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

Modelagem de Influência de Sócios das Redes Sociais pelos PageRank e Índice W-Entropia

Zheng Jianya

Dissertação apresentada como requisito parcial
para conclusão do Mestrado em Informática

Orientador
Prof. Dr. Li Weigang

Brasília
2012

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

Coordenador: Prof. Dr. Mauricio Ayala Rincón

Banca examinadora composta por:

Prof. Dr. Li Weigang (Orientador) — CIC/UnB
Prof.ª Dr.ª Gisele Lobo Pappa — DCC/UFMG
Prof.ª Dr.ª Genaína Nunes Rodrigues — CIC/UnB
Prof. Dr. Guilherme Caribé de Carvalho — ENM/UnB

CIP — Catalogação Internacional na Publicação

Jianya, Zheng.

Modelagem de Influência de Sócios das Redes Sociais pelos PageRank e Índice W-Entropia / Zheng Jianya. Brasília : UnB, 2012.
171 p. : il. ; 29,5 cm.

Dissertação (Mestrado) — Universidade de Brasília, Brasília, 2012.

1. Índice W-Entropia, 2. Redes sociais, 3. PageRank, 4. Teoria da informação

CDU 10/0002960

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

Modelagem de Influência de Sócios das Redes Sociais pelos PageRank e Índice W-Entropia

Zheng Jianya

Dissertação apresentada como requisito parcial
para conclusão do Mestrado em Informática

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

Prof.ª Dr.ª Gisele Lobo Pappa
DCC/UFMG

Prof.ª Dr.ª Genaína Nunes Rodrigues
CIC/UnB

Prof. Dr. Guilherme Caribé de Carvalho
ENM/UnB

Prof. Dr. Mauricio Ayala Rincón
Coordenador do Mestrado em Informática

Brasília, 09 de Março de 2012

Dedicatória

Ao meu pai Zheng Peiquan e à minha mãe Geng Kaiying, que, mesmo estando longe, nunca deixaram de me apoiar.

À minha esposa Sun Yajing pela sua compreensão, amor e carinho.

Ao meu tio Qu Fanyao e à sua família pelo apoio dado durante esta minha jornada.

Agradecimentos

Agradeço imensamente ao meu orientador, Prof. Dr. Li Weigang, pela confiança, pelos seus conhecimentos e pelo incentivo à pesquisa, que contribuíram para a concretização deste trabalho.

Agradeço aos professores da UnB, que me apoiaram sempre que foi necessário. Em especial aos doutores: Alba; Mauricio Ayala; Maria Emilia; Marcelo Ladeira, os quais tive o prazer de assistir suas aulas.

Agradeço a minha grande amiga Déborah Mendes e ao amigo Liu Yang por todo o apoio.

Agradeço aos colegas de pesquisa do grupo TransLab pelo apoio.

Resumo

As redes sociais desempenham um papel cada vez mais importante na comunicação das pessoas, e devido a este fato é necessário que sejamos capazes de medir a influência das pessoas nas redes sociais. Cada plataforma possui a sua lista de classificação para mostrar quem são os membros mais populares, mas esta medida é muito incompleta e unidimensional e a variação dos resultados entre as diferentes listas são sempre discrepantes. Da mesma forma, alguns pesquisadores têm proposto algoritmos computacionais diferentes para avaliar e medir esta influência, mas estes estudos são geralmente muito simples para expressar as características da transmissão de informações.

Este trabalho apresenta uma pesquisa a respeito de como medir a influência dos membros das redes sociais, aplicando o PageRank e a W-Entropia, mais precisamente. Dada uma única rede social, o algoritmo PageRank calcula a importância de cada pessoa com base na ligação intrínseca entre os membros, esse algoritmo é justo e dificilmente os resultados serão manipulados. Dadas diversas redes sociais, a W-Entropia, que utiliza a teoria de Shannon, pode medir o desequilíbrio entre plataformas diferentes durante a transmissão de informações, alcançando assim um resultado mais preciso.

Seguindo essa metodologia, o trabalho desenvolveu o Sistema W-Entropia para medir a influência das pessoas. Este sistema consiste de três partes: a parte do *crawler*, encarregada de coletar os dados e convertê-los para o formato exigido, a parte de cálculo, responsável por calcular a influência da pessoa e a parte de exibição, que exibe a lista de classificação na internet.

De acordo com o experimento, o algoritmo PageRank apresentou uma boa performance dentro de uma única plataforma, já que ele pode efetivamente eliminar a interferência de usuários inativos e obter um valor mais justo de influência. A W-Entropia obtida responde ao desequilíbrio entre plataformas diferentes durante a transmissão das informações. Com a utilização da entropia, o resultado coincidiu melhor com a lei de propagação de informações.

Palavras-chave: Índice W-Entropia, Redes sociais, PageRank, Teoria da informação

Abstract

As social networks play more and more important role in people's daily communication, it is necessary to measure a person's influence in social networks. Currently, every platform has its ranking list to show who the most popular member is. But this measurement is inaccurate and the results between different lists are always different. Similarly, some researchers have proposed various computation algorithms, but these studies are usually too simply to express the features of transmission of information.

This work presents a research that applied PageRank algorithm and W-Entropy index which is based on the theory of information to measure influence more precisely. For a single social network, PageRank calculates the importance of each person with the intrinsic link between members, this algorithm is fair and not easily manipulated. For multi-social networks, Shannon's theory can measure the unbalance between different platforms during the transmission of information, thus achieving the accurate result.

According to the methodology, this work developed W-Entropy system to measure people's influence. This system consists of three parts: the crawler part is in charge of collecting the data and converting them to the requirement format; the computation part is responsible for calculating the people's influence; the display part is for displaying the ranking list in the Internet.

With the experiment result, PageRank algorithm is with a good performance for a single platform, it can effectively remove the interference of inactive users and get a fair influence value. The W-Entropy index obtained from Shannon's entropy responses to unbalance between different platforms during the transmission of information. With the entropy, the result more coincided with the law of information propagation.

Keywords: W-Entropy index, Social networks, PageRank, Information theory

Sumário

1 Introdução **1**
- 1.1 Motivações . . . 2
- 1.2 Objetivos . . . 2
- 1.3 Metodologia . . . 3
  - 1.3.1 Algoritmo PageRank . . . 3
  - 1.3.2 Teoria da informação . . . 3
- 1.4 Organização do Trabalho . . . 4

2 Estado da Arte **6**
- 2.1 Contexto Geral . . . 7
  - 2.1.1 Definição . . . 7
  - 2.1.2 Avaliação . . . 7
- 2.2 Famecount . . . 8
  - 2.2.1 Introdução . . . 8
  - 2.2.2 Metodologia . . . 9
  - 2.2.3 Definição . . . 9
  - 2.2.4 Avaliação . . . 9
- 2.3 Klout . . . 10
  - 2.3.1 Pontuação Klout . . . 10
  - 2.3.2 Metodologia . . . 10
  - 2.3.3 Avaliação . . . 12

3 Revisão Bibliográfia **13**
- 3.1 Cadeia de Markov . . . 13
  - 3.1.1 Introdução . . . 13
  - 3.1.2 Definição Formal . . . 14
  - 3.1.3 Aplicações . . . 14
- 3.2 Algoritmo PageRank . . . 15
  - 3.2.1 Introdução . . . 15
  - 3.2.2 Algoritmo . . . 15
  - 3.2.3 Cálculo do PageRank . . . 18
  - 3.2.4 Propriedades de convergência . . . 19
- 3.3 Teoria da Informação . . . 20
  - 3.3.1 Definição . . . 20
  - 3.3.2 Entropia como conteúdo da informação . . . 20
  - 3.3.3 Representação através de grafos de um processo de Markov . . . 22

3.3.4 Incerteza e Entropia . . . 23
3.3.5 A Entropia de uma fonte de informação . . . 27

4 Modelagem e Arquitetura 28
4.1 Modelagem do Sistema . . . 28
4.2 Arquitetura . . . 28
4.3 Modelo PageRank . . . 29
4.3.1 Definição . . . 29
4.3.2 Cálculo . . . 30
4.3.3 O Processo do PageRank . . . 32
4.4 Modelo W-Entropia . . . 34
4.4.1 Definição do Índice W-Entropia . . . 35
4.4.2 Análise das propriedades do W-Entropia . . . 36

5 Sistema W-Entropia 38
5.1 Ambiente de Desenvolvimento . . . 38
5.1.1 Eclipse Galileo+Biblioteca Jsoup . . . 38
5.1.2 MySQL+PhpMyAdmin . . . 38
5.1.3 PHP . . . 39
5.2 Modelagem do Sistema . . . 39
5.2.1 Crawler . . . 40
5.2.2 Módulos para cálculo . . . 42
5.2.3 Módulo de exibição . . . 46

6 Estudo de Caso 48
6.1 Plano de Estudos . . . 48
6.2 O Cálculo do PageRank dos Jogadores do Flamengo no *Twitter* . . . 49
6.2.1 Introdução . . . 49
6.2.2 Relação do *Twitter* . . . 49
6.2.3 Preparando os dados . . . 50
6.2.4 A iteração do cálculo . . . 52
6.2.5 Resultado . . . 54
6.3 O Cálculo de PageRank do *ScienceNet* . . . 56
6.3.1 Introdução . . . 56
6.3.2 As relações do *ScienceNet* . . . 56
6.3.3 Modelo PageRank Personalidade . . . 57
6.4 O Cálculo do W-Entropia do *ScienceNet* . . . 59
6.5 O Cálculo da Influência em Diversas Plataformas . . . 61
6.5.1 Determinação da Distribuição dos Pesos no Ranking . . . 61
6.5.2 W-Entropia Análise Propriedade no Ranking . . . 62
6.5.3 Comparação de Classificação W-Entropia com Famecount . . . 63

7 Conclusão e Trabalhos Futuros 65
7.1 Conclusão . . . 65
7.2 Trabalhos Futuros . . . 66

Appendices 68

A O Resultado do W-Entropia no ScienceNet **68**

Referências **73**

Lista de Figuras

2.1 O logo do *Famecount* 8
2.2 A lista de classificação do *Famecount* 8
2.3 O Alcance Verdadeiro do *Klout* 11
2.4 A Amplificação do *Klout* 11
2.5 A Rede do *Klout* 12

3.1 Conexão simples 16
3.2 Conexão complexa 16
3.3 Taxas de convergência para um banco de dados com *links* de tamanho total e da metade do tamanho (Page et al. 1998) 19
3.4 Valores para os três exemplos. 21
3.5 Figura do exemplo B 22
3.6 Figura do exemplo C 23
3.7 A figura da incerteza 24
3.8 Entropia, no caso de duas possibilidades com probabilidade p e $(1 - p)$ . . 25

4.1 A arquitetura do modelo 29
4.2 O grafo direcionado para a estrutura da rede social (Page et al. 1998) . . . 33
4.3 O gráfico da Entropia h e $h * m$ com m (Weigang et al. 2011a) 36

5.1 Arquitetura do sistema 40
5.2 O processo de trabalho do *crawler* 41
5.3 O fluxograma do módulo computacional PageRank 44
5.4 O fluxograma do módulo computacional W-Entropia 46
5.5 A arquitetura do módulo de exibição 47

6.1 Relacionamentos do *Twitter* 49
6.2 As relações entre os jogadores do Flamengo 51
6.3 As relações entre os membros do *ScienceNet* 56
6.4 Comparação da Influência Entre o Texas Holdem Poker e Barack Obama . 63
6.5 A Comparação dos Parâmetros Entre Famecount e W-Entropia Classificação 64

Lista de Tabelas

2.1 Top10 do Twitter e do Facebook(Jan/2012) 7

4.1 As relações entre as pessoas 33
4.2 Conjuntos de dados com $n = 3$. 36
4.3 Os valores dos parâmetros de seis conjuntos para todos os termos 37

5.1 A Estrutura da Tabela no Banco de Dados 47

6.1 As contas dos jogadores de futebol do Flamengo 50
6.2 O valor PR para os jogadores de futebol do Flamengo Futebol Clube . . . 55
6.3 O valor PR dos blogueiros do *ScienceNet* 58
6.4 As informações conflito nas três listas 58
6.5 Classificação dos três itens do *ScienceNet* 60
6.6 O Índice W-Entropia dos autores do *ScienceNet* 60
6.7 Comparando o Ranking Famecount e W-Entropia 62

A.1 O resultado do W-Entropia no ScienceNet 68

Capítulo 1

Introdução

Com o rápido desenvolvimento da Internet, as redes sociais passaram a desempenhar um papel muito importante na comunicação da sociedade atual. Uma rede social é um serviço online, uma plataforma ou um site que tem como objetivo a construção e reflexão de redes sociais ou relações sociais entre as pessoas, que, por exemplo, compartilham interesses e/ou atividades. Uma rede social consiste em uma representação de cada usuário (muitas vezes um perfil), seus laços sociais e uma variedade de serviços adicionais. A maioria das redes sociais são online e fornecem meios para os usuários interagam através da Internet, tais como e-mails e mensagens instantâneas. Sites de redes sociais permitem aos usuários compartilhar ideias, atividades, eventos e interesses dentro das suas redes individuais.

Os principais tipos de redes sociais são aqueles que contêm categorias (como pessoas que estudam ou trabalham contigo), meios para se conectar com os amigos (geralmente com páginas auto-descritas) e um sistema de recomendação ligado à confiança. Existem muitas redes sociais populares atualmente, como exemplo o *Facebook* e o *Twitter*, que são amplamente utilizados no ocidente, o *Orkut* e o *Hi5* na América do Sul e América Central, e o *Mixi*, o *Multiply*, o *Orkut*, o *Wretch* e o *Cyworld* na Ásia e Ilhas do Pacífico e o *Facebook* , o *Twitter*, o *LinkedIn* e o *Google+* são muito populares na Índia e no Paquistão. *Sina*, *Sohu*, *Tencent* são redes sociais muito famosas na China, e além delas, existem várias redes sociais para área específicas, por exemplo, o *ScienceNet*, sendo esta uma rede social feita principalmente para os cientistas.

O uso de redes sociais em um contexto corporativo apresenta o potencial de ter um grande impacto no mundo dos negócios e trabalho. Redes sociais conectam as pessoas a um baixo custo, o que pode ser muito interessante para os empresários e as pequenas empresas que procuram expandir suas bases de contato. Estas redes muitas vezes agem como uma ferramenta de gestão de relacionamento com os clientes para as empresas que vendem produtos e serviços. As empresas também podem usar redes sociais para a publicidade na forma de *banners* e anúncios de texto. Uma vez que as empresas operam globalmente, as redes sociais facilitam para que as empresas mantenham contato com os clientes de todo o mundo.

Aplicações para sites de redes sociais têm se estendido para o mundo dos negócios, e as marcas estão criando os seus próprios aplicativos, um setor conhecido como "rede da marca". É a ideia de que uma marca pode construir a sua relação com o consumidor, conectando os seus consumidores à imagem da marca em uma plataforma que lhes fornece

conteúdo, elementos de participação e um sistema de classificação ou pontuação. "Rede da marca" é uma nova forma de capitalizar sobre as tendências sociais, como ferramentas de *marketing*.

1.1 Motivações

Alguns pesquisadores já notaram o potencial enorme do *e-marketing* pelas redes sociais. A página da Coca Cola no *Facebook* possui mais de 36 milhões de fãs (Facebook 2012) e a *StarBucks* possui 2 milhões de seguidores no *Twitter* (Twitter 2012). Isso significa que eles podem enviar as informações do produto a milhões de pessoas sem pagar nada.

Para desenvolver este potencial enorme, é necessário medir a influência do indivíduo na rede social, ajudando a empresa a procurar um porta-voz para propagar informações sobre seus produtos, e também concluir se a sua estratégia de *marketing* está funcionando, auxiliando a gestão a tomar decisões.

Hoje em dia, o método principal para classificar as pessoas de acordo com sua influência nas redes sociais é medir o número de fãs ou seguidores. A maioria das redes sociais publica uma lista de classificação para mostrar quem são as pessoas/marcas mais influentes em sua plataforma, mas esta medida é unilateral e fácil de ser manipulada. Recentemente, surgiram alguns pesquisadores e instituições que estudam algoritmos para medir a influência do indivíduo a partir de diferentes redes sociais, mas existem alguns problemas nos algoritmos simples: seus resultados não podem refletir o desequilíbrio da transmissão da informação entre várias redes.

Acima de tudo, é necessário criar uma medida de influência individual nas redes sociais, a nova medida deve ser justa, razoável e mais abrangente, deve refletir a influência do indivíduo a partir de várias redes sociais.

1.2 Objetivos

Objetivo Geral

Modelar a influência dos indivíduos nas redes sociais e propor um método para medir a influência de maneira justa e compreensível.

Objetivos Específicos

1. Entender os procedimentos propostos por outros pesquisadores no domínio do problema de medir a influência nas redes sociais. Analisar e comparar as vantagens e desvantagens desses métodos.
2. Definição de um modelo computacional para medir a influência dentro de uma rede social. Este modelo é baseado na estrutura entre os relacionamentos existentes dos indivíduos e o a adequação do algoritmo.
3. Definição de um modelo computacional para medir a influência entre diferentes redes sociais. Este modelo integra as principais informações e leva em conta como ocorre a propagação da informação, fornecendo assim, um resultado mais preciso para a influência de um usuário.

4. Implementar um sistema para medir a influência automaticamente. O sistema deve ser capaz de realizar as funções de selecionar dados, calcular com dois tipos de modelos de computação e publicar os resultados.
5. Realização de quatro estudos de caso, conforme os resultados adquiridos pela pesquisa, para atestar e/ou realizar possíveis ajustes ao modelo.

1.3 Metodologia

A metodologia envolve o uso do algoritmo PageRank (Page et al. 1998) e do Índice W-Entropia que aplica a teoria da informação (Shannon 1949) (Shannon 2001).

1.3.1 Algoritmo PageRank

O PageRank é um algoritmo de análise de links usado pelo mecanismo de busca Google, que atribui um peso numérico a cada elemento de um conjunto de links com o objetivo de "medir" a sua importância relativa dentro do conjunto.

A definição do valor PageRank para o elemento u pode ser expressa como:

$$PR(u) = \frac{1-d}{N} + d \sum_{v \in B_u} \frac{PR(v)}{L(v)} \tag{1.1}$$

O valor PageRank para um elemento u é dependente dos valores PageRank para cada elemento v dentro do conjunto B_u (este conjunto contém todos os elementos que se ligam ao elemento u) dividido pelo número $L(v)$ de links do elemento v. A variável d é um fator de amortecimento que geralmente é definida com um valor em torno de 0,85.

1.3.2 Teoria da informação

A entropia é uma medida de desordem ou mais precisamente, de imprevisibilidade. Shannon denotou a entropia H de uma variável aleatória discreta X com os possíveis valores $\{x_1, ..., x_n\}$ como:

$$H(X) = E(I(X)) \tag{1.2}$$

Aqui, E é o valor esperado e I é a informação contida em X.

$I(X)$ é uma variável aleatória. Se p denota a função massa de probabilidade de X, então a entropia pode ser escrita explicitamente como:

$$H(x) = \sum_{i=1}^{n} p(x_i) I(x_i) = \sum_{i=1}^{n} p(x_i) \log_b \frac{1}{p(x_i)} = -\sum_{i=1}^{n} p(x_i) \log_b p(x_i) \tag{1.3}$$

onde b é a base do logaritmo utilizado.

No caso de $p_i = 0$ para algum i, o valor do $log_b 0$ é considerado como 0, o que é consistente com o limite:

$$\lim_{p \to 0+} p \log p = 0 \tag{1.4}$$

1.4 Organização do Trabalho

Esse trabalho se trata de uma pesquisa a respeito do impacto nas redes sociais e propõe um modelo computacional para calculá-la baseando-se no PageRank e na teoria da informação. Segue abaixo um breve resumo de cada capítulo.

Capítulo 1 : Introdução
Inicialmente, esse capítulo trata de uma visão geral dos problemas da influência nas redes sociais, contendo objetivos gerais e objetivos específicos que foram atingidos no decorrer do trabalho. Uma breve descrição das metodologias e tecnologias empregadas também foram abordadas.

Capítulo 2: Estado da Arte
Numa etapa seguinte, foram efetuadas pesquisas e levantamento dos estudos. Pesquisou-se a respeito da Pontuação *Klout*, o Índice-Fame do *Famecount* e outros, esses índices também medem o impacto nas redes sociais e cada um possui suas vantagens e desvantagens, nesse capítulo eles são analisados.

Capítulo 3. Metodologia
Esse capítulo apresenta as metodologias utilizadas para o desenvolvimento deste trabalho. As metodologias utilizadas foram o algoritmo PageRank do Google e a teoria da informação estudada por Shannon.

Capítulo 4: Modelagem
Nesse capítulo são apresentados os modelos computacionais utilizados para calcular os impactos dos indivíduos. O modelo PageRank é responsável por calcular o impacto dos indivíduos em uma única rede social através das relações entre eles. O outro modelo é o W-Entropia que é responsável por calcular os impactos considerando o desequilíbrio durante a transmissão das informações entre redes diferentes.

Capítulo 5: Implementação
Esse capítulo descreve a estrutura do sistema W-Entropia, esse sistema é a implementação do modelo computacional. Ele possui três partes principais: um *crawler* para coletar informações, um núcleo para calcular o impacto de cada indivíduo com os dados obtidos pelo *crawler* e uma terceira parte para tornar público esses resultados.

Capítulo 6. Estudo de caso
Foram realizados quatro estudos diferentes, utilizando dados do *Twitter*, do *ScienceNet*, do *Facebook* e do *Google*. O primeiro estudo com os jogadores de futebol do Flamengo Futebol Clube mostra detalhadamente o processo do modelo PageRank. O estudo do *ScienceNet* mostra a colaboração do modelo PageRank com o modelo W-Entropia para medir o impacto pelas multi-características do indivíduo numa rede social. O estudo do *Facebook*, *Twitter* e *Google* utiliza o modelo W-Entropia para calcular o impacto entre várias plataformas.

Capítulo 7. Conclusão e trabalhos futuros.
Esse capítulo contém as conclusões do trabalho e as sugestões para trabalhos futuros.

Capítulo 2

Estado da Arte

A rápida ascensão das redes sociais online também atrai cada vez mais os interesses de pesquisadores. Mislove et al. (2007) confirmou que as propriedades das leis das potências, de pequeno mundo e de liberdade de escala do mundo real também existem no mundo online. Garton et al. (1997) argumentou sobre a utilidade de um estudo sobre as redes sociais para o estudo da comunicação mediada por computador, revisou alguns conceitos básicos de análise de redes sociais, descreveu como coletar e analisar dados destas redes e demonstrou que estes dados podem ser, e tem sido, utilizados para estudar a comunicação mediada por computador. Jamali (2006) propuseram um roteiro para um estudo a respeito de trabalhar em diferentes aspectos da análise das redes sociais. Kleinberg (2007) discutiu o desafio da mineração de dados nas redes sociais.

