Métricas de análise de links e qualidade de conteúdo: um estudo de caso na Wikipédia

Raíza Tamae Sarkis Hanada

Métricas de análise de links e qualidade de conteúdo: um estudo de caso na Wikipédia

Raíza Tamae Sarkis Hanada

Orientadora: **Profa. Dra. Maria da Graça Campos Pimentel**

Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação - ICMC-USP, como parte dos requisitos para obtenção do título de Mestre em Ciências - Ciências de Computação e Matemática Computacional. *VERSÃO REVISADA*

USP – São Carlos
Junho de 2013

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados fornecidos pelo(a) autor(a)

Hanada, Raíza
H233m Métricas de análise de links e qualidade de conteúdo: um estudo de caso na Wikipédia / Raíza Hanada; orientadora Maria da Graça Pimentel. -- São Carlos, 2013.
67 p.

Dissertação (Mestrado - Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2013.

1. Recuperação de Informação. 2. Análise de Links. 3. Wikipédia. I. Pimentel, Maria da Graça, orient. II. Título.

Agradecimentos

Agradeço à minha família e aos meus amigos por todo o apoio dado nessa etapa da minha vida. Obrigada por, mesmo longe, permanecerem tão perto de mim.

Agradeço ao ICMC-USP por ter me acolhido e fornecido um ambiente propício para o desenvolvimento desta pesquisa, e à professora Maria da Graça Campos Pimentel pela orientação dada. A infra-estutura da instiuição e a orientação recebida se tornaram fundamental para meu crescimento como acadêmica.

Agradeço a colaboração da Universidade Federal do Amazonas, em particular do professor Marco Antônio Pnheiro de Cristo e dos alunos do grupo Banco de Dados e Recuperação de Informação. A parceria estabelecida foi essencial para o desenvolvimento desta pesquisa. Da mesma forma, quero agradecer o InWeb, da Universidade Federal de Minas Gerais, pelo fornecimento de seus dados.

Agradeço à FAPEAM pelos recursos fornecidos. Os recursos foram importantes para que a minha permanência na cidade de São Carlos, realizando este trabalho o qual tem, certamente, um imenso valor para o desenvolvimento da ciência.

Por fim, agradeço a Deus por ter me guiado em todos os momentos.

Resumo

Muitos links entre páginas na Web podem ser vistos como indicadores de qualidade e importância para as páginas que eles apontam. A partir desta ideia, vários estudos propuseram métricas baseadas na estrutura de links para inferir qualidade de conteúdo em páginas da web. Contudo, até onde sabemos, o único trabalho que examinou a correlação entre tais métricas e qualidade de conteúdo consistiu de um estudo limitado que deixou várias questões em aberto. Embora tais métricas sejam muito bem sucedidas na tarefa de ranquear páginas que foram fornecidas como respostas para consultas submetidas para máquinas de busca, não é possível determinar a contribuição específica de fatores como qualidade, popularidade e importância para os resultados. Esta dificuldade se deve em parte ao fato de que a informação sobre qualidade, popularidade e importância é difícil de obter para páginas da web em geral. Ao contrário de páginas da web, estas informações podem ser obtidas para artigos da Wikipédia, uma vez que qualidade e importância são avaliadas por especialistas humanos, enquanto a popularidade pode ser estimada com base nas visualizações dos artigos. Isso torna possível a verificação da relação existente entre estes fatores e métricas de análise de links, nosso objetivo neste trabalho. Para fazer isto, nós implementamos vários algoritmos de análise de links e comparamos os rankings obtidos com eles com os obtidos considerando a avaliação humana feita na Wikipédia com relação aos fatores qualidade, popularidade e importância. Nós observamos que métricas de análise de links são mais relacionadas com qualidade e popularidade que com importância e a correlação é moderada.

Palavras-chaves: Recuperação de Informação, Análise de Links, Wikipédia.

Abstract

Many links between Web pages can be viewed as indicative of the quality and importance of the pages pointed to. Accordingly, several studies have proposed metrics based on links to infer web page content quality. However, as far as we know, the only work that has examined the correlation between such metrics and content quality consisted of a limited study that left many open questions. In spite of these metrics having been shown successful in the task of ranking pages which were provided as answers to queries submitted to search machines, it is not possible to determine the specific contribution of factors such as quality, popularity, and importance to the results. This difficulty is partially due to the fact that such information is hard to obtain for Web pages in general. Unlike ordinary Web pages, the content quality of Wikipedia articles is evaluated by human experts, which makes it feasible to verify the relation between such link analysis metrics and the quality of Wikipedia articles, our goal in this work. To accomplish that, we implemented several link analysis algorithms and compared their resulting rankings with the ones created by human evaluators regarding factors such as quality, popularity and importance. We found that the metrics are more correlated to quality and popularity than to importance, and the correlation is moderate.

