PMKT
REVISTA BRASILEIRA DE PESQUISAS DE MARKETING, OPINIÃO E MÍDIA

Editor-Chefe: Fauze Najib Mattar | Editor Adjunto: Claudio Alberto de Moraes | Sistema de avaliação: Triple Blind Review | Publicação: Associação Brasileira de Empresas de Pesquisa (ABEP) | Idiomas: Português e Inglês

AMiner - Metadados de Pesquisas Acadêmicas por meio da Inteligência Artificial

AMiner-Metadata of Academic Research through Artificial Intelligence

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

Submissão: 14 setembro 2019
Aprovação: 18 setembro 2019

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.

Abstract

[1] Mestre em Administração pela Faculdades Metropolitanas Unidas com Especialização em Pesquisa de Marketing e Comportamento de Consumo. Professor na Universidade de São Caetano do Sul. Consultor de Marketing Digital. Endereço: Rua Coronel Oscar Porto, 70, Paraíso, 04003-000, São Paulo, SP, Brasil. E-mail: norbertofatecsp@hotmail.com

[2] Mestre em Administração pela Faculdades Metropolitanas Unidas com Especialização em Pesquisa em Inovação e Organizações Inovadoras. Professor na Universidade Anhanguera. E-mail: giulianorainatto@yahoo.com.br

[3] Mestre em Administração pela Faculdades Metropolitanas Unidas com Especialização em Pesquisa de Marketing e Comportamento de Consumo. Consultor de Negócios. E-mail: genesiorenovato@yahoo.com.br

[4] Mestre em Administração pela Faculdades Metropolitanas Unidas com Especialização em Pesquisa de Marketing e Comportamento de Consumo. E-mail: jucil.faria@gmail.com

In this article, we introduce a new on-line academic research and mining system, the AMiner. It is the second generation ArnetMiner system. A free, Artificial Intelligence (AI) -based search engine that aims to surpass Google Scholar is expanding its article corpus to cover about 10 million research articles on administration, computer science, and neuroscience, among other equally important areas. . Since its launch in 2008, a number of other AI-based academic search engines have come together, promising to rank and rank academic papers using a more sophisticated understanding of their content and context. AMiner's academic research algorithms and data are available to researchers through an application programming interface (API). This paper is organized as follows. First we 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 of AMiner. Section 5 lists the datasets we built. Finally, Section 6 makes a conclusion.

Keywords: *Bibliometrics, AMiner, Metadata, Artificial Intelligence.*

Como citar (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), 72-86. Recuperado de www.revistapmkt.com.br

Como citar (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, 72-86, 2019. Disponível em: www.revistapmkt.com.br. Acesso em:

1 Introdução

A bibliometria é o uso de métodos estatísticos para analisar os dados dos índices de citação. Podem ser analisados para determinar a popularidade e o impacto de artigos, autores e publicações específicas. A análise de citações é um método bibliométrico comumente usado que é baseado na construção do grafo de citações, uma representação de rede ou gráfico das citações entre documentos. Muitos campos de pesquisa usam métodos bibliométricos para explorar o impacto de seu campo, o impacto de um conjunto de pesquisadores, o impacto de determinado artigo ou para identificar documentos particularmente impactantes dentro de um campo específico de pesquisa

O ArnetMiner (também conhecido como AMiner) é um serviço on-line gratuito usado para indexar, pesquisar e extrair grandes dados científicos. O sistema foi projetado para pesquisar e executar operações de mineração de dados em publicações acadêmicas na Internet, usando a análise de redes sociais para identificar conexões entre pesquisadores, conferências e publicações. Isso permite fornecer serviços como descoberta de especialistas, pesquisa geográfica, análise de tendências, recomendação de revisores, pesquisa de associação, pesquisa de curso, avaliação de desempenho acadêmico e modelagem de tópicos. O ArnetMiner foi criado como um projeto de pesquisa em análise de influência social, ranking de redes sociais e extração de redes sociais. O ArnetMiner é comumente usado na academia para pesquisar trabalhos acadêmicos, literatura escolar, jornais de universidades e artigos variados. e desenhar correlações estatísticas sobre pesquisa e pesquisadores. Ele atraiu mais de 10 milhões de acessos IP independentes de 220 países e regiões. O produto foi usado na plataforma SciVerse da Elsevier, e em conferências acadêmicas como SIGKDD, ICDM, PKDD, WSDM. O ArnetMiner extrai automaticamente o perfil do pesquisador da Web. Ele coleta e identifica as páginas relevantes e usa uma abordagem unificada para extrair dados dos documentos identificados. Também extrai publicações de bibliotecas digitais on-line usando regras heurísticas. Integra os perfis dos pesquisadores extraídos e as publicações extraídas. Emprega o nome do pesquisador como o identificador. Uma estrutura probabilística foi proposta para lidar com o problema da ambiguidade do nome na integração. Os dados integrados são armazenados em uma base de conhecimento da rede de pesquisadores (RNKB). Os outros produtos principais da área são o Google Scholar, o Scirus da Elsevier e o projeto de código aberto CiteSeer.

O ArnetMiner publicou vários conjuntos de dados para fins de pesquisa acadêmica, incluindo Open Academic Graph, DBLP (um conjunto de dados aumentando citações nos dados DBLP do Digital Bibliography & Library Project), Desambiguação de Nomes e Análise de laços sociais. A variedade de sites acadêmicos de redes sociais, incluindo o Google Scholar, Microsoft Academic, Semantic Scholar, ResearchGate e Academia.edu ganharam grande popularidade ao longo da última década. O objetivo comum desses sistemas de rede social acadêmica é fornecer aos pesquisadores uma plataforma integrada para consultar informações e recursos acadêmicos, compartilhar suas próprias conquistas e se conectar com outros pesquisadores.