E-Marketing é um campo importante de pesquisa nas redes sociais. Hartline et al. (2008) identificou uma família de estratégias de marketing ideais para as redes sociais. Senecal (2004) investigou com consumidores o uso de fontes de recomendação online e sua influência sobre as escolhas de produtos online. Medir a influência do indivíduo nas redes sociais é um dos campos de pesquisa mais ativos nos últimos anos. Gill (2004) revisou a forma de medir a influência da blogosfera na opinião pública e nos meios de comunicação. Anagnostopoulos (2008) estudou a relação entre a influência e a correlação em redes sociais. Kempe et al. (2003) propôs um algoritmo para maximizar a expansão da influência através de uma rede social. Tang (2009) propôs a "Topical Affinity Propagation" para modelar a influência tópica em nível social em redes grandes. Goyal et al. (2010) focou na probabilidade de que o indivíduo possa construir modelos de influência a partir de um grafo social e um log de ações. Katona et al. (2011) estudou o processo de difusão em uma rede social online, dada as conexões individuais entre os membros. Crandall et al. (2008) desenvolveu técnicas para identificar e modelar as interações entre influência social e seleção, utilizando dados de comunidades online, onde ambas se modificam. Trusov et al. (2010) fez uma pesquisa sobre a determinação da influência de usuários nas redes sociais da Internet.

Este capítulo apresenta três métodos de se classificar as pessoas nas redes sociais e mostra quais são as vantagens e desvantagens de cada um.

2.1 Contexto Geral

Cada rede social possui a sua lista de classificação das pessoas. Esta lista é ordenada por uma característica do indivíduo, por exemplo, os fãs do *Facebook* e os seguidores do *Twitter*. Pessoas que tem mais seguidores ou fãs estão no topo da lista. A página do *Facebook* está em primeiro lugar na lista do Facebook com 57.869.909 fãs. Observando outra lista, a cantora *Lady Gaga* está em primeiro lugar na lista do *Twitter* com 17.876.980 seguidores, os dados foram coletados em janeiro de 2012.

2.1.1 Definição

A influência de um indivíduo por esta medida depende de uma característica especial, como número de seguidores, número de fãs, visitas e assim em diante.

$$Inf(u) = P(u) \tag{2.1}$$

Onde u é o indivíduo e P é a característica especial.

Tabela 2.1: Top10 do Twitter e do Facebook(Jan/2012)

	Facebook		Twitter	
1	Facebook	57.869.909	Lady Gaga	17.876.980
2	Texas Hold'em Poker	55.171.290	Justin Bieber	16.288.284
3	Eminem	50.499.033	Katy Perry	13.723.406
4	YouTube	49.032.561	Shakira	12.463.952
5	Rihanna	48.770.414	Kim Kardashian	12.456.244
6	Lady Gaga	46.314.254	Britney Spears	12.302.999
7	Shakira	43.720.295	Barack Obama	11.761.713
8	Michael Jackson	43.421.521	Rihanna	11.670.594
9	Family Guy	40.567.187	Taylor Swift	10.320.518
10	Justin Bieber	39.006.445	Selena Gomez	9.546.144

2.1.2 Avaliação

Estas medidas fornecem informações essenciais sobre os indivíduos, mas existem duas deficiências:

1. Esta medida está de acordo com os dados de uma plataforma específica, cada plataforma tem sua lista de classificação. Considere o presidente Barack Obama como exemplo: ele possui mais de 11,6 milhões de seguidores, estando em sétimo lugar na lista do *Twitter*. No *Facebook* ele possui 24,4 milhões de fãs, sendo metade do número de fãs da cantora Rihanna, que fica abaixo dele na lista do *Twitter*. Isto mostra como as listas são diferentes entre si.

2. Esta medida é fácil de ser manipulada. Nas redes sociais existem muitas contas de usuários inativos que se registram somente para seguir outros usuários. Esse tipo de seguidor não deveria trazer nenhuma influência para o usuário que está sendo seguido.

2.2 Famecount

2.2.1 Introdução

O *Famecount* (Famecount 2012) é um *website* que gera estatísticas dos famosos no *Facebook*, no *Twitter* e no *YouTube*. Ele obtém dados diretamente destes serviços através de um aplicativo e então organiza os dados para produzir uma lista de classificação.

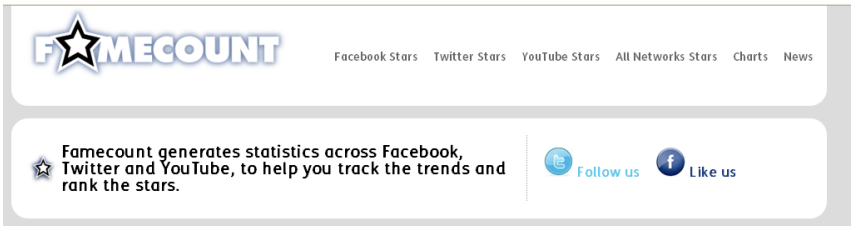

Figura 2.1: O logo do *Famecount*

O formato para exibir os dados é apresentado na figura seguinte:

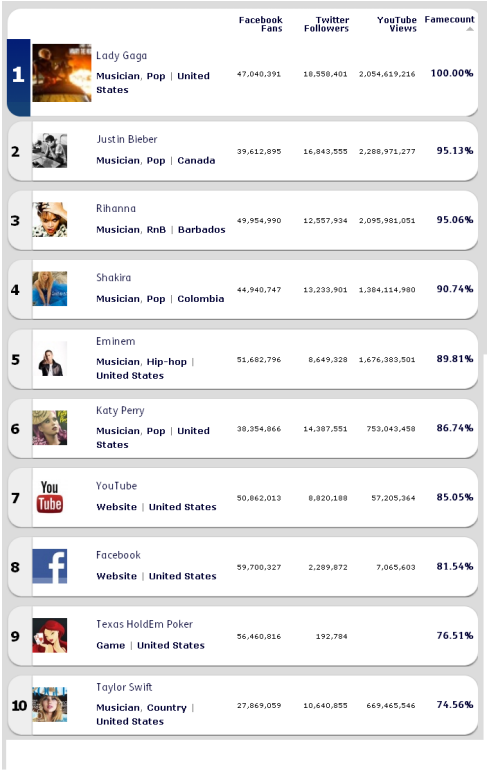

Figura 2.2: A lista de classificação do *Famecount*

Essa figura mostra as informações das três plataformas(*Facebook*, *Twitter* e *Youtube*), na última coluna é exibido o Índice-Fama.

2.2.2 Metodologia

O Índice-Fama é uma medida da popularidade de uma pessoa ou organização nas três redes sociais. Ele lista qualquer pessoa que tenha uma presença oficial no *Facebook* ou em qualquer uma dessas redes sociais. O índice calcula a popularidade de cada usuário relativa ao usuário mais popular de cada rede social. Ela forma uma composição de pontos, juntando duas ou três características, colocando pesos numéricos para refletir a utilização de cada plataforma. Também é fatorado em níveis de interação e comprometimento de cada usuário nas redes sociais, o que é gera um pequeno peso no cálculo. O índice é ajustado para se tornar mais logaritmo(para que aqueles com menos fãs não tenham pontuações *Famecount* insignificantes), a base é modificada para criar uma pontuação de no máximo 100% e ele é atualizado diariamente. Por conta do índice ser calculado em relação ao maior Índice-Fama, o Índice-Fame de um usuário pode cair quando sua popularidade relativa aos outros diminui, mesmo que seu número de fãs ou seguidores esteja crescendo(e vice-versa).

2.2.3 Definição

A partir da metodologia acima, podemos assumir a seguinte definição para o Índice-Fame:

$$Fame(u) = p_1 * \frac{N_1(u)}{N_1(max)} + p_2 * \frac{N_2(u)}{N_2(max)} + p_3 * \frac{N_3(u)}{N_3(max)} + I(u) \tag{2.2}$$

Nessa definição, N_i representa uma rede social específica, $N_i(u)$ representa uma característica de um indivíduo u na rede social N_i, $N_i(max)$ representa a característica de maior valor em N_i, p_n representa o valor do peso de cada rede social e $I(u)$ é o valor do comprometimento e participação do indivíduo, $p_1 + p_2 + p_3 + I(max) = 1$.

$$Fame\text{-}index(u) = \log_{Fame(max)} Fame(u) \tag{2.3}$$

2.2.4 Avaliação

A medida do Índice-Fame integra informações das três principais redes sociais e também considera o nível de interação do indivíduo. Assim, o Índice-Fame é uma maneira abrangente de medir a influência em redes sociais, e para tornar o resultado mais adequado foi utilizado o logaritmo para ajustar o índice. Assim, o resultado do Índice-Fame é mais acurado do que as demais listas de classificação que só levam em conta uma característica.

Mas ainda existem alguns problemas nessa medida:

1. Apesar do *Famecount* integrar dados de diversas redes, ele não considera o desequilíbrio da transmissão de informações. A forma como a informação é propagada não pode ser calculada simplesmente por adição. Por exemplo, considere dois grupos de pessoas: todas as pessoas do primeiro grupo conhecem a pessoa **A** , porém ninguém do segundo grupo a conhece. Com a medida do *Famecount*, o Índice-Fame da pessoa **A** é 50%. Imagine outra pessoa **B**, metade das pessoas do primeiro e do segundo grupo a conhecem, então o Índice-Fame da pessoa **B** também é 50%. Não existe nenhuma diferença entre essas duas pessoas, porém, na realidade não é assim

informação se propaga como a transmissão de uma doença contagiosa, e ela irá se propagar no grupo até atingir um limite máximo.

2. As redes sociais que o *Famecount* levou em consideração não são adequadas, e favorecem somente os artistas. Muitas pessoas não possuem um canal no *YouTube*, exceto os artistas. Por exemplo, o escritor Paulo Coelho tem um bom desempenho no *Facebook* e no *Twitter*, mas recebeu pontuação zero no item *YouTube*, por isso ele está em uma baixa posição na lista do *Famecount*.

2.3 Klout

A Pontuação *Klout* é uma medida da influência dos indivíduos nas redes sociais, e foi desenvolvida por uma empresa de São Francisco - Estados Unidos. A análise é feita sobre os dados obtidos a partir de sites como *Twitter* e *Facebook*. Ela mede o tamanho da rede de uma pessoa, o conteúdo criado e como outras pessoas interagem com esse conteúdo. Os pesquisadores do *Klout* tem sido alvo de críticas substanciais tanto pelo o seu modelo de negócio, quanto pelo seu princípio de funcionamento.

2.3.1 Pontuação Klout

Indivíduos que se inscrevem para o *Klout* ou que estão ligados a aqueles que o fazem, recebem uma "Pontuação *Klout*". Os índices variam de 1 a 100, uma maior pontuação corresponde a uma maior avaliação pelo *Klout* da sua amplitude e da força de sua influência online. A Pontuação *Klout* é dividida em medidas, também variando de 1 a 100, que *Klout* chama "Alcance verdadeiro" (*True Reach*), "Probabilidade de Amplificação" (*Amplification Probability*) e "Índice da rede" (*Network Score*).

A precisão da Pontuação *Klout* tem sido questionada por diferentes pesquisadores, e é usada por comerciantes de mídia social como um barômetro da sua influência.

2.3.2 Metodologia

A medida do *Klout* para a influência usa valores de dados do *Twitter*, tais como: contar o número de pessoas que um indivíduo segue, a contagem de seguidores, o número de *retweets*, o número de membros das listas, quantos *spams* ou contas fantasmas estão seguindo você, quão influentes as pessoas que *retwitam* você são e se o seu conteúdo é original. Essas informações são misturadas com os dados do *Facebook*, como comentários, "curtições" e o número de amigos em sua rede para chegar a uma "Pontuação *Klout*", que mede a influência online do usuário.

Alcance Verdadeiro

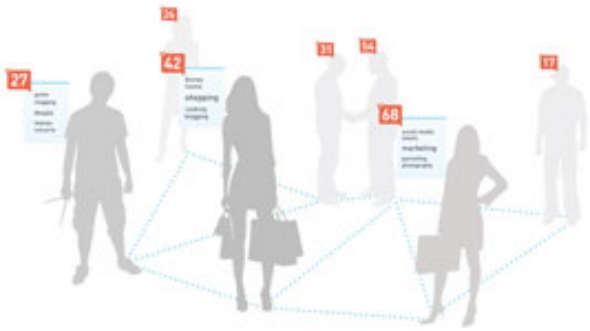

Figura 2.3: O Alcance Verdadeiro do *Klout*

O "Alcance Verdadeiro" (*True Reach*) é o número da influência de uma pessoa. O *Klout* filtra *spams* e *bots* e foca nos usuários que estão agindo sobre o conteúdo das pessoas. Uma pessoa possui um Alcance Verdadeiro alto se, quando esta pessoa posta uma mensagem, outras pessoas tendem a respondê-la ou compartilhá-la.

Amplificação

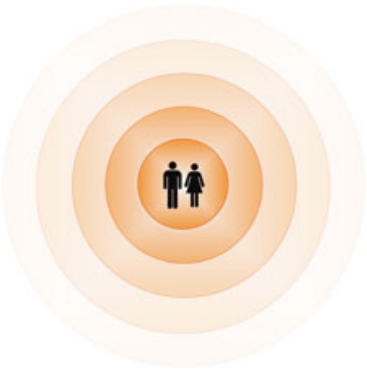

Figura 2.4: A Amplificação do *Klout*

A "Amplificação" (*Amplification*) é o quanto uma pessoa influência as outras. Leva em consideração o número de pessoas que costumam responder ou compartilhar as mensagens postadas pelo indivíduo em consideração. Se as pessoas costumam reagir ao conteúdo de um indivíduo, então ele ou ela tem uma pontuação de Amplificação alta.

Rede

Figura 2.5: A Rede do *Klout*

A "Rede" (*Networking*) indica a influência das pessoas no Alcance Verdadeiro. Leva em consideração quantas vezes as pessoas mais influentes reagem ao conteúdo do usuário em questão. Quando o fazem, eles estão aumentando a pontuação da Rede desta pessoa.

O modelo de negócio gira, então, em torno de conectar empresas com indivíduos de alta influência. As empresas pagam para entrar em contato com os indivíduos com Pontuação *Klout* elevada na esperança de que o recebimento de mercadoria grátis, brindes e outras regalias irão influenciá-los a espalhar publicidade positiva a respeito destas empresas. De acordo com *CEO* do *Klout*, Joe Fernandez, cerca de 50 destas parcerias foram estabelecidas desde novembro de 2011.

2.3.3 Avaliação

A Pontuação *Klout* é considerado o melhor método para se medir a influência atualmente. A base de dados do *Klout* contém as principais redes sociais existentes, como *Facebook*, *Twitter*, *LinkedIn*, *Foursquare*, *Google+*, *Blogger*, *Youtube* e outras. Sua metodologia combina análise estatística e semântica para obter o resultado da influência de um indivíduo. A Pontuação *Klout* do impacto é abrangente e objetiva.

A desvantagem do *Klout* é que o valor do impacto é absoluto, ele não pode refletir o impacto em um campo específico. Ele não permite que você saiba quem é o mais influente em alguma área, por exemplo, não permite saber quem é o político mais influente.

Capítulo 3

Revisão Bibliográfia

Este capítulo descreve as técnicas e os conceitos que serviram para a formalização deste trabalho. A cadeia de Markov, que é a base utilizada pelo PageRank e pela teoria da informação, será mencionada primeiro. O Algoritmo PageRank será descrito em detalhes na segunda seção. A última seção irá descrever a teoria da informação.

3.1 Cadeia de Markov

Uma cadeia de Markov (Pankin 1987), em homenagem Andrey Markov, é um sistema matemático que sofre a transição de um estado para outro, entre um número finito ou enumerável de estados possíveis. É um processo aleatório caracterizado como sem memória: o próximo estado depende apenas do estado atual e não da sequência de eventos que o precederam. Este tipo específico de "perda de memória" é chamada de propriedade de Markov.

3.1.1 Introdução

Formalmente, uma cadeia de Markov é um processo aleatório discreto com a propriedade de Markov. Um processo de tempo aleatório discreto envolve um sistema que está em um determinado estado a cada passo, com o estado mudando aleatoriamente entre as etapas. A propriedade de Markov afirma que a distribuição de probabilidade condicional para o sistema no próximo passo (e de fato, em todas as etapas futuras) depende apenas do estado atual do sistema, e não do estado do sistema em etapas anteriores.

As mudanças de estado do sistema são chamadas transições e as probabilidades associadas com as várias mudanças de estado são chamadas de probabilidades de transição (Usatenko 2009). O conjunto de todos os estados e probabilidades de transição caracteriza completamente uma cadeia de Markov (Meyn et al. 2009). Por convenção, assume-se que todos os possíveis estados e transições foram incluídos na definição dos processos, por isso há sempre um próximo estado e o processo continua infinitamente.

3.1.2 Definição Formal

Uma cadeia de Markov é uma sequência de variáveis aleatórias $X_1, X_2, X_3, \cdots$, com a propriedade de Markov, ou seja, dado o estado atual, os estados futuros e passados são independentes. Formalmente,

$$Pr(X_{n+1} = x \mid X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n) = Pr(X_{n+1} = x \mid X_n = x_n) \tag{3.1}$$

Os valores possíveis de x_i formam um conjunto contável S chamado de espaço de estados da cadeia.

Se o espaço de estados é finito, a distribuição de probabilidade de transição pode ser representada por uma matriz, chamada de matriz de transição, com o (i, j)ésimo elemento de P igual a:

$$p_{ij} = Pr(X_{n+1} = j \mid X_n = i) \tag{3.2}$$

Uma vez que cada linha p possui soma igual a 1 e todos os elementos são não negativos, P é uma matriz estocástica direita.

3.1.3 Aplicações

Existem várias aplicações da cadeia de Markov, aqui são apresentados dois temas relacionados a este trabalho.

Ciência da informação

As Cadeias de Markov são utilizadas em todo processamento de informações. O famoso artigo de Claude Shannon escrito em 1948 chamado "*A mathematical theory of communication*", que em uma única etapa criou o campo da teoria da informação, abre com a introdução do conceito de entropia através de Markov e, a modelagem do idioma inglês é apresentada como exemplo. Tais modelos idealizados podem capturar muitas das regularidades estatísticas dos sistemas. Mesmo sem descrever a estrutura completa do sistema com perfeição, com tal modelo de sinais pode se tornar possível e muito eficaz a compressão de dados por meio de técnicas de codificação da entropia, como a codificação aritmética. Elas também permitem a estimação do estado eficaz e o reconhecimento de padrões.

Aplicações na Internet

O PageRank de uma página web, como o usado pelo Google, é definido por uma cadeia de Markov. Ele é a probabilidade de que outras páginas consigam se ligar à uma página i na distribuição estacionária da cadeia de Markov. Se N é o número de páginas ligadas e uma página i tem K_i links, então ele tem probabilidade de transição $\frac{\alpha}{k_i} + \frac{1-\alpha}{N}$ para todas as páginas que estão ligadas e $\frac{1-\alpha}{N}$ para todas as páginas que não estão vinculadas. O parâmetro α é considerado como cerca de 0,85 (Page et al. 1998).

Os modelos de Markov também têm sido utilizados para analisar o comportamento da navegação web de usuários. A transição de um link por um usuário em um determinado site pode ser modelada utilizando modelos de primeira ou de segunda ordem de Markov e pode ser usada para fazer previsões sobre a navegação futura e para personalizar a página *web* para um usuário individual.

3.2 Algoritmo PageRank

PageRank é um algoritmo de análise de links, seu nome é em homenagem a Larry Page e é usado pela ferramenta de busca da Internet Google, que atribui um peso numérico a cada elemento de um conjunto de hiperlinks de documentos, tais como a *World Wide Web*, com o propósito de "medir" a sua importância relativa dentro do conjunto. O algoritmo pode ser aplicado a qualquer coleção de entidades com citações e referências recíprocas. O peso numérico que ele atribui a qualquer determinado elemento E é referido como o PageRank de E e denotado pelo $PR(E)$.

3.2.1 Introdução

O PageRank (Page et al. 1998)(Franceschet 2011) resulta de um algoritmo matemático baseado no grafo, o *webgraph*, criado por todas as páginas *World Wide Web* como nós e hiperlinks como bordas. O valor da classificação indica a importância de uma página específica. Um hiperlink para uma página conta como um voto a favor para aquela página. O PageRank de uma página é definido recursivamente e depende do número e valor do PageRank de todas as páginas que apontam para ela, chamados de "*incoming links*"(links recebidos). Uma página que está ligada a muitas páginas com PageRank alto recebe uma alta classificação para si. Se não há links para uma página web então não há suporte para esta página.

3.2.2 Algoritmo

O PageRank é uma distribuição de probabilidade utilizada para representar a probabilidade de que uma pessoa aleatoriamente clique em *links* que chegam em qualquer página particular. O PageRank pode ser calculado para coleções de documentos de qualquer tamanho. Supõe-se em diversas pesquisas que a distribuição é dividida igualmente entre todos os documentos da coleção no início do processo computacional. Os cálculos do PageRank exigem várias passagens, chamado de "iterações", através da coleção para ajustar os valores aproximados do PageRank, melhor refletindo o valor teórico da verdade.

A probabilidade é expressa como um valor numérico entre 0 e 1. A probabilidade 0,5 é geralmente expressa como uma "chance de 50%" de algo acontecer. Assim, um PageRank de 0,5 significa que há 50% de chance de uma pessoa clicar em um link aleatório e ser direcionada para o documento com o PageRank 0,5.

Algoritmo Simplificado

Tomemos como exemplo, um pequeno universo de quatro páginas da web: **A**, **B**, **C** e **D**. A aproximação inicial do PageRank será dividida igualmente entre estes quatro documentos. Assim, cada documento começaria com um PageRank estimado de 0,25.

Na forma original do PageRank os valores iniciais eram simplesmente 1, significando que a soma de todas as páginas era o número total de páginas na web naquela época. Versões posteriores do PageRank (fórmulas abaixo) assumem uma distribuição de probabilidade entre 0 e 1. Aqui uma distribuição de probabilidade simples será usada, por isso o valor inicial de 0,25.

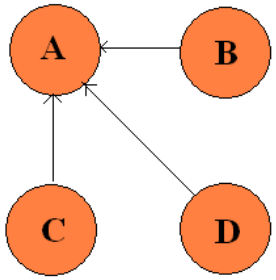

Figura 3.1: Conexão simples

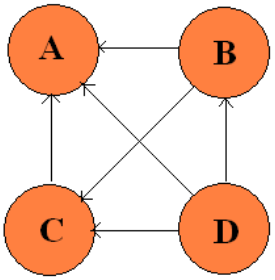

Figura 3.2: Conexão complexa

Se as páginas **B**, **C** e **D** só possuem links para **A**, elas conferem cada uma, um valor de PageRank igual a 0,25 para **A**. Todos os PR neste sistema simplista se reúnem para **A**, porque todos os links são direcionados para **A**.

$$PR(A) = PR(B) + PR(C) + PR(D) \tag{3.3}$$

Dessa forma, o PageRank do **A** vale 0,75.

Suponha que a página **B** possui um link para a página **C** e também para a página **A**, enquanto a página **D** possui links para todas as três páginas. O valor dos "votos" dos links é dividido entre todos os links externos em uma página. Assim, a página **B** fornece um voto de valor 0,125 para a página **A** e um voto de valor 0,125 para a página **C**. Apenas um terço do PageRank de **D** é contado para o PageRank de **A**(aproximadamente 0,083).

$$PR(A) = \frac{PR(B)}{2} + \frac{PR(C)}{1} + \frac{PR(D)}{3} \tag{3.4}$$

Em outras palavras, o PageRank conferido por um link externo é igual ao valor do PageRank do próprio documento dividido pelo número normalizado de links externos L (presume-se que os *links* para URLs específicas sejam contados apenas uma vez por documento).

$$PR(A) = \frac{PR(B)}{L(B)} + \frac{PR(C)}{L(C)} + \frac{PR(D)}{L(D)} \tag{3.5}$$

No caso geral, o valor PageRank para qualquer página u pode ser expresso como:

$$PR(u) = \sum_{v \in B_u} \frac{PR(v)}{L(v)} \tag{3.6}$$

Ou seja, o valor de PageRank para uma página u é dependente dos valores PageRank de cada página v dentro do conjunto B_u (este conjunto contém todas as páginas que ligam para a página u), dividido pelo número $L(v)$ de links da página v.

