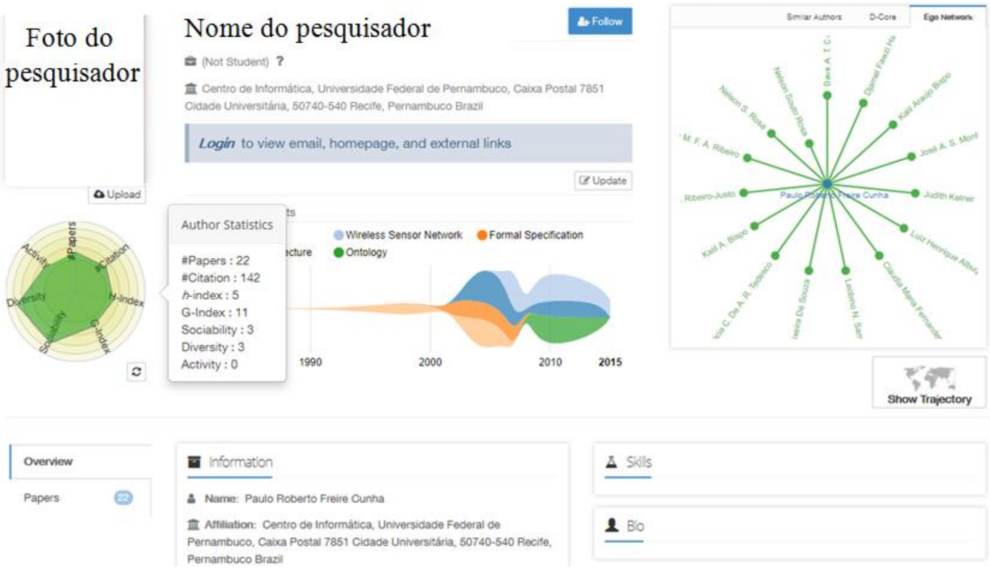


Foto do pesquisador

Nome do pesquisador

(Not Student) ?

Centro de Informática, Universidade Federal de Pernambuco, Caixa Postal 7851 Cidade Universitária, 50740-540 Recife, Pernambuco Brazil

Login to view email, homepage, and external links

Author Statistics

#Papers : 22
#Citation : 142
h-index : 5
G-Index : 11
Sociability : 3
Diversity : 3
Activity : 0

Information

Name: Paulo Roberto Freire Cunha

Affiliation: Centro de Informática, Universidade Federal de Pernambuco, Caixa Postal 7851 Cidade Universitária, 50740-540 Recife, Pernambuco Brazil

Figure 3 - Example researcher profile.

3.2 Disambiguation name

We define the researcher profile scheme, extending the FOAF ontology, as shown in Figure 2. In the scheme, 24 properties and two relationships are defined.We collected more than 200 million of existing online data libraries publications including DBLP, ACM DL, CiteSeerX and others. In each data source, the authors are identified by their names. To integrate the researcher profiles and publication of data, we use the name of the researcher and the author's name as the identifier. This process inevitably has the ambiguous problem. A few years ago, we proposed a probabilistic framework based on Random Fields Hidden Markov (HMRH), which is able to capture dependencies between observations (here each item is seen as an observation). The disambiguation problem is released when assigning a tag to each paper, with each tag representing a real researcher. More recently, we proposed a comprehensive framework to further address the problem of disambiguation of names. The overview of the framework is shown in Figure 4.

