Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Aplicação do Follow Model e Classificação de Entidades de uma Rede Social Acadêmica

Ícaro Araújo Dantas

Monografia apresentada como requisito parcial
para conclusão do Bacharelado em Ciência da Computação

Orientador
Prof. Dr. Li Weigang

Brasília
2015

Universidade de Brasília — UnB
Instituto de Ciências Exatas
Departamento de Ciência da Computação
Bacharelado em Ciência da Computação

Coordenador: Prof. Dr. Homero Luiz Piccolo

Banca examinadora composta por:

Prof. Dr. Li Weigang (Orientador) — CIC/UnB
Prof. Dr. Jorge Carlos Lucero — CIC/UnB
Prof. Dr. Flávio de Barros Vidal — CIC/UnB

CIP — Catalogação Internacional na Publicação

Dantas, Ícaro Araújo.

Aplicação do Follow Model e Classificação de Entidades de uma Rede Social Acadêmica / Ícaro Araújo Dantas. Brasília : UnB, 2015.
77 p. : il. ; 29,5 cm.

Monografia (Graduação) — Universidade de Brasília, Brasília, 2015.

1. Analise de Rede, 2. Follow Model, 3. Medidas de Centralidade

CDU 004.4

Endereço: Universidade de Brasília
Campus Universitário Darcy Ribeiro — Asa Norte
CEP 70910-900
Brasília–DF — Brasil

Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Aplicação do Follow Model e Classificação de Entidades de uma Rede Social Acadêmica

Ícaro Araújo Dantas

Monografia apresentada como requisito parcial
para conclusão do Bacharelado em Ciência da Computação

Prof. Dr. Li Weigang (Orientador)
CIC/UnB

Prof. Dr. Jorge Carlos Lucero
CIC/UnB

Prof. Dr. Flávio de Barros Vidal
CIC/UnB

Prof. Dr. Homero Luiz Piccolo
Coordenador do Bacharelado em Ciência da Computação

Brasília, 10 de dezembro de 2015

Dedicatória

Dedico a minha mãe, Sandra Maria, ao meu pai, Silvestre Dantas e ao meu irmão Mateus.

Agradecimentos

Agradeço a minha família pelo apoio e presença em minha vida; à Empresa Júnior de Computação-CJR, pelas experiências e oportunidades profissionais; a todos meus amigos que estiveram do meu lado; ao professor Li Weigang, pelos seus ensinamentos e por ter me apresentado as oportunidades que existem em áreas de pesquisa;

Resumo

Modelos de medidas de centralidade são úteis em diversas áreas do conhecimento, na computação algumas de suas aplicações são: interpretação de linguagem natural, Análise de de redes sociais, sistemas de classificação automática, ferramentas de busca, inteligência de mercado, dentre outras. Esse trabalho sugere a criação de um novo modelo para a classificação de objetos relacionados ao ambiente acadêmico, e sugere também que podemos utilizar o Follow Model como uma forma de representação para os modelos. Para o desenvolvimento do trabalho e testes dos modelos criados é elaborada uma Micro Rede Social Acadêmica, sendo essa um subgrafo da rede de arquivos disponíveis no Google Acadêmico.

Palavras-chave: Analise de Rede, Follow Model, Medidas de Centralidade

Abstract

Centrality models are nowadays largely used in several scientific areas, in computer science there are some applications as: natural language, social network analisys, automatic classification systems, search engines, business intelligence, and others. This monograph suggests the creation of a new model for the classification of objects related to the academic environment , and also suggests that we can use the Follow Model as a form of representation for models . For the development of work and testing of the models is an elaborate Micro Academic Social Network , this being a subgraph of the network files available in Google Scholar

Keywords: Network analitics, Follow Model, Centrality Measures

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Desde o surgimento da Internet, o potêncial dessa ferramenta vem crescendo junto com o enorme número de aplicações que, hoje em dia, funcionam nesta plataforma, crescimento esse incentivado principalmente por pesquisadores como Timothy John Berners-Lee [2] criador e principal divulgador da WEB. A utilização se tornou tão grande que no último censo apresentado pela WorldWideWebSize.com, feito no dia 13 de Novembro de 2015, aponta a existência de 4.73 bilhões de páginas WEB conhecidas [1].

Junto com a quantidade de conteúdo, cresceu também a quantidade de usuários da Internet que como mostra TSH Teo et al. [23] estão atras dos mais diversos conteúdos e para atender toda essa demanda, mecanismos de buscas cada vez mais sofisticados e eficientes são necessários. Olhando para esse cenário de crescimento podemos listar dois desafios tecnologicos:

1. O desenvolvimento de sistemas capazes de fornecer exatamente a informação que o usuário deseja
2. Executar essa consulta de forma eficiente, ou seja, impossibilitando que o usuário sinta-se desconfortável devido a espera por resultado.

Para solucionar o primeiro problema, sobre a classificação da informação, foram criados diversos modelos matemáticos que quando aplicados sobre alguma estrutura eles retornam como resultado uma classificação, baseados em critérios pré definidos. Esse modelos são chamados de Modelos de Centralidade, aplicados em estruturas que possam ser representadas por grafos e possuem a capacidade de retornar diversar informações, como por exemplo, qual o nó que possui maior influência dentro da rede ou ainda, quem provavelmente iniciou a divulgação de alguma informação dentro de um sistema fechado. Como exemplo podemos citar o modelo criado por L Page et al [20], chamado de PageRank que é um dos mecânismos utilizados pela Google para fazer a classificação dos resultados de busca.

Além disso surgem também formas otimizadas de como executar esses modelos matemáticos por meio de algoritmos de forma otimizada, onde a complexidade do algoritmo seja reduzida a complexidade do problema. Brandes [4] demonstra vários algoritmos que executam diversas variações do modelo chamado *Bettweenness Centrality* introduzido por Freeman [13] em 1977. Outro modelo mais recente que trata das redes sociais, como um todo, foi criado por Sandes [8] chamado de Follow Model e otimizado posteriormente por

Jinaya [17]. Em seu trabalho Sandes demonstra como executar consultas de forma rápida em uma rede social, mostrando os resultados obtidos no desafio proposto pelo WISE 2012, onde 19 queries deveriam ser executadas, com a menor latência possível e a maior vazão.

Como exemplo para aplicação desses modelos de centralidade e busca otimizada, podemos pegar o cenário acadêmico onde, em geral, existe uma busca frequênte por material ciêntifico como livros, artigos, teses, dissertações, documentos técnicos e artigos de trabalho, de qualidade para poder-se embazar o início de uma pesquisa. Segundo Khabsa [19] existem atualmente cerca de 114 milhões de documentos científicos acessíveis na WEB. O mesmo estudo mostra que o Google Scholar, uma ferramenta de busca de materiais acadêmicos, fornece o acesso a 100 milhões de documentos, cerca de 90% do total. Além dele existe ainda o Microsoft Acadêmic Search, Scopus, o portal de periódicos da Capes, dentre outros. Essa análise de dados acadêmicos que envolvem classifição é assunto de estudo da Bibliometria [22].

Fica claro que o problema de classificação em tempo ótimo é algo que pode ser aplicado em diversas áreas, basta que exista um usuário que deseja encontrar alguma informação em meio de várias outras. Por isso esse trabalho propõe a utilização de modelos de centralidade específicos para o cenário acadêmico.

1.1 Motivação

Com o crescimento de material acadêmico disponível na WEB, fica cada vez mais difícil de encontrar informações de referência para uma pesquisa sem um mecanismo inteligênte por trás. Nota-se uma oportunidade de melhorar os mecanismos existentes para seleção de informações fornecidas pelos sistemas de busca, salvando tempo e possivelmente ajudando com a qualidade da pesquisa do usuário. Para isso vê-se no Follow Model uma potêncial ferramenta que pode otimizar a velocidade de busca com as técnicas já conhecidas e demonstradas por Sandes [8] e Jinaya [17].

Apesar de existirem bastantes pesquisas sobre análise de material acadêmico, não se têm conhecimento de modelos que façam uso de características de uma rede heterogênea, ou seja, uma rede que exista mais de um tipo de ator. Por isso esse trabalho busca estudar, implementar e analisar uma forma de classificação que utilize qualidades de diversos níveis da rede para chegar em um resultado melhor, podendo também tirar proveito do Follow Model.

1.2 Objetivos

1.2.1 Principal

Como objetivo principal deste trabalho temos a criação de um modelo que utilize caracteristicas de uma rede heterogênea para classificação de autores e artigos utilizando-se também o Follow Model para garantia de boa performance.

1.3 Específico

Além do objetivo principal este trabalho também possui os seguintes objetivos específicos:

1. Estudar formas de classificação de autores e artigos ciêntificos
2. Mostrar a possíbilidade da utilização do Follow Model em de classificação já existentes
3. Criar modelo de classificação
4. Criar micro rede social acedêmica para realização do estudo de caso
5. Discutir os resultados obtidos, verificando as vantagens e desvantagens do novo modelo criado.

1.4 Descrição dos Capítulos

A organização desse texto reflete as diversas etapas cumprídas para alcançar os objetivos específicos listados na seção anterior.

O Capítulo 2 explica o que é uma Rede Social Online, quais os conceitos matemáticos que representam essa entidade. Capítulo 3 apresenta conceitos da Bibliometria utilizados na análise de material científico. Além disso será apresentadas técnicas de classificação utilizadas atualmente.

No Capítulo 4 será introduzido os conceitos do Follow Model. Diremos como funciona as representações das relações dentro de uma rede, as operações existentes que podem ser aplicadas ao modelo e a técnicas de consulta utilizadas que fazem com que o modelo tenha um bom tempo de execução. No final do capítulo será mostrado também como aplicar o Follow Model nos modelos listados no Capítulo 3.