Fator de amortecimento

O algoritmo PageRank afirma que até mesmo um visitante imaginário que clique aleatoriamente em *links* irá eventualmente parar de clicar. A probabilidade, em qualquer etapa, que a pessoa vá continuar é um fator de amortecimento d. Vários estudos têm testado diferentes fatores, mas é geralmente assumido que o fator de amortecimento será um valor em torno de 0,85.

O fator de amortecimento é subtraído de 1 (e em algumas variações do algoritmo, o resultado é dividido pelo número de documentos N na coleção) e este termo é então adicionado ao produto do fator de amortecimento e a soma dos valores PageRank recebidos. Isto é,

$$PR(A) = \frac{1-d}{N} + d\left(\frac{PR(B)}{L(B)} + \frac{PR(C)}{L(C)} + \frac{PR(D)}{L(D)} + \cdots\right) \tag{3.7}$$

O Google recalcula as pontuações PageRank cada vez que rastreia a *Web* e reconstrói seu índice. À medida que o Google aumenta o número de documentos em sua coleção, a aproximação inicial do PageRank diminui para todos os documentos.

A fórmula usa um modelo de um visitante aleatório que fica entediado depois de vários cliques e muda para uma página aleatória. O valor de PageRank de uma página reflete a chance de que o visitante aleatório vá acessar aquela página clicando em um link. Pode ser entendida como uma cadeia de Markov em que os estados são páginas e as transições são todas igualmente prováveis e são os links entre as páginas.

Se uma página não possui links para outras páginas, torna-se um sorvedouro e, portanto, encerra o processo aleatório de visitas. Se o visitante aleatório chega a uma página de sorvedouro, ele seleciona uma outra URL aleatoriamente e continua visitando novamente.

Ao calcular o PageRank em páginas que não possuem links externos é assumido que elas ligam-se a todas as outras páginas na coleção. Sua pontuação de PageRank é portanto, dividida igualmente entre todas as outras páginas. Em outras palavras, para ser justo com as páginas que não são sorvedouros, estas transições aleatórias são adicionadas a todos nós na *Web*, com uma probabilidade residual normalmente de $d = 0{,}85$, estimada a partir da frequência com que um visitante usa em média o recurso de favoritos do seu navegador.

Então, a equação é a seguinte:

$$PR(p_i) = \frac{1-d}{N} + d \sum_{p_j \in M(p_i)} \frac{PR(p_j)}{L(p_j)} \tag{3.8}$$

onde $p_1, p_2, \cdots, p_N$ são as páginas em consideração, $M(p_i)$ é o conjunto de páginas que apontam para p_i, $L(p_j)$ é o número de links externos na página p_j e N é o número total de páginas.

Os valores de PageRank são as entradas do autovetor dominante da matriz de adjacência modificada. Isso torna o PageRank uma métrica particularmente elegante, o autovetor é:

$$R = \begin{vmatrix} PR(p_1) \\ PR(p_2) \\ \vdots \\ PR(p_N) \end{vmatrix} \tag{3.9}$$

onde R é a solução da equação :

$$R = \begin{vmatrix} \frac{1-d}{N} \\ \frac{1-d}{N} \\ \vdots \\ \frac{1-d}{N} \end{vmatrix} + d \begin{pmatrix} \ell(p_1,p_1) & \ell(p_1,p_2) & \cdots & \ell(p_1,p_N) \\ \ell(p_2,p_1) & \ddots & & \vdots \\ \vdots & & \ell(p_i,p_j) & \\ \ell(p_N,p_1) & \cdots) & & \ell(p_N,p_N) \end{pmatrix} R \tag{3.10}$$

onde a função de adjacência $\ell(p_i, p_j)$ é 0 se a página p_j não aponta para p_i, e é normalizada tal que, para cada j

$$\sum_{i=1}^{N} \ell(p_i, p_j) = 1 \tag{3.11}$$

Isto é, os elementos de cada coluna somam até 1, então a matriz é uma matriz estocástica.

3.2.3 Cálculo do PageRank

O cálculo do PageRank é bastante simples se ignorarmos as questões de escala. Seja S qualquer vetor sobre páginas da *Web*, então o PageRank pode ser calculado da seguinte forma:

$R_0 \leftarrow S$
$loop:$
$R_{i+1} \leftarrow AR_i$
$g = ||R_i||_1 - ||R_{i+1}||_1$
$R_{i+1} \leftarrow R_{i+1} + gE$
$\delta \leftarrow ||R_{i+1} - R_i||_1$
$while \delta < \epsilon$

Note que o fator g aumenta com a taxa de convergência e mantém $||R||_1$. Uma normalização alternativa é multiplicar R pelo fator apropriado. O uso de g pode ter um pequeno impacto na influência de E.

3.2.4 Propriedades de convergência

Como pode ser visto a partir do gráfico da figura 3.3 o PageRank em um grande banco de dados com 322.000.000 *links* converge para uma tolerância razoável em aproximadamente 52 iterações. A convergência de metade dos dados leva cerca de 45 iterações. Este gráfico sugere que a escala PageRank funciona muito bem até mesmo para coleções extremamente grandes, já que o fator de escala é aproximadamente linear no $\log(N)$.

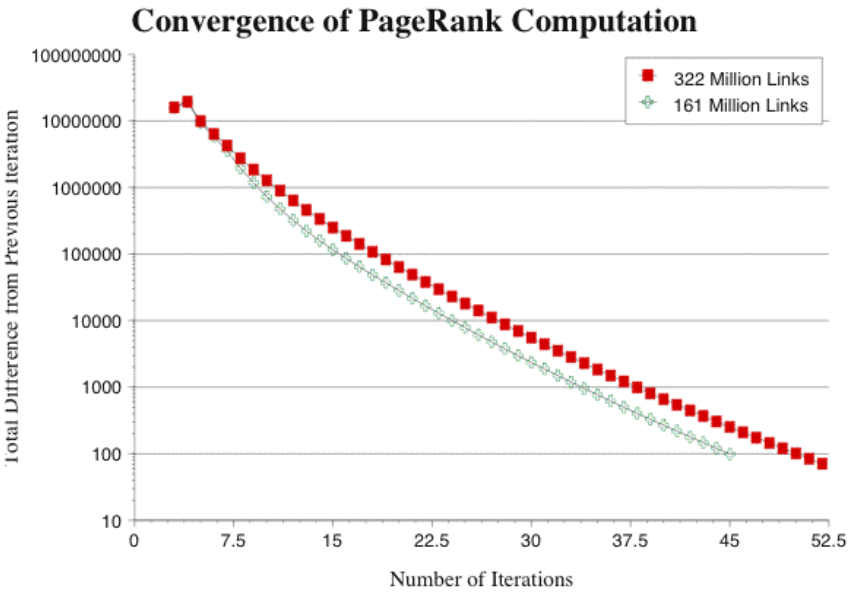

Figura 3.3: Taxas de convergência para um banco de dados com *links* de tamanho total e da metade do tamanho (Page et al. 1998)

Uma das ramificações interessantes de que o cálculo PageRank converge rapidamente é o fato da Web ser um grafo em expansão. Para entender isso melhor, apresentamos um breve panorama da teoria de passeios aleatórios em grafos. Um passeio aleatório em um grafo é um processo estocástico, onde em qualquer passo de tempo dado, estando em um nó específico do grafo, são escolhidas arestas externas uniformemente de forma aleatória para determinar o nó para visitar no próximo passo de tempo. Um grafo é dito expansor se for o caso em que cada subconjunto (não muito grande) de nós S possui uma vizinhança, que é maior do que algum fator α $|S|$ vezes, α é chamado de fator de expansão. Um grafo tem um bom fator de expansão se, e somente se, o maior autovalor for suficientemente maior do que o segundo. Um passeio aleatório em um grafo é dito ser "*rapidly-mixing*" se ele rapidamente converge para uma distribuição limitante no conjunto de nós do grafo (tempo logarítmico no tamanho do gráfico). Um passeio aleatório pode ser dito "*rapidly-mixing*" em um grafo, somente se o grafo for um expansor ou tiver uma separação de autovalor.

Implementação no MATLAB

O seguinte programa foi desenvolvido utilizando o MATLAB.

```
function [v] = rank(M, d, v_quadratic_error)
N = size(M, 2);
v = rand(N, 1);
v = v ./ norm(v, 2);
last_v = ones(N, 1) * inf;
M_hat = (d .* M) + (((1 - d) / N) .* ones(N, N));
while(norm(v - last_v, 2) > v_quadratic_error)
        last_v = v;
        v = M_hat * v;
        v = v ./ norm(v, 2);
end
```

3.3 Teoria da Informação

Na teoria da informação, a entropia é uma medida da incerteza associada a uma variável aleatória. Neste contexto, o termo geralmente refere-se à entropia de Shannon, que quantifica o valor esperado da informação contida em uma mensagem.

A entropia de Shannon é uma medida do conteúdo de informação média (Shannon 2001). Um conteúdo está em falta quando não se sabe o valor da variável aleatória. O conceito foi introduzido por Claude E. Shannon em seu artigo " A Teoria Matemática da Comunicação " de 1948.

3.3.1 Definição

Nomeada em homenagem ao teorema de Boltzmann-H, Shannon denotou a entropia H de uma variável aleatória discreta X com os possíveis valores $x_1, \cdots, x_n$ como,

$$H(X) = E(I(X)) \tag{3.12}$$

Aqui, E é o valor esperado, e I é o conteúdo de informação de X.

$I(X)$ é uma variável aleatória. Se p denota a função de massa de probabilidade de X então a entropia pode ser explicitamente escrita como

$$H(x) = \sum_{i=1}^{n} p(x_i) I(x_i) = \sum_{i=1}^{n} p(x_i) \log_b \frac{1}{p(x_i)} = -\sum_{i=1}^{n} p(x_i) \log_b p(x_i) \tag{3.13}$$

Onde b é a base do logaritmo usado.

No caso em que $p_i = 0$ para algum i, o valor do $\log_b 0$ é considerado 0, fato que é consistente com o limite:

$$\lim_{p \to 0+} p \log p = 0 \tag{3.14}$$

3.3.2 Entropia como conteúdo da informação

Essa seção apresenta os casos matemáticos em que apenas define-se abstratamente um processo estocástico que gera uma sequência de símbolos.

A) Suponha que temos cinco letras **A**, **B**, **C**, **D** e **E**, que são escolhidas, cada uma com probabilidade de 20%, considerando que escolhas sucessivas são independentes. Isto levaria a uma sequência como a seguinte:
B D C B C E C C C A D C B D D A A E C E E A
A B B D A E E C A C E E B A E E C B C E A D

B) Usando as mesmas cinco letras, considere que as probabilidades sejam de 40%, 10%, 20%, 20% e 10%, respectivamente, com escolhas sucessivas independentes. Uma mensagem típica gerada é:
A A A C D C B D C E A A D A D A C E D A
E A D C A B E D A D D C E C A A A A A D

C) Uma estrutura mais complicada é obtida se símbolos sucessivos não forem escolhidos de forma independente e se suas probabilidades dependem das letras anteriores. No caso mais simples deste tipo, uma escolha depende apenas da letra anterior e não daquelas que vieram antes. A estrutura estatística pode ser descrita por um conjunto de probabilidades de transição $p_i(j)$. A probabilidade de que a letra i seja seguida pela letra j. Os índices i e j tem um alcance sobre todos os símbolos possíveis. Uma segunda maneira equivalente de especificar a estrutura é fornecer o "digrama" de probabilidades $p(i,j)$, ou seja, a frequência relativa do digrama i e j. A frequência de letras $p(i)$ (a probabilidade da letra i), as probabilidades de transição $p_i(j)$ e as probabilidades de digrama $p(i,j)$ estão relacionadas pela seguinte fórmula:

$$p(i) = \sum_j p(i,j) = \sum_j p(j,i) = \sum_j p(j)p_j(i) \tag{3.15}$$

$$p(i,j) = p(i)p_i(j) \tag{3.16}$$

$$\sum_j p_i(j) = \sum_i p(i) = \sum_{i,j} p(i,j) = 1 \tag{3.17}$$

Como um exemplo específico suponha que há três letras **A**, **B** e **C** com as tabelas de probabilidade:

$p_i(j)$		j		
		A	B	C
i	A	0	$\frac{4}{5}$	$\frac{1}{5}$
	B	$\frac{1}{2}$	$\frac{1}{2}$	0
	C	$\frac{1}{2}$	$\frac{2}{5}$	$\frac{1}{10}$

i	$p(i)$
A	$\frac{9}{27}$
B	$\frac{16}{27}$
C	$\frac{2}{27}$

$p(i,j)$		j		
		A	B	C
i	A	0	$\frac{4}{15}$	$\frac{1}{15}$
	B	$\frac{8}{27}$	$\frac{8}{27}$	0
	C	$\frac{1}{27}$	$\frac{4}{135}$	$\frac{1}{135}$

Figura 3.4: Valores para os três exemplos.

Uma mensagem típica desta fonte é a seguinte:
A B B A B A B A B A B A B A B B B A B B B B B A B A B A B A B A B A B B
B A C A C A B B A B B B B A B B A B A C B B B A B A

O próximo aumento na complexidade envolveria frequências trigrama, mas não mais. A escolha de uma letra dependeria das duas letras precedentes, mas não da mensagem antes desse ponto. Um conjunto de frequências trigrama $p(i, j, k)$ ou, equivalentemente, um conjunto de probabilidades de transição $p_i j(k)$ seria necessário. Continuando, desta forma obtém-se processos estocásticos sucessivamente mais complicados. No caso geral n-grama, um conjunto de n-gramas probabilidades $p(i_1, i_2, \cdots, i_n)$ ou de probabilidades de transição $p_{i_1,i_2,\cdots,i_n}(i_n)$ é necessário para especificar a estrutura estatística.

3.3.3 Representação através de grafos de um processo de Markov

Processos estocásticos do tipo descrito acima são conhecidos matematicamente como processos discretos de Markov e têm sido amplamente estudados na literatura. O caso geral pode ser descrito da seguinte forma: existe um número finito de possíveis "estados" de um sistema: $S_1, S_2, \cdots, S_n$. Além disso, há um conjunto de probabilidades de transição: $p_i(j)$ é a probabilidade de que, se o sistema está no estado S_i, ele vá para o próximo estado S_j. Para tornar esse processo de Markov uma fonte de informação, precisamos apenas supor que uma letra é produzida para cada transição de um estado para outro. Os estados corresponderão ao "resíduo de influência" das letras anteriores.

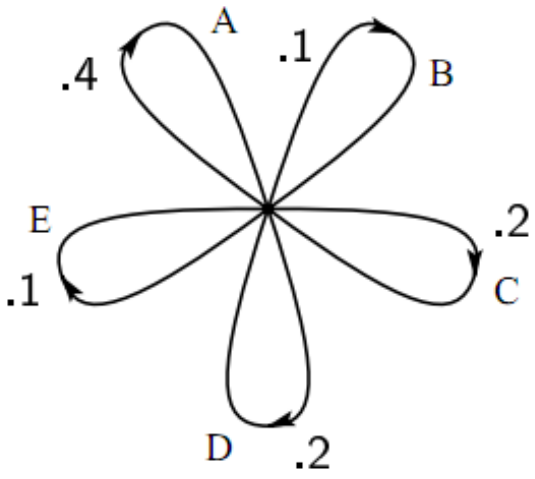

Figura 3.5: Figura do exemplo B

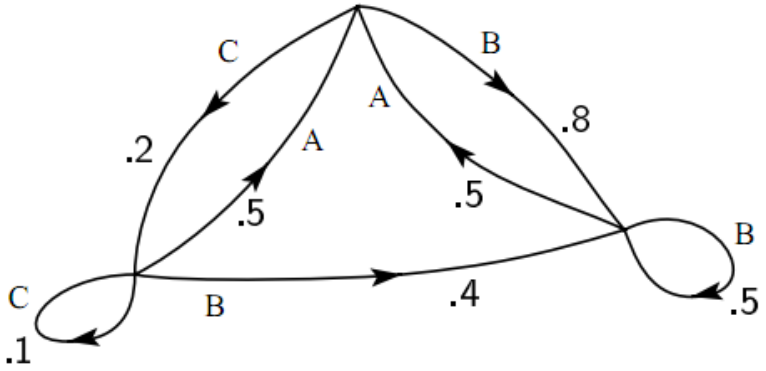

Figura 3.6: Figura do exemplo C

As situações podem ser representadas através de grafos como os das Figuras 3.5 e 3.6 acima mostram. Os "estados" são os pontos de junção no grafo e as probabilidades e letras produzidas para uma transição são dadas ao lado da linha correspondente. Na figura do exemplo B, há apenas um estado, já que letras sucessivas são independentes. Na figura do exemplo C, existem os mesmos números de letras e estados.

Se um exemplo de trigrama fosse construído, haveria no máximo n^2 estados correspondentes para cada um dos pares possíveis de letras anteriores à que está sendo escolhida.

3.3.4 Incerteza e Entropia

Representamos uma fonte de informação discreta como um processo de Markov. Podemos definir uma quantidade que irá medir, em alguns casos, quanta informação é "produzida" por tal processo, ou melhor, qual taxa de informação é produzida?

Suponha que temos um conjunto de eventos possíveis, cujas probabilidades de ocorrência são $p_1, p_2, \cdots, p_n$. Essas probabilidades são conhecidas, mas é tudo que sabemos sobre qual evento irá ocorrer. Podemos encontrar uma medida de quanta "escolha" está envolvida na seleção do evento ou de como estamos incertos do resultado?

Se houver uma medida deste tipo, digamos $H(p_1, p_2, \cdots, p_n)$, é razoável exigir dela as seguintes propriedades:

1. H deve ser contínua em p_i.

2. Se todo p_i for igual a $p_i = \frac{1}{n}$, então H deve ser uma função monotônica crescente de n. Com eventos igualmente prováveis, há mais escolha, ou incerteza quando há mais eventos possíveis.

3. Se uma escolha pode ser dividida em duas escolhas sucessivas, a original H deve ser a soma ponderada dos valores individuais de H. O significado disto é ilustrado na figura 3.7.

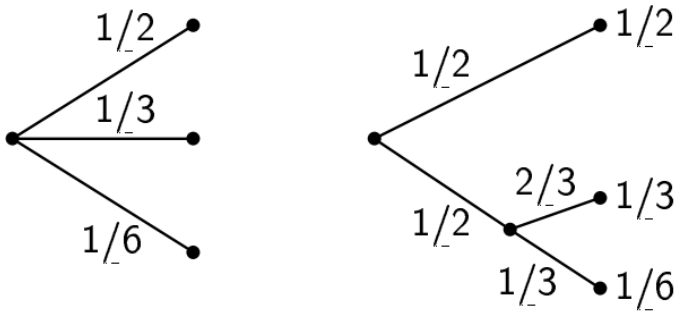

Figura 3.7: A figura da incerteza

Na esquerda temos três probabilidades $p_1 = 1/2, p_2 = 1/3, p_3 = 1/6$. Na direita nós primeiro escolhemos entre duas possibilidades, cada uma com valor $1/2$, e caso a segunda ocorra, fazemos uma outra escolha com probabilidades $2/3$ e $1/3$. Os resultados finais têm as mesmas probabilidades de antes. Exigimos, neste caso especial, que

$$H(\frac{1}{2}, \frac{1}{3}, \frac{1}{6}) = H(\frac{1}{2}, \frac{1}{2}) + \frac{1}{2}H(\frac{2}{3}, \frac{1}{3}) \tag{3.18}$$

O coeficiente é $1/2$ porque esta segunda opção só ocorre na metade das vezes.

O único H que satisfaz as três hipóteses acima é da forma:

$$H = -\sum_{i=1}^{n} p_i \log p_i \tag{3.19}$$

Quantidades da forma $H = -\sum p_i \log p_i$ desempenham um papel central na teoria da informação como medida de escolha, informação e incerteza. A forma de H vai ser reconhecida como a da entropia como é definida em certas formulações da mecânica estatística, onde p_i é a probabilidade de um sistema estar na célula i. H é, então, por exemplo, o H do famoso teorema de Boltzmann. Chamaremos $H = -\sum p_i \log p_i$ de entropia do conjunto de probabilidades $p_1, \cdots, p_n$. Se x é uma variável de chance, iremos escrever $H(x)$ como a sua entropia, assim x não é um argumento de uma função, mas um rótulo para um número, para diferenciá-lo a partir de $H(y)$ digamos, a entropia da variável de chance y.

A entropia, no caso de duas possibilidades com probabilidades p e $q = 1 - p$, ou seja,

$$H = -(p \log p + q \log q) \tag{3.20}$$

é plotado na figura como uma função de p. A quantidade H possui uma série de propriedades interessantes que posteriormente comprovam-na como uma medida razoável de escolha ou de informação.

1. $H = 0$, se e somente se todo p_i exceto um forem iguais a 0. Assim, somente quando estamos certos do resultado é que H desaparece. Caso contrário, H é positivo.

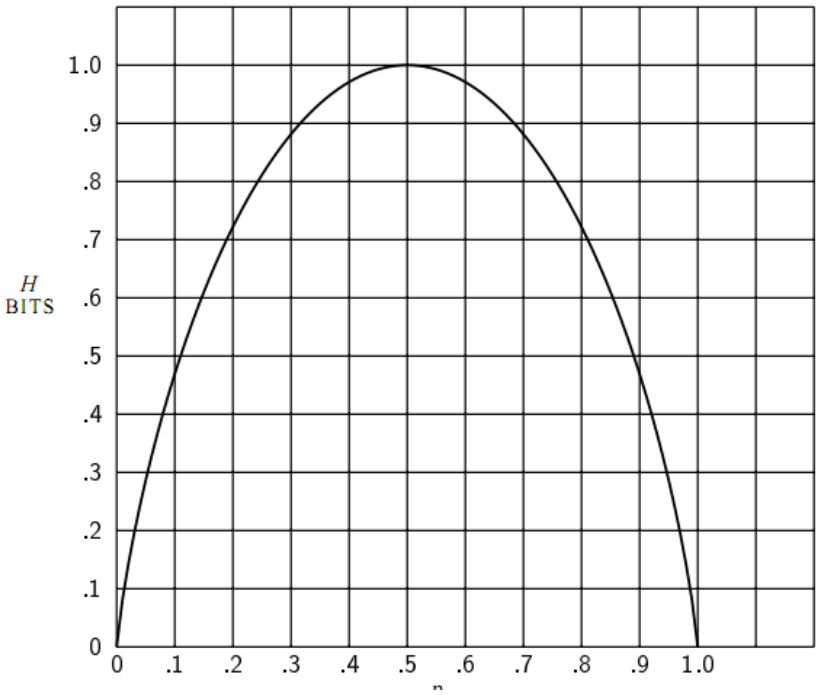

Figura 3.8: Entropia, no caso de duas possibilidades com probabilidade p e $(1-p)$

2. Para um dado n, H é máximo e igual a $\log n$ quando todo p_i for igual (por exemplo, $1/n$). Esta é intuitivamente a situação mais incerta.