Keywords: Information Retrieval, Link Analysis, Wikipedia.

Sumário

Glossário de Termos

Acurácia: Métrica usada para avaliar desempenho em Aprendizado de Máquina. Consiste na razão entre as previsões corretas e o total de previsões realizado. [Bishop, 2006] (pág. 33).

Aprendizagem de Máquina: Sub-campo da inteligência artificial dedicado ao desenvolvimento de algoritmos e técnicas para detecção de padrões em dados e uso destes padrões para definição de comportamento futuro, em geral, com intuito de aperfeiçoar o desempenho em uma tarefa. [Bishop, 2006] (pág. 33).

Autoridade: Especialista em certo tópico. No contexto de Análise de Links, é o nome dado à propriedade de uma página ser citada por várias outras páginas conhecidas por terem a propriedade de serem bons *hubs* (catálogos). [Kleinberg and Lawrence, 2001] (pág. 1).

Análise de Links: Técnica de análise de dados usada para avaliar relacionamentos entre nós de uma rede. [Baeza-Yates and Ribeiro-Neto, 2011] (pág. 7).

Bibliometria: Campo da Ciência da Informação que aplica métodos estatísticos e matemáticos para quantificar processos de comunidação escrita. No contexto deste trabalho, estes processos se referem principalmente à estimativa de certas propriedades de um documento a partir de referências feitas por este documento ou para este documento. [Baeza-Yates and Ribeiro-Neto, 2011] (pág. 7).

Citações: Menção de uma informação extraída de outra fonte. [Smith, 2004] (pág. 8).

Classificação: Tarefa de identificar a que conjunto de categorias (sub-população) pertencente uma nova observação, com base em um conjunto de dados

de treino contendo observações cujas categorias são conhecidas. [Bishop, 2006] (pág. 33).

Correlação Estatística: Grau de correspondência entre duas variáveis. É positiva ou direta se as variáveis mudam na mesma direção. É negativa ou indireta quando elas mudam em direções opostas. [Bishop, 2006] (pág. 3).

Domínio: Nome que representa um recurso IP. No contexto deste trabalho, este recurso refere-se normalmente a um Web site. Por exemplo, na URL `jogos.uol.com.br`, o domínio é o site `uol.com.br`. [Berlt et al., 2010] (pág. 14).

Dump: Registro da estrutura e dados de banco de dados normalmente na forma de instruções em SQL. Neste trabalho, se refere à forma como cópias da Wikipédia são disponibilizadas na Web. (pág. 39).

Hipergrafo: Generalização de um grafo, em que as arestas ligam qualquer número positivo de vértices. [Berlt et al., 2010] (pág. 14).

Hub: Catálogo. No contexto de Análise de Links, é o nome dado à propriedade de uma página citar várias outras páginas conhecidas por terem a propriedade de serem boas autoridades. [Kleinberg and Lawrence, 2001] (pág. 31).

Host: Nome que representa normalmente um computador dentro de um domínio. Por exemplo, na URL `jogos.uol.com.br`, o host é representado pelo nome `jogos.uol.com.br`. [Berlt et al., 2010] (pág. 14).

Importância: Qualidade do que tem muito valor ou interesse. (pág. 1).

Inlink: Qualquer link recebido por um nó Web de outro nó Web. (pág. 8).

Lei de potência: Relação matemática entre duas quantidades na qual a primeira quantidade varia como uma potência da segunda. [Bishop, 2006] (pág. 43).

Link: Referência para dados que o usuário pode seguir diretamente ou que pode ser seguido automaticamente. A referência é feita para o conteúdo completo de um nó Web ou para um elemento específico do nó. (pág. 1).

Links informacionais: Links para outras páginas que se julga importante que os usuários consultem, em razão de seu conteúdo. [Bishop, 2006] (pág. 14).

Links navegacionais: Links criados com o intuito de possibilitar aos usuários saltarem entre seções de um site. [Bishop, 2006] (pág. 14).

Maquinas de Busca: Sistema de software projetado para se buscar informações na Web. [Baeza-Yates and Ribeiro-Neto, 2011] (pág. 1).

Outlink: Qualquer link apontando de um nó Web para outro nó Web. (pág. 8).

Popularidade: Propriedade ou condição de ser amplamente admirado ou suportado. No contexto deste trabalho, uma página é considerada popular se seu conteúdo é acessado por muitas pessoas. (pág. 1).

Qualidade: Propriedade ou condição de uma coisa ou pessoa pela qual se distingue de outras; grau de conformidade com um padrão e adequação de uso. (pág. 1).

Ranking: Relacionamento entre um conjunto de itens tal que, para dois itens quaisquer, o primeiro é posicionado acima, abaixo ou no mesmo nível do segundo. [Baeza-Yates and Ribeiro-Neto, 2011] (pág. 1).