Chegamos então ao Capítulo 5 será demonstrada a criação da rede utilizada para testes seguido do Capitulo 6 onde é demonstrado o modelo proposto e o estudo de caso feito.

A monografia encerra com o Capítulo 7 onde é feita a conclusão e sugestão de trabalhos futuros.

Capítulo 2

Redes Sociais Onlinde e a Micro Rede Social Acadêmica

2.1 Redes Sociais Online

A necessidade de comunicação da espécie humana é algo que existe desde o início de sua existência. Precisamos da comunicação para avisar situações de riscos, indicar vontades, documentar fatos, dizer aos outros o que está se passando em nossa vida que gostaríamos de compartilhar. Essa necessidade inerente a nós faz com que a cada dia os meios de comunicação estejam mais e mais evoluídos, como ocorre com a telefonia, serviços de e-mail e também as redes sociais onnline Atualmente temos como exemplos de redes sociais mais famosas o Facebook, Tweeter e Instagram.

Segundo Ellison [11], sites de redes sociais são serviços Web que permitem ao individuo:

1. Construir um perfil público ou semi-público em um ecossistema fechado.
2. Manter uma lista de outros usuários com os quais ele mantem algum tipo de relação ou conexão.
3. Visualizar e percorrer a sua lista de conexões e aquelas feitas por outros usuários da rede.

Segundo o autor o que faz uma rede social especial não é o fato dela permitir que o usuário conheça pessoas desconhecidas, mas sim a capacidade que ele tem de articular e fazer visível a sua rede social. Essa características possibilitam as pessoas que relações que jamais seriam criadas sem uma rede social se tornem possíveis.

Outro comportamento percebido após a criação das redes sociais online (OSNs), é a relação que elas possuem com as redes sociais offline, essas segundas são os relacionamentos, ditos como *reais* que existem fora da internet.

Com o crescimentos das Redes Sociais Online, cada vez maior é o interesse em extrair destas características que permitam ao analista obter alguma vantagem, seja ela mercadológica ou acadêmica.

Nesse trabalho consideraremos uma rede social algo mais simplificado, que para existir precisa ter duas caracteristicas:

- Um conjunto de objetos

- Um conjunto de relações entre os objetos

Trasendo essas caracteristicas para o mundo acadêmico, podemos listar como objetos de uma rede social os autores e artigos publicados, e uma relação é criada quando um artigo cita outro, ou quando um autor participa de algum artigo. Por enquanto essas caracteristicas bastam, e serão explicadas mais a fundo nos capítulos seguintes. Agora entraremos em uma representação matemática dessas redes sociais chamada Grafos.

2.2 Grafos

As Redes Sociais Online são comumente representadas através de grafos, devido a isso daremos uma introdução a Teoria dos Grafos focando nos conceitos que utilizaremos no decorrer deste trabalho.

Um grafo é uma tupla $G = (V, E)$, onde V é o conjunto de vértices e, E é o conjunto de arestas cada um representado por uma combinação em $V^{(2)}$ que é o conjunto de todos os pares ordenados de V. Os vértices de um grafo são elementos que podem representa qualquer objeto em estudo e as relações se darão a partir desse objeto escolhido, por exemplo.

Digamos que estamos estudando os estados brasileiros e que a relação entre eles será a existência de caminho entre os estador, se existir um caminho então existirá uma aresta ligando-os. O grafo está representado na Figura 2.1.

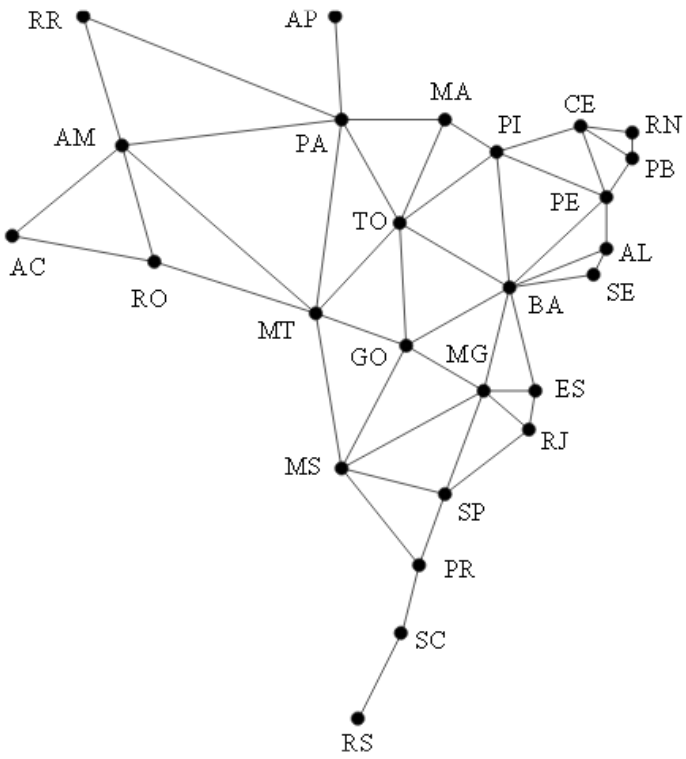

Figura 2.1: Caminhos entre os estados brasileiros. [12]

Outra possibilidade e considerarmos que os vértices são perfis em uma rede social, como o Tweeter. Nesse caso podemos considerar as relações como o fato de um perfil seguir o outro. Nesse caso podemos ter três tipos de relações, vértice a é seguidor de b, vértice a é seguidor por b e, ambos os vértices seguem um ao outro.

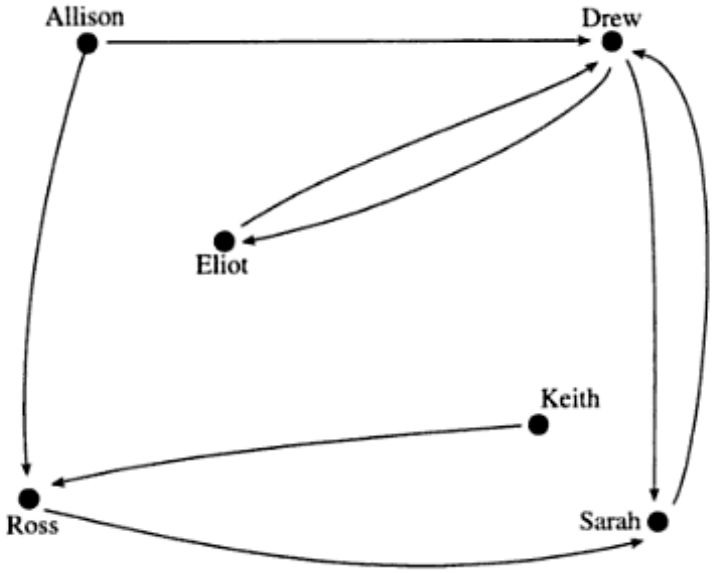

Figura 2.2: Representação de um grafo [24]

Veja que na Figura 2.2 as arestas possuem direção indicando qual perfil está seguindo e qual está sendo seguido. No caso de ser uma relação reciproca então haverá uma seta para cada lado. O grafo está representado na Figura 2.2

O **complemento** de um grafo $G = (V, E)$ é representado por $\overline{G} = (V, V^{(2)} \backslash E)$. Um grafo é dito **completo** se $E = V^{(2)}$ e **vazio** se $E = \emptyset$

2.2.1 Vizinhança

A **vizinhança** de um conjunto X de vértices de um grafo G é o conjunto de todos os vértices que têm algum vizinho em X. A vizinhança é denotada por.

$$\Gamma_G(X)$$

ou simplesmente por $\Gamma(X)$. A vizinhança de um vértice v é o conjunto $\Gamma(v)$, ou simplesmente por $\Gamma(v)$.

Muitas vezes para ajudar na representação das vizinhanças de um grafos utilizamos a **Matriz de Adjacência**, que é uma matriz NxN, onde N é a quantidade de vértices do grafo estudado e as colunas são preenchidas com 1's e 0's, onde 1 significa que existe uma aresta entre os vértices em questão e 0 que não existe tal aresta. Peguemos como exemplo o grafo da Figura 2.2, a matriz adjacência deste grafo seria.

Perceba na tabela que Allison é adjacente a Ross, mas o contrário não é verdade. Isso se da por que o grafo é direcionado e levamos em consideração a direção do caminho. Caso o grafo fosse não direcionado a matriz de adjacência será uma **Matriz Diagonal**

Tabela 2.1: Matriz de Adjacência da Figura 2.2

	Allison	Ross	Eliot	Drew	Keith	Sarah
Allison	-	1	0	1	0	0
Ross	0	-	0	0	0	1
Eliot	0	0	-	1	0	0
Drew	0	0	1	-	0	0
Keith	0	1	0	0	-	0
Sarah	0	0	0	1	0	-

Tabela 2.2: Matriz de Diagonal simples

	A	B	C
A	-	1	0
B	1	-	1
C	0	1	-

onde a parte inferior a diagonal da matriz é o reflexo da parte superior. Na Tabela 2.2 matriz diagonal simples.

2.2.2 Grau de um vértice

O grau de um vértice v é dado pelo número de arestas que incidem nele. O grau de um vértice será aqui denotado como.

$$g_G(v)$$

ou ainda $g(v)$. O grau mínimo de um grafo G é $\delta(G) := ming(v : v\epsilon V)$ e, o grau máximo dado por $\Delta(G) := maxg(v : v\epsilon V)$. Um grafo é regular se todos os vértices possuem o mesmo grau k, nesse caso chamamos o grafo de **k-regular**.