3. Suponha que há dois eventos, x e y, com m possibilidades para o primeiro e n para o segundo. Seja $p(i,j)$ a probabilidade de ocorrência de i para o primeiro e j para o segundo ao mesmo tempo. A entropia do evento conjunto é:

$$H(x,y) = -\sum_{i,j} p(i,j) \log p(i,j) \tag{3.21}$$

enquanto

$$H(x) = -\sum_{i,j} p(i,j) \log \sum_{j} p(i,j) \tag{3.22}$$

$$H(y) = -\sum_{i,j} p(i,j) \log \sum_{i} p(i,j) \tag{3.23}$$

É facilmente mostrado que:

$$H(x,y) \leq H(x) + H(y) \tag{3.24}$$

com igualdade somente se os eventos forem independentes (por exemplo, $p(i,j) = p(i)p(j)$). A incerteza de um evento conjunto é menor ou igual à soma das incertezas individuais.

4. Qualquer mudança em direção a equalização das probabilidades $p_1, p_2, \cdots, p_n$ aumenta o valor de H. Assim, se $p_1 < p_2$ e nós aumentarmos p_1, diminuindo p_2 na mesma quantidade, p_1 e p_2 serão mais próximas, então H aumenta. Mas geralmente, se realiza qualquer operação de "média" no p_i da forma

$$p_i' = \sum_j a_{ij} p_j \tag{3.25}$$

onde $\sum_i a_{ij} = \sum_j a_{ij} = 1$, e todo $a_{ij} \geq 0$, então H aumenta (exceto no caso especial em que esta transformação representa não mais do que uma permutação dos p_i, com H permanecendo a mesma).

5. Suponha que existem dois eventos x e y como no item três, não necessariamente independentes. Para qualquer valor particular i que x pode assumir, existe uma probabilidade condicional $p_i(j)$ em que y possui o valor j. Isso é dado por:

$$p_i(j) = \frac{p(i,j)}{\sum_j p(i,j)} \tag{3.26}$$

A entropia condicional de y,$H_x(y)$ foi definida como a média de entropia de y para cada valor de x, pesando de acordo com a probabilidade de sair aquele x em particular. Isso é:

$$H_x(y) = -\sum_{i,j} p(i,j) \log p_i(j) \tag{3.27}$$

Essa quantidade mede o quão incertos estamos de y em média quando sabemos x. Substituindo o valor de $p_i(j)$ obtém-se

$$\begin{aligned} H_x(y) &= -\sum_{i,j} p(i,j) \log p(i,j) + \sum_{i,j} p(i,j) \log \sum_j p(i,j) \\ &= H(x,y) - H(x) \end{aligned} \tag{3.28}$$

ou

$$H(x,y) = H(x) + H_x(y) \tag{3.29}$$

A incerteza (ou entropia) do evento conjunto x, y é a incerteza de x mais a incerteza de y quando se sabe x.

6. Do item três e do item cinco sabemos

$$H(x) + H(y) \geq H(x,y) = H(x) + H_x(y) \tag{3.30}$$

Logo

$$H(y) \geq H_x(y) \tag{3.31}$$

A incerteza de y nunca aumenta pelo conhecimento de x. Ela irá diminuir, a não ser que x e y sejam eventos independentes, em tal caso ela não irá se modificar.

3.3.5 A Entropia de uma fonte de informação

Considere uma fonte discreta do estado de tipo finito considerado acima. Para cada estado possível i haverá um conjunto de probabilidades $p_i(j)$ de produzir os vários símbolos possíveis j. Assim, há uma entropia H_i para cada estado. A entropia da fonte será definida como a média ponderada destes H_i, de acordo com a probabilidade de ocorrência dos estados em questão:

$$\begin{aligned} H &= \sum_i P_i H_i \\ &= -\sum_{i,j} P_i p_i(j) \log p_i(j) \end{aligned} \tag{3.32}$$

Esta é a entropia da fonte por símbolo de texto. Se o processo de Markov está avançando a uma taxa de tempo definido, há também uma entropia por unidade de tempo:

$$H' = \sum_i f_i H_i \tag{3.33}$$

onde f_i é a frequência média (ocorrências por segundo) do estado i. Claramente,

$$H' = mH \tag{3.34}$$

onde m é o número médio de símbolos produzidos por segundo. H ou H' mede a quantidade de informação gerada pela fonte por símbolo ou por segundo.

Capítulo 4

Modelagem e Arquitetura

Esse capítulo apresenta o modelo para se medir a influência e descreve sua arquitetura.

4.1 Modelagem do Sistema

Para lidar com as desvantagens dos métodos mencionados no capítulo 2, este trabalho propõe dois modelos computacionais para medir a influência do indivíduo em redes sociais. Esses modelos são: o Modelo PageRank, que mede a influência em uma única rede através das relações dos membros, e o Modelo W-Entropia, que calcula a influência com as informações das diferentes redes ou diferentes características em algumas redes sociais.

Dentro de uma rede social específica, as pessoas podem ser consideradas como um vértice e as relações entre elas podem ser pensadas como arestas dirigidas. Um *link* para uma pessoa conta como um voto de apoio. O algoritmo PageRank pode ser empregado para calcular a importância dos membros dessa rede. Essa medida é objetiva, difícil de ser manipulada e justa.

Quando se possui informações de diferentes redes sociais ou de diferentes propriedades de uma única rede, a teoria da informação é introduzida para calcular a influência nestas situações. A teoria da informação pode medir o desequilíbrio quando a informação é transmitida a partir de diferentes plataformas. Este método é mais adequado a lei da transmissão da informação.

4.2 Arquitetura

Esses dois modelos cooperam um com o outro para medir a influência do indivíduo nas redes sociais. Primeiro, os dados dos relacionamentos são processado pelo Modelo PageRank, então o resultado e as informações independentes serão processadas pelo Modelo W-Entropia. Finalmente, a lista de classificação será produzida.

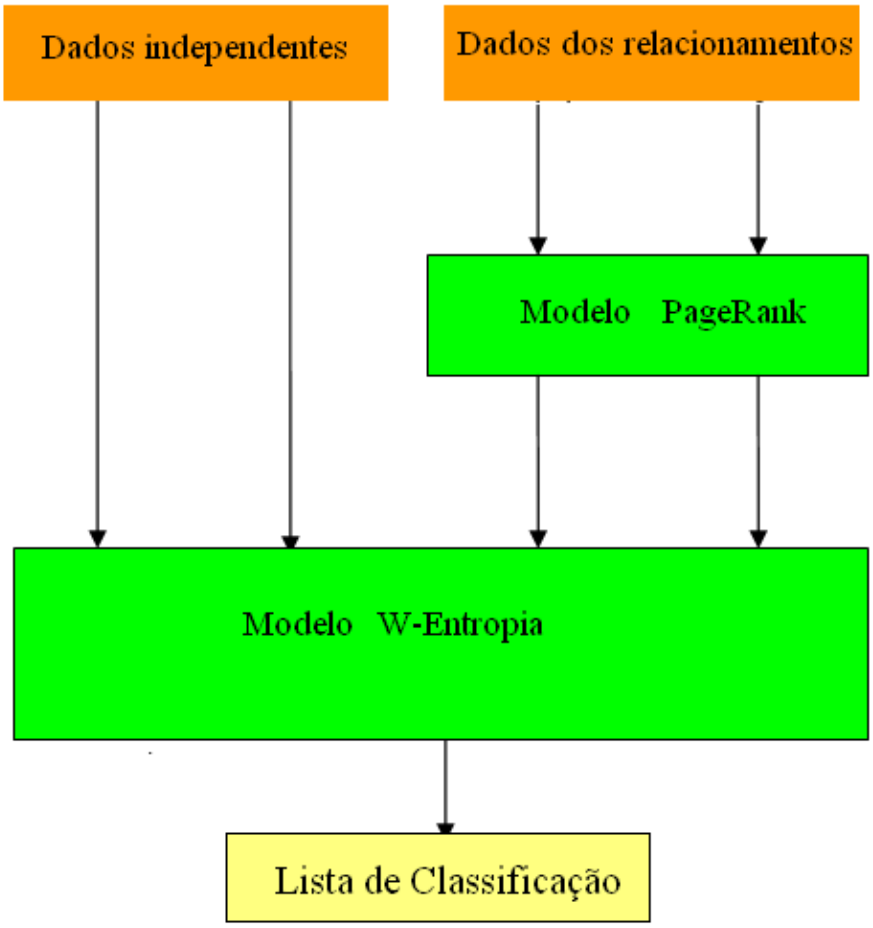

Figura 4.1: A arquitetura do modelo

Como mostra a figura 4.1, existem dois tipos diferentes de dados. Os dados independentes são as propriedades dos diferentes membros que não possuem relações entre si, por exemplo, o número de visitas, o tempo de registro e assim em diante. Os dados das relações são as propriedades em que os indivíduos possuem relações, por exemplo, relação de seguidores no *Twitter*, relação de fãs no *Facebook* e relação de recomendações no *ScienceNet*.

4.3 Modelo PageRank

4.3.1 Definição

Neste modelo, os membros podem ser considerados como vértices e as relações entre eles podem ser consideradas arestas dirigidas Leavitt et al. (2009). Um grafo dirigido pode ser construído para representar a estrutura da rede social. Com a concepção de PageRank, um *link* para uma pessoa conta como um voto de apoio. O PageRank de um indivíduo em uma rede social é definido recursivamente e depende do número e valor da métrica de todas as pessoas que se ligam a ele. Uma pessoa que está ligada a muitas

pessoas com um valor alto de ligações recebe uma alta classificação para si. Se não há links para uma pessoa, então não há suporte para ela.

Assim, podemos apresentar o algoritmo do PageRank:

$$PR(p_i) = \frac{1-d}{N} + d \sum_{p_j \in M(p_i)} \frac{PR(p_j)}{L(p_j)} \tag{4.1}$$

onde $p_1, p_2, \cdots, p_N$ são as pessoas levadas em consideração, $M(p_i)$ é o conjunto de pessoas que se ligam a p_i, $L(p_j)$ é o número de ligações que saem de p_j, e N é o número total de pessoas dentro da rede social escolhida. Os valores de PageRank são as entradas do autovetor dominante da matriz de adjacência modificada. Isso torna o PageRank uma métrica particularmente elegante, é o autovetor:

$$R = \begin{vmatrix} PR(p_1) \\ PR(p_2) \\ \vdots \\ PR(p_N) \end{vmatrix} \tag{4.2}$$

onde R é a solução da equação

$$R = \begin{vmatrix} \frac{1-d}{N} \\ \frac{1-d}{N} \\ \vdots \\ \frac{1-d}{N} \end{vmatrix} + d \begin{pmatrix} \ell(p_1,p_1) & \ell(p_1,p_2) & \cdots & \ell(p_1,p_N) \\ \ell(p_2,p_1) & \ddots & & \vdots \\ \vdots & & \ell(p_i,p_j) & \\ \ell(p_N,p_1) & \cdots & & \ell(p_N,p_N) \end{pmatrix} R \tag{4.3}$$

onde a função de adjacência $\ell(p_i, p_j)$ é 0 se a pessoa p_j não se liga a p_i, e é então normalizada, para cada j.

$$\sum_{i=1}^{N} \ell(p_i, p_j) = 1 \tag{4.4}$$

Os elementos de cada coluna somam até 1, então a matriz é uma matriz estocástica.

4.3.2 Cálculo

Para resumir, o PageRank pode ser calculado de forma iterativa ou algébrica. As operações básicas matemáticas realizadas no método iterativo e no método das potências são idênticos.

Método Iterativo

No primeiro caso, em $t = 0$, uma distribuição de probabilidade inicial é assumida, geralmente

$$PR(p_i; 0) = \frac{1}{N} \tag{4.5}$$

Em cada passo de tempo, o cálculo, conforme detalhado acima, rende

$$PR(p_i;t+1) = \frac{1-d}{N} + d \sum_{p_j \in M(p_i)} \frac{PR(p_j;t)}{L(p_j)} \tag{4.6}$$

ou em notação matricial

$$\mathbf{R}(t+1) = d\mathcal{M}\mathbf{R}(t) + \frac{1-d}{N}\mathbf{1} \tag{4.7}$$

onde $\mathbf{R}_i(t) = PR(p_i;t)$ e $\mathbf{1}$ é o vetor de colunas de comprimento N contendo apenas uns.

A matriz $\mathbf{M}$ é definida como

$$\mathcal{M}_{ij} = \begin{cases} 1/L(p_j), & \text{se } j \text{ se liga a } i \\ 0, & \text{senão} \end{cases} \tag{4.8}$$

i.e.,

$$\mathcal{M} := (K^{-1}A)^T \tag{4.9}$$

onde A denota a matriz de adjacência do grafo e K é a matriz diagonal com os graus na diagonal.

A computação termina quando para algum pequeno ϵ:

$$|\mathbf{R}(t+1) - \mathbf{R}(t)| < \epsilon \tag{4.10}$$

ou seja, quando a convergência é assumida.

Algébrica

Neste último caso, para $t \to \infty$, a equação

$$\mathbf{R}(t+1) = d\mathcal{M}\mathbf{R}(t) + \frac{1-d}{N}\mathbf{1} \tag{4.11}$$

pode ser escrita como

$$\mathbf{R} = d\mathcal{M}\mathbf{R} + \frac{1-d}{N}\mathbf{1} \tag{4.12}$$

A solução é dada por

$$\mathbf{R} = (\mathbf{I} - d\mathcal{M})^{-1}\frac{1-d}{N}\mathbf{1} \tag{4.13}$$

com a matriz identidade $\mathbf{I}$.

A solução existe e é única para $0 < d < 1$. Isto pode ser visto observando que $\mathcal{M}$ é por construção uma matriz estocástica e, portanto, tem um autovalor igual a um, por causa do teorema de Perron-Frobenius.

Método das potências

Se a matriz $\mathcal{M}$ é uma probabilidade de transição, ou seja, coluna estocástica com nenhuma coluna composta de apenas zeros e $\mathbf{R}$ é uma distribuição de probabilidade (ou

seja, $|\mathbf{R}| = 1, \mathbf{ER} = 1$ onde $\mathbf{E}$ é uma matriz de uns), a equação

$$\mathbf{R} = d\mathcal{M}\mathbf{R} + \frac{1-d}{N}\mathbf{1} \tag{4.14}$$

é equivalente a

$$\mathbf{R} = \left(d\mathcal{M} + \frac{1-d}{N}\mathbf{E}\right)\mathbf{R} =: \widehat{\mathcal{M}}\mathbf{R} \tag{4.15}$$

Assim, o PageRank $\mathbf{R}$ é o autovetor principal da $\widehat{\mathcal{M}}$. Uma maneira rápida e fácil de calcular isso é usando o método das potências: começando com um vetor arbitrário $x(0)$, o operador $\widehat{\mathcal{M}}$ é aplicado em seguida, ou seja,

$$x(t+1) = \widehat{\mathcal{M}}x(t) \tag{4.16}$$

até

$$|x(t+1) - x(t)| < \epsilon \tag{4.17}$$

Note que na primeira equação do método, a matriz sobre o lado direito do parêntese pode ser interpretada como

$$\frac{1-d}{N}\mathbf{I} = (1-d)\mathbf{P}\mathbf{1}^t \tag{4.18}$$

onde $\mathbf{P}$ é uma distribuição de probabilidade inicial. No caso atual

$$\mathbf{P} := \frac{1}{N}\mathbf{1} \tag{4.19}$$

Finalmente, se $\mathcal{M}$ tem colunas com apenas valores iguais a zero, elas devem ser substituídas com o vetor de probabilidade inicial $\mathbf{P}$. Em outras palavras

$$\mathcal{M}' := \mathcal{M} + \mathcal{D} \tag{4.20}$$

onde a matriz $\mathcal{D}$ é definida como

$$\mathcal{D} := \mathbf{P}\mathbf{D}^t \tag{4.21}$$

com

$$\mathbf{D}_i = \begin{cases} 1, & \text{se } L(p_i) = 0 \\ 0, & \text{senão} \end{cases} \tag{4.22}$$

Neste caso, os dois cálculos acima usam $\mathcal{M}$ e só resultam no mesmo PageRank se os seus resultados forem normalizados.

$$\mathbf{R}_{\text{power}} = \frac{\mathbf{R}_{\text{iterative}}}{|\mathbf{R}_{\text{iterative}}|} = \frac{\mathbf{R}_{\text{algebraic}}}{|\mathbf{R}_{\text{algebraic}}|}. \tag{4.23}$$

4.3.3 O Processo do PageRank

Esta seção irá descrever o algoritmo através de um exemplo simples. Imagine que há sete pessoas em uma rede social e as relações entre elas são da seguinte forma:

Tabela 4.1: As relações entre as pessoas

ID da Pessoa	Segue
1	2,3,4,5,7
2	1
3	1,2
4	2,3,5
5	1,3,4,6
6	1,5
7	5

Com essas relações, podemos obter um grafo direcionado como este a seguir, onde as setas representam as relações:

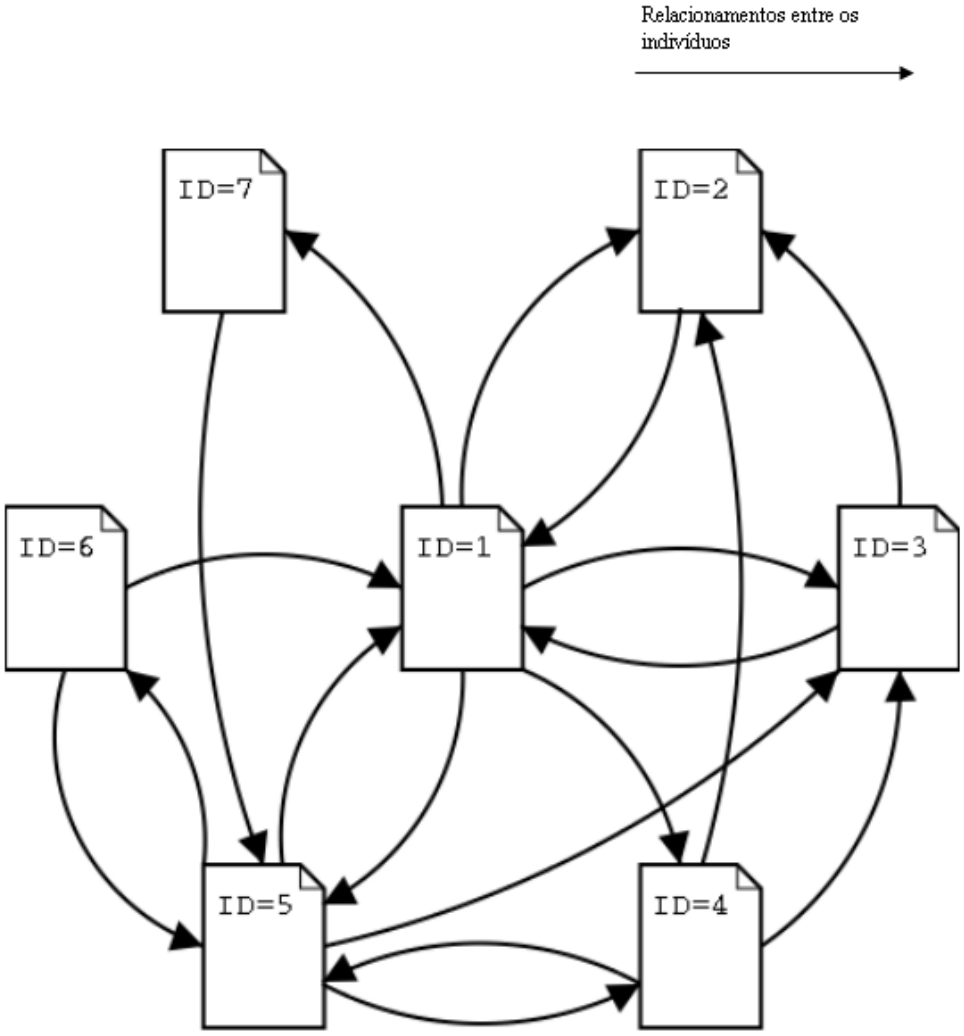

Figura 4.2: O grafo direcionado para a estrutura da rede social (Page et al. 1998)

A partir do grafo, podemos obter a matriz de adjacência:

$$A = \begin{vmatrix} 0 & 1/5 & 1/5 & 1/5 & 1/5 & 0 & 1/5 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1/3 & 1/3 & 0 & 1/3 & 0 & 0 \\ 1/4 & 0 & 1/4 & 1/4 & 0 & 1/4 & 0 \\ 1/2 & 0 & 0 & 0 & 1/2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{vmatrix} \tag{4.24}$$

A partir da matriz A, transpor a matriz M

$$M = A' = \begin{vmatrix} 0 & 1 & 1/2 & 0 & 1/4 & 1/2 & 0 \\ 1/5 & 0 & 1/2 & 1/3 & 0 & 0 & 0 \\ 1/5 & 0 & 0 & 1/3 & 1/4 & 0 & 0 \\ 1/5 & 0 & 0 & 0 & 1/4 & 0 & 0 \\ 1/5 & 0 & 0 & 1/3 & 0 & 1/2 & 1 \\ 0 & 0 & 0 & 0 & 1/4 & 0 & 0 \\ 1/5 & 0 & 0 & 0 & 0 & 0 & 0 \end{vmatrix} \tag{4.25}$$

A autovalor-dominante da matriz M é:

$$\text{não}NormalizadoR = \begin{vmatrix} 0,69946 \\ 0,38286 \\ 0,32396 \\ 0,24297 \\ 0,41231 \\ 0,10308 \\ 0,13989 \end{vmatrix} \tag{4.26}$$

depois de normalizado, o PageRank de cada pessoa é:

$$NormalizadoR = \begin{vmatrix} 0,303514 \\ 0,166134 \\ 0,140575 \\ 0,105431 \\ 0,178914 \\ 0,044728 \\ 0,060703 \end{vmatrix} \tag{4.27}$$

Este é o resultado final da influência das pessoas.

4.4 Modelo W-Entropia

É o modelo para medir a influência dos indivíduos com informações de diferentes redes sociais. O conceito de W-Entropia foi proposto pelo Weigang et al. (2011b) e baseou-se na teoria da informação, que permite que o desequilíbrio de informações entre diferentes plataformas seja medido. Este valor segue a lei da transmissão da informação.

4.4.1 Definição do Índice W-Entropia

Seja P_j a pessoa mais popular na rede social, j uma característica (número de seguidores, número de fãs, número de visitas), X_j o valor dessa característica do indivíduo X e p_j é a média da rede social j, $p_j = X_j/P_j$; Outras redes sociais podem ser representadas como $p_{j+1} = X_{j+1}/P_{j+1}$ e assim em diante. Os pesos das diferentes redes sociais são $a_1, a_2, \cdots, a_n, \sum a_j = 1$. Portanto, a média do indivíduo X é:

$$m = \sum_{j=1}^{n} a_j p_j, j = 1, \cdots, n \tag{4.28}$$

Este é um método muito simples e intuitivo, a transmissão das informações entre as diversas redes sociais não é equilibrada, por isso o valor m é usado como a média geral de cada indivíduo. Se alguém possui uma influência muito diferente entre as redes sociais, esta fórmula se tornará muito menos eficaz e, portanto, produzirá resultados que não refletem as condições reais.

A entropia da informação pode ser empregada para quantificar a distribuição desequilibrada da transmissão de informações entre as diferentes redes sociais. A entropia é definida como um coeficiente de correção para a transmissão entre as redes sociais.