Diversas questões dentro das redes sociais acadêmicas foram investigadas nesses sistemas. No entanto, a maioria dos problemas é investigada separadamente por meio de processos independentes. Como tal, não existe um processo congruente ou uma série de métodos para a mineração de redes sociais acadêmicas diferentes. A falta de tais métodos pode ser atribuída a dois motivos:

1) Falta de informação baseada em semântica. As informações de perfil de usuário obtidas apenas do usuário que inseriu suas informações ou extraídas por heurística são, por vezes, incompletas ou inconsistentes. Os usuários não preenchem informações pessoais apenas porque não estão dispostos a fazê-lo;

2) Falta de uma abordagem de modelagem unificada para mineração efetiva da rede social. Tradicionalmente, diferentes tipos de fontes de informação na rede social acadêmica foram modelados individualmente e, portanto, as dependências entre eles não podem ser capturadas. No entanto, podem existir dependências entre dados sociais. Serviços de busca de alta qualidade precisam considerar as dependências intrínsecas entre as diferentes fontes de informação heterogêneas.

Na AMiner, nosso objetivo é responder a quatro questões:

1) Como extrair automaticamente o perfil do pesquisador da Web existente?
2) Como integrar as informações extraídas (ou seja, perfis e publicações dos pesquisadores) de diferentes fontes?
3) Como modelar os diferentes tipos de fontes de informação em um modelo unificado?
4) Como fornecer serviços de pesquisa avançados em uma rede construída?

Para responder às perguntas acima, uma série de novas abordagens são implementadas dentro do sistema AMiner. A arquitetura geral do sistema é mostrada na Figura 1.

O sistema consiste principalmente de cinco componentes:

1) **Serviços** - Fornece vários serviços com base nos resultados de modelagem: pesquisa de perfil, descoberta de especialistas, análise de conferência, pesquisa de curso, pesquisa de subgráfico, navegador de tópicos, classificações acadêmicas e gerenciamento de usuários;
2) **Modelagem -** Utiliza um modelo probabilístico generativo para modelar simultaneamente os diferentes tipos de fontes de informação. O sistema estima uma mistura de distribuição de tópicos associada às diferentes fontes de informação;
3) **Acesso e Armazenagem -** Fornece armazenamento e indexação para os dados extraídos e integrados na base de conhecimento da rede de pesquisadores. Especificamente, para armazenamento, emprega Jena, uma ferramenta para armazenar e recuperar dados ontológicos; para indexação, emprega o método de indexação de arquivos invertidos, um método estabelecido para facilitar a recuperação da informação;
4) **Integração -** Junta e integra os perfis dos pesquisadores extraídos e as publicações extraídas. O aplicativo emprega o nome do pesquisador como o identificador. Um modelo probabilístico e uma estrutura abrangente foram desenvolvidos para lidar com o problema da ambiguidade do nome na integração. Os dados integrados são então armazenados, classificados e indexados em uma base de conhecimento de rede de pesquisa.
5) **Extração -** O foco está na extração automática de perfis de pesquisadores da Web. Primeiro, o serviço coleta e identifica as páginas relevantes (por exemplo, páginas iniciais ou páginas introdutórias) da Web e usa uma abordagem unificada para extrair dados dos documentos identificados. Também extrai publicações de bibliotecas digitais on-line usando regras heurísticas. Além disso, é adotada uma abordagem simples, mas muito eficaz, para o perfil de usuários da Web, aproveitando o poder do big data.

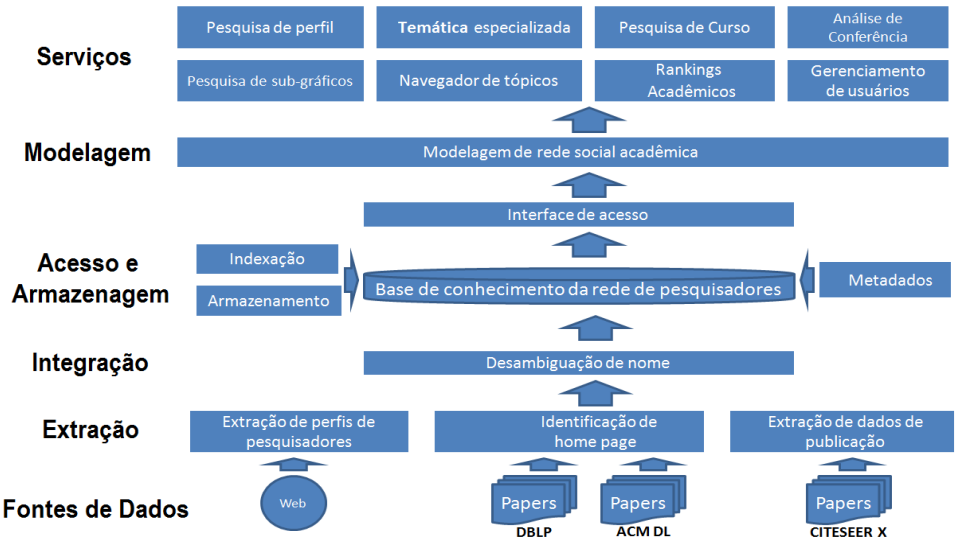

Figura 1- A arquitetura geral do sistema AMiner.

Para vários recursos do sistema, por exemplo, extração de perfis, desambiguação de nomes, modelagem de tópicos acadêmicos, pesquisa de perícia e mineração de redes sociais acadêmicas, propomos algumas novas abordagens para superar as desvantagens que existem nos métodos convencionais.

O restante deste trabalho está organizado da seguinte forma. 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.

2 Trabalhos Relacionados

Anteriormente várias questões nas redes sociais acadêmicas foram investigadas e alguns sistemas foram desenvolvidos conforme veremos a seguir.