2.2.3 Caminhos

Caminho é qualquer grafo que tenha a forma $(v_1, v_2, ..., v_n, v_iV_{i+1} : 1 \leq i \leq n)$. Ou seja, é um conjunto de vértices que admite uma permutação entre os vértices tal que

$$v_1v_2, v_2v_3, ..., v_{n-1}v_n = A(C)$$

.

Dizemos que v_1 e v_n são os extremos do caminho. Tendo, mais uma vez, a Figura 2.2 como exemplo, podemos encontrar os seguintes caminho no grafo.

1. Allison, Drew, Sarah
2. Allison, Ross, Sarah, Drew
3. Keith, Ross, Sarah, Drew, Eliot, Drew, Sarah

Subgrafos

Um **subgrafo** de um grafo G é qualquer grafo H tal que $V(H) \subseteq V(G)$ e $E(H) \subseteq E(G)$. Um subgrafo H de G é **próprio** se $V(H) \neq V(G)$ ou $E(H) \neq E(G)$. O subgrafo de G **induzido** por um subconjunto X de V(G) é o grafo (X, B) em que B é o conjunto de todas as arestas de G que têm ambas as pontas em X. Esse subgrafo é denotado por.

$$G[X]$$

2.2.4 *Brokerage*

Em alguns casos, os nós de uma rede pode pertencer a diferentes grupos. O índice de quebra basicamente nos indica quais nós são mais importantes para que a troca de mensagens, dentro de seu próprio grupo e, principalmente, entre grupos distintos.

Gould and Fernandez [15] estudaram, previamente, esse indicador levando em consideração subgrafos da forma $a-x-b$, onde a, x e b são nós. Dependendo do questionamento os nós podem ou não pertencer a grupos diferentes. Infelizmente em seus estudos só foram considerados nós, a e b, que estão diretamente ligados por x. No entanto sabemos que a informação pode trafegar por diferentes nós entre a e b antes de ser transmitida de um para outra. Para indicar o índice de quebra de todos esses nós intermediários, utilizamos Q-measure, modelo que será explicado nos próximos capítulos.

Outro conceito ligado a quebra, são os "buracos estruturais"(*Structural holes*). Segundo Burt [6], em uma rede social esses buracos são definidos como regiões desconexas ou pouco conectadas entre outros dois grupos fortemente conectados.

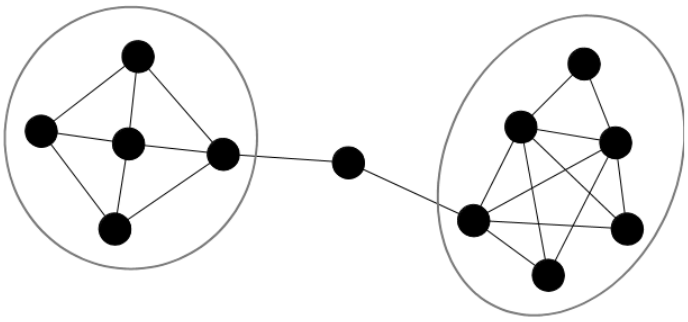

Figura 2.3: Grafo com alto grau de particionamento. [6]

Capítulo 3

Bibliometria e Modelos de Classificação

3.1 Bibliometria

Segundo Guedes [16] a Ciência da Informação é regida por um conjunto de leis e princípios estabelecidos, e a esse conjunto se da o nome de Bibliometria. Segundo ela a expressão Bibliometria foi utilizada pela primeira vez por E. Wyndham Hulme. Pao [21] define Bibliometria como a área de estudo que utiliza matematica e estatística para quantificar objetos ciêntificos, publicações, autores, palavras-chave, livros, periódicos, citações, e usuários. Por último tem a definição de Pitchard [22] que define Bibliometria como toda forma de estudo que possui o objetivo de dar valores aos processos de comunicação escrita.

A leis bibliometricas no geral seguem o ditado "poucos com muito e muitos com poucos" [16]. Isso indica que é comum um cenário onde um cientista com várias citações e trabalhos, recebem maior reconhecimento do que cientistas não tão produtivos. Abaixo discutiremos algumas leis que regem os principios da bibliometria.

Lei de Bradford

Essa lei tenta estimar o grau de relevência de um periódico, em dada área do conhecimento [3]. Bradford sugere que um núcleo de periódicos surge da seguinte forma. Digamos que um primeiro artigo sobre um novo assunto é publicado por um periódico, e aceito pela comunidade, isso fará com que esse periódico atraia mais cientistas e trabalhos, ao mesmo tempo outros periódicos começam a publicas sobre o assunto. Se o assunto continuar emergindo surgirá então um núcleo de periódicos que são os mais produtivos em um dado assunto. Para a ciência da informação a Lei de Bradford é útil em uma politica de aquisição de periódicos, possibilitando estimar o impacto de um periódico e seus custos.

Lei de Lokta

Essa lei vêm para estimar a produtividade de autores, e tem como ideia principal a seguinte premissa de que alguns pesquisadores de uma dada área produzem muito, ao passo que muitos pesquisadores produzem pouco. Exitem estudos que indicam que o número de cientistas que escrevem dois artigos é igual a um quarto do número de pessoas que escrevem um, assim como a quantidade que escreve três é 1/9 da quantidade

de autores que escrevem um. Mais uma vez vemos a relação da premissa básica, onde existem "muitos com pouco e poucos com muito".

Fator de Impacto

Esse é um fator utilizado para medir a influência de um artigo ou periódico publicado. A hipotese utilizada aqui é de que trabalhos citados mais recentemente são mais relevantes para a comunidade do que artigos menos citados. Mas essa premissa tem alguns problemas e causam preocupação aos pesquisadores, pois digamos que existem alguns artigos antigos que são muito citados, mas atualmente não são referenciados com tanta frequência mas continuam com um número muito grande de citações que sempre os mantem no topo da classificação. Devido a esse comportamento ciêntistas sugerem que seja definido um período de tempo no qual as citações devem ser consideradas, digamos, por exemplo, citações que ocorreram a mais de dois anos não irão entrar na contagem. Dessa forma a rotatividade dos trabalhos que estão no topo é maior, no entanto o fator de impacto continua a receber muitas criticas.

Teoria Epidêmica

Goffman [14] faz um paralelo entre a transmição de uma epidemia e a transmição de conhecimento. Goffman diz que para ocorrer uma epidemia são necessários dois elementos:

1. Uma população específica
2. Esposição ao material infeccioso onde a os membros da população estão em uma das categorias
    - infectado
    - sucetível
    - imune

Segundo o autor conhecendo-se caracteristicas expecíficas do cenário é possível prever se a epidêmia irá se proliferar, em quanto tempo, qual será seu ápice e sua duração. O trabalho então faz um paralelo com a passagem de conhecimento. Quando uma pessoa transmite informação a partir de um meio de comunicação a pessoa pode ou não aceitar aquele conhecimento, caso ela aceite ela é tido como infectada e isso fará com que ela transmita para outra pessoa.

3.2 Modelos de Classificação

Como dito no Capítulo 1 existem vários modelos que são capazes de abstrair as principais características de uma estrutura que possa ser representada como um grafo. Neste trabalho não trataremos somente os modelos julgados como mais importântes para o alcance dos objetivos. Nessa seção serão tratados modelos que possam indica, dentro de uma área prédeterminada, qual é o objeto mais indicado de acordo com alguma característica. Esses modelos seguem o conceito chamado *Eagenvector Centrality*, que indicam qual é o nó que iniciou o compartilhamento de alguma caracteristica na rede.

3.3 Katz Centrality

Leo Katz [18] guiou sua pesquisa para que pudesse desenvolver um método que respondesse a seguinte pergunta, "Quem é a pessoa que possui a informação desejada?". Para isso ele propõe uma abordagem diferente da adotada até a época, que era considerar, “quantos indicaram x?” propondo considerar “quem escolheu x?”. Dessa forma nós damos pesos aos votos das pessoas o que pode mudar a configuração de uma classificação dependendo de como esses votos estão distribuídos. Peguemos como exemplo a matriz Z abaixo.

Tabela 3.1: Matriz de adjacência para o problema de Katz

	A	B	C	D	E
A	-	1	0	1	1
B	0	-	0	0	1
C	0	1	-	0	
D	1	0	0	-	1
E	0	0	1	0	-

Nesta tabela indicamos que o ator A indicou B, D e E, o ator B indicou E, e assim por diante. Indicar a si mesmo não é possível e está indicado por ‘-‘. Podemos construir então um vetor s, contendo o número de indicações dada para cada ator, o que nos da s = {1, 2, 1, 1, 3}, onde A recebeu uma indicação, B recebeu duas, C e D receberam uma e E recebeu 3 indicações.

O vetor s é usado para representar o peso do voto de cada ator, e o usamos para recalcular a importância de cada nó considerando os pesos em s. O calculo é feito com a seguinte fórmula.

$$(I - Z^T)t = s$$

Onde I é a matriz identidade, Z é a matriz representada na tabela 3.1 acima, s o vetor de pesos e t o vetor resultante.

Ao recalcular temos t ={ 1, 2, 3, 1, 4}, perceba que o objeto C agora possui peso 3 e o nó D peso 4, isso ocorre pois C foi indicado por E, que possui 3 pontos, logo esses pontos foram tranferidos para C atrávez do voto que lhe foi dado e E possui agora peso 4, pois B passou a ter peso 2, logo o peso atual de E é $1 + 2 + 3$ que são os pesos de A, B e D respectivamente.

Leo Katz [18] ainda adiciona o conceito de atenuação para diferentes camadas, onde a influência de um nó sobre o outro é dissipada a medida que sua distância para o nó em analise aumenta. Sendo a a constante de atenuação e Zk a matriz vizinhança a k passos do nó em analise, temos a função s.

$$(\frac{1}{a^k}I - Z_k^T)t = s$$

Peguemos como exemplo a = 0.5. Temos em C que o ator B votou em E e o peso desse voto será a1.sb = 4. Perceba também que o ator C indica B que por sua vez indicou E, logo C influência em E da seguinte forma a2.sc = 0.25.