Antes de calcular a entropia, é necessário ajustar o conjunto de valores $p_1, p_2, \ldots, p_n$; a soma de todos os termos deste conjunto deve que ser igual a 1.

$$\begin{cases} q_j = p_j/(n+1), & j = 1, 2, \cdots, n \\ q_{n+1} = 1 - \sum p_j, & j = 1, 2, \cdots, n \end{cases} \tag{4.29}$$

$q_1, q_2, \cdots, q_n$ representam os valores numéricos da transmissão de informações entre diferentes redes sociais. Por outro lado, q_{n+1} é um percentual que representa a ausência das informações que estão sendo transferidas entre as diferentes redes.

$$h(q_1, q_2, \cdots, q_n, q_{n+1}) = -\sum q_j \log_{n+1} q_j \qquad j = 1, 2, \ldots, n+1 \tag{4.30}$$

A variável h apresentada na fórmula pode assumir qualquer valor entre 0 e 1. Quando a informação do indivíduo está sendo transmitida de forma uniforme entre as redes sociais, $h = 1$. Quando a informação do indivíduo está sendo transmitida de forma desigual, onde a maioria dos $p_1, p_2, \cdots, p_n$ são iguais a 0, então $h = 0$.

Com base nas fórmulas acima, o W-Entropia, que é o impacto de cada usuário nas redes sociais pode ser definido como:

$$W\text{-}Entropia = h * m \tag{4.31}$$

A fim de simplificar esta fórmula para efeitos de aplicação, o valor da fórmula foi dimensionado em relação à $W\text{-}Entropia_{max}$, que é o valor do índice máximo, e depois multiplicado por 100, isso resulta na seguinte equação:

$$\text{Índice} \quad W\text{-}Entropia \quad = 100 * W\text{-}Entropia/W\text{-}Entropia_{max} \tag{4.32}$$

4.4.2 Análise das propriedades do W-Entropia

O coeficiente h para o desequilíbrio durante a transmissão da informação é definido como a entropia da informação. Este coeficiente deve conter os seguintes atributos: quando todos os elementos forem iguais a 1, significa que as informações deste indivíduo estão sendo transmitidas de forma uniforme entre as redes sociais, de modo que o coeficiente de modificação é definido como 1. Por outro lado, quando todos os termos são iguais a 0, isto significa que a transmissão é desigual, pois o coeficiente de modificação é definido como igual a 0. O valor dos elementos variam de 0 a 1, portanto, o valor do coeficiente modificado também varia entre 0 e 1.

Para verificar a validade e eficácia do coeficiente modificado h, os seguintes parâmetros foram utilizados: $n = 3$ e seis conjuntos de dados foram calculados:

Tabela 4.2: Conjuntos de dados com $n = 3$.

Set 1	p_1	[0,0,1,0,2,,,,,1]	p_2	[0,0,0,,,,,0]	p_3	[0,0,0,,,,,0]
Set 2	p_1	[1,1,1,,,,,1]	p_2	[0,0,1,0,2,,,,,1]	p_3	[0,0,0,,,,,0]
Set 3	p_1	[1,1,1,,,,,1]	p_2	[1,1,1,,,,,1]	p_3	[0,0,1,0,2,,,,,1]
Set 4	p_1	[1,0,9,0,8,,,,,0]	p_2	[1,1,1,,,,,1]	p_3	[1,1,1,,,,,1]
Set 5	p_1	[0,0,0,,,,,0]	p_2	[1,0,9,0,8,,,,,0]	p_3	[1,1,1,,,,,1]
Set 6	p_1	[0,0,0,,,,,0]	p_2	[0,0,0,,,,,0]	p_3	[1,0,9,0,8,,,,,0]

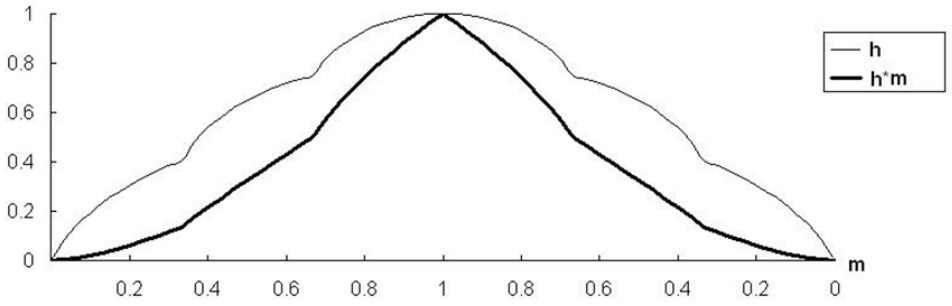

Figura 4.3: O gráfico da Entropia h e $h * m$ com m (Weigang et al. 2011a)

Com base nos dados, pode-se observar que nos três primeiros conjuntos de dados Set1, Set2 e Set3, a tendência do desequilíbrio durante a transmissão passou de 0 a 1, de modo que o coeficiente de modificação h também assumiu um valor que varia de 0 a 1. Os últimos três conjuntos de dados foram usados para ilustrar um cenário oposto, onde a tendência foi de 1 a 0 e o coeficiente modificado também variou de 1 a 0. Assim, $h * m$ e o índice W-Entropia terão a mesma tendência monotônica. A Tabela 4.2 apresenta todos os valores dos elementos obtidos e a Figura 4.3 é um gráfico de h em função de m. Estes resultados ilustram a propriedade de generalização do coeficiente modificado.

Tabela 4.3: Os valores dos parâmetros de seis conjuntos para todos os termos

	p1	q1	p2	q2	p3	q3	q4	m	h	h*m	
1	0	0	0	0	0	0	1	0	0	0	0
2	0,5	0,125	0	0	0	0	0,875	0,1667	0,2718	0,0453	4,53
3	1	0,25	0	0	0	0	0,75	0,3333	0,4056	0,1352	13,52
4	1	0,25	0,5	0,125	0	0	0,625	0,5	0,6494	0,3247	32,47
5	1	0,25	1	0,25	0	0	0,5	0,6667	0,75	0,5	50,00
6	1	0,25	1	0,25	0,5	0,125	0,375	0,8333	0,9528	0,7940	79,40
7	1	0,25	1	0,25	1	0,25	0,25	1	1	1	100
8	0,5	0,125	1	0,25	1	0,25	0,375	0,8333	0,9528	0,7940	79,40
9	0	0	1	0,25	1	0,25	0,5	0,6667	0,75	0,5	50,00
10	0	0	0,5	0,125	1	0,25	0,625	0,5	0,6494	0,3247	32,47
11	0	0	0	0	1	0,25	0,75	0,3333	0,4056	0,1352	13,52
12	0	0	0	0	0,5	0,125	0,875	0,1667	0,2718	0,0453	4,53
13	0	0	0	0	0	0	1	0	0	0	0

É possível observar que os valores de m na terceira e na décima primeiro da tabela são ambos 0,3333. Embora os valores de p_1, p_2 e p_3 sejam diferentes para essas linhas, os valores de $h = 0,4056$ e $h * m = 0,1352$ são os mesmos devido à simetria em m. Se os valores de p_1, p_2 e p_3 forem $0,3333, 0,3333, 0,3333$, então o valor de m também é 0,3333, neste caso, $h =$ 0,6037 e $h * m = 0,2012$, que são maiores do que os valores para a terceira e décima primeira linha. Este resultado suporta a validade do método de coeficiente modificado.

Capítulo 5

Sistema W-Entropia

Este capítulo descreve a implementação dos módulos mencionados no capítulo anterior. A fim de realizar essa função, o sistema W-Entropia foi desenvolvido.

5.1 Ambiente de Desenvolvimento

Antes de apresentar o protótipo, será feita uma breve descrição do ambiente de desenvolvimento utilizado, a fim de prover um melhor entendimento do protótipo implementado.

5.1.1 Eclipse Galileo+Biblioteca Jsoup

O Eclipse SDK (*Software Development Kit*) é formada pela Plataforma Eclipse, *Java Developments Tools* e o *Plugin Development Environment*. A Plataforma Eclipse é um ambiente de desenvolvimento multi-linguagem que compreende um ambiente de desenvolvimento integrado(IDE) e um sistema de plugin extensível. É escrito em sua maioria em Java.

O *Jsoup* é uma biblioteca Java para trabalhar com o HTML do mundo real. Ele fornece uma *API*(*Application Programming Interface*) muito conveniente para a extração e manipulação de dados, usando o melhor do DOM(*Document Object Model*), CSS(*Cascading Stylesheet*) e métodos como o *Jquery*.

5.1.2 MySQL+PhpMyAdmin

O MySQL é um sistema de gerenciamento de banco de dados que é executado como um servidor e fornece acesso a multiusuários a uma série de bancos de dados. O projeto de desenvolvimento do MySQL tem feito o seu código fonte disponível sob os termos da Licença Pública Geral GNU, bem como sob uma variedade de licenças. Isso tornou o MySQL uma escolha popular de banco de dados para uso em aplicações *web*, e é um componente central do amplamente usado software LAMP, LAMP é um acrônimo para "Linux, Apache, MySQL, Perl / PHP / Python".

MySQL é escrito em C e C++, ele funciona em várias plataformas de sistemas diferentes, por exemplo, Windows, Linux, Mac OS e outros sistemas operacionais. Muitas

linguagens de programação com *APIs* específicas incluem bibliotecas para acessar bancos de dados usando o MySQL.

O PhpMyAdmin é uma ferramenta de código aberto escrita em PHP destinada a lidar com a administração do MySQL com o uso de um navegador *web*. Ele pode executar várias tarefas, como criar, modificar ou excluir bancos de dados, tabelas, campos ou linhas; executar instruções SQL, ou gerenciar usuários e permissões.

5.1.3 PHP

PHP é uma linguagem de propósito geral originalmente projetada para o desenvolvimento *web* para produzir páginas dinâmicas. Está entre uma das primeiras línguas desenvolvidas com *scripting* ao lado do servidor, que é incorporada em um documento HTML, ao invés de chamar um arquivo externo para processar dados. Em última análise, o código é interpretado por um servidor *web* com um módulo do processador PHP que gera a página *web* resultante. O PHP pode ser implementado na maioria dos servidores *web* e também utilizado como um *shell* autônomo em quase todos os sistemas operacionais e plataformas de forma gratuita.

5.2 Modelagem do Sistema

O sistema W-Entropia foi desenvolvido para medir a influência do indivíduo em redes sociais. O sistema pode obter as informações de propriedade especial de redes sociais pela parte do *crawler* e calcular a influência de indivíduos pela parte do cálculo. Finalmente, a parte de exibição é usada para mostrar as informações do *ranking*.

Da Figura 5.1, vemos que o sistema possui três partes: O módulo do*crawler*, responsável por coletar as informações da rede social. Este módulo deve mudar a sua configuração para obter as diferentes propriedades das diversas redes sociais. Além da obtenção dos dados, este módulo também se encarregará de processar os dados originais para o formato desejado, a fim de executar a etapa seguinte.

O módulo de cálculo tem duas partes: os dados que possuem a propriedade hiperlink são processados pelo módulo PageRank em primeiro lugar. Esta parte calcula a importância de cada indivíduo pela estrutura do relacionamento. Após esta etapa, os dados de todas as redes sociais ou todas as propriedades são enviadas para o módulo W-Entropia. O módulo W-Entropia é a resposta para o cálculo do resultado final da influência.

O módulo de exibição é um website (http://www.wentropia.com/) que foi desenvolvido para exibir o *ranking* final na internet.

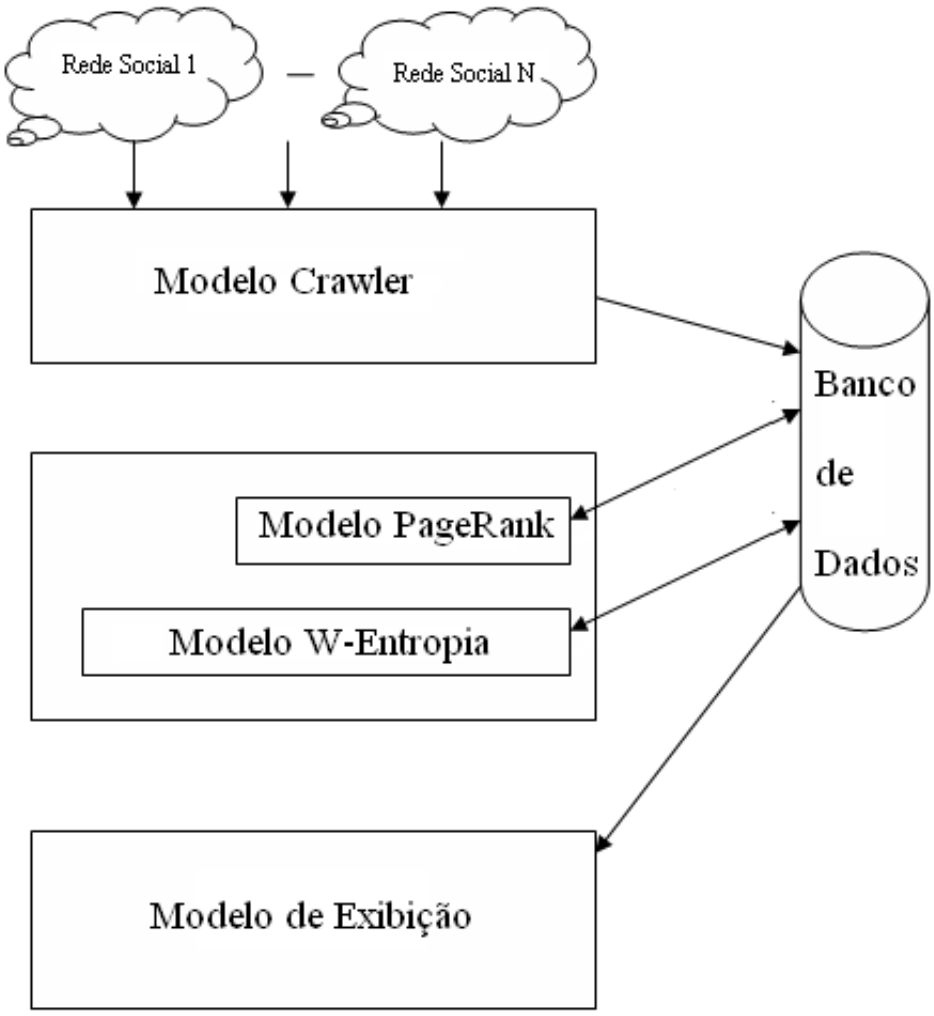

Figura 5.1: Arquitetura do sistema

5.2.1 Crawler

Um *crawler* da *web* (Kobayashi 2000)(Brin 1998) é um programa de computador que navega na *World Wide Web* de uma forma metódica e automatizada ou de forma ordenada. Este processo é chamado de *Web crawling* ou *spidering*. Muitos sites, em particular de ferramentas de busca, usam o *spidering* como um meio de fornecer dados atualizados. As informações das redes sociais são enormes, por isso é necessário desenvolver um *crawler* para visitar as páginas automaticamente.

O processo do *crawler*

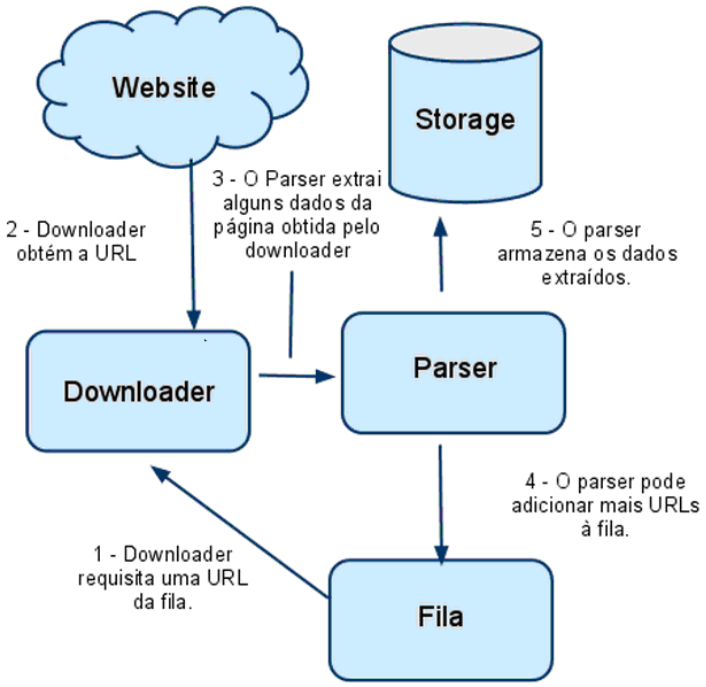

Figura 5.2: O processo de trabalho do *crawler*

1. O *downloader* é aplicado para as urls das unidade de urls. A unidade mantém uma lista de urls que esperam visitas pelo *crawler*. Inicialmente, esta lista é preenchida manualmente, mas quando o *crawler* começar a trabalhar, o módulo de *parser* irá adicionar urls a esta lista automaticamente.

2. O *downloader* rastreia as urls uma por uma. Este passo envolve a " Política de polidez ". Os *crawlers* podem recuperar dados de forma muita mais rápida e em maior profundidade do que os pesquisadores humanos, então eles podem ter um impacto paralisante sobre o desempenho de um site. O rastreador deve respeitar o protocolo de exclusão de robôs, também conhecido como o protocolo *robots.txt*, que é um padrão que os administradores utilizam para indicar quais as partes do seus servidores *web* não devem ser acessadas por *crawlers*.

3. O Módulo *Parser* extrai os dados úteis da página baixada. Embora o *downloader* baixe uma página por completo, é aproveitada apenas uma pequena parte dela. Portanto, o *parser* extrai os dados de acordo com a regra certa. Neste sistema, os dados necessários são: o número de fãs, o número de seguidores, o número de visitas e outras medidas relacionadas.

4. O *Parser* pode adicionar mais urls para a lista de urls. Durante o rastreamento, o analisador pode encontrar outras urls na página processada, portanto, essas urls são adicionadas à lista de urls. Há um problema de eficiência neste passo - evitar a adição da mesma url mais do que uma vez. Neste sistema, a url da rede social está

sempre relacionada com a ID do usuário, portanto, uma lista de IDs de usuários foi criada com o papel principal de evitar a duplicações de urls.

5. A informação dos indivíduos é armazenada no banco de dados. Cada rede social possui uma tabela específica para armazenar seus dados.

Estratégias do *crawler*

De acordo com a demanda real, duas estratégias de *crawling* (Cho et al. 1998) foram aplicadas: Estratégia toplist e Estratégia clássica.

1. Estratégia toplist: Essa estratégia faz uma lista de rastreamento a partir da lista dos mais influentes da rede social. Foca no indivíduo mais popular, este método é eficiente e fácil de implementar. A desvantagem é que não pode recolher toda a informação da rede.

2. Estratégia clássica: Esta estratégia é a estratégia mais comum dos *crawlers* das ferramentas de busca. Primeiro elabora-se uma lista de urls manualmente, então o *crawler* começa a obter informações e obter novas urls enquanto rastreia a internet. Na rede social, esta estratégia pode ser modificada para construir uma lista de IDs para cada indivíduo. Quando o *crawler* obtém a informação das ID existentes, ele pode descobrir novas IDs, como, por exemplo, dos seguidores desses indivíduos. Esta estratégia pode rastrear quase toda a rede social. Mas a maior desvantagem da estratégia clássica é ser ineficiente, porque toda nova ID deve ser comparada com as IDs existentes, para verificar se esta ID já existe ou não, a complexidade deste processo é $O(N^2)$.

5.2.2 Módulos para cálculo

Essa parte possui dois módulos: Módulo PageRank e Módulo de cálculo W-Entropia.

Módulo PageRank

Esse módulo é responsável pelo cálculo da influência do indivíduo em uma única rede social. Ele constrói uma matriz quadrada com a relação entre diferentes membros. Para ser justo, cada membro tem a mesma influência inicial. Em seguida, é transformado o vetor valor PR com a matriz quadrada até obter o resultado final.

1. Na primeira verificação que faz, o formato dos dados estão em conformidade com a exigência de módulo de cálculo PR. Os dados devem ser construídos para uma matriz quadrada e a soma dos elementos de cada linha tem que se igualar exatamente. Isso garante que cada indivíduo tenha a mesma influência inicial.

2. Inicializar o vetor PR de acordo com o número de elementos. Com a ideia do algoritmo PageRank original, aqui a soma do valor PR de todos os elementos também deve ser igual a 1. No final deste passo, tem-se um vetor coluna com N linhas e todos os valores dos elementos é $1/N$, sendo N o número dos elementos.

3. O vetor PR é transformado com a matriz obtida a partir das relações dos indivíduos na rede social. Após essa transformação, o valor PR mudará dependendo das diferentes probabilidades da matriz de relacionamento.

4. Verifica-se se a matriz possui um sorvedouro. Normaliza-se o vetor PR novo e compara-se com o valor inicial 1.

5. Se o sorvedouro existe, esta etapa deve adicionar o valor PR perdido para o vetor. O sorvedouro irá reduzir o valor PR em cada iteração. Para garantir a justiça do resultado deste algoritmo, o algoritmo deve adicionar o valor perdido por igual a cada indivíduo.

6. Calcula-se o parâmetro de controle. Compara-se o novo valor PR com o último valor PR para verificar se o resultado convergiu ou não. Se o resultado for menor do que o limite, o algoritmo para o novo valor PR é o resultado final, ou então o algoritmo continua até que o resultado tenha convergido.

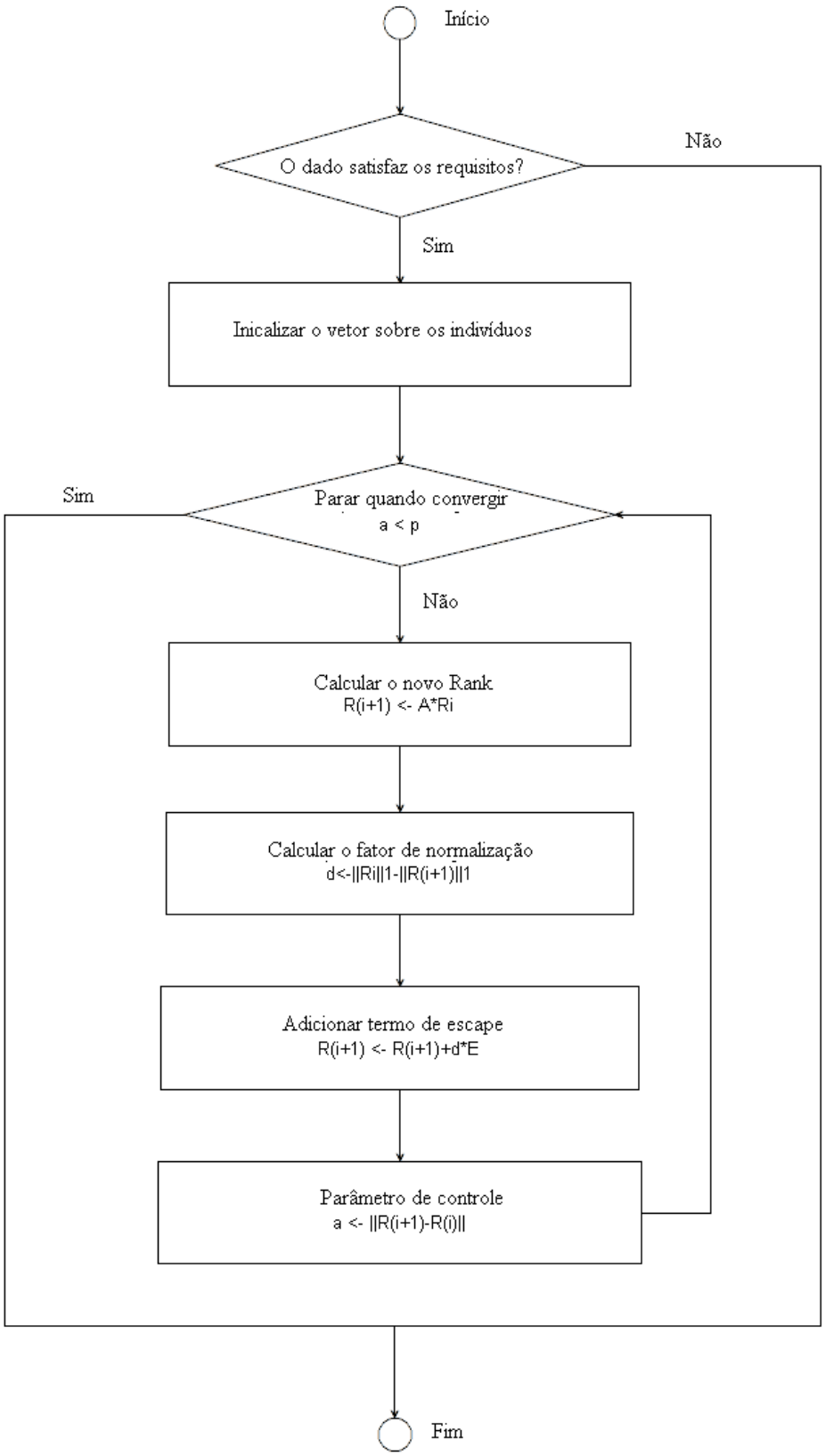

Figura 5.3: O fluxograma do módulo computacional PageRank

O módulo de cálculo W-Entropia

O módulo W-Entropia é responsável por calcular a influência abrangente entre várias redes sociais.