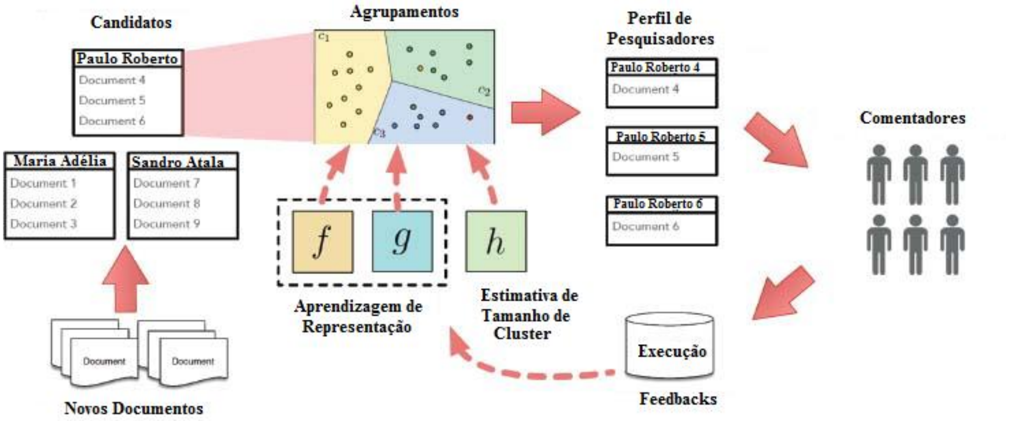


Figure 4 - Overview of the proposal for a comprehensive framework to further address the problem of name disambiguation.

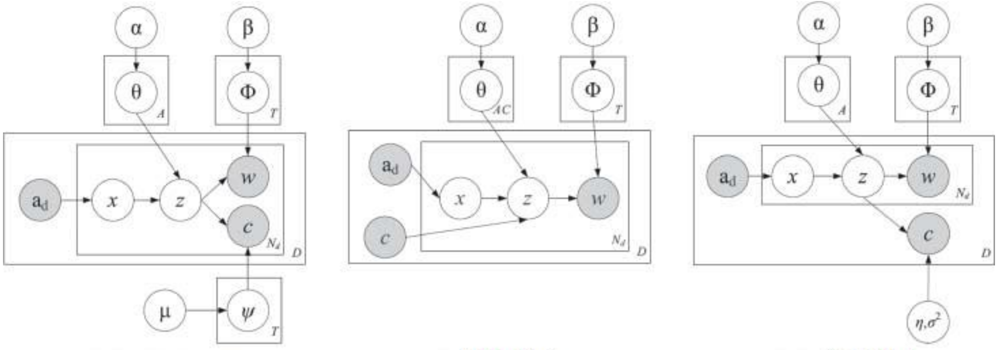
To improve accuracy, we engage in human annotators disambiguation process. The method has now been implanted in AMiner to deal with the problem of name disambiguation in the billion range, demonstrating its effectiveness and efficiency.

3.3 Modeling topics

In academic research, the representation of the text document content, authors of interests and conference themes is critical of any approach. Traditionally, the documents are represented on the assumption of the "bag words" (BOW). However, this representation can not use the "semantic" dependencies between words. Moreover, in the course of an academic research there are different types of information sources, so as to capture the dependencies between them becomes a challenging problem. Unfortunately, existing topics models such as Latent Semantic Indexing (pLSI) probability, the Dirichlet Allocation Latent (LDA) and the author-subject model can not be applied directly to the academic research context.

A unified approach to topics of modeling is proposed to simultaneously model document characteristics, authors, conferences and dependencies between them (for simplicity, we use the conference to designate conference, journal and book on model). The proposed model is called Author-Topic model-Conference (ACT). More specifically, different strategies can be used to model the distribution of topics (as shown in Figure 5) and consequently implemented the models may have different capacities of knowledge representation.

In Model 1 in Figure 5 (a) Each author is associated with weights on a mixture of topics. For example, each word token correlated to a role and, as a conference stamp associated with each word token is generated from a sample topic. In Model 2 in Figure 5 (b) each pair conferencing author is associated with a mixture of weights for the topics, and word tokens are generated from the sample topics. In Model 3 in Figure 5 (c), each author is associated topics, each word token is generated from a sample topic and then the conference is generated from the sampled topics of all tokens of words in a document.



(a) Modelo 1 (b) Modelo 2 (c) Modelo 3

Figure 5 - Different strategies employed to model the distribution of topics.