Google Scholar: fornece um mecanismo de pesquisa para identificar os hiperlinks de publicações que estão disponíveis publicamente ou podem ser obtidos por meio de bibliotecas institucionais. O Google Acadêmico não é um site de rede social no sentido geral, mas ainda assim se tornou uma plataforma importante para pesquisar recursos acadêmicos, acompanhar as pesquisas mais recentes, promover as próprias conquistas e acompanhar o impacto acadêmico.

***Microsoft Academic*:** emprega tecnologias de aprendizado de máquina, análise semântica e mineração de dados para ajudar os usuários a explorar informações acadêmicas mais poderosamente.

***Semantic Scholar*:** é projetado para ser um mecanismo de busca "inteligente" para ajudar os pesquisadores a encontrar melhores publicações acadêmicas mais rapidamente. Em comparação com o Google Acadêmico e a Microsoft Academic, o Semantic Scholar pode destacar rapidamente os artigos mais importantes e identificar as conexões entre eles.

***ResearchGate*:** tem o objetivo de conectar pesquisadores geograficamente distantes e permitir que eles se comuniquem continuamente. Os usuários registrados do site têm um perfil de usuário e podem compartilhar sua produção de pesquisa, incluindo artigos, dados, capítulos de livros, patentes, propostas de pesquisa, algoritmos, apresentações e código-fonte de *software*. Os usuários também podem acompanhar as atividades de outras pessoas e participar de discussões com eles.

Academia.edu: é um site de rede social acadêmica com fins lucrativos. Ele permite que seus usuários criem um perfil, compartilhem seus trabalhos, monitorem seu impacto acadêmico, selecionem áreas de interesse e sigam a pesquisa que evolui em campos específicos.

Embora a maioria dos sistemas acima tenha integrado uma quantidade gigantesca de recursos acadêmicos e fornecido meios abundantes de pesquisa e consulta de funções de rede social, eles não realizaram análise sistemática de nível semântico ou mineração. Consequentemente, nosso objetivo principal é fornecer uma abordagem de modelagem unificada para obter uma compreensão maior e mais profunda da conexão semântica em redes acadêmicas grandes e heterogêneas, compostas por autores, artigos, conferências, periódicos e organizações. Como resultado, o sistema pode fornecer pesquisa especializada e pesquisa centrada no pesquisador.

3 Metodologia

Nesta seção, apresentamos em detalhes os desafios da mineração de dados de redes sociais acadêmicas por meio do sistema AMiner e apresentamos os métodos e soluções.

3.1 Extração de perfil

Definimos o esquema do perfil do pesquisador, estendendo a ontologia da FOAF, como mostra a Figura 2. No esquema, 24 propriedades e duas relações são definidas.

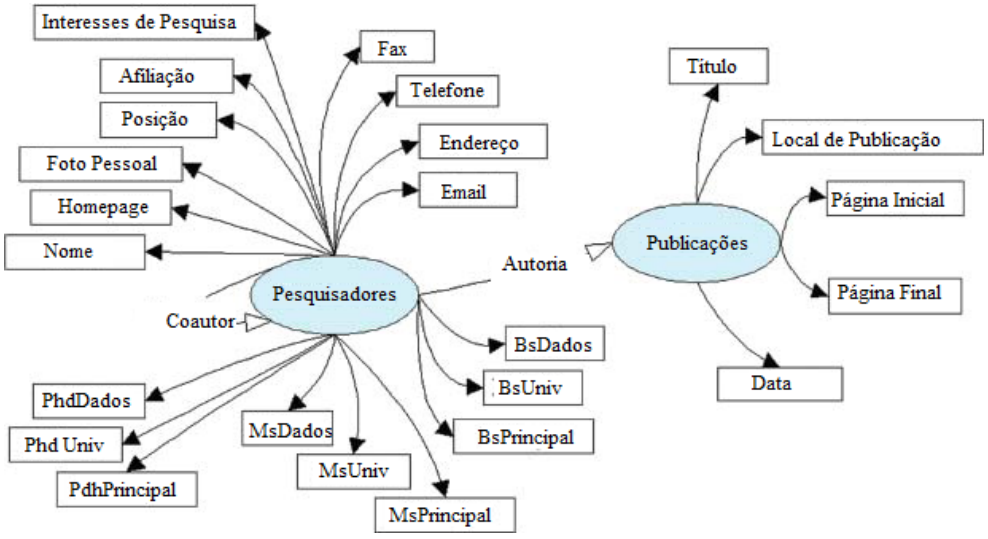

Figura 2 - Esquema do perfil do pesquisador, estendendo a ontologia da FOAF.

Certamente não é uma tarefa trivial extrair a rede de pesquisa da Web. Os pesquisadores de diferentes universidades, institutos ou empresas têm diferentes modelos de página, perfil, e *feeds* de dados. Portanto, um método de extração ideal deve considerar o processamento de todos os tipos de modelos e formatos. A abordagem que propomos consiste em três etapas:

1) **Identificação de página relevante** - Dado o nome de um pesquisador, primeiro obtemos uma lista de páginas da Web por um mecanismo de pesquisa (a API do Google é usada) e depois identificamos a página inicial ou a página de introdução usando um classificador. Definimos um conjunto de recursos, como, por exemplo, se o título da página contém o nome da pessoa e se o endereço da URL (em parte) contém o nome da pessoa e se emprega SVM para a classificação;

2) **Pré-processamento** - Separa-se o texto em *tokens* e atribuímos *tags* possíveis a cada *token*. Os *tokens* formam as unidades básicas e as páginas formam as sequências de unidades na etapa de marcação a seguir;
3) **Marcação** - Dada uma sequência de unidades, determinamos a sequência de *tags* correspondente mais provável usando um modelo de marcação treinado. O tipo de *tag* corresponde à propriedade definida na Figura 2. Definimos cinco tipos de *tokens* (palavra padrão, palavra especial, *token* de imagem, termo e sinal de pontuação) e usamos heurística para identificar *tokens* na Web. Depois disso, atribuímos várias *tags* possíveis a cada *token* com base no tipo de *token* e, em seguida, um modelo de CRF treinado é usado para encontrar a melhor atribuição de *tag* com a maior probabilidade.