Recuperação de Informação: Área da Computação que lida com o armazenamento de documentos e a recuperação automática de informação associada a eles. [Baeza-Yates and Ribeiro-Neto, 2011] (pág. 7).

Regressão: Técnica estatística que permite explorar e inferir a relação de uma ou mais variáveis dependentes com um conjunto de variáveis independentes específicas. [Bishop, 2006] (pág. 61).

Reputação: Crenças ou opiniões que se têm sobre alguém ou algo. (pág. 1).

Spam de Links: Uso de links entre páginas com o intuito de manipular deliberadamente a relevância ou proeminência dos recursos indexados de forma inconsistente com os objetivos do sistema de busca. [Baeza-Yates and Ribeiro-Neto, 2011] (pág. 14).

Visibilidade: Condição do que é percebido sem grande esforço. (pág. 1).

Wikipedista: Editor de artigo na Wikipedia. (pág. 21).

Wikiprojeto: Conjunto de páginas destinado à gestão de um tema ou família de temas dentro da Wikipédia, assim como o grupo de editores que usa essas páginas para colaborar no trabalho enciclopédico. (pág. 21).

Lista de Abreviaturas

HITS *Hyperlink-Induced Topic Search*

HTML *HyperText Markup Language*

ICMC Instituto de Ciências Matemáticas e de Computação

RI *Recuperação de Informação*

INWEB Instituto Nacional de Ciência e Tecnologia para WEB

RFC *Request for Comments*

SQL *Structured Query Language*

TREC *Text REtrieval Conference*

USP Universidade de São Paulo

URL *Uniform Resource Locator*

XML *Extensible Markup Language*

Lista de Figuras

Lista de Tabelas

Capítulo

1

Introdução

1.1 Contextualização

A Web se expandiu ao longo das últimas décadas para se tornar um enorme repositório de documentos, no qual informações são constantemente modificadas, inseridas e removidas de forma descentralizada e desordenada. Apesar disso, foram desenvolvidas com sucesso ferramentas capazes de encontrar informações específicas nesse repositório. Essas ferramentas são conhecidas como máquinas de busca.

Máquinas de busca permitem que usuários forneçam um conjunto de termos por meio de consultas, representando sua necessidade de informação, e encontrem páginas a princípio relevantes de acordo com os termos dados. Um algoritmo de *ranking* é responsável por atribuir uma nota de relevância às páginas e ordená-las de forma que as mais relevantes fiquem no topo e, portanto, sejam as mais visíveis. Esses valores são estimados agregando várias evidências, entre as quais algumas que exploram a estrutura de links da Web. Para essas, em particular, a Web é vista como um grafo direcionado, no qual os nós representam as páginas e as arestas representam os *hyperlinks*. O estudo das propriedades e das relações que podem ser inferidas a partir desse grafo é chamado Análise de Links.

De acordo com uma intuição comum em Análise de Links, quando um autor de uma página cria um link para outra página, ele endossa o conteúdo da página apontada, sugerindo que ela é de boa *qualidade* e pode ser uma autoridade em um determinado assunto [Kleinberg, 1999]. Neste trabalho, nos referimos a essa intuição como "Hipótese de Qualidade". Segundo essa hipótese, a quali-

dade do conteúdo de uma página pode ser inferida por quantas e quais páginas a citam ou, de forma equivalente, pela *reputação* que ela possui. Essa hipótese contribui para a ideia de que, quanto mais uma página é apontada por outras, maior é a probabilidade de ela ser preferida por um usuário dentre diversas outras páginas alternativas sobre o mesmo tópico. Como consequência, páginas que são muito referenciadas podem ser melhor ranqueadas por algoritmos que procuram respostas para consultas submetidas à máquina de busca. Da mesma forma, páginas novas ou menos populares provavelmente terão menor *visibilidade* ainda que seu conteúdo seja muito relevante para um certo tópico de interesse.

Na literatura, são reportadas diversas métricas para ordenar páginas que fazem uso das ideias de Análise de Links [Baeza-Yates et al., 2006]. Os resultados desses métodos geralmente são avaliados por especialistas humanos que verificam o quão relevante as páginas retornadas são para o conjunto de consultas fornecido. Embora essas estratégias de avaliação confirmem que Análise de Links é útil para ranquear páginas de acordo com a sua relevância, elas não mostram que a "Hipótese da Qualidade" é verdadeira. Em outras palavras, como questionam Amento et al. [2000], "métricas de análise de links são úteis pra estimar qualidade de conteúdo?". A grande dificuldade em responder essa pergunta reside na necessidade de se conhecer, para uma página Web, a sua qualidade, a qual é uma informação muito difícil de obter. Como resultado e, de acordo com a pesquisa bibliográfica realizada, essa hipótese não foi verificada a contento até o momento.