3.4 Expert search

When looking for academic resources and formulate a query, the user tries to find authors with expertise, Jobs and conferences related to the areas of research interest. In AMiner system, we present a framework of specialized research topic level. Unlike traditional Web search engines that perform the recovery and classification at the document level, we study the problem of specialization in research topic level in relation to different heterogeneous networks. A unified topic model, called Citation-Tracing-Topic (CTT), is proposed to model both aspects threads of different objects in the academic network. Based on models of learned topics, we investigate the problem of specialization search in three dimensions: classification, trace analysis of quotes and topics of graphic research. Specifically, we propose a method of random walk in topic level to classify different objects. In citation analysis traits we try to figure out how a study influences its follow-up study. Finally, we developed a graphics topics search function, based on the modeling and analysis of topics tracking quotes. Figure 6 gives an example of the result of experts found for the query "Social Network Analysis".

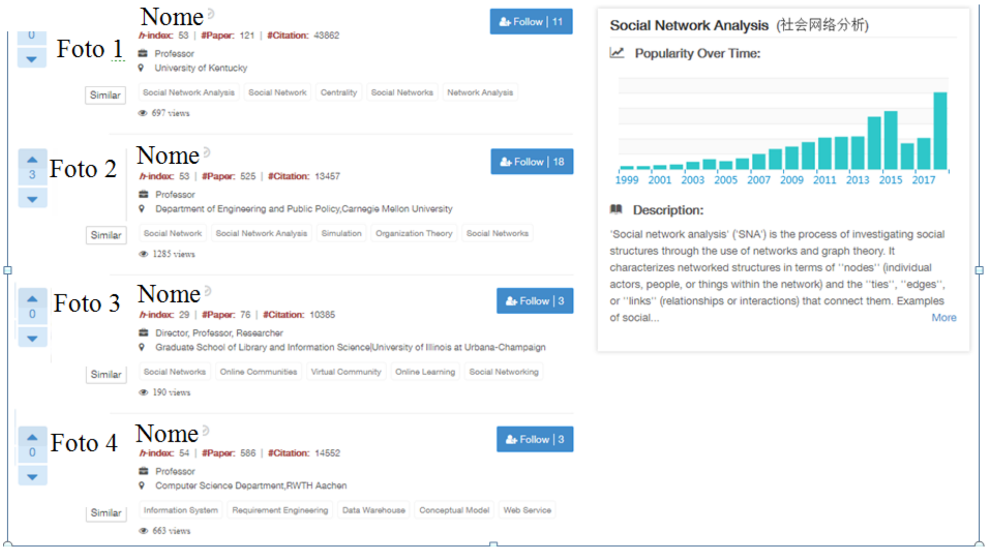


Foto 1
Nome
h-index: 53 | #Paper: 121 | #Citation: 43862
Professor
University of Kentucky
Similar | Social Network Analysis | Social Network | Centrality | Social Networks | Network Analysis
697 views
Follow | 11

Foto 2
Nome
h-index: 53 | #Paper: 525 | #Citation: 13457
Professor
Department of Engineering and Public Policy,Carnegie Mellon University
Similar | Social Network | Social Network Analysis | Simulation | Organization Theory | Social Networks
1285 views
Follow | 18

Foto 3
Nome
h-index: 29 | #Paper: 76 | #Citation: 10385
Director, Professor, Researcher
Graduate School of Library and Information Science|University of Illinois at Urbana-Champaign
Similar | Social Networks | Online Communities | Virtual Community | Online Learning | Social Networking
190 views
Follow | 3

Foto 4
Nome
h-index: 54 | #Paper: 586 | #Citation: 14552
Professor
Computer Science Department,RWTH Aachen
Similar | Information System | Requirement Engineering | Data Warehouse | Conceptual Model | Web Service
663 views
Follow | 3

Social Network Analysis (社会网络分析)

Popularity Over Time:

Description:

'Social network analysis' ('SNA') is the process of investigating social structures through the use of networks and graph theory. It characterizes networked structures in terms of ''nodes'' (individual actors, people, or things within the network) and the ''ties'', ''edges'', or ''links'' (relationships or interactions) that connect them. Examples of social...