3.4 Page Rank

PageRank foi criado por Sergey [5] com o objetivo de de medir a importância relativa de uma página na web, baseando-se no fato de que as páginas da web podem ser representadas a partir de um grafo a medida que as páginas podem ser representadas como nós e as arestas seriam as representações de um hiperlink de uma página para outra. O modelo segue a seguinte formula:

$$PR(a) = c \sum_{v=B_u} \frac{PR(v)}{N_v} + cE(u)$$

Na equação o a que esta sendo analisado, v é um nó que pertence ao conjunto B_U, conjunto de nós que referenciam a, N_v é o número de paginas total que o nó v referencia.

A ideia do modelo é básica e funcionna seguindo a mesma ideia de Leo Katz, no entanto nesse novo modelo, o peso de um nó não passa integralmente para aqueles que ele tem relação, mas sim é dividido igualmente entre todos aqueles que ele possui uma aresta conectando.

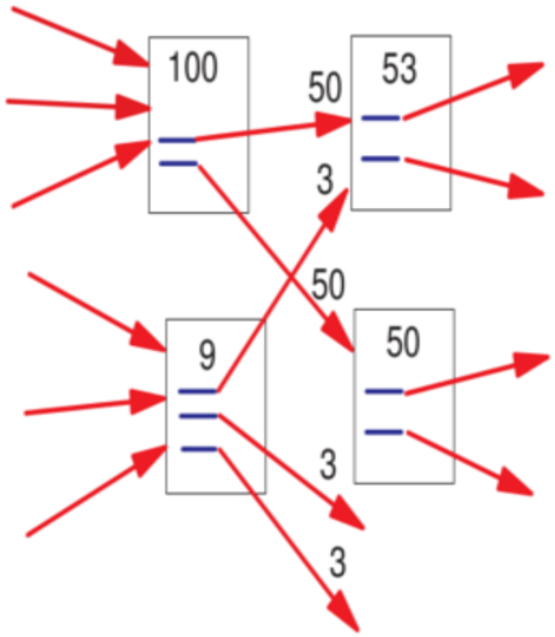

Figura 3.1: Peso da passagem para o PageRank. [5]

Perceba dentro da rede pode existir nós que não referenciam nenhum outro nó, ou talvez uma referência circular, onde uma página A possui referência para B e B para A. Devido a esses problemas que existe a constante c e o vetor $E(u)$ na formula apresentada. A constante c é a constante de atenuação e possui o mesmo funcionamento da constante na fórmula de Katz, já o vetor $E(u)$ é uma inicialização onde todos os nós começam com o peso definido neste vetor, ou também possuiram no minimo esse peso.

3.5 Inventor Rank

Aqui apresentamos o primeiro modelo que trabalha tirando proveito de uma rede heterogênea como mostrado na Figura 3.2, diferente do PageRank e AuthorRank que trabalham somente em redes homogêneas como na Figura **??** e 3.1, onde os nós são todos do mesmo tipo. Em seu trabalho Du et al. (2015) apresenta uma rede heterogênea de patentes e inventores, bem similar a MRSA. A rede é constituída de duas sub-redes. G_I é a rede de coinventores, onde os relacionamentos são não direcionados e representam as coautorias, além disso os relacionamentos possuem peso, que é o número de vezes que os inventores trabalharam juntos. G_IP é uma rede de patentes que não possuem relacionamentos entre si, mas se relacionam com os nós da rede G_I da seguinte forma. Se um inventor da rede G_I participou em alguma patente de G_IP então um relacionamento é criado entre as duas redes. Esse relacionamento possui um peso m_ip, dado em função da posição do inventor na lista de autoria da patente.

$$m_ip = \frac{1}{ordem_p(i)}$$

Onde ordem é a posição em que o inventor i aparece na lista de autoria da patente p.

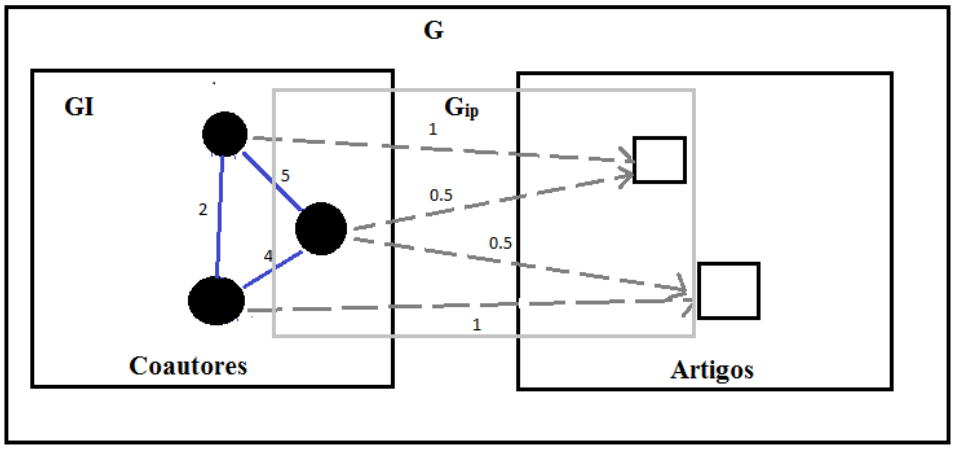

Figura 3.2: Representação da rede heterogênea estudada por Du et al. 2015 [10]

O cálculo de classificação dos nós é formalizado a partir de três regas.

- **Regra 1**: Inventores bem classificados tendem a fazer patentes com outros inventores bem classificados.

$$R_i(k) = a_ii[\sum r = 1^n M_ii(k,r)R_i(r)w_ii + R_i(k)(1 - w_ii)]$$

  Nessa função temos k como o inventor em estudo, r é coinventor de k, M_ii uma matriz composta pelo número de coautorias que cada inventor possui com outro inventor.

- **Regra 2**:Inventores bem classificados geralmente produzem patentes bem classificados.

$$R_p(j) = \frac{a_p i[\sum k = 1^n M_p i(j,k) R_i(k)]}{R_m ax}$$

Aqui k é o inventor da patente j, e $M_p i$ é a matriz composta pela posição de cada inventor na lista de autoria da patente.

- **Regra 3**:Patentes bem classificadas são feitas por inventores bem classificados.

$$R_i(r) = a_i p[\sum_{j=1}^{n} M_i p(r,j) R_p(j) w_i p + R_i(r)(1 - w_i p)]$$

Nessa última função j é a patente do inventor r. $M_i p$ é a matriz simétrica de $M_p i$. Os fatores α, presentes nas três funções são utilizados para denotar a importância da regra no momento do cálculo de classificação, e as constantes w são constantes de atenuação, assim como no algoritmo do PageRank.

A execução do modelo é feita a partir de um ciclo das regras, ou seja, a regra 1 é aplicada a todos os nós possiveis, em seguida aplicamos a regra 2 e por último aplicamos a regra 3. A execução desse ciclo é repetido até que a rede se estabilize e não haja mais mudança de classificação dentro da rede. Ao final teremos a classificação de todos os nós.

Capítulo 4

Follow Model

Follow Model foi criado por Sandes et al [8] com a intenção de representar relações em redes sociais online (OSNs), como Twitter e Weibo. As relações que existem nessas redes são definidas pelo fato de quem segue quem. Para o modelo criado existem três possíveis relações, sendo elas: a é *follower* de b, o que nos leva imediatamente a segunda relação onde b é *followee* (seguido por) de a. A terceira e última relação é no caso em que ambos os nós são seguidores um do outro, sendo a *follower* de b e b *follower* de a, nesse caso dizemos que os nós são *r-friends*. Essas relações são representadas por arestas de um grafo $G = (V, E)$ da seguinte forma. Sendo $a, b \in V$ nós de nossa rede, se $(a, b) \in E$, então a é *follower* de b e b é *followee* de a. Caso também exista uma relação $(b, a) \in E$, os nós também serão chamados de *r-friends*. As funções abaixo demonstram as relações aqui descritas.

- $f_in(a) = v|(v, a) \in E$, é o conjunto de *followers* de a, onde $v \in V*$, V -> V*, $V* \subset V$
- $f_out(a) = v|(a, v) \in E$, é o conjunto de *followee* de a, onde $v \in V*$, V -> V*, $V* \subset V$
- $f_r(a) = f_out(a) \cap f_in(a)$, é o conjunto de todos *r-friends* de a, V -> V*, $V* \subset V$

Para cada função f_in, f_out, f_r, podemos também ter variações como, $f_in^p(a) = p(v)|(v, a) \in E$, onde p(v) é um atributo do nó v, como por exemplo, nome do artigo, classificação, etc. Outra variação é $f_in^w(a) = w(v)|(v, a) \in E$, onde w(v) é um atributo do relacionamento entre a e v, como o peso que um relacionamento pode ter. Além da facilidade de representar relações entre componentes de uma rede, o *Follow Model* possui outras propriedades, como a relação inversa, composição e extensão, que nos ajudam a aumentar a representatividade que o modelo pode ter.

4.1 Aplicações

Jianaya et al [17] utilizaram modelos dos *Follow Model* para efetuar consultas no desafio WISE 2012. E demonstram como o modelo possui bom desempenho para execução das consultas.

4.1.1 Algoritmo *Aggregate-Rank-Delete*

Este algoritmo foi criado por Sandes et al [9] para encontrar os Top-X membros considerando tempo e outras propriedade restritas as Redes Sociais Online. Esse algoritmo se baseia em dois processos, Indexação e *Aggregate-Rank-Delete*.