Essa verificação é mais difícil ainda quando se observa que a qualidade não é o único fator que contribui para que uma página seja citada. Outros fatores, investigados na literatura, são a sua *importância*, a sua *visibilidade* e a sua *popularidade* [Cho and Roy, 2004]. Esses três fatores, somados a qualidade e reputação, constituem conceitos centrais para este trabalho e merecem uma discussão mais cuidadosa.

Dado um conjunto de opções alternativas, o termo qualidade se refere a um nível de excelência que, quando atribuído a uma das opções, permite distingui-la das demais. No contexto da Web, a qualidade de uma página está associada a aspectos tão diversos quanto que informação é fornecida, por quem, qual a sua utilidade para os usuários, como ela é apresentada (em termos de usabilidade da página), quão acessível ela é em termos de clareza e objetividade, etc. Contudo, a percepção de qualidade é subjetiva, o que implica que diferentes usuários podem atribuir diferentes graus de qualidade a uma mesma página. Nesse sentido, seria mais correto dizer que, além de boa qualidade, uma página muito citada possui boa reputação, uma vez que reputação se refere à opinião que o público tem

sobre algo. Um conceito correlato ao de reputação é o de popularidade. Uma página popular é aquela que atrai muitos usuários, tais como as páginas de entrada de grandes portais de notícia ou páginas com conteúdo de grande apelo. Pode-se dizer que páginas populares são páginas que possuem boa reputação perante uma grande audiência. Por sua vez, páginas de visibilidade destacada são aquelas que podem ser notadas pelos usuários sem muito esforço, como as vinculadas a um evento incomum, como uma tragédia. Finalmente, páginas importantes são aquelas às quais se atribui grande valor ou interesse, como páginas de serviços providos pela Receita Federal.

Embora distintos, todos esses fatores estão relacionados entre si, uma vez que um conteúdo de maior qualidade pode se tornar mais popular e, portanto, mais visível. Da mesma forma, uma página pode ganhar grande visibilidade mediante uma campanha publicitária e, então, se tornar popular. Uma página pode se tornar popular também, independente de sua qualidade, por ser foco de um serviço do governo que atinge um grande número de cidadãos e, como resultado da sua necessidade, ganhar qualidade.

1.2 Motivação

Para páginas Web comuns é difícil determinar de forma confiável e independente sua qualidade, importância, popularidade, reputação e visibilidade. Na Wikipédia[1], porém, artigos são explicitamente avaliados por sua comunidade, pelo menos em relação a fatores como qualidade e importância. Além disso, a comunidade mantém estatísticas detalhadas sobre as visitas que seus artigos recebem ao longo do tempo, o que pode ser usado para estimar popularidade.

A disponibilidade dessas informações permite realizar um estudo em larga escala sobre qual a correlação entre as métricas de análise de links e os fatores qualidade, importância e popularidade. De acordo com uma pesquisa bibliográfica realizada, o único estudo que analisou tal correlação foi bastante limitado ao verificar apenas qualidade e popularidade, usando um número reduzido de páginas relacionadas com poucos tópicos e avaliadas por poucos especialistas humanos [Amento et al., 2000]. Tal limitação motivou seus autores a afirmarem que um estudo em maior escala seria necessário para obter resultados mais conclusivos.

Com este trabalho, esperamos compreender melhor a natureza das métricas de análise de links, de modo a fundamentar investigações que levam a propostas de novas métricas para estimativa automática de qualidade de conteúdo e de ranqueamento de páginas. Essas são áreas promissoras, dado seu impacto em várias outras aplicações como busca [Bendersky et al., 2011; Vadrevu and Veli-

[1] http://pt.wikipedia.org

pasaoglu, 2011] e recomendação [Surdeanu et al., 2011; Suryanto et al., 2009]. Além disso, a avaliação automática de qualidade é vantajosa para a própria Wikipédia, uma vez que a classificação de qualidade de seus artigos ainda é um processo essencialmente manual. Embora muitos trabalhos tenham investigado esse problema [Dondio and Barrett, 2007; Lim et al., 2006; Stvilia et al., 2005; Wang and Iwaihara, 2011; Wöhner and Peters, 2009], não são reportados estudos que consideram métricas baseadas em links externos à Wikipédia, uma outra contribuição deste trabalho. Finalmente, as conclusões obtidas para Wikipédia podem, em alguma medida, ser úteis pra Web em geral, uma vez que comprovadas semelhanças entre aspectos em suas estruturas de links.

1.3 Formulação do Problema

Embora haja na literatura vários trabalhos que sugerem que métricas de análise de links medem fatores como qualidade [Amento et al., 2000; Berlt et al., 2010; Kleinberg, 1999], reputação [Berlt et al., 2010; Page et al., 1999], importância [Amento et al., 2000], popularidade [Amento et al., 2000] e visibilidade [Borodin et al., 2005; Kleinberg and Lawrence, 2001; Lempel and Moran, 2001], nenhum deles verificou tais afirmações de forma direta. Isso se deve aos fatos de que (1) a maioria desses trabalhos procura melhorar os resultados da tarefa de ranking e essas métricas, independentemente do que realmente estimam, são úteis nessa tarefa; e (2) não existem meios simples de se obter, de forma confiável e em larga escala, estimativas desses fatores.