1. Obter o valor máximo de cada item do banco de dados. Definir o valor máximo como a base para calcular a porcentagem do indivíduo em cada item. Há um princípio flexível neste passo, se o valor máximo de um item for distribuído de forma muito desigual, caso, por exemplo, o primeiro lugar seja maior do que o segundo e outros, então este valor máximo pode ser substituído manualmente para um valor mais próximo do segundo, para tornar o resultado mais razoável.

2. Começar a iteração para calcular a informação de um indivíduo em todos os itens. Estes valores são temporários, eles não são adequados para o algoritmo e por isso serão ajustados nas seguintes etapas.

3. Os valores percentuais podem refletir o fundamento de um indivíduo. De acordo com a importância de diferentes redes sociais, é atribuído um peso diferente para todas. Este passo calcula a média de todos os itens como um fator da influência.

4. Como mencionado anteriormente, os valores percentuais não são adequados para o algoritmo. Esta etapa é responsável por ajustar estes valores para o formato dos requisitos do algoritmo. Para um valor de porcentagem existente, um novo valor é obtido usando o valor antigo dividido pela soma do número de itens com 1. O um nesta equação representa a ausência de informações que está sendo transferida entre as diferentes redes e o seu valor de porcentagem é obtido por 1 menos o valor obtido pela primeira equação.

5. Com todos os dados obtidos no passo 4, a entropia de cada indivíduo pode ser calculada pela fórmula da entropia. Durante este cálculo, pode ser encontrado o valor de percentagem igual a 0. O valor do logarítmico não pode ser 0, então nesta situação (em que valor é 0), o valor atribuído a esta parte é 0.

6. Essa etapa é responsável por calcular o W-Entropia com a média multiplicando a entropia m. Esse W-Entropia é expresso como um decimal.

7. Para tornar o resultado mais adequado ao hábito humano, o resultado final é o Índice W-Entropia de valor relativo que é obtido por cada W-Entropia dividido pelo valor máximo de W-Entropia e, finalmente, multiplicado por 100.

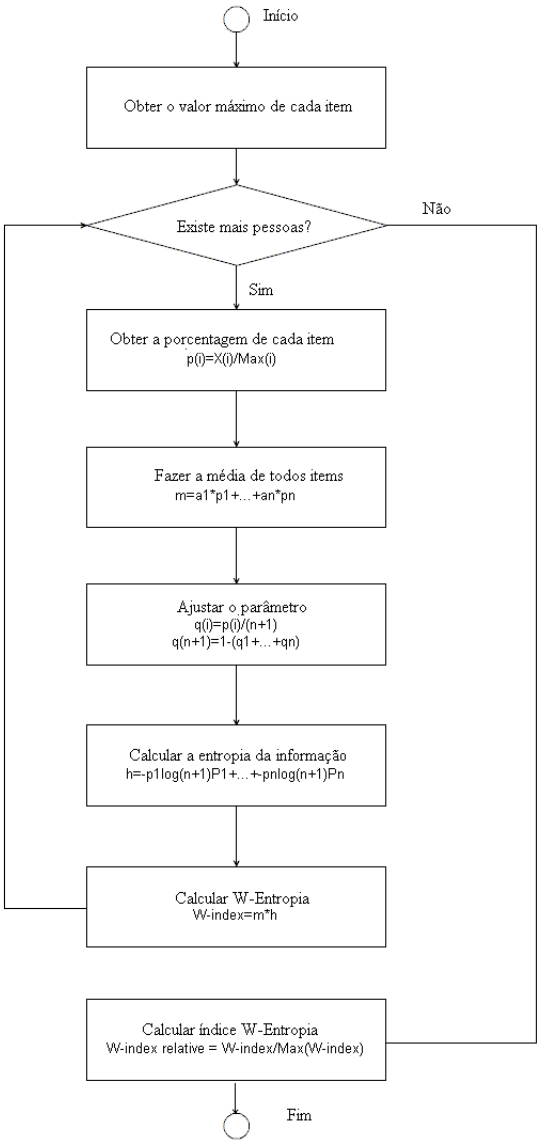

Figura 5.4: O fluxograma do módulo computacional W-Entropia

5.2.3 Módulo de exibição

Esse módulo é responsável por exibir o resultado do sistema. Para publicar o *ranking* na internet, um *website* foi construído. Esse site foi desenvolvido com a linguagem de programação PHP e baseado na arquitetura B/S(*Browser/Server*) .

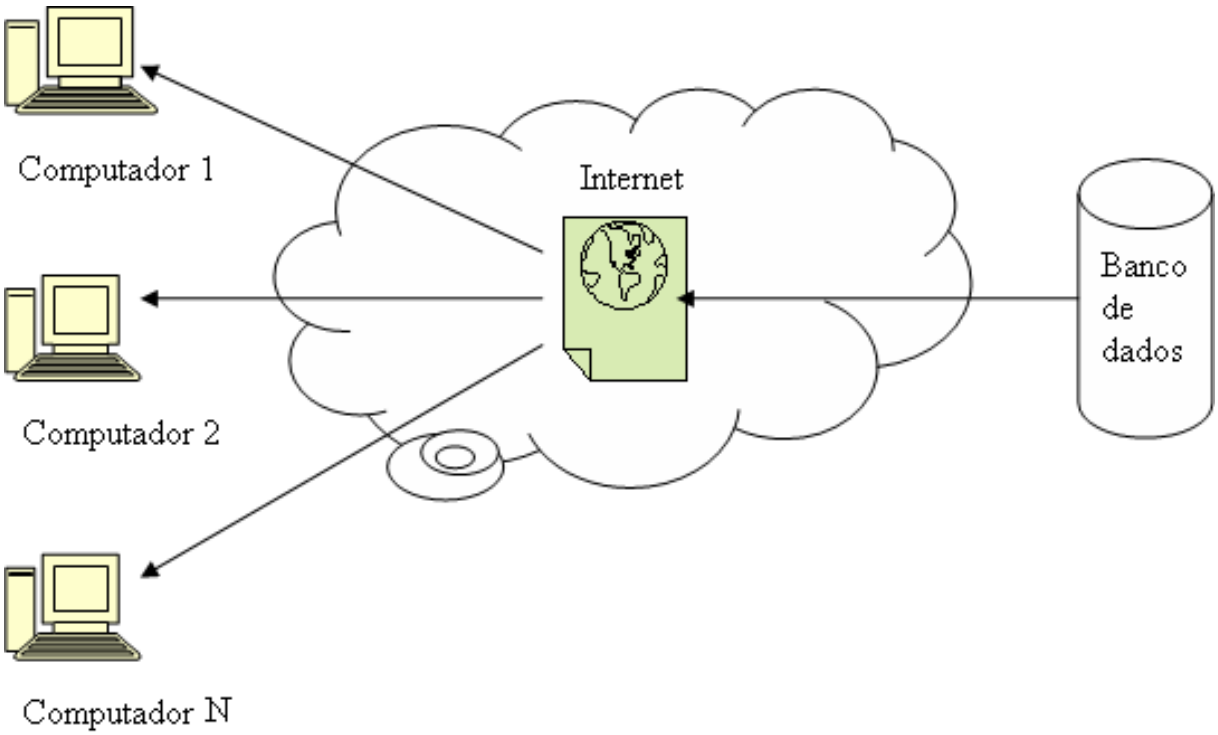

Figura 5.5: A arquitetura do módulo de exibição

Ao lado do servidor, todos os dados são armazenadas no banco de dados MySQL, a estrutura do banco de dados é apresentada na tabela a seguir. A linguagem PHP é aplicada para processar as operações lógicas, como por exemplo, a conexão com o banco de dados, lidar com a solicitações do cliente e outros.

Tabela 5.1: A Estrutura da Tabela no Banco de Dados

items	type
id	int
name	varchar(50)
image	varchar(200)
plat_1	varchar(50)
dado_plat_1	bigint
plat_2	varchar(50)
dado_plat_2	bigint
...	...
plat_n	varchar(50)
dado_plat_n	bigint
entropy	float
W-index	float

Capítulo 6

Estudo de Caso

O objetivo deste capítulo é aplicar o Sistema W-Entropia com a realização de quatro estudos. Com os dados de diferentes redes sociais, foi calculada e analisada a influência dos membros. Neste capítulo, foram conduzidos quatro estudos correspondendo a situações diferentes e os resultados foram analisados, comparando-os com as listas de classificação existentes.

6.1 Plano de Estudos

Esta seção é uma apresentação dos quatro estudos:

O primeiro calcula a influência dos membros do Twitter Weng et al. (2010). Para demonstrar os detalhes do algoritmo PageRank, levou-se em consideração os jogadores de futebol do Flamengo Futebol Clube. As informações foram coletadas pelo Sistema W-Entropia e um grafo direcionado foi construído pela relação de "seguir" entre os jogadores. Com a aplicação do algoritmo PageRank, as informações detalhadas de cada iteração serão mostradas neste estudo. Com a análise do resultado que foi encontrado é possível observar a vantagem deste método.

O segundo é o cálculo da influência dos membros da rede social *ScienceNet* com o algoritmo PageRank. Neste estudo, foram considerados 5.265 usuários desta rede social. O relacionamento entre os membros do *ScienceNet* é diferente do *Twitter*. Enquanto no *Twitter* a relação entre os usuários é de "N-para-um", já no ScienceNet, os artigos dos blogs foram considerados como unidades, e a relação existente entre os diferentes membros é "recomendar" ou não, o artigo do blog, e quantos artigos do blog foram recomendados. Este estudo não é apenas uma avaliação do Sistema W-Entropia, ele também introduz o método para analisar outros tipos de relacionamentos das redes sociais.

O terceiro também realiza um estudo da plataforma ScienceNet. Existe um pouco de discrepância entre o resultado obtido com o algoritmo PageRank no segundo estudo e as duas listas de classificação publicadas pelo ScienceNet. Uma dessas listas é uma classificação pelo número de visualizações totais dos artigos, a outra é uma lista de classificação pelo número médio de visitas por artigo. Com estes três dados diferentes, o algoritmo W-Entropia foi adotado para calcular a influência dos membros de uma forma mais abrangente. A partir do resultado deste estudo, a vantagem de se utilizar o Sistema W-Entropia para se lidar com vários dados diferentes pode ser notada.

O último estudo é um estudo de multiplataformas, que inclui o *Facebook*, o *Twitter* e o *Google Search Result*. Com as informações obtidas a partir de diferentes plataformas, foi calculada a influência dos indivíduos nas principais redes sociais e o resultado foi analisado, comparando com as listas já existentes.

6.2 O Cálculo do PageRank dos Jogadores do Flamengo no *Twitter*

6.2.1 Introdução

O Clube de Regatas do Flamengo (Flamengo 2012) é uma agremiação poliesportiva brasileira com sede na cidade do Rio de Janeiro fundada para disputas de remo em 17 de novembro de 1895. Criado no bairro de mesmo nome,mudou-se para o bairro da Gávea na primeira metade do século XX.

Apesar de ter sido fundado como um "Clube de Regatas", o esporte mais tradicional no Flamengo é o futebol. Único clube carioca campeão da Copa Intercontinental (Mundial Interclubes), e juntamente com o Corinthians a quarta equipe que mais vezes conquistou o Campeonato Brasileiro com 5 títulos cada um. É a equipe mais vitoriosa do Campeonato Carioca com 32 títulos.

6.2.2 Relação do *Twitter*

O *Twitter* Kwak et al. (2010) é um serviço de rede social e um serviço de microblogging que permite aos seus usuários enviar e ler mensagens de texto de até 140 caracteres, conhecidos como "tweets". Foi criado em março de 2006 por Jack Dorsey e lançado em julho do mesmo ano. O serviço rapidamente ganhou popularidade em todo o mundo, com mais de 300 milhões de usuários a partir de 2011, gerando mais de 300 milhões de tweets e lidando com mais de 1,6 bilhões de buscas por dia. Ele foi descrito como "os SMS da Internet". O Twitter permite aos usuários a capacidade de atualizar seu perfil usando seu telefone móvel, através de mensagens de texto ou aplicativos lançados para certos *smartphones/tablets*.

No *Twitter*, os usuários podem se inscrever para receber os tweets de outros usuários – a pessoa é "seguida" pelos seus assinantes, que são conhecidos como "seguidores". No *Twitter*, o usuário está no centro das relações.

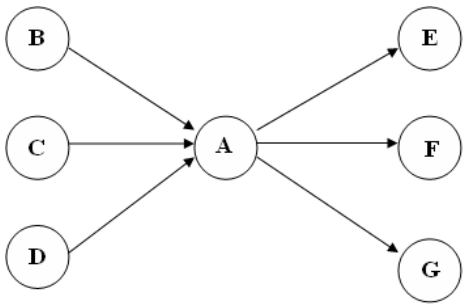

Figura 6.1: Relacionamentos do *Twitter*

A figura 6.1 exemplifica esse tipo de relacionamento, onde os usuários B, C e D seguem o usuário A, e o usuário A também segue outros usuários, os usuários E, F e G. Cada usuário pode seguir alguém e pode também ser seguido por alguém, estas são as duas únicas opções de relacionamento.

6.2.3 Preparando os dados

Este estudo serviu para medir a influência com o conjunto de dados dos jogadores de futebol do Flamento Futebol Clube com base nas suas contas do *Twitter*. Dos 23 jogadores, apenas 17 possuem conta no *Twitter*.

Tabela 6.1: As contas dos jogadores de futebol do Flamengo

Número	Nome	Twitter
1	Alex Sandro da Silva	@alex_silva03
2	Carlos Renato de Abreu	@Renatoo_abreu
3	Dario Bottinelli	@Bottinelli18Fc
4	David Braz de Oliveira Filho	@DavidBraz_14
5	Diego Maurício Machado de Brito	@FC_DMauricio_49
6	Gonzalo Antonio Fierro Caniullán	@FierroGonzalo
7	Guilherme Ferreira Pinto	@negueba_19
8	Jael Ferreira Vieira	@Jael_Gol
9	Leonardo da Silva Moura	@leomoura2
10	Paulo Victor Mileo Vidotti	@27paulovictor
11	Rafael Galhardo de Souza	@galhardo22
12	Rodrigo Oliveira da Silva Alvim	@FC_RAlvim21
13	Ronaldo de Assis Moreira	@10Ronaldinho
14	Thiago Neves Augusto	@_ThiagoNeves07
15	Thomás Jaguaribe Bedinelli	@jaguaribethomas
16	Vander Luiz Silva Souza	@Fc_VanderFla
17	Welinton Souza Silva	@fcwelinton3

Podemos obter a figura:

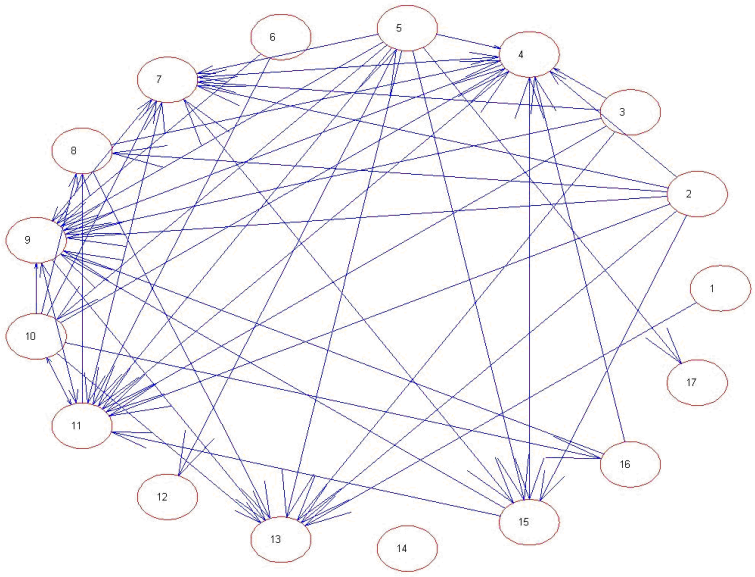

Figura 6.2: As relações entre os jogadores do Flamengo

A matriz de adjacência desta figura: (1 significa que o número da linha segue o número da coluna, 0 significa que não possuem relação).

$$\begin{vmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 1 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{vmatrix}$$

A classificação de cada pessoa contribuiu igualmente para os outros que ela seguia. Finalmente, a matriz foi transformada da seguinte forma:

$$\left|\begin{array}{cccccccccccccccc}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1/7 & 0 & 0 & 1/7 & 1/7 & 1/7 & 0 & 1/7 & 0 & 1/7 & 0 & 1/7 & 0 & 0 \\
0 & 0 & 0 & 1/5 & 0 & 0 & 1/5 & 0 & 1/5 & 0 & 1/5 & 0 & 1/5 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1/4 & 0 & 1/4 & 0 & 1/4 & 0 & 0 & 0 & 1/4 & 0 & 0 \\
0 & 0 & 0 & 1/9 & 0 & 0 & 1/9 & 0 & 1/9 & 1/9 & 1/9 & 1/9 & 1/9 & 0 & 1/9 & 0 & 1/9 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/2 & 0 & 1/2 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1/4 & 0 & 0 & 0 & 0 & 1/4 & 0 & 1/4 & 0 & 0 & 0 & 1/4 & 0 & 0 \\
0 & 0 & 0 & 1/2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/2 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1/3 & 0 & 0 & 0 & 1/3 & 0 & 1/3 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1/7 & 0 & 0 & 1/7 & 1/7 & 1/7 & 0 & 1/7 & 0 & 1/7 & 0 & 0 & 1/7 & 0 \\
0 & 0 & 0 & 1/5 & 0 & 0 & 1/5 & 1/5 & 1/5 & 1/5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1/4 & 0 & 0 & 1/4 & 0 & 1/4 & 0 & 1/4 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1/2 & 0 & 0 & 0 & 0 & 1/2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{array}\right|$$

Pode-se obter a matriz M com a transposição da matriz anterior:

$$\left|\begin{array}{cccccccccccccccccc}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1/7 & 1/5 & 0 & 1/9 & 0 & 1/4 & 1/2 & 0 & 1/7 & 1/5 & 0 & 0 & 0 & 1/4 & 1/2 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1/7 & 1/5 & 1/4 & 1/9 & 0 & 0 & 0 & 1/3 & 1/7 & 1/5 & 0 & 0 & 0 & 1/4 & 0 & 0 & 0 \\
0 & 1/7 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/7 & 1/5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1/7 & 1/5 & 1/4 & 1/9 & 1/2 & 1/4 & 0 & 0 & 1/7 & 1/5 & 0 & 0 & 0 & 1/4 & 1/2 & 0 & 0 \\
0 & 0 & 0 & 0 & 1/9 & 0 & 0 & 0 & 0 & 0 & 1/5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1/7 & 1/5 & 1/4 & 1/9 & 1/2 & 1/4 & 0 & 1/3 & 1/7 & 0 & 0 & 0 & 0 & 1/4 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1/9 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1/7 & 1/5 & 0 & 1/9 & 0 & 0 & 1/2 & 1/3 & 1/7 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1/7 & 0 & 1/4 & 1/9 & 0 & 1/4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/7 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1/9 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{array}\right|$$

6.2.4 A iteração do cálculo

Inicia-se a computar o valor PR seguindo o algoritmo PR, com o $pr1[1/17, \ldots, , , 1/17]$, O d é a diferença entre $sum(prN)$ e $sum(pr(N+1))$, c é a diferença entre o vetor $pr(N+1)$ e prN,

$$pr = \left| \begin{array}{c} 0,0588 \\ 0,0588 \\ 0,0588 \\ 0,0588 \\ 0,0588 \\ 0,0588 \\ 0,0588 \\ 0,0588 \\ 0,0588 \\ 0,0588 \\ 0,0588 \\ 0,0588 \\ 0,0588 \\ 0,0588 \\ 0,0588 \\ 0,0588 \end{array} \right| \xrightarrow[c=2,9412e-005]{d=0,1765} \left| \begin{array}{c} 0,0104 \\ 0,0104 \\ 0,0104 \\ 0,1455 \\ 0,0692 \\ 0,0104 \\ 0,1063 \\ 0,0390 \\ 0,1602 \\ 0,0287 \\ 0,1386 \\ 0,0169 \\ 0,1533 \\ 0,0104 \\ 0,0547 \\ 0,0188 \\ 0,0169 \end{array} \right| \xrightarrow[c=-2,7756e-005]{d=0,1806} \left| \begin{array}{c} 0,0106 \\ 0,0106 \\ 0,0106 \\ 0,1228 \\ 0,0275 \\ 0,0106 \\ 0,1572 \\ 0,0439 \\ 0,1449 \\ 0,0460 \\ 0,1612 \\ 0,0183 \\ 0,1092 \\ 0,0106 \\ 0,0827 \\ 0,0147 \\ 0,0183 \end{array} \right| \xrightarrow[c=2,9394e-005]{d=0,1382}$$

$$pr = \left| \begin{array}{c} 0,0081 \\ 0,0081 \\ 0,0081 \\ 0,1429 \\ 0,0264 \\ 0,0081 \\ 0,1533 \\ 0,0485 \\ 0,1570 \\ 0,0434 \\ 0,1657 \\ 0,0112 \\ 0,1023 \\ 0,0081 \\ 0,0827 \\ 0,0147 \\ 0,0112 \end{array} \right| \xrightarrow[c=-2,2651e-005]{d=0,1216} \left| \begin{array}{c} 0,0072 \\ 0,0072 \\ 0,0072 \\ 0,1428 \\ 0,0183 \\ 0,0072 \\ 0,1610 \\ 0,0477 \\ 0,1584 \\ 0,0432 \\ 0,1702 \\ 0,0101 \\ 0,1038 \\ 0,0072 \\ 0,0853 \\ 0,0134 \\ 0,0101 \end{array} \right| \xrightarrow[c=-1,9859e-005]{d=0,1210} \left| \begin{array}{c} 0,0071 \\ 0,0071 \\ 0,0071 \\ 0,1439 \\ 0,0172 \\ 0,0071 \\ 0,1617 \\ 0,0484 \\ 0,1594 \\ 0,0432 \\ 0,1714 \\ 0,0092 \\ 0,1016 \\ 0,0071 \\ 0,0861 \\ 0,0133 \\ 0,0092 \end{array} \right| \xrightarrow[c=-3,5628e-005]{d=0,1178}$$

$$pr = \begin{vmatrix} 0,0069 \\ 0,0069 \\ 0,0069 \\ 0,1445 \\ 0,0161 \\ 0,0069 \\ 0,1624 \\ 0,0484 \\ 0,1599 \\ 0,0431 \\ 0,1721 \\ 0,0088 \\ 0,1019 \\ 0,0069 \\ 0,0863 \\ 0,0131 \\ 0,0088 \end{vmatrix} \xrightarrow[c=-4,2946e-005]{d=0,1176} \begin{vmatrix} 0,0069 \\ 0,0069 \\ 0,0069 \\ 0,1446 \\ 0,0158 \\ 0,0069 \\ 0,1626 \\ 0,0485 \\ 0,1600 \\ 0,0431 \\ 0,1723 \\ 0,0087 \\ 0,1017 \\ 0,0069 \\ 0,0864 \\ 0,0131 \\ 0,0087 \end{vmatrix} \xrightarrow[c=1,4104e-005]{d=0,1173} \begin{vmatrix} 0,0069 \\ 0,0069 \\ 0,0069 \\ 0,1447 \\ 0,0156 \\ 0,0069 \\ 0,1600 \\ 0,0485 \\ 0,1600 \\ 0,0431 \\ 0,1724 \\ 0,0087 \\ 0,1017 \\ 0,0069 \\ 0,0864 \\ 0,0131 \\ 0,0087 \end{vmatrix} \xrightarrow[c=-8,1785e-006]{d=0,1172}$$

$$pr = \begin{vmatrix} 0,0069 \\ 0,0069 \\ 0,0069 \\ 0,1447 \\ 0,0155 \\ 0,0069 \\ 0,1627 \\ 0,0485 \\ 0,1601 \\ 0,0431 \\ 0,1724 \\ 0,0086 \\ 0,1017 \\ 0,0069 \\ 0,0865 \\ 0,0131 \\ 0,0086 \end{vmatrix}$$

6.2.5 Resultado

Depois de realizar 10 iterações, o resultado já converge. Finalmente, o impacto dos jogadores dentro deste conjunto de dados é listado na tabela abaixo.