More

Figure 6 - Example of results of experts found with the query "Social Network Analysis".

3.5 Mining Academic Social Network

Based on AMiner system, this set of academic social networking mining functions centered research includes analysis of social influence, relationship mining, similarity analysis, collaborative recommendation and evolution of the community.

Social Influence Analysis- In large social networks, people are influenced by others for various reasons. We propose a propagation model for Affinity Topics (TAP) to differentiate and quantify the social influence. The TAP can get results of any modeling topics and existing network infrastructure to perform the spread of influence in the discussion level. Recently design one end of the tip structure we call DeepInf to the characteristics of representation and learning to predict the social influence. Each user is represented by a local subnet in which it is embedded. A graphical neural network is used to learn the representation of the subnet that, in turn, effectively integrates the user-specific resources and network structures. The structure of DeepInf is shown in Figure 7.

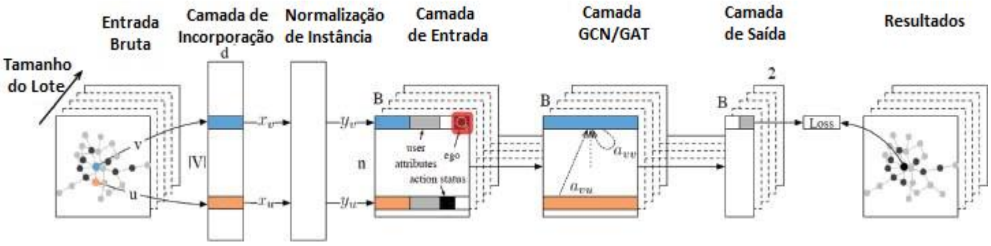


Figure 7 - DeepInf structure.

Mining Social Networking -Infer the type of social relationship between two users is a very important task in mining social networking. We propose a framework called two-stage model Factors Graph Probabilistic with time constraint (TPFG) to infer counselor aide relations in coauthors network. The main idea is to leverage a model of probabilistic factors chart with time constraint to decompose the joint probability of the unknown directors of all authors. In addition, we developed a framework called TranFG to classify the type of social relations between different heterogeneous resources. The structure incorporates social theories in a factor graph model which effectively

Similarity analysis -Estimate the similarity between the vertices is a key issue in the analysis of social networks. We propose a sampling-based method for estimating the similar top-k vertices. The method is based on the new idea of the sampling method for random walks known as Panther. Given a particular network as a starting point, the Panther randomly generates multiple paths of a predefined length and then the similarity between two vertices can be modeled to estimate the possibility that the two vertices appear in the same ways.

Collaboration recommendation -interdisciplinary collaborations generated a huge impact on society. However, it is often difficult for researchers to establish these collaborations between domains. We analyzed the collaboration of data between fields of research publications and topics we propose a model of Learning between Domains (CTL) for collaborative recommendation. To deal with sparse connections, CTL consolidates the collaboration between existing domains through layers of threads, instead of using author layers. This alleviates the problem of scarcity. To cope with additional knowledge, the CTL models topic distributions source fields and target separately, as well as the correlation between domains. To deal with the asymmetry of topics, the CTL models only topics relevant to collaboration between domains.

Community development -As social networks are very dynamic, it is interesting to study how people in different networks form clusters and how the various clusters evolve over time. We study coevolution mining objectsqualifiedin a special type of heterogeneous network, star network call. Then, as we examine the objectsqualifiedThey influence each other in the evolution of the network. It proposes a development based on Hierarchical Dirichlet Process Mixture Model that detects objects coevolutionqualified in the form of a cluster of evolution qualifiedin dynamic networks stellar. An efficient inference algorithm is provided to learn the model.