Não é possível garantir que uma página é muito citada porque tem conteúdo de qualidade. Para aplicações diretamente interessadas na qualidade do conteúdo ou em estimativas de sua popularidade e importância, isso pode ser um problema. Além disso, uma melhor compreensão dessas métricas poderia facilitar a criação de métodos melhores e mais robustos. Fica clara a necessidade da compreensão de qual é a relação existente entre ligações por meio de links e fatores como qualidade, importância, popularidade, reputação e visibilidade. Dentre esses fatores, qualidade, importância e popularidade são medidos de forma independente na Wikipédia. Logo, considerando o cenário exposto, pretendemos investigar qual a correlação entre esses fatores e métricas de análise de links na Wikipédia.

1.4 Perguntas da Pesquisa

Considerando que estamos interessados nos fatores importância, popularidade e qualidade do conteúdo, este trabalho pretende responder as seguintes perguntas:

- Como esses fatores se relacionam entre si?

- Na Wikipédia, qual a relação entre esses fatores e as métricas de Análise de Links?
- Considerando cada fator, quais as métricas mais adequadas para medi-los?
- Em quais pontos os resultados obtidos durante a pesquisa divergem dos reportados na literatura?
- Em que medida as conclusões obtidas para Wikipédia podem ser estendidas à Web em geral?

1.5 Objetivos

Este trabalho tem como objetivo principal investigar qual é a relação existente entre métricas de análise de links e os fatores qualidade, popularidade e importância de artigos da Wikipédia.

Os objetivos específicos deste trabalho são:

1. Obter coleções de dados com informações necessárias para investigar as relações entre os fatores de interesse e métricas de análise de links. Uma das bases deve ser representativa da Web, externa à Wikipédia, porém conectada a esta. A segunda base deve conter artigos da Wikipédia com, pelo menos, informações relativas a sua qualidade de conteúdo, popularidade, importância e links para a coleção externa;
2. Implementar diferentes métricas de análise de links;
3. Verificar a correlação estatística entre as diferentes métricas de análise de links e os fatores qualidade, popularidade e importâcia nos artigos da Wikipédia. Discutir em que medida esses resultados são válidos para Web em geral;
4. Discutir que métricas são mais úteis para a Wikipédia em termos de previsão dos fatores qualidade, importância e popularidade de um artigo.

1.6 Metodologia

Para a realização desta pesquisa, foi importante ter conhecimento acerca de quais trabalhos existentes na literatura podiam contribuir para o seu desenvolvimento. Esse conhecimento pôde ser obtido por meio de um levantamento bibliográfico sobre o estado da arte do problema em questão.

Também foi necessário ter conhecimento a respeito de quais dados de interesse poderiam ser extraídos para o desenvolvimento da pesquisa e quais métodos de comparação e de avaliação seriam utilizados. Os dados utilizados foram

qualidade do conteúdo, importância relativa e popularidade dos artigos da Wikipédia. As notas de qualidade e importância dos artigos utilizada estão de acordo com os critérios estabelecidos pela própria comunidade da Wikipédia, expostos na Seção 2.4. A Seção 4.2 apresenta a maneira como os dados foram capturados. As métricas de análise de links adotadas para esta pesquisa foram Indegree, Outdegree, PageRank e variações desses métodos. Explicações sobre os métodos utilizados são fornecidas na Seção 2.2.

Para a extração do grafo de links, foram usadas coleções representativas da Web brasileira e da Wikipédia. O tamanho das coleções e a grande quantidade de dados nelas contidos tornaram este trabalho desafiador. As métricas de análise de links foram aplicadas sobre o grafo obtido. As coleções são comparadas e descritas no Capítulo 4.

Uma vez que os dados de interesse foram extraídos, foi possível realizar a comparação desses dados com os resultados obtidos pelas métricas de análise de links. Métricas de comparação de rankings foram utilizadas na comparação dos métodos.

1.7 Resultados

A partir dos experimentos executados, foi possível observar que métricas de análise de links são mais correlacionadas com popularidade do que com qualidade e importância. Em diversos casos, métricas simples como Indegree (mais frequente) e Outdegree (apenas para qualidade dos artigos) apresentaram desempenhos comparáveis ao PageRank, métrica mais complexa. Para importância e popularidade, as variações de hipergrafo apresentaram resultados satisfatórios, podendo ser boas opções para medir esses fatores, dado que necessita-se de pouco processamento para essas métricas. Detalhes das conclusões alcançadas são apresentados no Capítulo 6.