Indexação

Consideremos que $S = (e_a, t_1), (e_b, t_m), ...(e_n, t_m)$ represente uma série de eventos, assim como *retweet* ou menções. O índice da variável e indica um evento produzido por um usuário da rede, e t o momento que o evento ocorreu. O conjunto $C(e_k)$, representa todos os eventos do usuário k dentro do conjunto S, e $|C(e_k)|$ é o número de ocorrências. $S[t_i...t_j]$ indica o grupo de eventos que ocorreram no intervalo t_i e t_j. Por exemplo, uma mensagem a, do usuário A, foi reenviada no momento t_1 e no momento t_2 pelos seus *followers*; uma mensagem do usuário B, foi reenviada no momento t_3. Existe então o conjunto de eventos $S = (e_A, t_1), (e_A, t_2), (e_B, t_3)$. C é usado para representar o evento de um usuário específico, por exemplo, o evento do usuário A pode ser representado como $C(e_A) = t_1, t_2$ e $|C(e_A)| = 2$ enquanto que B é representado por $C(e_B) = t_3$ e $|C(e_B)| = 1$.

Consultar o número de ocorrências de um evento e_k durante o período $[t_1, t_2]$ se torna uma atividade dispendiosa quando a quantidade de dados é muito grande. Uma maneira simples de manter essa consulta independente de todos usuários e_k ordenamos a lista por tempo, $C(e_k) = t_1, t_2, ...t_m$, onde $t_1 < t_2 < ... < t_m$. Com essa lista podemos, de maneira eficiente, consutar o número de eventos dentro de um período.

A figura 4.1 demonstra como é aplicada a indexação de dados. O evento e_k ocorre treze vezes. Se nós queremos saber o número de ocorrências desse evento durante um período $[t_a, t_b]$, tudo que precisamos é subtrair o ordinal da primeira ocorrência depois de t_a do ordinal da última ocorrência antes de t_b.

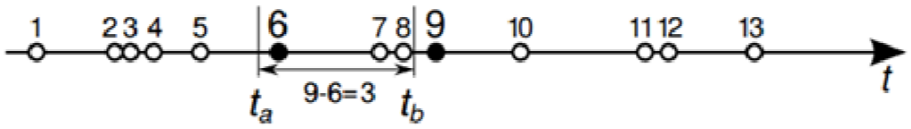

Figura 4.1: Etapa de indexação

4.1.2 *Aggregate-Rank-Delete*

Aggregate-Rank-Delete é a segunda etapa para que as consultas tenham uma boa eficiência, e permitem que pesquisas mais robustas sejam feitas. Como o nome sugere existem três etapas nesse processo: agregação, classificação e limpeza.

Aggregation

Considerando que temos um intervalo $[t_a...t_b]$ que será usado como condição para nossa consulta, e esse intervalo por ser uma hora, um dia, um ano e assim por diante, podemos usar esse intervalo Δt, para dividir nossa linha em diversos pedaços de tempo, sempre escolhendo um pedaço menor do que o tempo estipulado para a consulta. Com isso o custo computacional deve aumentar, mas o tempo da consulta irá tender para valores muito menores.

Em todo intervalo $S[t_s, t_{S+\Delta t}]$ nós podemos obter dois números sobre o evento e_k, $min(e_k, t_S)$ e $max(e_k, t_S)$. Para um dado intervalo $[t_i, ..., t_j]$, $min(e_k, t_S)$ é o menor número de ocorrências durante um período, enquanto $t_j \in [t_S, t_{S+\Delta t}]$. Assim como $max(e_k, t_S)$ é o maior número de ocorrências durante esse intervalo.

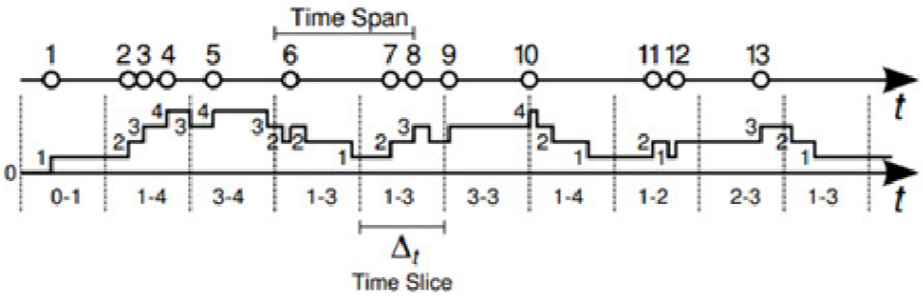

Figura 4.2: Etapa de agregação

A figura 4.2 demonstra a etapa de agregação. O intervalo escolhido está demonstrado na parte superior da figura, enquanto que o tamanho do corte está representado na parte inferior. Pegando o quinto corte, temos que $min(e_k, t_S) = 1$ por que até que o evento 7 ocorra temos dentro do nosso intervalo a ocorrência do evento 6. O $max(e_k, t_S) = 3$ que é a soma dos evento, 6, 7 e 8.

Classificação

Na etapa de agregação consegue-se obter todas as informações desejadas em um dado intervalo pré determinado. Para isso vimos que precisamos de um grande espaço de memória que seja capaz de armazenar toda essa informação. A etapa de classificação foi criada justamente para diminuir esse custo de memória, eliminando dados desnecessários.

Para todo corte de tempo, os usuários são ordenados pelo $min(e_k, t_S)$ e então obtemos uma lista $(e_1, e_2, ..., e_x, e_{x+1}, ..., e_n)$. Digamos que queremos pesquisar os Top-X onde X é a quantidade que um evento deve ocorrer em um dado período. Note que podemos excluir todos os dados que não atendam a um minimo definido, diminuindo o tamanho da nossa pesquisa.

Limpeza

Nessa etapa, elimina-se a quantidade de dados eliminando os dados desnecessários detectados na etapa de classificação. Com isso o conjunto de dados de interesse são reduzidos significativamente e a consulta alcança desempenhos satisfatórios.

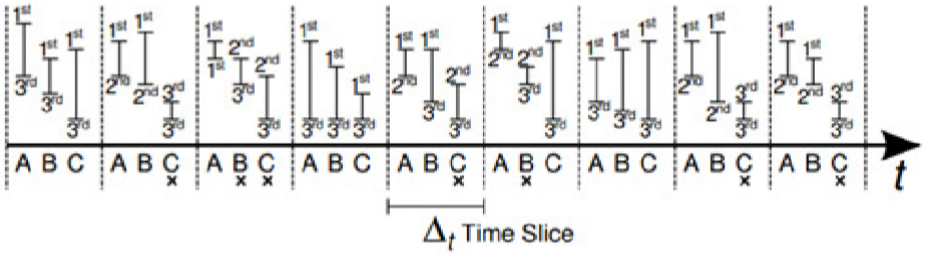

Figura 4.3: Etapa de limpeza

A figura 4.3 mostra o funcionamento da etapa de limpeza, onde X é igual a 1. Para cada pedaço de tempo são armazenados o maior valor de $min(e_A, t_S)$. Olhando para a quinta parte, marcada pelo Δt, vemos que o usuário A possui o maior valor de $min(e_A, t_S)$ dentre os usuário A, B e C, logo A é escolhido para ser analisado posteriormente. No caso de B, seu valor de $max(e_B, t_S)$ é maior que o minimo de A e também será guardado para analise. Já o usuário C não alcançou nenhum resultado que possa ser de interesse então ele será eliminado dos dados de interesse.

4.2 Conversões dos Modelos

A conversão desses modelos para o *follow model* é incentivada para que a interpretação de cada formula fique mais fácil, além de ajudar o programador no momento em que estão sendo montadas as consultas no banco, para o cálculo dos modelos. Os modelos convertidos ficam da seguinte forma.

Conversão do PageRank

O PageRank, calculado de acordo com a Equação 1, pode ser interpretado conforme a equação abaixo:

$$PR(i) = c\left[\frac{s(f_{in}^{p}(i))}{|f_{out}(f_{in}(i))|}\right]$$

Aqui simplesmente substitui-se o somatório, por uma representação com *Follow Model*, onde f_in^p (i) representa a sequência de classificações dos nós que possuem caminho para i, e $s(f_in^p(i))$ é uma função de soma para esses valores. $f_out(f_in(i))$ é o número de nós para onde os seguidores de i possuem caminho.

Conversão do *InventorRank*

Como apresentado na anteriormente, o *InventorRank* é baseado em três regras. A primeira delas é calculada de acordo com a equação a baixo.

$$R_i(k) = a_i i[\sum r = 1^n M_i i(k,r) R_i(r) w_i i + R_i(k)(1 - w_i i)]$$

A equação expressa uma relação de coautoria, e podemos utilizar $f_r(k)$ para representar a sequência de coautores do autor k. A equação fica da seguinte forma.

$$R_i(k) = a_i i\{[s(f_r^w(i).f_r^p(i).w_{ii}) + R_i(k)(1 - w_{ii}).|f_r(k)|]\}$$

Devemos detalhar que a multiplicação entre $f_r^w(i)$ e $f_r^p(i)$ e w_{ii}, deve ocorrer elemento por elemento, ou seja, o elemento na posição 1 de $f_r^w(i)$, vezes o elemento na posição 1 de $f_r^p(i)$, vezes w_{11} e assim por diante. As outras regras permanecem da mesma forma, pois não possuem relações que podem ser representadas pelo *Follow Model*.