Recentemente, revisitamos o problema do perfil do usuário da Web no Big Data e propomos uma abordagem simples, mas muito eficaz, chamada MagicFG, para criar perfis de usuários da Web, aproveitando o poder do Big Data. Para evitar a propagação de erros, a abordagem integra a identificação de página e a extração de perfil em uma estrutura unificada.

Para melhorar o desempenho do perfil, apresentamos o conceito de credibilidade contextual. O quadro proposto também apoia a incorporação do conhecimento humano. Define o conhecimento humano como declarações lógicas de Markov e as formaliza em um modelo de gráficos de fatores. O método MagicFG foi implantado no sistema AMiner para criar perfis de milhões de pesquisadores. A Figura 3 dá um exemplo de perfil de pesquisador.

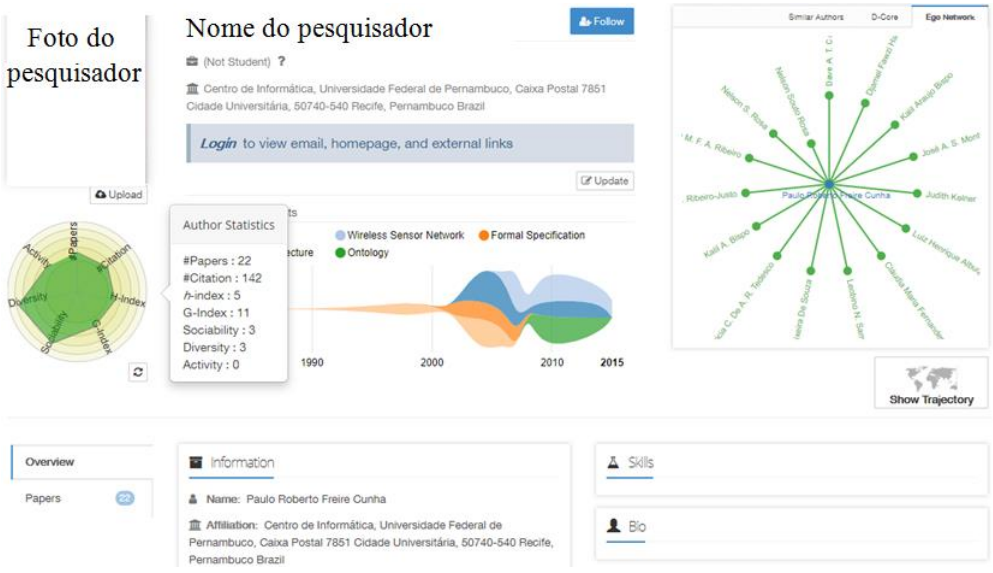

Figura 3 - Exemplo de perfil de pesquisador.

3.2 Desambiguação do nome

Definimos o esquema do perfil do pesquisador, estendendo a ontologia da FOAF, como mostra a Figura 2. No esquema, 24 propriedades e duas relações são definidas. Coletamos mais de 200 milhões de publicações de bibliotecas de dados on-line existentes, incluindo DBLP, ACM DL, CiteSeerX e outras. Em cada fonte de dados, os autores são identificados por seus nomes. Para integrar os perfis do pesquisador e os dados de publicação, usamos o nome do pesquisador e o nome do autor como o identificador. Este processo tem inevitavelmente o problema ambíguo. Há alguns anos, propusemos uma estrutura probabilística baseada em Campos Aleatórios de Markov Ocultos (HMRH), que é capaz de capturar dependências entre

observações (aqui cada artigo é visto como uma observação). O problema de desambiguação é lançado ao atribuir uma *tag* a cada papel, com cada *tag* representando um pesquisador real.

Mais recentemente, propusemos uma estrutura adicional abrangente para abordar o problema da desambiguação de nomes. A visão geral do *framework* é mostrada na Figura 4.

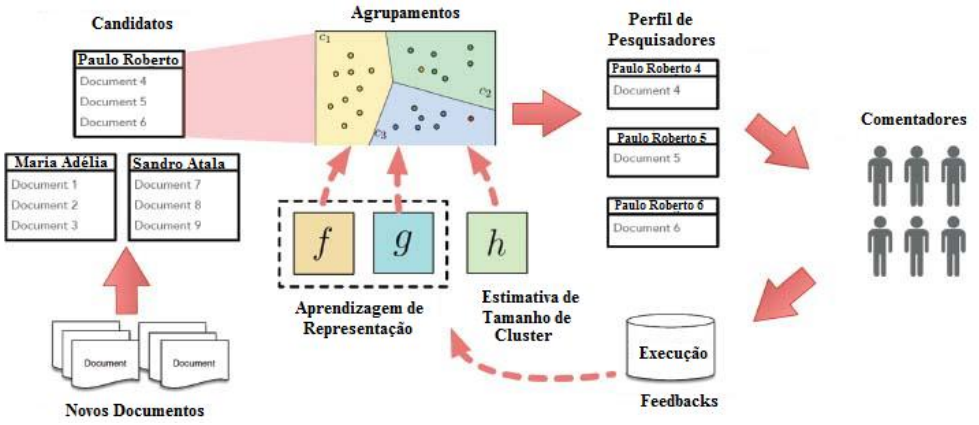

Figura 4 - Visão geral da proposta de uma estrutura adicional abrangente para abordar o problema da desambiguação de nomes.

Para melhorar a precisão, envolvemos anotadores humanos no processo de desambiguação. O método foi agora implantado na AMiner para lidar com o problema da desambiguação de nomes na escala de bilhões, o que demonstra sua eficácia e eficiência.