1.8 Estrutura da dissertação

Esta dissertação está organizada nos seguintes capítulos: o Capítulo 2 apresenta conceitos relevantes para a pesquisa, como métricas de análise de links e o processo de atribuição de notas de qualidade e importância da Wikipédia; o Capítulo 3 descreve trabalhos relacionados a esta pesquisa, bem como trabalhos usados como guias para os procedimentos observados durante a resolução do problema; o Capítulo 4 descreve e compara as coleções obtidas; o Capítulo 5 mostra os experimentos realizados e os resultados obtidos; e, por fim, no Capítulo 6, são apresentadas as conclusões obtidas e direções futuras desta pesquisa.

Capítulo

2

Conceitos relacionados

Neste capítulo, são apresentadas informações básicas necessárias para a compreensão do trabalho realizado. Em particular, são discutidos os conceitos de análise de links, são sumarizadas métricas de análise de links implementadas neste trabalho e métricas de comparação de ranking e são apresentados os procedimentos para atribuição de notas de qualidade e de importância a artigos da Wikipédia.

2.1 Análise de links

Conforme discutido previamente, algoritmos de ranking usados por máquinas de busca se beneficiam de diversas estratégias para inferir relevância das páginas. Uma dessas estratégias, chamada de Análise de Links, consiste na análise do grafo da Web para estimar a importância global de cada página. A principal intuição é de que um link de uma página p_1 para p_2 pode representar um voto para a reputação de p_2, que corresponde a dizer que p_2 tem conteúdo com qualidade (Hipótese da Qualidade).

A análise de links surgiu como uma sub-área da área de Recuperação de Informação (RI). O principal objetivo de trabalhos da área de Recuperação de Informação é investigar mecanismos que permitam realizar uma busca automática de informações relevantes a uma determinada consulta. Na Web, o conteúdo descentralizado e desordenado da grande quantidade de documentos representa um grande desafio para a área de RI [Manning et al., 2008]. A análise de links surgiu da necessidade de encontrar novas evidências para recuperar documentos relevantes. Até então, as abordagens mais conhecidas consideravam basicamente os termos existentes nas páginas. Um dos grandes problemas encontra-

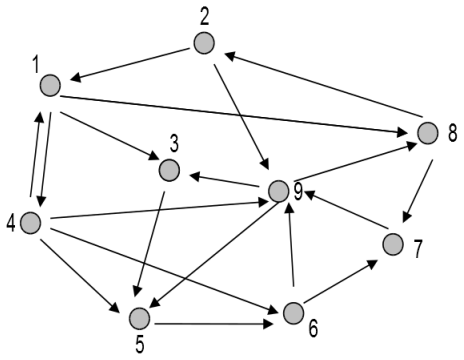

Figura 2.1: *Exemplo de uma representação de páginas Web em grafos*

dos com essas abordagens é que páginas nem sempre possuem conteúdo textual ou, quando possuem, nem sempre usam os mesmos termos das consultas para designar seus conteúdos [Manning et al., 2008]. Kleinberg [1999] mostrou que páginas interligadas podem possuir conteúdos semelhantes (localidade por tópico). Essa ideia se mostrou valiosa para RI, o que estimulou várias pesquisas. Grande parte dessas pesquisas têm o objetivo de avaliar a qualidade do conteúdo de uma página, sendo boa parte dos esforços utilizados para determinar as páginas mais relevantes para uma determinada consulta declarada por um usuário [Henzinger, 2000].

Na análise de links tradicional, a Web é modelada como um grafo direcionado $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, onde $\mathcal{V}$ é o conjunto de páginas e $\mathcal{E}$ é o conjunto de links. Existem $\mathcal{N}$ páginas em $\mathcal{V}$, sendo que cada uma delas é representada pelo nó v_a, onde a é o seu índice no grafo. Cada link de $\mathcal{E}$ é representado por $e_a b$, que indica que a página v_a aponta para a página v_b. A Figura 2.1 ilustra um exemplo dessa estrutura. Links de páginas de $\mathcal{V}$ que apontam para uma determinada página v_a são chamados de *inlinks* de v_a. Os *outlinks* da página v_a, por sua vez, são as páginas de $\mathcal{V}$ para as quais v_a aponta. Dessa forma, podemos dizer que na Figura 2.1 os inlinks da página v_1 são v_2 e v_4, enquanto seus outlinks são v_4, v_3 e v_8. Representando o conjunto de inlinks de uma página v_a por $i(v_a)$ e o conjunto de outlinks por $o(v_a)$, temos que $i(v_1) = \{v_2, v_4\}$ e $o(v_1) = \{v_3, v_4, v_8\}$. Quando um usuário sai de uma página para outra dizemos que ele está "navegando" de uma página para outra.