Capítulo 5

Micro Rede Social Acadêmica

5.1 Ambientação

Lembramos ao leitor o conceito simplicado de rede social que será usado nesse trabalho. Com odito no Capitulo 2 para existir uma rede social basta que tenhamos as seguintes características:

- Um conjunto de objetos
- Um conjunto de relacionamentos

Tendo em cista que cerca de 114 milhões de documentos ciêntificos como livros, artigos, teses, dissertações, documentos técnicos e artigos de trabalho, são hoje acessíveis pela Web [19] e todos esses documentos fazem referência uns aos outros, podemos considerar essa rede de documentos como uma rede social, onde os objetos são os próprios documentos e as relações são definidas pelas referências feitas entre eles

5.2 Construção da Rede

Atualmente o Google Acadêmico é uma das fontes mais conhecidas além de possuir um dos maiores acervos de material científico na Web, tornando-se uma das principais opções para busca de artigos. Para execução de nossos testes, criamos a Micro Rede Social Acadêmica (MRSA). A MRSA é um subgrafo da rede de documentos do Google Acadêmico que contêm somente documentos de uma área de pesquisa. Foram coletados um total de 195 artigos o que gerou a seguinte rede

5.2.1 Elementos e relações entre eles

Tendo como exemplo a rede coletada a partir do Google Acadêmico, temos como elementos básicos: autores, artigos, eventos e periódicos. Por enquanto nosso trabalho estuda somente artigos e autores, deixando os outros tópicos para trabalhos futuros. Os relacionamentos são formados por citações e coautorias, o que nos leva as seguintes relações:

Relações entre artigos

1. Citados por um artigo p são todos os artigos citados por p.

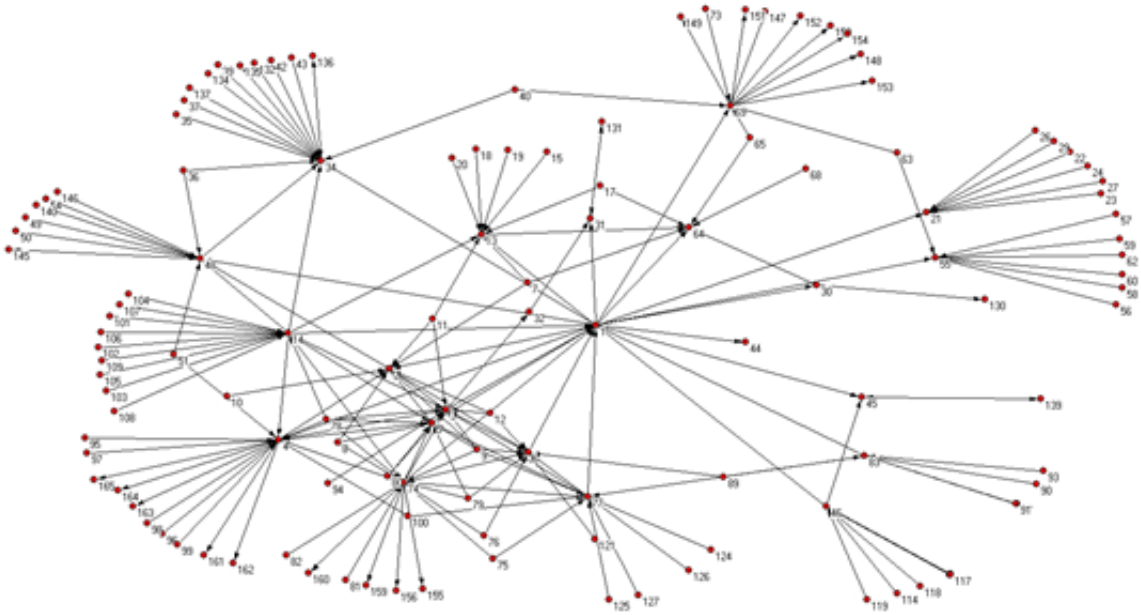

Figura 5.1: 195 artigos coletados do Google e o relacionamento entre eles.

2. Citadores de um artigo p são todos os artigos que citaram p.

Relações de coautoria entre autores Em muitos artigos tempo mais de um autor. Para esses autores a relação entre eles é dita como uma relação de Coautoria onde os autores $a_1, a_2, ..a_n$ que possuem participação no artigo P, são coautores.

Utilizando Follow Model

Followee Dizemos que um nó a é followee de b se b foi citado por a. Essa relação serve para as duas entidades da rede, autores e artigos, pois quando um artigo cita outro, consideramos também que os autores do artigo estão citando autores de outro artigo.

Followee Dizemos que um nó a é follower de b se a cita b. Essa relação serve para as duas entidades da rede, autores e artigos, assim como o caso anterior.

R-Friend Dizemos que dois nós a e b é R-Friend nas situações de coautoria. Essa relação só é valida para autores, pois entendemos que não há como um artigo A citar e ser citado por um artigo B, pois eles são publicados em épocas diferentes.

5.2.2 Construção de rede Homogênea

A partir dos dados coletados foram criadas quatro redes. Duas delas são homogêneas, ou seja, existe somente um tipo de dado em todos os nós que à compõem.

A primeira é uma rede onde o conjunto V é composto de artigos, e as arestas em E representam as relações de citações, ou seja, se $(A, B)\epsilon E$, isso nos indica que o artigo A citou B, e por consequência, que B foi citado por A. A rede está demonstrada na Figura 5.1.

A segunda rede (Figura 5.2) criada é formada pelos autores dos artigos representados na primeira rede. Então o conjunto V é formado por autores e o conjunto E assim como

na primeira é formada pelas citações entre eles. Digamos que um artigo A cita o artigo B. Consideramos então que os autores do artigo A citaram os autores do artigo B, e essa relação é representada como uma aresta $(A^{'}, B^{'})\epsilon E$, onde $A^{'}$ é autor do artigo A e $B^{'}$ autor do artigo B.

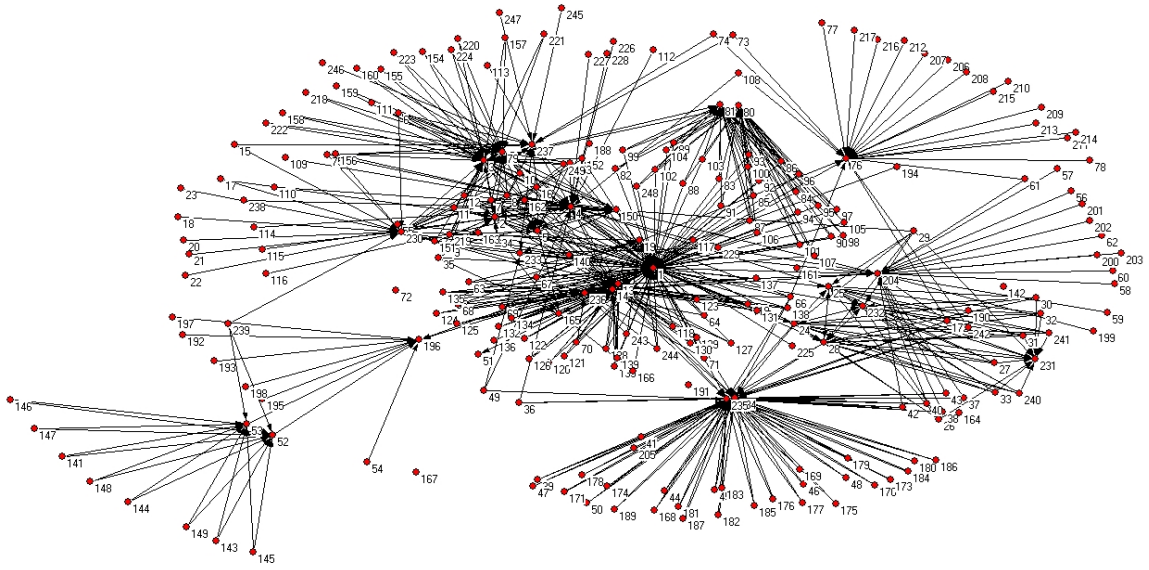

Figura 5.2: 249 autores e a relação de citações entre eles.

A última rede homogênea (Figura 5.3) também é formada por autores e as relações do grafo são as relações de coautoria entre eles.

5.2.3 Construção de rede heterogênea

A última rede é uma rede heterogênea. Nesta, existem dois tipos de nós, que são artigos e autores. Nosso grafo aqui é dado por $G = (V, V^{'}, E, E^{'}, E^{''}, C)$, onde V é o conjunto de artigos e E as relações entre eles, $V^{'}$ é o conjunto de autores e $E^{'}$ a relação entre eles e $E^{''}$ é o conjunto de relações entre autores e artigos. No conjunto $E^{''}$ as relações existentes são relações de participação. Por exemplo, se o autor $A^{'}$, participou do artigo A, então existe uma relação $(A^{'}, A)\epsilon E^{''}$. O conjunto C é o conjunto que representa relações de coautoria. Esta rede esta representada na Figura 5.4.