3.3 Modelagem de tópicos

Na pesquisa acadêmica, a representação do conteúdo de documentos de texto, interesses de autores e temas de conferências é uma questão crítica de qualquer abordagem. Tradicionalmente, os documentos são representados com base na suposição do "saco de palavras" (BOW). No entanto, essa representação não pode utilizar as dependências "semânticas" entre palavras. Além disso, no decorrer de uma pesquisa acadêmica existem diferentes tipos de fontes de informação, portanto, como capturar as dependências entre elas, torna-se um problema desafiador. Infelizmente, modelos de tópicos existentes, como a Indexação semântica latente (pLSI) probabilística, a Alocação de Dirichlet Latente (LDA) e o modelo Autor-Assunto não pode ser aplicado diretamente ao contexto da pesquisa acadêmica. Isso ocorre porque eles simplesmente não podem capturar todas as dependências intrínsecas entre documentos e conferências.

Uma abordagem unificada de modelagem de tópicos é proposta para modelar simultaneamente características de documentos, autores, conferências e dependências entre eles (para simplificar, usamos conferência para designar conferência, diário e livro no modelo). O modelo proposto é chamado modelo Autor-Conferência-Tópico (ACT). Mais especificamente, diferentes estratégias podem ser empregadas para modelar as distribuições de tópicos (como mostrado na Figura 5) e, consequentemente, os modelos implementados podem ter diferentes capacidades de representação do conhecimento.

No Modelo 1 da Figura 5 (a) cada autor está associado a uma mistura de pesos sobre tópicos. Por exemplo, cada *token* de palavra correlacionado a um papel e, da mesma forma, um carimbo de conferência associado a cada *token* de palavra, é gerado a partir de um tópico de amostra. No Modelo 2 da Figura 5(b) cada par autor-conferência está associado a uma mistura de pesos sobre os tópicos, e os *tokens* de palavras são gerados a partir dos tópicos da amostra. No Modelo 3 da Figura 5(c), cada autor é associado a tópicos, cada *token* de palavra é gerado

a partir de um tópico de amostra e, em seguida, a conferência é gerada a partir dos tópicos amostrados de todos os *tokens* de palavras em um documento.

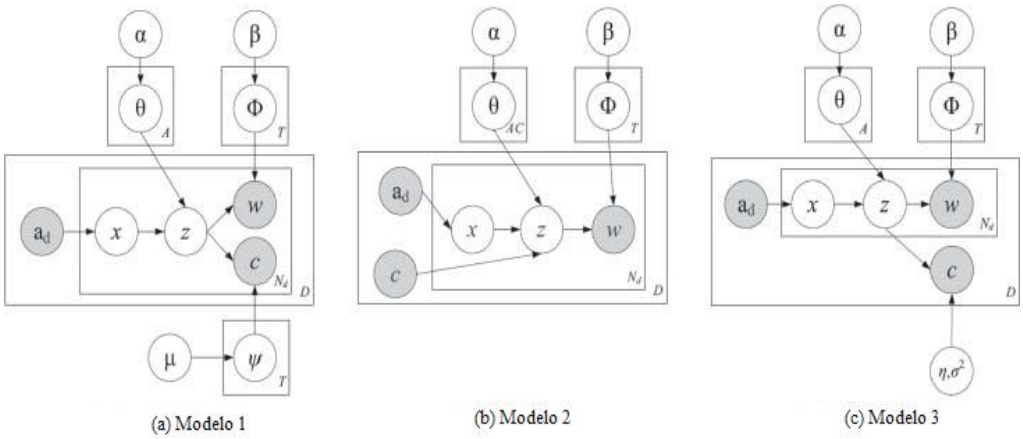

Figura 5 - Diferentes estratégias empregadas para modelar as distribuições de tópicos.

3.4 Pesquisa especializada

Ao procurar recursos acadêmicos e formular uma consulta, o usuário procura encontrar autores com conhecimentos específicos, trabalhos e conferências relacionados às áreas de interesse da pesquisa. No sistema AMiner, apresentamos uma estrutura de pesquisa de especialização em nível de tópico. Diferentemente dos tradicionais mecanismos de busca da Web que realizam a recuperação e a classificação no nível do documento, estudamos o problema de pesquisa de especialização em nível de tópico em relação a redes heterogêneas distintas. Um modelo de tópico unificado, chamado Citation-Tracing-Topic (CTT), é proposto para modelar simultaneamente aspectos tópicos de diferentes objetos na rede acadêmica.

Com base nos modelos de tópicos aprendidos, investigamos o problema de pesquisa de especialização em três dimensões: classificação, análise de rastreamento de citações e pesquisa de gráfico de tópicos. Especificamente, propomos um método de caminhada aleatória em nível de tópico para classificar diferentes objetos. Na análise de traços de citações, procuramos descobrir como um estudo influencia seu estudo de acompanhamento. Finalmente, desenvolvemos uma função de busca de gráficos tópicos, com base na modelagem de tópicos e análise de rastreamento de citações. A Figura 6 dá um exemplo do resultado de especialistas encontrados para a consulta "Social Network Analysis".