4 Application

The AMiner was developed to provide comprehensive research and mining services for social networks of researchers. In this system, we focus on: (1) create a profile based on semantics for each researcher, extracting informationdistributed Web; (2) integrating academic data (for example, bibliographic data and profiles of researchers) from multiple sources; (3) search for precisely the heterogeneous network; (4) analyze and discover interesting patterns of social network built researcher. The main search and analysis functions in AMiner are summarized in the following section.

Profile search - Enter a researcher name (for example, Paulo Roberto Freire). The system returns the profile based on semantics created for the researcher using information extraction techniques. On the profile page, the extracted and integrated information includes: contact information, photo, quote statistics, assessment of academic performance, research interest (time), educational background, personal social graph, research funding (currently only US and CN) publications and records (including citation information and documents that are automatically assigned to several different domains).

Specialized theme - Enter a query (for example, social network analysis). The system will return specialists in this topic. In addition, the system will suggest the best conference and major work on this topic. There are two classification algorithms: VSM and ACT. The first is similar to the conventional language model and the second is based on our Author-Conference-Topic model (ACT). Users can also provide feedback to the search results.

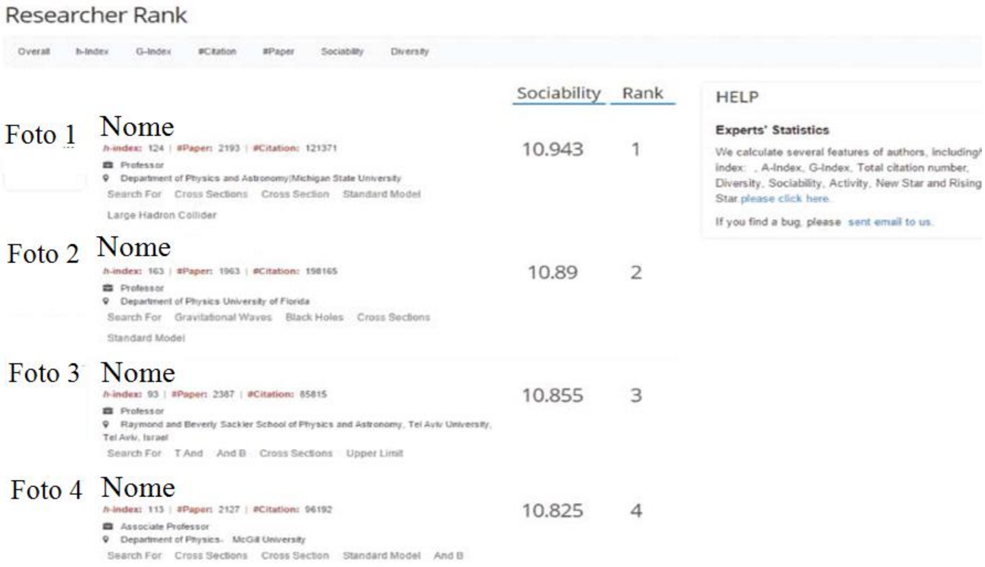
Review Conference - Enter a conference name (for example, KDD). The system will return those who are the most active researchers in this conference, as well as major works.

Survey course - Enter a query (eg, data mining). The system will return those who are teaching courses relevant to the query.

Research subgraphs - Enter a query (eg, data mining). The system will first tell you what topics are relevant to the query (for example, five topics "Data mining", "XML Data", "Data Mining / Query Processing", "Data Design / Web Database" and " web Mining "are relevant), and then display the most important sub-graph found in each relevant topic, expanded with a summary for the sub-graph.

Browser topics - Based on our Author-Conference-Topic model (ACT), automatically discovered 200 important topics of publications. For each topic, automatically assign a label to represent their meanings. In addition, the browser presents the most active researchers, conferences / most relevant work and the trend of evolution of the discovered topics.

Academic degrees - We defined eight measures to evaluate the achievement of the researcher. The measures include "h-index", "Citation" "Uptrend", "Activity", "Longevity", "Diversity," "Sociability" and "New Star". For each measure, we produce a ranking list in different domains. For example, one can search for those who have the highest citation numbers in the field of "social network analysis". Figure 8 gives an example of ranking of researchers for sociability index.