Smith [2004] avaliou quais motivos levam autores de páginas a criar links para outras páginas a fim de analisar qual a diferença existente entre os links Web e as citações. Em sua pesquisa, ele concluiu que os motivos que levam autores a criarem links são bem mais variados do que os motivos que levam pessoas a fazerem citações. Segundo ele, há oito motivos para fazer links na Web: download de arquivos, publicidade, informações sobre áreas geográficas,

páginas relacionadas, informação sobre o criador da página, reconhecimento do suporte, citação formal de um estudo e informações gerais sobre o link. Citações e links mostraram-se semelhantes quando o motivo para fazer um link refere-se a pesquisas, como e-journals, trabalhos técnicos, reportagens e publicações em conferência, o que totalizou apenas 20% dos links das páginas na Web investigadas no estudo. É possível que, devido a sua natureza enciclopédica, a Wikipédia faça parte dessa pequena porcentagem.

2.1.1 Vulnerabilidades encontradas em análise de links

Para pessoas que viram na Web uma oportunidade para ampliar seus negócios e, principalmente, fazer publicidade, surgiu uma questão importante: "Como tornar suas páginas e anúncios mais visíveis ao público?". Em sistemas de busca, quando um usuário faz uma pesquisa por um termo, páginas relevantes a esse termo são retornadas de acordo com uma pontuação atribuída pelo mecanismo de busca, de forma que as págnas consideradas mais relevantes aparecem no topo do ranking de respostas. Dada a importância da estruturas de links para a formação do ranking, muitas empresas se sentiram incentivadas a simular padrões de links de páginas que fazem com que seus conteúdos pareçam mais interessantes e/ou importantes do que são. Tais estratégias são conhecidas como *spam de links*.

Criadores de spam (*spammers*) utilizam três estratégias básicas para aumentar o ranking de suas páginas [Caverlee et al., 2007]. A primeira chama-se *hijacking*. Esse tipo de spam acontece quando spammers conseguem colocar em páginas confiáveis links para uma página que eles desejam que obtenha maior pontuação no ranking. Algumas formas de inserir esses links são por meio de *wikis* editáveis abertas ao público, *webblogs* reconhecidos ou em quadros de mensagens. Outra forma de fazer spams é conhecida como *honeypots*. Neste caso, há menos risco de exposição do spam, pois *spammers* criam sites de qualidade que apontam para páginas confiáveis, acumulando autoridade. Dentre as páginas apontadas, está a página de spam alvo que se mistura com as páginas de autoridade, o que torna difícil a sua detecção por sistemas de controle de spam. Por fim, outra forma existente de spam em links é a *collusion*. Esta acontece quando vários spammers constroem uma estrutura de links ao redor de páginas confiáveis, fazendo esforços coletivos para promoções das suas páginas mutuamente. Essas estratégias, quando modificadas ou combinadas, podem ser difíceis de identificar [Caverlee et al., 2007].

Outra vulnerabilidade das análises de links é a de que páginas populares acabam sendo beneficiadas ao serem pontuadas pelas funções de ranking, o que prejudica a pontuação de páginas novas ou menos conhecidas, ainda que

elas sejam mais relevantes. Isso acontece porque páginas que aparecem no topo do ranking atraem mais atenção do que as páginas que estão mais abaixo e, por isso, tornam-se conhecidas. Dessa forma, é mais comum que criadores de novos sites façam links para páginas populares do que para as demais, o que aumenta ainda mais a sua pontuação no ranking. Páginas novas ou desconhecidas, por não serem populares, são vistas por poucas pessoas e, por isso, recebem poucos links, diminuindo ainda mais a sua pontuação no ranking. Essa situação torna-se injusta, uma vez que páginas de qualidade podem ser ignoradas somente pelo fato de não terem tido a chance de serem reconhecidas por outras pessoas [Cho and Roy, 2004].

2.2 Métricas de análise de links

Na área de Recuperação de Informação, a análise de links é utilizada principalmente para avaliar a relevância do conteúdo de uma página por meio de algoritmos de ranking, de acordo com termos de busca fornecidos pelos usuários, ordenando-as de acordo com a nota obtida, dado que cada página possui uma nota determinada pelo algoritmo. O ideal é que não haja o retorno de muitas páginas relacionadas ao termo, mas sim de poucas páginas com muita relevância para o usuário que realiza a busca.

Nesta seção, serão detalhadas as métricas de análise de links utilizadas em nosso estudo. Em particular, focamos nas métricas mais conhecidas, ou seja, Indegree, Outdegree e PageRank. Nas descrições fornecidas a seguir, será considerada a utilização do grafo original da Web, contendo todos os links e todas as páginas. Ao longo da pesquisa, outros grafos serão utilizados, resultando em variações desses métodos tradicionais. As variações dos métodos e os grafos originados serão descritas ao final desta seção.

2.2.1 Indegree e Outdegree

Indegree [Bray, 1996] é um método simples baseado em grafos, no qual é verificada a quantidade de páginas I que apontam para uma determinada página v_a (ou seja, a quantidade de inlinks de v_a). A quantidade de outlinks O de uma página v_a é chamada de Outdegree.