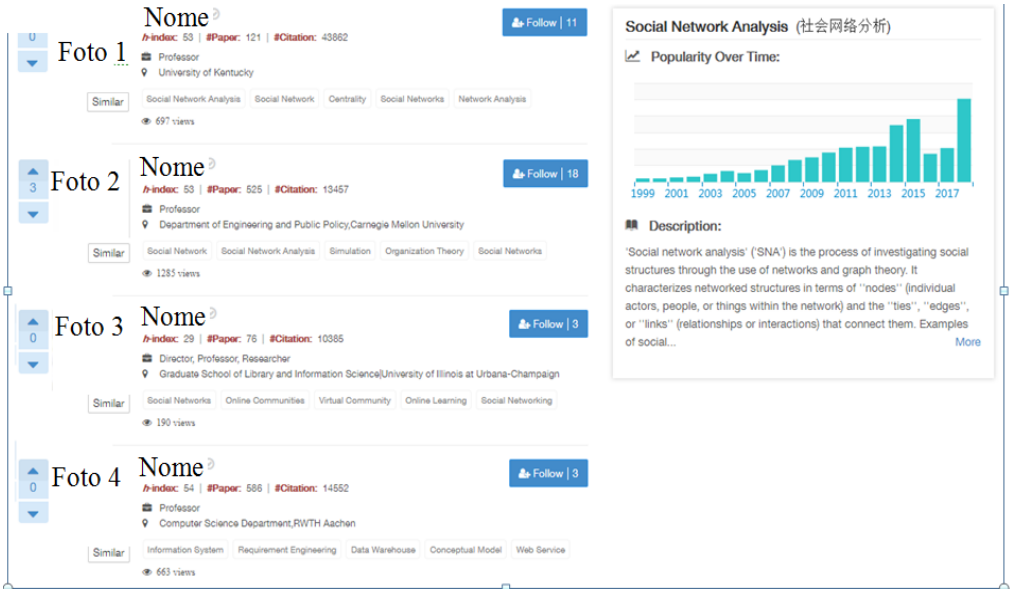

Figura 6 - Exemplo de resultados de especialistas encontrados com a consulta "Social Network Analysis".

3.5 Mineração da Rede Social Acadêmica

Com base no sistema AMiner, esse conjunto de funções de mineração de redes sociais acadêmicas centradas em pesquisadores inclui análise de influência social, mineração de relacionamento, análise de similaridade, recomendação de colaboração e evolução da comunidade.

Análise de Influência Social - Em grandes redes sociais, as pessoas são influenciadas por outros por vários motivos. Propomos um modelo de Propagação por Afinidade de Tópicos (TAP) para diferenciar e quantificar a influência social. O TAP pode obter resultados de qualquer modelagem de tópicos e da estrutura de rede existente para executar a propagação de influência no nível do tópico. Recentemente projetamos uma estrutura de ponta a ponta que chamamos de DeepInf para o aprendizado de representação de características e para prever a influência social. Cada usuário é representado por uma sub-rede local na qual ele está embutido. Uma rede neural gráfica é usada para aprender a representação da sub-rede que, por sua vez, efetivamente integra os recursos específicos do usuário e as estruturas de rede. A estrutura do DeepInf é mostrada na Figura 7.

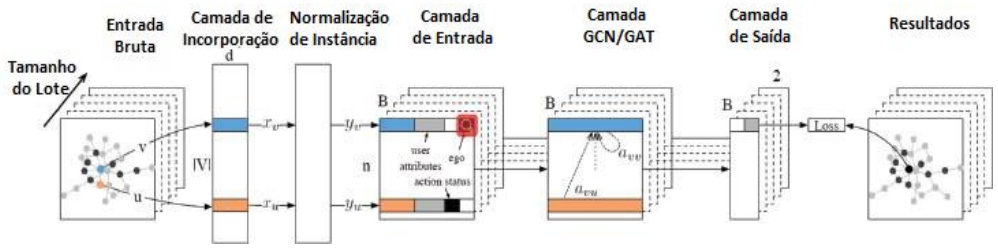

Figura 7 - Estrutura do DeepInf.

Mineração de Relacionamento Social - Inferir o tipo de relacionamento social entre dois usuários é uma tarefa muito importante na mineração de relacionamento social. Propomos um *framework* de dois estágios denominado Modelo Gráfico de Fatores Probabilísticos com restrição de Tempo (TPFG) para inferir relações de conselheiro-assessor na rede de coautores.

A ideia principal é alavancar um modelo de gráfico de fatores probabilísticos com restrição de tempo para decompor a probabilidade conjunta dos conselheiros desconhecidos sobre todos os autores. Além disso, desenvolvemos uma estrutura denominada TranFG para classificar o tipo de relações sociais entre diferentes recursos heterogêneos. A estrutura incorpora teorias sociais em um modelo de gráfico de fator que, efetivamente, melhora a precisão de prever os tipos de relações sociais em uma rede de destino ao emprestar conhecimento de outra rede de origem.

Análise de similaridade - Estimar a similaridade entre os vértices é uma questão fundamental na análise de redes sociais. Propomos um método à base de amostragem para estimar os topo-k vértices similares. O método baseia-se na nova ideia do método de amostragem por caminhos aleatórios conhecido como Panther. Dada uma rede particular como ponto de partida, o Panther gera aleatoriamente vários caminhos de um comprimento predefinido e, em seguida, a similaridade entre dois vértices pode ser modelada como estimativa da possibilidade de que os dois vértices apareçam nos mesmos caminhos.

Recomendação de colaboração - Colaborações interdisciplinares geraram um enorme impacto na sociedade. No entanto, geralmente é difícil para os pesquisadores estabelecerem essas colaborações entre domínios. Analisamos os dados de colaboração entre domínios de publicações de pesquisa e propomos um modelo de Aprendizado de Tópicos entre Domínios (CTL) para recomendação de colaboração. Para lidar com conexões esparsas, o CTL consolida as colaborações entre domínios existentes por meio de camadas de tópicos, em vez de utilizar camadas de autor. Isso alivia o problema de escassez. Para lidar com conhecimentos complementares, o CTL modela distribuições de tópico de domínios de origem e de destino separadamente, bem como a correlação entre domínios. Para lidar com a assimetria de tópicos, o CTL apenas modela tópicos relevantes para a colaboração entre domínios.