Tabela 6.2: O valor PR para os jogadores de futebol do Flamengo Futebol Clube

Rank	Nome	Valor PR
1	Rafael Galhardo de Souza	0,1724
2	Guilherme Ferreira Pinto	0,1627
3	Leonardo da Silva Moura	0,1601
4	David Braz de Oliveira Filho	0,1447
5	Ronaldo de Assis Moreira	0,1017
6	Thomás Jaguaribe Bedinelli	0,0865
7	Jael Ferreira Vieira	0,0485
8	Paulo Victor Mileo Vidotti	0,0431
9	Diego Maurício Machado de Brito	0,0155
10	Vander Luiz Silva Souza	0,0131
11	Rodrigo Oliveira da Silva Alvim	0,0086
12	Welinton Souza Silva	0,0086
13	Alex Sandro da Silva	0,0069
14	Carlos Renato de Abreu	0,0069
15	Dario Bottinelli	0,0069
16	Gonzalo Antonio Fierro Caniullán	0,0069
17	Thiago Neves Augusto	0,0069

Desse resultado chegamos as seguintes análises:

1. O Rafael Galhardo de Souza tem o maior valor PageRank da equipe: 0,1724. Vemos do grafo da Figura 6.2 que existem 9 pessoas ligadas a ele, e ele tem 1/3 do Leonardo da Silva Moura e um quarto de David Braz, Guilherme Ferreira e Thomás Jaguaribe. O PageRank final do Leonardo Moura é 0,1601, o valor PR do David Braz é 0,1447, o do Guilherme Ferreira é 0,1627 e o do Thomás Jaguaribe é 0,0865. Embora ele tenha conseguido metade do valor do Gonzalo Antonio, o valor final do PR dele é 0,0069, portanto, esse item não ajudou-o a obter mais pontos.

2. O Guilherme Pinto Ferreira tem o valor PageRank igual a 0,1627, sendo o segundo da equipe. Da figura das conexões, existem 8 pessoas ligadas a ele. Assim como o Rafael Galhardo, ele também conseguiu 1/3 do Leonardo Moura, 1/5 do Rafael Galhardo (o maior valor PR) e 1/4 do Thomás Jaguaribe.

3. O Leonardo da Silva Moura tem o terceiro maior valor PageRank da equipe: 0,1601. O interessante é que ele tem 10 pessoas ligadas a ele, é o maior número da equipe, porém ele não conseguiu a nota mais alta. Isso ocorreu porque o algoritmo PageRank considera a qualidade das ligações. O Carlos Renato, Dario Bottinelli e Gonzalo Antonio possuem o valor PageRank baixo: 0,0069, o do Welinton Souza é 0,0086, do Thomás jaguatibe é 0,0131. Assim, o resultado de Leonardo mostra que o número de ligações não influencia igualmente, a qualidade da ligação também é importante.

4. Utilizando o fator de amortecimento, as pessoas sem link também possuem influência.

6.3 O Cálculo de PageRank do *ScienceNet*

Este estudo integrou o modelo de computação PageRank para medir a influência do indivíduo no *ScienceNet*.

6.3.1 Introdução

O ScienceNet.cn (ScienceNet 2012), lançado em janeiro de 2007, é o portal *online* líder servindo a comunidade científica chinesa. Ele fornece notícias de ciência oportunas e confiáveis, uma plataforma interativa e uma seção de classificados ativos.

Além de possuir notícias científicas e informações valiosas classificadas, o ScienceNet.cn também abriga a comunidade mais ativa e de alto perfil virtual de cientistas da língua chinesa.

O ScienceNet.cn é co-patrocinado pela Academia Chinesa de Ciências (CAS), Academia Chinesa de Engenharia (CAE) e Fundação Nacional de Ciência Natural da China (NSFC).

Este site possui vários canais, tais como Artigos, Notícias, Blog, Fórum e etc. Na sua parte de blog, existem duas listas de classificação, uma lista mostra os blogs que possuem um maior número de visitas, já a outra lista mostra os blogs com o maior número de visitas médias por artigo. Estas duas listas de classificação não são consideradas justas. Portanto, esta plataforma foi escolhida para a aplicação do Sistema W-Entropia.

6.3.2 As relações do *ScienceNet*

A relação entre os membros do *ScienceNet* é de "recomendação". A diferença entre o *ScienceNet* e o *Twitter* é que o *ScienceNet* considera os artigos dos blogs como sendo unidades e as pessoas podem recomendar os artigos caso elas gostem. Um membro pode publicar mais de um artigo e a influência de todos os seus artigos constituem a influência total deste membro. A figura 6.3 seguinte ilustra as relações.

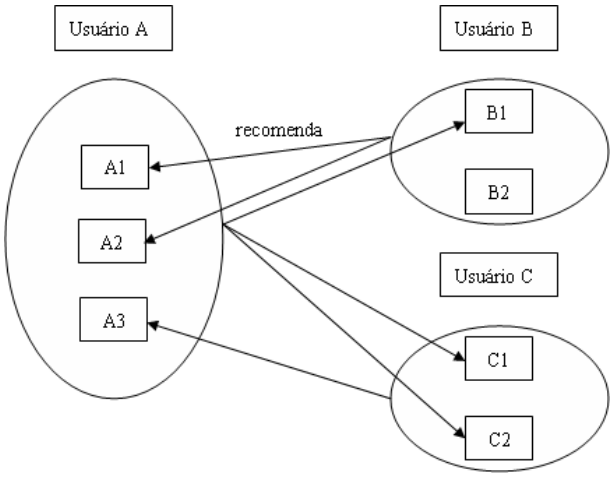

Figura 6.3: As relações entre os membros do *ScienceNet*

Nesta figura, cada usuário possui seu blog pessoal e cada blog pode possuir diversos artigos. O usuário A, por exemplo, possui 3 artigos em seu blog e todos os artigos foram recomendados por outros usuários. É possível perceber que o relacionamento existente é de recomendações, um usuário pode recomendar artigos de outros blogs, por isso o artigo está no centro das relações, porém quem realiza essas relações são os usuários.

6.3.3 Modelo PageRank Personalidade

Com a relação de recomendar, podemos usar o modelo PageRank para calcular o valor PageRank de cada indivíduo. Se **A** recomendar N artigos, cada recomendação de **A** possui um peso de $1/N$. Há uma situação especial, se um autor **B** escrever m artigos e se a pessoa **A** recomendar n deles ($n < m \& n < N$), então o peso da transferência **B** ao autor **A** é n/N.

Assim, o PageRank do autor **B** é:

$$PR(B) = \sum_{v \in B(B)} n \frac{PR(A)}{L(A)} \tag{6.1}$$

onde $B(B)$ é o conjunto dos recomendados de B, n é o número de vezes A recomendou um artigo de B e $L(A)$ é o número que A recomendou.

Computação

Neste estudo, 3.600 artigos foram selecionados da lista do número de recomendações. Estes artigos envolvem 5.265 pessoas. Com o conjunto de dados, um grafo dirigido com 5.265 nós foi construído com a relação de "recomendar". O cálculo é o mesmo que foi utilizado no primeiro estudo.

Resultados e Análises

Dentro destas 5.265 pessoas, apenas 481 pessoas (o número de autores) têm o valor PageRank significativo, as outras pessoas, que não são autoras, não possuem o *backlink* no grafo direcionado. Depois de iterar o cálculo, o PageRank de cada autor pode ser obtido. Por conta do grande número de dados, aqui só serão exibidos os 10 melhores autores:

Tabela 6.3: O valor PR dos blogueiros do *ScienceNet*

Classificação	UID	Valor PR	Erro relativo
1	1557	0,0395	0,18%
2	254303	0,0330	0,21%
3	51814	0,0295	0,06%
4	531950	0,0272	0,10%
5	111635	0,0260	0,09%
6	496920	0,0253	0,15%
7	40247	0,0251	0,05%
8	53483	0,0225	0,06%
9	2237	0,0188	0,04%
10	41757	0,0186	0,01%

Desta tabela,

1. O No.1557 obteve o primeiro lugar da lista. Publicou 170 artigos dos 3600 artigos. Não só o número de artigos ajudou-o a obter a maior pontuação. O número de recomendações também empenhou um papel importante no valor PageRank, 5.740 vezes recomendaram seus artigos.

2. O No.254303 ficou em segundo lugar da lista. Ele publicou 83 artigos dos 3600 artigos e obteve 4.136 recomendações.

3. O No.111635 obteve o quinto lugar da lista. Ele publicou 86 artigos dos 3600 artigos e obteve 2.747 recomendações.

4. O No.2237 obteve o nono lugar da lista. Ele publicou 24 artigos dos 3600 artigos. Seus trabalhos têm 3.262 recomendações.

5. Comparando o número 111635 e o 2237, o autor 2237 tem mais recomendações do que o número 111635, mas o seu valor PageRank é menor, este resultado mostra a importância da qualidade de *backlink* mais uma vez.

Avaliação

twitter-rela O *ScienceNet* tem dois critérios para classificar seus membros. Um deles é o número total de visitas e o outro é o número de visitas médias por artigo. Esta parte vai comparar a lista de classificação gerada pelo PageRank com as listas de classificação do *ScienceNet*, levando em consideração os dados em conflitos.

Tabela 6.4: As informações conflito nas três listas

UID	Nome	Rank PR	Valor PR	Rank por visitas	Número de visitas	Rank méida	Número médio
2374	Wang	No.59	0,0038	No.10	3582851	No.3	19904
415	Huang	No.249	0,0004	No.1	8548802	No.281	1142
565522	Shen	No.291	0,00019	No.234	378335	No.2	32833

1. Wang está em uma boa posição nas listas que o *ScienceNet* publicou, mas está numa posição ruim na classifição do PageRank (posição 59).

2. Huang está em primeiro lugar na lista de visitas com o valor 8548802, mas o PageRank dele é 0,0004, está em 249º lugar e na lista de média de visitas, está em 281º lugar com valor 1142.

3. Shen está em segundo lugar na lista por número médio com o valor 32833, mas o PageRank dele é 0,0035, estando em 291º lugar, o número de visitas dele também não é bom, está em 234º lugar com o valor 378335.

.

Para medir a influência dos membros de forma global, ele deve sintetizar vários parâmetros.

6.4 O Cálculo do W-Entropia do *ScienceNet*

Este estudo é baseado no anterior, foi considerado o valor PR dele e adicionadas as visitas totais e as visitas médias de todos os artigos do blog, estes são os três elementos que compõem a influência. A influência dos membros será calculada pelo algoritmo W-Entropia.

Preparando os dados

Há três itens empregados para calcular o W-Entropia. Foi considerado o valor PageRank com o peso de 40%, o número de visitas totais com o peso de 30% e o número médio de visitas por artigo com o peso de 30%.

O primeiro passo é obter o valor médio destes três itens.

$$m = \sum_{i=1}^{n} a_i * p_i \tag{6.2}$$

E então ajustar o parâmetro p_i para q_i

$$\begin{cases} q_j = p_j/(n+1), & j = 1, 2, \cdots, n \\ q_{n+1} = 1 - \sum p_j, & j = 1, 2, \cdots, n \end{cases} \tag{6.3}$$

Então calcula-se a entropia de cada indivíduo

$$h(q_1, q_2, \cdots, q_n, q_{n+1}) = -\sum q_j \log_{n+1} q_j \qquad j = 1, 2, \ldots, n+1 \tag{6.4}$$

A tabela abaixo apresenta os dez maiores valores em cada item.

Tabela 6.5: Classificação dos três itens do *ScienceNet*

Rank	UID	Valor PR	UID	Todas as visitas	UID	Visitas Médias
1	1557	0,0395	415	8545710	2374	77119
2	254303	0,0330	280034	6660327	2237	21209
3	51814	0,0295	2277	6384899	176	8363
4	531950	0,0272	176	6253178	65865	8306
5	111635	0,0260	1557	5412290	295006	7748
6	496920	0,0253	53483	5014640	265898	7567
7	40247	0,0251	126	4202693	1565	7273
8	53483	0,0225	200147	3700696	52239	6413
9	2237	0,0188	41757	3649381	287179	5881
10	41757	0,0186	2237	3586787	3377	5588

Considerando 0,0395 a base do valor PageRank, 8.545.710 a base de todas as visitas e 77.119 a base das visitas médias.

Resultados e Análises

O resultado do W-Entropia é o seguinte:

Tabela 6.6: O Índice W-Entropia dos autores do *ScienceNet*

Rank	Nome	ID	Índice W-Entropia
1	Wu Yishan	1557	100
2	Rao Yi	2237	68,92293842
3	Shi Yigong	46212	60,00704327
4	Chen An	53483	59,50526819
5	Cao Guangfu	40247	49,60618336
6	Wu Feipeng	51814	46,22369429
7	Wang Hongfei	176	45,42629305
8	Wang Feiyao	2374	44,54352492
9	Li Xuekuang	254303	42,90553618
10	Lv Zhe	111635	38,79296722

A partir do resultado do W-Entropia:

1. O autor 1557 continua em primeiro lugar. No item do valor PageRank, ele está em primeiro lugar. No item de visitas totais, ele está em 5º lugar, embora a média de visitas por artigo não esteja muito alta, ele conseguiu o maior índice W-Entropia

2. O autor 2237 está em segundo lugar. No item do valor PageRank, ele está em 9º lugar. No item de visitas totais, ele está em 11º lugar e no item de média de visitas por artigo ele está em 2º lugar, por conta da sua distribuição equilibrada ele conseguiu o segundo lugar.

3. O autor 2374 mencionado acima está no 8º lugar. Este resultado é muito melhor do que seu valor de PageRank. Ele conseguiu bons resultados nos itens de visitas

totais e visitas médias, assim, com a teoria da informação, ele conseguiu um bom índice.

6.5 O Cálculo da Influência em Diversas Plataformas

Este estudo aplica o índice W-Entropia usando dados obtidos do *Twitter*, *Facebook* e *Google* para calcular o impacto de alguns indivíduos nas redes sociais. O número de seguidores do *Twitter*, Fãs do *Facebook* e resultados do *Google*, são de 03/11/2011. Os resultados do *Google* também foram incluídos nesse estudo pois eles representam a quantidade de informação contida na Internet a respeito de um indivíduo.

6.5.1 Determinação da Distribuição dos Pesos no Ranking

Para estudar uma distribuição de peso adequado, quatro grupos de peso são projetados nesta pesquisa. A sequência da distribuição segue a seguinte ordem, FacebookP (P1), TwitterP (P2) e MGoogleP (ou YouTubeP ou GoogleP, P3), respectivamente. Peso G1: 45%, 30% e 25%; Peso G2: 30%, 45% e 25%; Peso G3: 35%, 35% e 30%; Peso G4: 33,34%, 33,33% e 33,33%. A Tabela embaixa apresenta os índices W-entropia que foram calculados a partir destes quatro grupos de peso e assim comparar com o índice Famecount. Os dados reais, FacebookP, TwitterP e MGooleP são coletados de 20 pessoas ou marcas.

1. MGoogleP é usado como o terceiro parâmetro e possui maior efetividade em comparação com YouTubeP. Alguns membros, como a marca Facebook, YouTube, Barak Obama, Britney Spears e Michael Jackson estão bem classificadas no W-Entropia. Por exemplo, o presidente Obama está no décimo segundo lugar por Famecount, em sétimo lugar no W-entropia (usando Peso G2), devido aos seus grandes números de resultados, 981 milhões, em busca no Google.

2. O coeficiente de correlação entre índice Famecount e o índice W-Entropia de peso G1 é 0,8807, a partir do peso G2 é 0,9094, a partir do peso G3 é 0,8871 e de Peso G4 é 0,8799. Tendo o índice do Famecount como referência, a distribuição do peso G2 possui maior coeficiente de correlação, esta é uma razão para a escolha do Peso G2 do W-Entropia para classificação.

3. Com as distribuições de pesos diferentes, os índices de W-Entropia são diferentes. Com base nos resultados da Tabela 6.7, as sequências de classificação de índice W-Entropia são alterados,mas não significativamente com a mudança dos grupos de peso. Pelas razões acima citadas é utilizado o peso G2 como peso padrão.

Tabela 6.7: Comparando o Ranking Famecount e W-Entropia

	Famecount		Weight G1		Weight G2		Weight G3		Weight G4	
Nome	No.	Índice	No.	Índice	No.	Índice	No.	Índice	No.	Índice
Lady Gaga	1	100	1	100	1	100	1	100	1	100
Rihanna	2	95,55	4	92,86	4	94,97	4	93,73	5	93,60
Bieber	3	94,92	2	96,26	3	95,96	3	96,12	3	96,12
Shakira	4	91,38	8	83,44	8	85,37	8	83,38	7	82,14
Eminem	5	90,08	9	76,39	9	82,78	9	78,60	9	77,84
KatyPerry	6	87,05	5	92,14	6	92,40	7	92,27	6	92,28
YouTube	7	85,44	3	95,87	2	98,65	2	98,19	2	98,71
Facebook	8	81,88	7	86,34	5	94,49	6	92,41	8	81,88
T. Poker	9	76,17	15	36,09	15	51,49	15	42,69	15	41,85
Taylor Swift	10	74,66	11	63,39	12	63,54	13	62,47	13	61,79
, C. Ronaldo	11	74,48	14	53,23	14	59,57	14	54,90	14	53,85
B. Obama	12	72,91	6	91,35	7	90,80	5	92,70	4	93,67
S. Gomez	13	72,62	13	60,09	13	63,38	12	62,74	12	62,40
M. Jackson	14	71,62	12	61,71	10	73,99	10	70,59	10	72,19
B. Spears	15	71,31	10	73,31	11	65,90	11	69,58	11	69,25

Em aplicações práticas, o método AHP (Saaty 1990) ou outros metodos podem ser usados para analise de distribuição de peso e atribuição dos valores mais apropriados.

6.5.2 W-Entropia Análise Propriedade no Ranking

Para entender melhor a propriedade de W-Entropia, os índices do Poker Texas Hold'em e o presidente Barack Obama são estudados de acordo com índice Famecount e W-Entropia. Para unificar os dados originais, YouTubeP é usado como parâmetro terceiro pelos dois métodos de classificação.

Texas Hold'em Poker é uma marca que tem mais de 57 milhões de fãs no Facebook, mas não tem o mesmo sucesso ou influência no Twitter e YouTube. FacebookP dela é 0,9311, TwitterP é 0,0102 e YouTubeP é 0; o média m é calculada como 0,4220, veja a tabela 6.7. Em Famecount, o índice dela é 76,17 e fica em nono lugar. em W-Entropia, a entropia h, como o coeficiente de distribuição, é 0,4037, o índice W-Entropia dela é 36,98 e classificado em quinto lugar.

O presidente Obama tem mais de 25,19 milhões de fãs no Facebook, 12,48 milhões seguidores no Twitter e 170 milhões de visualizações no YouTube. Seu FacebookP é 0,4094, TwitterP é 0,6500 e YouTubeP é 0,0737; a média m dele é calculado como 0,3977, menor que o Texas Hold'em Poker, veja a tabela 6.7. Podemos dizer que o Texas Hold'em poker tem mais influência do que o presidente Obama? No ranking Famecount, o índice do Obama é 72,91 e ficou em décimo segundo lugar. Na W-Entropia, a entropia h é 0,6066, o índice é 50,23 e em décimo primeiro lugar maior que o Texas Hold'em Poker.

A figura 6.4 ilustra a comparação dos parâmetros e índices entre Texas Hold'em Poker e Barack Obama. Comparando-se os parâmetros dos dois casos, a média m de Texas Hold'em Poker é maior do que Obama, mas a informação (número de fãs ou seguidores ou pontos de vista) entre Facebook, Twitter e YouTube não está bem distribuída. A entropia

h de Obama é maior do que h do Poker Texas Hold'em. Como resultado, o presidente Obama está com um maior índice de W-Entropia e melhor ranking. É claro que mostra a melhor distribuição da informação através das redes de multi-sociais, a maior influência do membro como Obama.

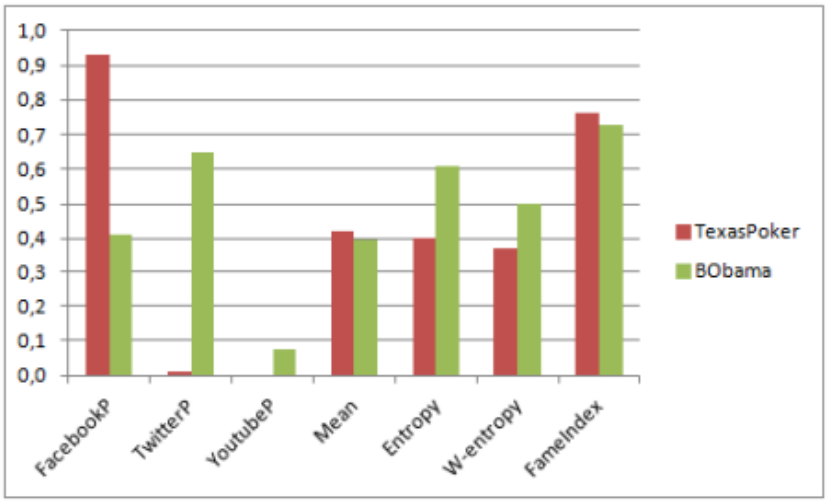

Figura 6.4: Comparação da Influência Entre o Texas Holdem Poker e Barack Obama

Da mesma forma, no caso do músico Selena Gomez, ela tem mais de 29 milhões de fãs no Facebook com FacebookP 0,4377, 10 milhões de seguidores no Twitter com TwitterP 0,5357 e YouTubeP 0,3040, então m é calculado como 0,4337, veja a table 6.4. Em Famecount, o índice é 72,62 e, em décimo terceiro lugar. Na W-Entropia, a entropia h é 0,6990, o índice é 60,36 e classificado em décimo lugar.