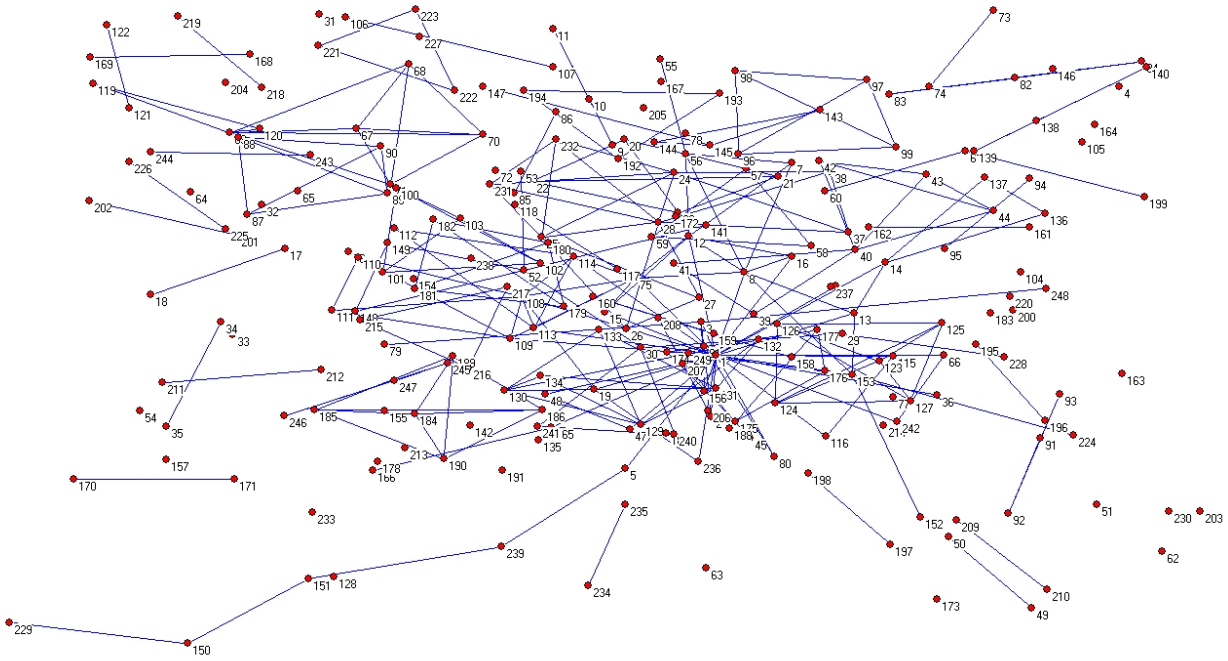

Figura 5.3: Coautores.

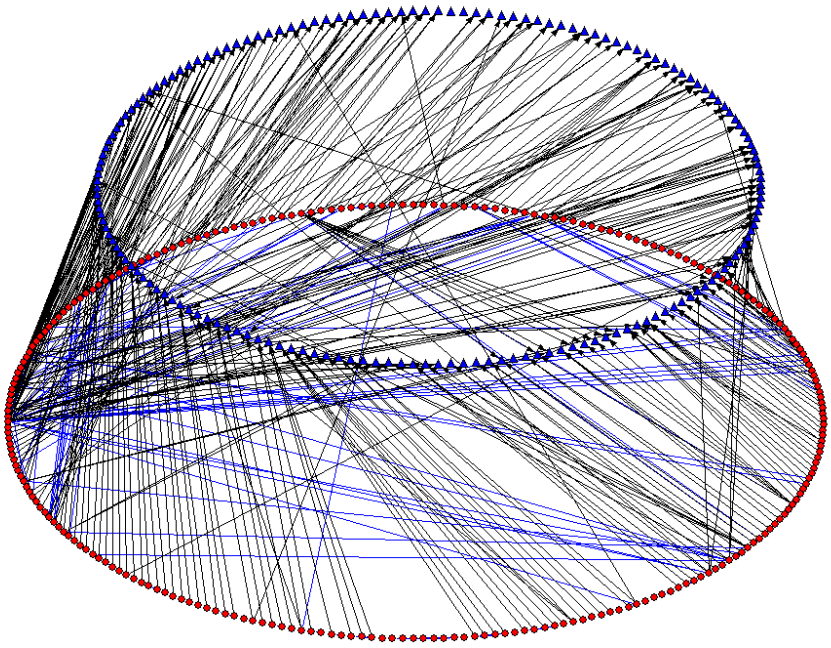

Figura 5.4: Rede heterogênea. Triângulos azuis representam artigos. Círculos vermelhos representam autores.

Capítulo 6

IRank e Estudo de Caso

Essa seção ira apresentar o modelo proposto, chamado IRank, além de estudos anteriores feitos com os modelos de classificação apresentados no Capítulo 3 e alguns conceitos que foram importantes na hora de definir o novo modelo.

6.1 Estudos Anteriores Feitos com MRSN

Weigang [25] demonstra em seu trabalho o funcionamento dos algoritmos, PageRank, AuthorRank e InventorRank em uma MRSN formada por artigos da área de gerência de trafego aéreo. Outro estudo também relatando o comportamento desses algoritmos foi feito por Ícaro [7]. Nesses trabalhos é relatado que os três algoritmos funcionam de forma similar para a classificação dos objetos da rede, tendo em vista que os resultados são estremamente parecidos ao final da classificação. No entando fica clara a superioridade do InventorRank quanto a resistência de inclusão de novas caracteristicas ao seu modelo. Quando é incluído o fator de impacto dos periódicos dentro dos modelos, o PageRank se comporta de forma completamente diferente, enquanto que o InventorRank mantém praticamente os mesmo resultados. Os pesquisadores acreditam que isso ocorra devido ao fato do InventorRank utilizar mais de uma regra para gerar sua classificação. No entando em estudos posteriores feitos com uma rede de topologia distinta, onde os artigos costuman ter um número elevado de autores, percebeu-se uma fragilidade do InventorRank, que é a alta classificação que o modelo atribui a trabalhos que possuem vários autores, chegando a classificar nas primeiras posições, artigos que não possuiam nenhuma citação, mas sim um número extremamente elevado de participantes, problema que também se reflete no PageRank pelo mesmo motivo Devido a esses problemas relatados nos modelos foi pensado um novo modo de classificação que aproveita-se todas as caracteristicas de rede em todas as suas camadas.

6.2 O RankI

O RankI é um modelo criado para classificação de artigos e autores, que explora caracteristicas das duas camadas, tendo uma visão geral do sistema para obter um resultado mais justo ao final da classificação. O modelo segue duas regras básicas:

1. Um autor reconhece um bom trabalho

2. Um autor é reconhecido pelo seu trabalho

A primeira regra visa retirar um problema de interpretação que ocorre no PageRank. Como o PageRank funciona em redes homogêneas, para fazer a classificação de autores é necessário que existam somente autores e a relação é criada considerando que um autor cita outro autor, quando na verdade isso não acontece. No ambiente acadêmico se preza muito pela veracidade dos dados fornecidos em um trabalho e o trabalho não será reconhecido por que as pessoas que o fez, já fizeram bons trabalhos no passado, mas sim por que o novo trabalho possui qualidade e é digna de reconhecimento. Por issa a relação de citação entre autores é ficticia e o que ocorre na realidade é o reconhecimento do artigo publicado. Levando isso em consideração a primeira regra possui a seguinte formula matemática.

$$RP(j) = c \sum_{v=f_{in}(j)} C_{(j,v)} \frac{RP(v)}{\|f_{out}(v)\|} + cE(j)$$

Onde RP é a função de classificação, j é o artigo que está sendo classificado e v é um artigo que pertence ao conjunto dos *followers* de j. A função funciona básicamente como o PageRank reconhecendo a qualidade de um artigo de acordo com o número de citações que ele recebe, com uma diferença. O peso do artigo não passa integralmente para o artigo citado, além de ser dividido entre todos os citados ele tirasse também o peso dos autores de v que já foram coautores dos autores de j. $C_{(j,v}$ é uma matriz de adjacência onde na posição (j, v) tem armazenao a seguinte função:

$$C_{(j,v)} = \frac{N(v) - NC(j, v)}{N(v)}$$

Onde $N(v)$ é o número de autores do artigo v e $NC(j, v)$ é o número de coautores dos autores de j que estão em v. Essa modificação na função de classificação ocorre devido a Teoria Epidemica de Goffman [14]. Como o author havia citado, é possível que as pessoas sejam infectadas por idéias, devido a isso é retirado o peso dos coautores, pois esses são considerados como infectados e estão predispostos a citar trabalhos de pessoas que já tiveram parceria anteriormente.

A segunda regra do modelo é regida pela seguinte fórmula;

$$RI(j) = c \sum_{v=A(j)} P_{(j,v)} RP(v) + cE(j)$$

Onde RI é a função de classificação do autor, RP a função de classificação do autor, $A(j)$ é o conjunto de artigos do autor j. Mas uma vez possuimos uma matriz de adjacência $P(j, v)$ a porcentagem de merito que ele possui no trabalho de acordo com a posição que ele aparece na lista de autores, ou seja, se ele for o primeiro autor receberá mais pontos do que a pessoa que está no final da lista de autoria.

6.3 Estudo de Caso

Nessa seção o experimento feito é descrito em detalhes, para demonstrar como o modelo RankI se sai em comparação com os outros algoritmos

A MRSA, representada na Figura 5.4, foi utilizada para todos, ela que possui no total, 1380 autores e 195 artigos. Os artigos e autores possuem relações de citação, onde um artigo cita outro e um autor cita outro. Há também as relações entre autores e os artigos nos quais tiveram participação, além da relação de coautoria entre autores.

Para todos os testes as constantes dos modelos seguiram o seguinte padrão. Para a constante c do PageRank, AuthorRank e RankI foi utilizado o valor 0.15. Para os parâmetros α_{ii}, α_{ip} e α_{pi}, foram usados os valores 0.4, 0.4 e 0.2 respectivamente. As constantes w_{ii} e w_{ip} foram valoradas, ambas, com 0.5.