Evolução Comunitária - Como as redes sociais são bastante dinâmicas, é interessante estudar como as pessoas nas redes formam *clusters* diferentes e como os vários *clusters* evoluem com o tempo. Estudamos a coevolução de mineração de objetos qualificados em um tipo especial de rede heterogênea, chamada de rede de estrelas. Em seguida, examinamos como os objetos qualificados influenciam uns aos outros na evolução da rede. Foi proposta uma evolução baseada no Modelo de Mistura de Processos Hierárquicos Dirichlet que detecta a coevolução de objetos qualificados sob a forma de uma evolução de *cluster* qualificado em redes estelares dinâmicas. Um algoritmo de inferência eficiente é fornecido para aprender o modelo proposto.

4 Aplicação

O AMiner foi desenvolvido para fornecer serviços abrangentes de pesquisa e mineração para redes sociais de pesquisadores. Neste sistema, nos concentramos em: (1) criar um perfil baseado em semântica para cada pesquisador, extraindo informações da Web distribuída; (2) integrar dados acadêmicos (por exemplo, os dados bibliográficos e os perfis dos pesquisadores) de múltiplas fontes; (3) pesquisar com precisão a rede heterogênea; (4) analisar e descobrir padrões interessantes da rede social pesquisadora construída. As principais funções de pesquisa e análise no AMiner são resumidas na seção seguinte.

Pesquisa de perfil - Digite um nome de pesquisador (por exemplo, Paulo Roberto Freire). O sistema retornará o perfil baseado em semântica criado para o pesquisador usando técnicas de extração de informações. Na página de perfil, as informações extraídas e integradas incluem: informações de contato, foto, estatísticas de citação, avaliação de desempenho acadêmico, interesse de pesquisa (temporal), histórico educacional, gráfico social pessoal, financiamento de pesquisa (atualmente apenas EUA e CN) e registros de publicações (incluindo informações de citação e os documentos que são atribuídos automaticamente a vários domínios diferentes).

Temática especializada - Insira uma consulta (por exemplo, análise de redes sociais). O sistema retornará especialistas neste tópico. Além disso, o sistema irá sugerir a melhor conferência e os principais trabalhos sobre este tópico. Existem dois algoritmos de classificação: VSM e ACT. O primeiro é semelhante ao modelo de linguagem convencional e o segundo baseia-se no nosso modelo de Autor-Conferência-Tópico (ACT). Os usuários também podem fornecer *feedbacks* para os resultados da pesquisa.

Análise de conferência - Digite um nome de conferência (por exemplo, KDD). O sistema retornará aqueles que são os pesquisadores mais ativos nesta conferência, bem como os principais trabalhos.

Pesquisa de curso - Insira uma consulta (por exemplo, mineração de dados). O sistema retornará aqueles que estão ministrando cursos relevantes para a consulta.

Pesquisa de subgráficos - Insira uma consulta (por exemplo, mineração de dados). O sistema primeiro informará quais tópicos são relevantes para a consulta (por exemplo, cinco tópicos "Data mining", "Dados XML", "Mineração de Dados / Processamento de Consultas", "Design de Dados / Banco de Dados da Web" e "Web Mining" são relevantes) e, em seguida, exibirá o subgráfico mais importante descoberto em cada tópico relevante, ampliado com um resumo para o subgráfico.

Navegador de tópicos - Com base em nosso modelo Autor-Conferência-Tópico (ACT), descobrimos automaticamente 200 tópicos importantes das publicações. Para cada tópico, atribuímos automaticamente um rótulo para representar seus significados. Além disso, o navegador apresenta os pesquisadores mais ativos, as conferências / trabalhos mais relevantes e a tendência de evolução dos tópicos descobertos.

Graus Acadêmicos - Definimos oito medidas para avaliar a realização do pesquisador. As medidas incluem "h-index", "Citation", "Uptrend", "Activity", "Longevity", "Diversity", "Sociability" e "New Star". Para cada medida, produzimos uma lista de classificação em diferentes domínios. Por exemplo, pode-se pesquisar aqueles que têm os números de citação mais altos no domínio de "Social network analysis". A Figura 8 dá um exemplo de *ranking* de pesquisadores por índice de sociabilidade.

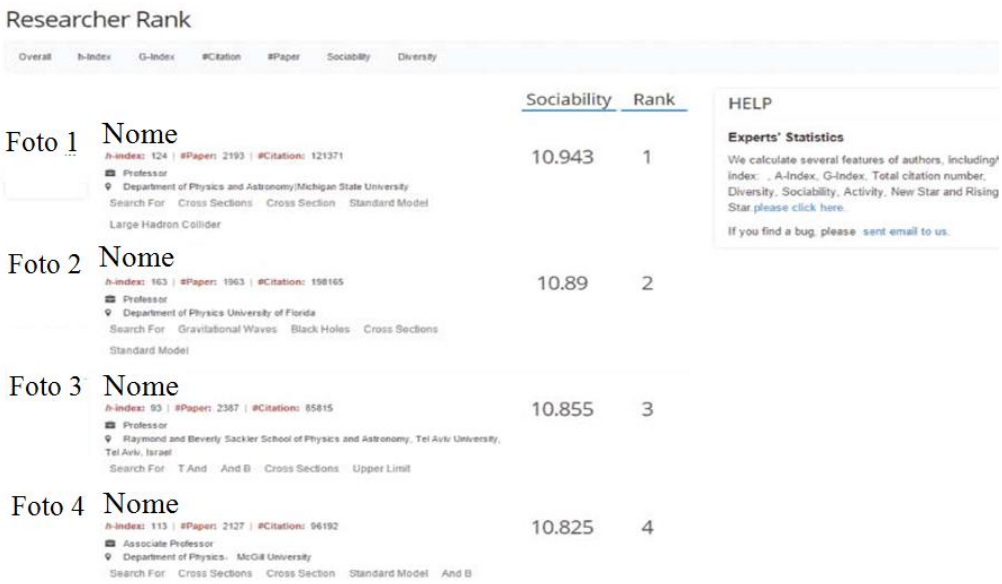

Figura 8 - Exemplo de *ranking* de pesquisadores por índice de sociabilidade.