6.5.3 Comparação de Classificação W-Entropia com Famecount

Famecount é um site de sucesso para apresentar as estrelas de redes sociais na Internet ocidental com o objetivo de comércio e entretenimento. W-Entropia é um método científico para medir a influência das pessoas e marcas na Internet. Junto com os 20 membros ou marcas na figura 6.5, observa-se também que o primeiro é mais geral e analisou sem a mudança exata eo índice W-Entropia é alterado conforme a tendência dos dados reais. Este cenário pode ser observado na figura 6.5, onde, significa (m), (h) a entropia, Famecount e W-Entropia índices juntamente com 20 membros são apresentados.

A classificação Famecount é baseada na média dos parâmetros do Facebook, Twitter e YouTube. Somente usando esta média não reflete na distribuição de informação de pessoas ou de marcas através de multi-plataformas. Este cenário pela W-Entropia é muito mais realista e abrangente, o que é demonstrado pelo estudo comparativo dos casos de Poker Texas Hold'em com Barack Obama e Coca-Cola com Ricardo Kaká etc, ver figuras 6.4 e 6.5.

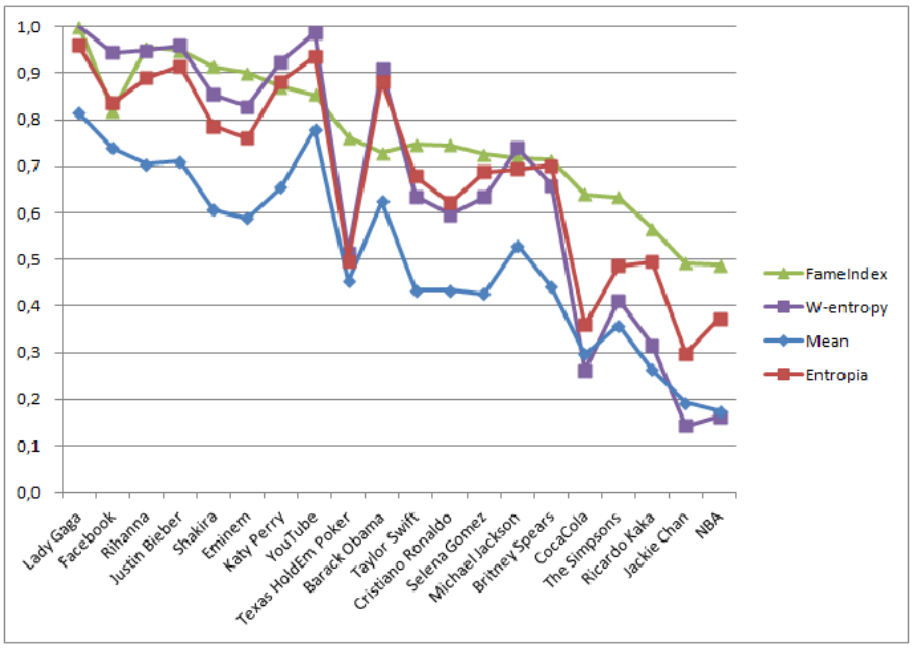

Figura 6.5: A Comparação dos Parâmetros Entre Famecount e W-Entropia Classificação

Um aspecto importante da Famecount é que os índices de outros membros são relativos aos membros mais populares. Isto pode não ser adequado para a investigação científica sobre a medição da influência dos membros.

Capítulo 7

Conclusão e Trabalhos Futuros

7.1 Conclusão

Consideram-se atingidos os objetivos deste trabalho, visto que foi desenvolvido um sistema para medir a influência do indivíduo na rede social, utilizando-se técnicas de PageRank e o W-Entropia. Um protótipo baseado em dados reais do *Twitter* e do *ScienceNet* foi implementado e diversos experimentos foram realizados para analisar seu comportamento e performance.

Foi criado um modelo de computação que mede a influência de um único indivíduo dentro de uma rede social com o modelo Pagerank. O cálculo se baseia na estrutura da relação entre os indivíduos, assim, o resultado é mais justo. O modelo W-Entropia pode calcular a influência de um indivíduo entre várias plataformas, com a entropia este modelo pôde medir o desequilíbrio da transmissão das informações, então o resultado final é mais completo e razoável.

Para implementar esse modelo de computação, foi necessário desenvolver um *crawler* para coletar informações da internet. O *crawler* processa essas informações para o formato que satisfaz o modelo de computação. O processo de cálculo é executado *offline*, finalmente transmite o resultado para o modelo de exibição, para mostrar na internet.

Por fim, destacam-se as principais realizações deste trabalho:

— Modelo PageRank: Aplicar o algoritmo PageRank para medir a influência do indivíduo na rede social e obter o resultado das ligações internas entre os indivíduos. Este método pode evitar a manipulação humana. Além disso, no início do algoritmo, cada pessoa possui um valor igual de influência, esta propriedade garante que este algoritmo seja justo para todos.

— Modelo W-Entropia: Empregando a teoria da informação pode-se medir o desequilíbrio das informações durante a transmissão entre várias plataformas. Com este desequilíbrio, a influência de uma pessoa entre diferentes plataformas pode ser medida com mais precisão.

— Realizar o sistema W-Entropia: Este sistema consiste de três partes: o *crawler*, a computação e o modelo de exibição. Este sistema pode obter informações, calcular a influência e exibir o ranking de influência.

— Aplicar o sistema em dados reais: Neste trabalho foram realizados quatro estudos diferentes, a partir dos quais foi possível demonstrar a eficácia do Sistema W-Entropia. O

resultado obtido se mostrou mais preciso e mais coerente do que as listas de classificação existentes nas redes sociais.

7.2 Trabalhos Futuros

O Linkedin é uma ótima ferramenta para a realização de contatos profissionais e para que uma empresa possa expandir sua rede de contatos. No Brasil são mais de 4 milhões de participantes e para atingir essas pessoas o Linkedin possui a ferramenta Linkedin Ads. O perfil desta rede social é mais sério e os anúncios geralmente tem o objetivo de divulgar um serviço de uma empresa.

Atualmente existe um outro tipo de rede social, que é organizada por relacionamentos(chamados "conexões"), como por exemplo o *LinkedIn*. Ao contrário das redes sociais que já foram mencionadas, os relacionamentos entre os usuários possuem diferentes níveis, o cargo profissional de uma pessoa também possui diferentes níveis de influência sobre as outras pessoas. Então, para medir a influência nesse tipo de rede social, é necessário um novo modelo de análise para calcular os relacionamentos e pesos para cada membro.

Appendices

Apêndice A

O Resultado do W-Entropia no ScienceNet

Tabela A.1: O resultado do W-Entropia no ScienceNet

Rank	ID	Número de visitas	Número médio	Valor PR	Valor W-Entropia
1	1557	5413417	2430	0.03946996	100
2	2237	3589401	14771	0.01876348	68.92293842
3	46212	1015311	44143	0.00505712	60.00704327
4	53483	5019292	1312	0.02246953	59.50526819
5	40247	2859835	3115	0.02513438	49.60618336
6	51814	631180	8646	0.02946048	46.22369429
7	176	6253921	5823	0.00578556	45.42629305
8	2374	3582851	19904	0.00386243	44.54352492
9	254303	1028937	2078	0.03302538	42.90553618
10	111635	1667328	2549	0.02595729	38.79296722
11	415	8548802	1142	0.00038254	38.05218196
12	531950	490909	2821	0.02718757	30.23012982
13	200147	3701286	1806	0.01062706	28.19068553
14	51597	3356013	4498	0.00915356	28.04411405
15	280034	6661939	596	0.00081533	26.54358088
16	4699	1657416	5163	0.01461514	25.48392085
17	2277	6387328	1149	0.00053791	25.42238428
18	2984	2825525	2500	0.01089298	23.8083064
19	378335	131333	32833	0.00026825	23.67353495
20	496920	290394	703	0.02526153	21.80705325
21	126	4203233	1856	0.00348680	19.65540988
22	480705	567542	1762	0.01784382	16.65149875
23	2211	2155221	2350	0.00900306	16.50759155
24	89391	934021	1489	0.01483248	15.12094828
25	40615	1093298	1641	0.01150212	12.40296379
26	71964	2239151	4878	0.00295323	12.06452226

27	829	1062931	2271	0.01053593	11.92128558
28	2638	1914509	2782	0.00575822	11.54579053
29	455154	573425	1018	0.01375972	11.04279864
30	39626	2114836	2035	0.00500675	10.77567406
31	41757	434	48	0.01858659	10.70937824
32	224810	1575187	832	0.00824535	10.31852698
33	5190	885667	5826	0.00617884	9.936242295
34	45	1482828	3530	0.00520346	9.544298888
35	1565	2249622	5420	0.00049424	9.340656038
36	279992	1528286	1810	0.00642090	9.260103515
37	249679	554044	3820	0.00905579	9.219439749
38	2321	2748351	1982	0.00116942	8.96438203
39	265898	585438	7317	0.00526522	8.77028974
40	41174	1641453	3568	0.00373119	8.689944237
41	3377	596375	8770	0.00360988	8.329822888
42	290052	1221194	2523	0.00619309	8.31334452
43	295006	934453	8343	0.00222520	8.082687275
44	45671	1227627	3697	0.00477546	8.027228478
45	200071	1379720	1918	0.00554657	7.769440287
46	453185	598644	633	0.01043368	7.462968856
47	287179	627653	4581	0.00570591	6.978797529
48	2344	933940	4225	0.00451829	6.914519624
49	38899	1888427	3784	0.00086717	6.750446813
50	915	1161205	4899	0.00257706	6.545574312
51	475	2145106	1711	0.00145055	6.541441024
52	43772	434	48	0.01368605	6.401850106
53	216720	491658	2092	0.00794728	6.200847943
54	542699	24455	12227	0.00184254	6.19668825
55	40049	605924	3155	0.00616440	6.075633731
56	412323	541432	2274	0.00721807	5.954826993
57	39472	850093	4208	0.00387272	5.921731093
58	460310	808020	1715	0.00621741	5.738980251
59	39416	569600	3218	0.00583732	5.68389156
60	39070	1800218	1564	0.00171275	5.38098414
61	71485	797334	4454	0.00321041	5.296786682
62	55503	472065	1761	0.00727370	5.262611122
63	226	2278585	907	0.00049192	5.260402508
64	1248	313925	11211	0.00023331	5.225783484
65	65865	488195	4605	0.00435413	5.155046982
66	296014	485871	5339	0.00356250	5.010408584
67	1750	681438	811	0.00676199	4.885833173
68	2361	698094	3635	0.00379846	4.79711943
69	39731	1739223	1481	0.00131299	4.718621889
70	362400	1757258	666	0.00190460	4.631179833

71	330732	1082908	2560	0.00272175	4.512882497
72	39061	645614	6329	0.00144102	4.479725093
73	393255	326619	7423	0.00189101	4.371832588
74	38667	629660	4919	0.00246039	4.338507829
75	356017	455534	9110	0.00008360	4.270119669
76	39446	534539	2686	0.00461102	4.134859577
77	55745	1013430	1419	0.00358055	4.104043859
78	2068	822135	4216	0.00181666	4.015483424
79	63234	828321	5050	0.00112098	4.002531161
80	56669	1201079	1631	0.00235483	3.925624103
81	216627	416453	5479	0.00238579	3.863432571
82	210844	779314	1808	0.00377743	3.744094087
83	3075	1209880	2647	0.00115532	3.69629659
84	454867	244784	5208	0.00318608	3.670322316
85	320333	332208	1092	0.00628497	3.462726537
86	28418	554391	1470	0.00458076	3.330437955
87	3779	434	48	0.00933862	3.322543045
88	206819	761005	2615	0.00251521	3.300914793
89	91358	10163	10163	0.00018498	3.236048831
90	359436	828015	3000	0.00168932	3.147383395
91	324673	750913	2068	0.00273072	3.072711081
92	640160	35575	7115	0.00188268	3.033360073
93	2317	1321675	2259	0.00022176	2.997469891
94	27691	1095515	3139	0.00021306	2.87933815
95	52239	284839	5585	0.00145990	2.77757467
96	39465	558010	5580	0.00026655	2.73152098
97	502444	207567	1904	0.00491285	2.688160674
98	385748	290884	5103	0.00154389	2.595406225
99	214181	613997	2102	0.00250126	2.532174631
100	43510	535945	2791	0.00219182	2.499978799
101	2644	724015	3497	0.00077107	2.466862521
102	5281	890643	1113	0.00207997	2.448027023
103	293156	610499	1387	0.00293911	2.381402673
104	200056	580450	1802	0.00264733	2.357415613
105	45849	193316	7159	0.00015899	2.34995347
106	2024	1153094	1673	0.00040546	2.339769022
107	5793	397160	1975	0.00321956	2.285168451
108	296123	353784	1538	0.00383467	2.278555552
109	281175	371550	3472	0.00195440	2.245501032
110	448631	89113	452	0.00627224	2.20588533
111	302992	634178	1933	0.00199831	2.178527804
112	39523	963711	2289	0.00036872	2.170016984
113	38036	420034	1585	0.00323977	2.146203928
114	70942	299436	3219	0.00232176	2.133986582

115	215715	382185	1144	0.00381645	2.125832595
116	284259	526224	4616	0.00018893	2.08850205
117	39346	601120	1832	0.00204870	2.067785677
118	299127	389635	1198	0.00360306	2.056685078
119	2396	293116	5428	0.00045622	2.033701913
120	45640	124861	6243	0.00053956	1.985752873
121	54025	255099	2452	0.00291127	1.961514233
122	439042	209410	894	0.00464120	1.957121498
123	39437	794021	2473	0.00046153	1.90775686
124	39356	887989	1368	0.00091876	1.861407604
125	458	653610	2246	0.00111128	1.856567119
126	3598	896197	1978	0.00029628	1.798115801
127	612799	57547	4110	0.00218436	1.790445907
128	42818	505382	1701	0.00210183	1.78822633
129	290140	141720	3543	0.00218067	1.770712132
130	278905	433559	1086	0.00296822	1.766216537
131	260340	395840	1552	0.00260384	1.714727444
132	502764	46057	903	0.00499272	1.674508614
133	438991	426878	1350	0.00256991	1.67355708
134	627086	60568	1211	0.00455351	1.658779173
135	279177	523883	1102	0.00221778	1.598828838
136	590130	168904	1373	0.00359870	1.564647839
137	290937	373125	1615	0.00228613	1.525891193
138	95499	439554	1564	0.00199381	1.508458975
139	76913	434	48	0.00592612	1.502919699
140	537167	85687	4080	0.00151174	1.498256623
141	402046	216811	4336	0.00065487	1.476793862
142	279096	73293	5637	0.00030062	1.435914128
143	392525	249562	1862	0.00244583	1.426028306
144	472757	243376	4125	0.00057900	1.412202565
145	614814	100845	4384	0.00101910	1.405501485
146	40992	391577	2175	0.00145700	1.399729632
147	39358	433563	3468	0.00018577	1.353351682
148	647453	5270	5270	0.00067426	1.31323403
149	41701	507506	1120	0.00164230	1.284161457
150	39377	234218	4182	0.00028794	1.25385051
151	540428	77392	797	0.00396698	1.240331214
152	639148	5716	5716	0.00021889	1.234196692
153	433169	11475	5737	0.00015959	1.224935817
154	40115	153089	2097	0.00219352	1.184899841
155	190	801618	1045	0.00024403	1.14447233
156	72483	495996	2285	0.00041774	1.142276747
157	266190	733067	874	0.00058434	1.117160481
158	45134	574024	1728	0.00044241	1.090051121

159	219944	339897	2023	0.00115481	1.083150733
160	714	398528	797	0.00194324	1.062745709
161	303939	434	48	0.00480994	1.04102571
162	279293	557347	683	0.00130025	1.039563266
163	339326	726953	1086	0.00023975	1.021545389
164	50350	589157	1083	0.00073594	1.006495274
165	40560	115059	4261	0.00016188	0.95979186
166	273197	553938	1643	0.00029483	0.949229084
167	107667	417329	1806	0.00069856	0.946136036
168	111883	395014	1908	0.00065043	0.920580684
169	85508	37216	1772	0.00246124	0.917701722
170	3017	634908	1213	0.00023427	0.905185352
171	541954	215658	413	0.00275960	0.886321492
172	616896	58624	2442	0.00166935	0.885513749
173	2402	234337	2519	0.00070952	0.862226236
174	51231	306058	2914	0.00010194	0.852157511
175	504218	101546	3760	0.00026240	0.819102243
176	465938	8953	4476	0.00017946	0.81019422
177	100463	346486	2038	0.00043572	0.786809622
178	267	391986	1875	0.00034611	0.772544391
179	105489	525205	1065	0.00044876	0.764871945
180	2055	286277	1974	0.00067962	0.763235507
181	49924	564841	1260	0.00011458	0.750157396
182	201176	399094	2046	0.00010213	0.732579407
183	47157	89737	3589	0.00021444	0.723893644
184	40168	407126	1296	0.00056788	0.700118406
185	277237	357102	1266	0.00079629	0.698159966
186	43242	190846	2936	0.00015807	0.684910543
187	232802	329624	633	0.00141954	0.674306914
188	234554	517237	1286	0.00008550	0.671714658
189	585031	91234	2225	0.00111160	0.668304196
190	575926	241833	562	0.00185702	0.661715967
191	54593	270118	854	0.00143243	0.656743434
192	228000	576755	924	0.00010439	0.654388765
193	216405	41307	3755	0.00014301	0.651658148
194	268546	15406	3851	0.00018345	0.645160208
195	66009	434	48	0.00365607	0.641200928
196	241229	520062	717	0.00043110	0.63463183
197	203132	252679	1093	0.00120721	0.62639142
198	64458	488930	929	0.00034747	0.624125847
199	535993	39573	719	0.00260137	0.618185399
200	44406	72530	3296	0.00021449	0.610020862

Referências

Anagnostopoulos, A. e Kumar, R. e. M. M. (2008). Influence and correlation in social networks. In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 7–15. ACM. 6

Brin, S., P. L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, 30(1-7):107–117. 40

Cho, J., Garcia-Molina, H., and Page, L. (1998). Efficient crawling through url ordering. *Computer Networks and ISDN Systems*, 30(1-7):161–172. 42

Crandall, D., Cosley, D., Huttenlocher, D., Kleinberg, J., and Suri, S. (2008). Feedback effects between similarity and social influence in online communities. In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 160–168. ACM. 6

Facebook (2012). Facebook. `http://www.facebook.com`. 2

Famecount (2012). Famecount. `http://www.famecount.com`. 8

Flamengo (2012). Flamengo. `http://www.flamengo.com.br`. 49

Franceschet, M. (2011). Pagerank: Standing on the shoulders of giants. *Communications of the ACM*, 54(6):92–101. 15

Garton, L., Haythornthwaite, C., and Wellman, B. (1997). Studying online social networks. *Journal of Computer-Mediated Communication*, 3(1):0–0. 6

Gill, K. (2004). How can we measure the influence of the blogosphere. In *WWW 2004 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics*. Citeseer. 6

Goyal, A., Bonchi, F., and Lakshmanan, L. (2010). Learning influence probabilities in social networks. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 241–250. ACM. 6

Hartline, J. Mirrokni, V., and Sundararajan, M. (2008). Optimal marketing strategies over social networks. In *Proceeding of the 17th international conference on World Wide Web*, pages 189–198. ACM. 6

Jamali, M. e Abolhassani, H. (2006). Different aspects of social network analysis. In *Web Intelligence, 2006. WI 2006. IEEE/WIC/ACM International Conference on*, pages 66–72. Ieee. 6

Katona, Z., Zubcsek, P., and Sarvary, M. (2011). Network effects and personal influences: The diffusion of an online social network. *Journal of Marketing Research*, 48(3):425–443. 6

Kempe, D., Kleinberg, J., and Tardos, É. (2003). Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 137–146. ACM. 6

Kleinberg, J. (2007). Challenges in mining social network data: processes, privacy, and paradoxes. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 4–5. ACM. 6

Kobayashi, M. e Takeda, K. (2000). Information retrieval on the web. *ACM Computing Surveys (CSUR)*, 32(2):144–173. 40

Kwak, H., Lee, C., Park, H., and Moon, S. (2010). What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600. ACM. 49

Leavitt, A., Burchard, E., Fisher, D., and Gilbert, S. (2009). The influentials: New approaches for analyzing influence on twitter. *Web Ecology Project, http://tinyurl.com/lzjlzq*. 29

Meyn, S., Tweedie, R., and Glynn, P. (2009). *Markov chains and stochastic stability*, volume 2. Cambridge University Press Cambridge. 13

Mislove, A., Marcon, M., Gummadi, K., Druschel, P., and Bhattacharjee, B. (2007). Measurement and analysis of online social networks. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, pages 29–42. ACM. 6

Page, L., Brin, S., Motwani, R., and Winograd, T. (1998). The pagerank citation ranking: Bringing order to the web. viii, 3, 14, 15, 19, 33

Pankin, M. (1987). Markov chain models: Theoretical background. *Retrieved October*, 22:2008. 13

Saaty, T. (1990). How to make a decision: the analytic hierarchy process. *European journal of operational research*, 48(1):9–26. 62

ScienceNet (2012). Sciencenet. `http://www.sciencenet.cn`. 56

Senecal, Nantel, J. (2004). The influence of online product recommendations on consumers' online choices. *Journal of Retailing*, 80(2):159–169. 6

Shannon, C. (1949). Communication theory of secrecy systems. *Bell system technical journal*, 28(4):656–715. 3

Shannon, C. (2001). A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, 5(1):3–55. 3, 20

Tang, J. e Sun, J. e. W. C. e. Y. Z. (2009). Social influence analysis in large-scale networks. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 807–816. ACM. 6

Trusov, M., Bodapati, A., and Bucklin, R. (2010). Determining influential users in internet social networks. *Journal of Marketing Research*, 47(4):643–658. 6

Twitter (2012). Twitter. `http://twitter.com`. 2

Usatenko, O. (2009). *Random Finite-valued Dynamical Systems: Additive Markov Chain Approach*. Cambridge Scientific Publishers. 13

Weigang, L., Jianya, Z., and Daniel, L. (2011a). Analysis of w-entropy index: the impact of members on social networks. In *The IADIS International Conference WWW/INTERNET, Rio de janeiro, Brazil, November 6-7 , 2011, Proceedings*, pages 171–178. viii, 36

Weigang, L., Jianya, Z., and Daniel, L. (2011b). W-index the impact of members on social networks. In *Web Information Systems and Mining: International Conference, Wism 2011, Taiyuan, China, September 24-25, 2011, Proceedings*, pages 226–233. 34

Weng, J., Lim, E., Jiang, J., and He, Q. (2010). Twitterrank: finding topic-sensitive influential twitterers. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 261–270. ACM. 48