	Sociability	Rank
Nome *h*-index: 124 | #Paper: 2193 | #Citation: 121371 Professor Department of Physics and Astronomy|Michigan State University Search For Cross Sections Cross Section Standard Model Large Hadron Collider	10.943	1
Nome *h*-index: 163 | #Paper: 1963 | #Citation: 198165 Professor Department of Physics University of Florida Search For Gravitational Waves Black Holes Cross Sections Standard Model	10.89	2
Nome *h*-index: 93 | #Paper: 2387 | #Citation: 85815 Professor Raymond and Beverly Sackler School of Physics and Astronomy, Tel Aviv University, Tel Aviv, Israel Search For T And And B Cross Sections Upper Limit	10.855	3
Nome *h*-index: 113 | #Paper: 2127 | #Citation: 96192 Associate Professor Department of Physics- McGill University Search For Cross Sections Cross Section Standard Model And B	10.825	4

HELP

Experts' Statistics

We calculate several features of authors, including' index: , A-Index, G-Index, Total citation number, Diversity, Sociability, Activity, New Star and Rising Star.please click here.

If you find a bug, please sent email to us.

Figure 8 - Example of ranking of researchers for sociability index.

User Management -You can register as a user to: (1) modify the information of the extracted profile; (2) provide feedback on the search results; (3) the following researchers AMiner; and (4) create a page AMiner (which can be used to advertise conferences and workshops, or recruit students).

5 Data Set

The AMiner gathered a large set of academic data with over 130 million profiles of researchers and 233 million of Internet publications until June 2018, along with several subsets that were built for purposes other than research. Details of these subsets are as follows and can be found athttps://aminer.org/.

Quote Network -The citation information is extracted from DBLP, ACM DL and other sources. The data set contains articles 1,572,277 and 2,084,019 citation relations. Each item is associated with short, authors, year, place and title. The data set can be used for grouping with network information and side, studying the influence on the network quotes, finding the most influential documents, topics modeling analysis etc.

Academic Social Network - This data includes articles, quotation on paper, information about the author and the author's collaboration. The dataset contains 1,712,433 authors, articles 2092356, 8024869 and 4258615 citation relations collaborative relationships observed between the authors.

Councilor adviser - The data set consists of 2,792,833 and 815,946 authors relations coauthoring. To evaluate the performance of inferred adviser adviser relationships between co-authors, create smaller data on the truth of the terrain using the following method: (1) collect project advisor aide to information Mathematics Genealogy and Genealogy AI project; (2) manually track the counselor-supervisor of information from the initial pages of the researchers. Finally, we label 1,534 coauthor relations of which 514 are counselor aide relations.

Topic coauthor - It is a network of co-authors based on topics that contains 640 134 authors of 8 threads and 1,554,643 relations co-author. The eight topics are: Data Mining / Association Rules, Web Services, Bayesian Networks / Belief function, Web Mining / Fusion Information, Semantic Web / Logical descriptions, Machine Learning, Database Systems / XML data and Information Retrieval.

Topic article author - The set of data is collected for cross-domain recommendation purposes containing 33,739 authors associated with five topics, in addition to 139,278 relations co-author. The five topics are Data Mining (with 6,282 authors and 22862 co-authors of relationships), Medical Informatics (with 9,150 authors and 31851 relations co-authors), Theory (with 5,449 authors and 27,712 co-authors), Visualization (with 5,268 authors) and 19,261 relationships of co-authors) and Database (with 7,590 authors and co-authors of 37,592 relationships).

Quote theme - It is a network of quotations based on topics that contains 2,329,760 articles 10 topics and 12,710,347 relations quotes. The 10 topics are: Data Mining / Association Rules, Web Services, Bayesian Networks / Belief function, Web Mining / Information Fusion, Semantic Web / Logical Description, Machine Learning, Database Systems / XML Data Recognition Standards / Image Analysis, Information Retrieval and Natural Language System / translation Machine Statistics.