Gerenciamento de usuários - Pode-se registrar como um usuário para: (1) modificar as informações do perfil extraído; (2) fornecer *feedback* sobre os resultados da pesquisa; (3)

seguem pesquisadores da AMiner; e (4) criar uma página da AMiner (que pode ser usada para anunciar conferências e *workshops*, ou recrutar estudantes).

5 Conjunto de Dados

A AMiner reuniu um grande conjunto de dados acadêmicos com mais de 130.000.000 de perfis de pesquisadores e 233.000.000 de publicações da Internet até junho de 2018, juntamente com vários subconjuntos que foram construídos para diferentes fins de pesquisa. Os detalhes desses subconjuntos são os seguintes e podem ser encontrados em https://aminer.org/.

Rede de Citação - Os dados de citação são extraídos de DBLP, ACM DL e outras fontes. O conjunto de dados contém 1.572.277 artigos e 2.084.019 relações de citação. Cada artigo é associado com resumo, autores, ano, local e título. O conjunto de dados pode ser usado para agrupamento com informações de rede e laterais, estudando a influência na rede de citações, encontrando os documentos mais influentes, a análise de modelagem de tópicos etc.

Rede Social Acadêmica - Esses dados incluem artigos, citação em papel, informações sobre o autor e colaboração do autor. O conjunto de dados contém 1.712.433 autores, 2.092.356 artigos, 8.024.869 relações de citação e 4.258.615 relações de colaboração observadas entre os autores.

Conselheiro-assessor - O conjunto de dados é composto por 815.946 autores e 2.792.833 relações de coautoria. Para avaliar o desempenho de inferir relações de conselheiro-assessor entre coautores, criamos dados menores sobre a verdade do terreno usando o seguinte método: (1) coletar as informações do conselheiro-assessor do projeto Mathematics Genealogy e do projeto AI Genealogy; (2) rastrear manualmente as informações do conselheiro-orientador a partir das páginas iniciais dos pesquisadores. Finalmente, rotulamos 1.534 relações de coautor das quais 514 são relações de conselheiro-assessor.

Coautor do tópico - É uma rede de coautores baseada em tópicos que contém 640.134 autores de 8 tópicos e 1.554.643 relações de coautoria. Os oito tópicos são: Mineração de Dados / Regras de Associação, Serviços da Web, Redes Bayesianas / Função de Crença, Mineração da Web / Fusão de Informações, Web Semântica / Lógicas de Descrições, Machine Learning, Sistemas de Banco de Dados / Dados XML e Recuperação de Informações.

Autor de artigo de tópico - O conjunto de dados é coletado para fins de recomendação de domínio cruzado que contém 33.739 autores associados a 5 tópicos, além de 139.278 relações de coautoria. Os cinco tópicos são Mineração de Dados (com 6.282 autores e relacionamentos de 22862 coautores), Informática Médica (com 9.150 autores e 31851 relações de coautores), Teoria (com 5.449 autores e 27.712 coautores), Visualização (com 5.268 autores) e 19.261 relacionamentos de coautores) e Database (com 7.590 autores e 37.592 relacionamentos de coautores).

Citação-tema - É uma rede de citações baseada em tópicos que contém 2.329.760 artigos de 10 tópicos e 12.710.347 relações de citações. Os 10 tópicos são: Mineração de Dados / Regras de Associação, Web Services, Redes Bayesianas / Função de Crença, Web Mining / Information Fusion, Web Semântica / Lógicas de Descrição, Aprendizado de Máquina, Sistemas de Banco de Dados / Dados XML, Reconhecimento de Padrões / Análise de Imagens, Recuperação de Informações e Sistema de Linguagem Natural / Tradução Estatística de Máquinas.

Comunidade do Kernel - É uma rede de coautoria com 822.415 nós e 2.928.360 extremidades não direcionadas. Cada vértice representa um autor e cada borda representa um relacionamento de coautor.

Coautor dinâmico - O conjunto de dados contém 1.768.776 artigos publicados durante o período de 1986 a 2012 com 1.629.217 autores envolvidos. Cada ano é considerado como um registro de data e hora e há 27 registros de data e hora no total. Em cada registro de data e hora,

criamos uma borda entre dois autores se eles tiverem coautoria de, pelo menos, um artigo nos três anos mais recentes (incluindo o ano atual). Nós convertemos a rede de coautor não direcionada em uma rede direcionada considerando cada borda não direcionada como duas bordas direcionadas simétricas.

Temática especializada - Esse conjunto de dados é uma referência para a descoberta de especialistas, que contém 1.781 especialistas de 13 tópicos.

Pesquisa de associação - Esse conjunto de dados é usado para avaliar a eficácia de abordagens de pesquisa de associação que contém 8.369 pares de autores específicos para nove tópicos. Cada par de autores contém um autor de origem e um autor de destino.

Resultados do Modelo de Tópico para o Conjunto de Dados da AMiner - Há os resultados do modelo ACT no conjunto de dados da AMiner que contém os principais 1.000.000 de artigos e autores de 200 tópicos.

Coautor - Esta é uma rede de coautores no sistema AMiner que contém 1.560.640 autores e 4.258.946 relações de coautoria.

Desambiguação - Esse conjunto de dados é usado para estudar a desambiguação de nomes em uma biblioteca digital. Ele contém 110 autores e suas afiliações, bem como seus resultados de desambiguação (verdade fundamental).

6 Conclusão

Reconhecemos que a AMiner ainda está em fase de desenvolvimento, tanto na escala de recursos quanto na qualidade dos serviços. No entanto, no futuro, vamos explorar métodos inteligentes adicionais para extrair conhecimento profundo de redes científicas e implantaremos uma estrutura mais conveniente e personalizada para fornecer pesquisa acadêmica e encontrar serviços.

Referências

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.