6.3.1 Classificação de autores e artigos

O primeiro resultado obtido é sobre a classificação dos artigos de nossa rede indicado pela Tabela 6.1. Verificamos na tabela uma grande diferença entre as colunas do InventorRank com relação ao PageRank e o RankI. Isso se da pela caracteristica do InventorRank de atribuir muitos pontos aos trabalhos que possuem grande número de autores. O Artigo R listado na primeira posição de sua não possui nenhuma citação, mas é um dos artigos que mais possuem autores 20 no total, e seus autores possuem um grande número de coauthores que como veremos influência diretamente na classificação de um autor, e como a regra 3 implica que autores bem classificados geram artigos bem classificados, isso acaba se tornando um efeito cascata, pois o peso do número de autores se soma ao peso de número de coautores. Já as colunas do RankI e PageRank são mais homogêneas apresentando resultados parecidos. Isso se dá pois os dois modelos prezam pelo número de citações que um artigo recebe, com uma pequena diferença que é a de que o RankI diminui o peso de uma citação nos casos em que há coautores no artigo que cita. Essa pequena mudança foi suficiente para mudar a colocação de todos os artigos depois da segunda posição. O Artigo A se mantém em primeiro pois ele possui apenas quatro autores que não possuem coautoria com nenhum outro autor na rede, o que maximiza o valor que ele pode receber para cada citação, além disso o trabalho possui 10 citações dentro da rede de teste o que é um dos maiores valores. Em contra partida o Artigo M, terceiro lugar na classificação do PageRank é rabaixado para oitavo pelo RankI. Nesse artigo 4 de seus autores trabalharam em outros artigos da rede, totalizando 18 artigos trabalhados e dentre esses 18, 4 citaram o Artigo M, o que faz diminuir o valor agregado om as citações que esse recebe.

A Tabela 6.1 nos mostra novamente a discrepãncia do InventorRank com relação aos outros modelos, novamente reforçando o valor que esse modelo dá para o número de coautores de um artigo. Olhando agora para as colunas do PageRank e RankI, vemos que a diferença na classificação aumenta com relação a classificação que ocorreu, na Tabela 6.1. Essa diferênça se da pelo fato do PageRank ser um algoritmo para redes homogêneas, e na rede de autores, o autor terá um seguidos para autor de artigos que citaram seu trabalho, dessa forma artigos que possuem muitos coautores, no momento que citam algum trabalho, automaticamente passam o peso de todos os seus autores para os autores do trabalho citado. Essa é uma característica que não ocorre no RankI, nesse modelo a classificação do Author é de acordo com o classificação de seu trabalho, um bom trabalho vai gerar bons autores, prezando totalmente pelo merito de seu esforço. Além disso temos também a influência da posição do autor no seu trabalho, o que possibilita que pessoas que são sempre os responsáveis pela idéia, tenham classificações melhores dentro da rede.

Tabela 6.1: Top 10 Artigos, resultado obtido pelos modelos

Posição	RankI	InventorRank	PageRank
1	Artigo K	Artigo A	Artigo K
2	Artigo R	Artigo B	Artigo L
3	Artigo L	Artigo C	Artigo M
4	Artigo N	Artigo D	Artigo J
5	Artigo O	Artigo E	Artigo N
6	Artigo J	Artigo F	Artigo O
7	Artigo P	Artigo G	Artigo P
8	Artigo M	Artigo H	Artigo Q
9	Artigo S	Artigo I	Artigo R
10	Artigo Q	Artigo J	Artigo S

Tabela 6.2: Top 10 Autores classificados pelo PageRank, InventorRank e RankI

Posição	RankI	InventorRank	PageRank
1	Autor A	Autor R	Autor A
2	Autor B	Autor S	Autor F
3	Autor C	Autor	Autor B
4	Autor D	Autor U	Autor K
5	Autor E	Autor V	Autor L
6	Autor F	Autor Y	Autor M
7	Autor G	Autor X	Autor N
8	Autor H	Autor W	Autor O
9	Autor I	Autor Z	Autor P
10	Autor J	Autor A1	Autor Q

Capítulo 7

Conclusão e Trabalhos Futuros

7.1 Visão Geral do Trabalho

Esse trabalho discutiu sobre a necessidade de melhoria sobre a forma de classificação de documentos ciêntificos assim como a de seus autores, propondo a criação de um modelo que utilize características de uma rede heterogênea. Os experimentos mostraram que os modelos propostos como PageRank e InventorRank possuem problemas no momento da classíficação, apesar do InventorRank se dar muito bem com a rede testada por Ìcaro [7], vimos que com uma mudança de topologia o modelo não se mantém, gerando resultados que não possuem valor para uma suposta pessoa que necessite de uma classificação de autores e artigos. Vimos também que o PageRank tenta se basear no sucesso de um trabalho, considerando somente o número de vezes que o objeto foi citado dentro da rede, no entando também apresenta problemas por ser um modelo que utiliza dados de somente uma camada no momento da classificação. O RankI se mostrou valorozo, demonstrando que um modelo de classificação mais específico para um determinado tipo de ambiente, o ambiente acadêmico, pode-se alcançar resultados interessantes e mais verdadeiros quando estamos avaliando o merito dos trabalhos. Ficou claro também como é simples de utilizar o Follow Model nas medidas de centralidade, pois como esses modelos são usados em grafos, é fácil encaixar alguma regra que substitua os argumentos do modelo original para um modelo que utilize as representações do Follow Model

7.2 Trabalhos Futuros

Como o objetivo deve trabalho foi estudar um modelo de classificação mais adequado a realidade acadêmica e a utilização do Follow Model em modelos de centralidade. Uma possibilidade de trabalho futuro é o acrescimo de mais caracteristicas que podem ser consideradas na rede, ou o melhor aproveitamento das já utilizadas, por exemplo, seria interessante saber qual o grau de proximidade entre as pessoas que citam o trabalho uma das outras, como uma relação orientador-aluno ou colegas de pesquisa, estudantes de um mesmo departamento ou universidades diferentes, para podermos medir de forma melhor o alcance dos trabalhos realizados. Fica também como sugestão a verificação da otimização de performance que o Follow Model pode ter nos algoritmos PageRank, Inventor Rank e RankI, comparando seu desenpenho com métodos utilizados atualmente.

Referências

[1] WorldWideWebSize.com | The size of the World Wide Web (The Internet). 1

[2] Tim Berners-Lee, Robert Cailliau, Jean-François Groff, and Bernd Pollermann. World-wide web: the information universe. *Internet Research*, 2(1):52–58, 1992. 1

[3] S. C. Bradford. Sources of information on specific subjects. *Egineering*, 137:85–86, 1934. 9

[4] Ulrik Brandes. On variants of shortest-path betweenness centrality and their generic computation. *Social Networks*, 30(2):136–145, 2008. 1

[5] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, 30(1):107–117, 1998. vii, 12

[6] Ronald S Burt. Structural holes and good ideas1. *American journal of sociology*, 110(2):349–399, 2004. vii, 8

[7] Ícaro Araújo Dantas, Li Weigang, and Ahmed Abdelfattah Saleh. Construção de micro rede social acadêmica para análise a influência dos artigos e autores. 24, 28

[8] Edans FO De Sandes, Li Weigang, and Alba Cristina MA de Melo. Logical model of relationship for online social networks and performance optimizing of queries. In *Web Information Systems Engineering-WISE 2012*, pages 726–736. Springer, 2012. 1, 2, 15

[9] Edans FO De Sandes, Li Weigang, and Alba Cristina MA de Melo. Logical model of relationship for online social networks and performance optimizing of queries. In *Web Information Systems Engineering-WISE 2012*, pages 726–736. Springer, 2012. 16

[10] Yao Chang-qing Li Nan Du, Yong-ping. Using heterogeneous patent network features to rank and discover influenctial inventors. *Frontiers of Information Technology & Electronic Engineering*, 2015. vii, 13

[11] Nicole B Ellison et al. Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, 13(1):210–230, 2007. 4

[12] Paulo Feofiloff, Yoshiharu Kohayakawa, and Yoshiko Wakabayashi. Uma introdução sucinta à teoria dos grafos. *Disponıvel em http://www. ime. usp. br/˜ pf/teoriadosgrafos*, 2011. vii, 5

[13] Linton C Freeman. A set of measures of centrality based on betweenness. *Sociometry*, pages 35–41, 1977. 1

[14] William Goffman and VA Newill. Generalization of epidemic theory. *Nature*, 204(4955):225–228, 1964. 10, 25

[15] Roger V Gould and M Roberto. Formal approach to brokerage in. *Sociol Methodol*, 19:89–126, 1989. 8

[16] Vânia LS Guedes and Suzana Borschiver. Bibliometria: uma ferramenta estatística para a gestão da informação e do conhecimento, em sistemas de informação, de comunicação e de avaliação científica e tecnológica. *CINFORM–Encontro Nacional de Ciência da Informação*, 6, 2005. 9

[17] Zheng Jianya, Li Weigang, and Lorna Uden. Top-x querying in online social networks with mapreduce solution. In *The 8th International Conference on Knowledge Management in Organizations*, pages 397–410. Springer, 2014. 2, 15

[18] Leo Katz. A new status index derived from sociometric analysis. *Psychometrika*, 18(1):39–43, 1953. 11

[19] Madian Khabsa and C Lee Giles. The number of scholarly documents on the public web. *PLOS one*, 9(5):e93949, 2014. 2, 20

[20] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. 1999. 1

[21] Miranda Lee Pao. *Concepts of information retrieval*. Englewood, Colo.: Libraries Unlimited, 1989. 9

[22] Alan Pritchard. Statistical bibliography or bibliometrics? *Journal of documentation*, (25):348–349, 1969. 2, 9

[23] Thompson SH Teo, Vivien KG Lim, and Raye YC Lai. Intrinsic and extrinsic motivation in internet usage. *Omega*, 27(1):25–37, 1999. 1

[24] Stanley Wasserman and Katherine Faust. *Social network analysis: Methods and applications*, volume 8. Cambridge university press, 1994. vii, 6

[25] Dantas I. A.-Saleh A. A. Weigang, L. and D Li. Analytical queries within micro scholar social networks, research report in translab. *Report in TransLab*, 2015. 24