Kernel Community - It is a network of co-authored with 822,415 2,928,360 us and not directed ends. Each vertex is an author and each edge represents a co-author relationship.

Dynamic coauthor - The dataset contains 1,768,776 articles published during the period 1986 to 2012 with 1,629,217 authors involved. Each year is considered as a date and time stamp and there are 27 date and time stamps in total. On each date and time stamp, we created a border between two authors if they have co-authoring at least an article in the most recent three years (including the current year). We convert the coauthor network not directed in a targeted network considering each undirected edge as two edges directed symmetrical.

Thematic expert - This data set is a reference to the discovery of experts, containing 1,781 experts from 13 topics.

Association search - This data set is used to evaluate the effectiveness of association of research approaches containing 8,369 pairs of specific authors to nine topics. Each pair of authors contains a source author and a target author.

Topic Model results for the data set of AMiner - There are the results of the ACT model in AMiner data set that contains the main one million articles and authors 200 topics.

Coauthor - This is a network coauthors in AMiner system containing 1,560,640 4,258,946 authors and coauthor relations.

Disambiguation -This data set is used to study the disambiguation of names in a digital library. It contains 110 authors and their affiliations, as well as the results of disambiguation (fundamental truth).

6 Conclusion

We recognize that AMiner is still under development, both in the scale of resources and quality of services. However, in the future, we will explore additional intelligent methods to extract deep knowledge of scientific networks and will implement a more convenient structure and customized to provide academic research and find services.

Refference

Baepler, P., & Murdoch, C. J. (2010). Academic analytics and data mining in higher education. *International Journal for the Scholarship of Teaching and Learning*, *4*(2), 17.

Borkar, S., & Rajeswari, K. (2013). Predicting students academic performance using education data mining. *International Journal of Computer Science and Mobile Computing*, *2*(7), 273-279.

Cabanac, G., Frommholz, I., & Mayr, P. (2019, April). Bibliometric-Enhanced Information Retrieval: 8th International BIR Workshop. In *European Conference on Information Retrieval* (pp. 394-399). Springer, Cham.

Nasution, M. K., & Noah, S. A. (2011, June). Extraction of academic social network from on-line database. In *2011 International Conference on Semantic Technology and Information Retrieval* (pp. 64-69). IEEE.

Nasution, M. K., Noah, S. A. M., & Saad, S. (2016). Social network extraction: Superficial method and information retrieval. *arXiv preprint arXiv:1601.02904*.

Ovadia, S. (2014). ResearchGate and Academia. edu: Academic social networks. *Behavioral & social sciences librarian*, *33*(3), 165-169.

Parack, S., Zahid, Z., & Merchant, F. (2012, January). Application of data mining in educational databases for predicting academic trends and patterns. In *2012 IEEE International Conference on Technology Enhanced Education (ICTEE)* (pp. 1-4). IEEE.

Shah, T., & Pudi, V. (2019). Mining Intellectual Influence Associations. In *BIR@ ECIR* (pp. 100-111).

Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., & Su, Z. (2008, August). Arnetminer: extraction and mining of academic social networks. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 990-998). ACM.

Zhang, Y., Zhang, F., Yao, P., & Tang, J. (2018, July). Name Disambiguation in AMiner: Clustering, Maintenance, and Human in the Loop. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 1002-1011). ACM.

Wan, H., Zhang, Y., Zhang, J., & Tang, J. (2019). Aminer: Search and mining of academic social networks. *Data Intelligence*, *1*(1), 58-76.

Wolfram, D. (2016, June). Bibliometrics, information retrieval and natural language processing: natural synergies to support digital library research. In *Proceedings of the joint workshop on bibliometric-enhanced information retrieval and natural language processing for digital libraries (BIRNDL)* (pp. 6-13).

Wu, C. J., Chung, J. M., Lu, C. Y., Lee, H. M., & Ho, J. M. (2011, August). Using Web-mining for academic measurement and scholar recommendation in expert finding system. In *2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology* (Vol. 1, pp. 288-291). IEEE.