As fórmulas para Indegree, $In(v_a)$, e Outdegree, $Out(v_a)$, para uma página qualquer v_a são representadas a seguir:

$$In(v_a) = \sum_{q \in i(v_a)} 1. \tag{2.1}$$

$$Out(v_a) = \sum_{q \in o(v_a)} 1. \tag{2.2}$$

onde $i(v_a)$ é o conjunto com todas as páginas que apontam para v_a (inlinks de v_a), e $o(v_a)$ é o conjunto com todas as páginas para as quais v_a aponta (outlinks de v_a).

v_a	$i(v_a)$	$In(v_a)$	$o(v_a)$	$Out(v_a)$
v_1	$\{v_2, v_4\}$	2	$\{v_3, v_4, v_8\}$	3
v_2	$\{v_8\}$	1	$\{v_1, v_9\}$	2
v_3	$\{v_1, v_9\}$	2	$\{v_5\}$	1
v_4	$\{v_1\}$	1	$\{v_1, v_5, v_6, v_9\}$	4
v_5	$\{v_3, v_4, v_9\}$	3	$\{v_6\}$	1
v_6	$\{v_4, v_5\}$	2	$\{v_7, v_9\}$	2
v_7	$\{v_6, v_8\}$	2	$\{v_9\}$	1
v_8	$\{v_2,v_7\}$	2	$\{v_1, v_9\}$	2
v_9	$\{v_2,v_4,v_6,v_7\}$	4	$\{v_3,v_5,v_8\}$	3

Tabela 2.1: *Tabela com valores de Indegree e Outdegree para o exemplo da Figura 2.1*

A Tabela 2.1 exemplifica as métricas Indegree e Outdegree. Seja $\mathcal{V}$ um conjunto que possui todas as páginas de um grafo. Chamaremos de $\vec{i}$ o vetor composto por todas as páginas de $\mathcal{V}$ ordenadas pelos seus valores de Indegree, e de $\vec{o}$ o vetor composto por todas as páginas de $\mathcal{V}$ ordenadas pelos seus valores de Outdegree. Para o exemplo da Tabela 2.1, esses vetores assumem então os seguintes valores:

$$\vec{i} = (v_9, v_5, v_1, v_3, v_6, v_7, v_8, v_2, v_4)$$

$$\vec{o} = (v_4, v_1, v_9, v_2, v_6, v_8, v_3, v_5, v_6)$$

Na Tabela 2.1 é possível observar que o nó v_9 possui o seu valor de Indegree mais alto do que as demais, indicando que muitas páginas o consideram relevante. Juntamente com v_1 e v_4, v_9 também possui alto valor de Outdegree em relação aos outros nós, sugerindo que v_9 pode ser um polo de informações.

A grande vantagem desses métodos é a sua fácil implementação. A desvantagem é que esses métodos tratam todos os links com pesos iguais, desconsiderando qual papel que os links exercem no grafo e, portanto, não tratando diversos problemas existentes em análise de links.

2.2.2 PageRank

O método do PageRank foi proposto por Page et al. [1999], sendo amplamente utilizado na academia e conhecido por haver indícios de sua inserção no cálculo do Google[1] para ranqueamento de páginas [Henzinger, 2000].

As pontuações das páginas são computadas pelo algoritmo do PageRank por

[1] http://www.google.com

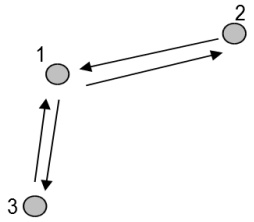

Figura 2.2: *Representação de uma estrutura de links.*

meio de modelos probabilísticos iterativos e é eficiente por não considerar todos os links das páginas com os mesmos pesos, como ocorre com Indegree e Outdegree. O peso é atribuído a cada link de acordo com a sua importância [Henzinger, 2000]. O método do PageRank se baseia na ideia de que um usuário, ao estar em um ambiente Web, pode navegar aleatoriamente entre as páginas apontadas pela sua página atual com uma determinada probabilidade entre eles. Assim, ele pode ir da página v_a para uma página qualquer v_b apontada por v_a através do link e_{ab}. À medida em que o usuário vai navegando pela Web, algumas páginas acabam sendo mais visitadas do que outras, aumentando o valor do seu PageRank. Páginas com grande probabilidade de visita possuem pontuações maiores do que as outras. Na Figura 2.2, considerando que o usuário está na página v_1, temos que a probabilidade de ele ir tanto à página v_2 quanto a v_3 é 0.5, enquanto que, se ele estivesse na página v_2 ou na v_3, a probabilidade de ele ir para a página v_1 seria 1.

Caso, durante a navegação, o algoritmo chegue em um ponto em que a página atual não aponta para nenhuma outra, há uma *operação de teletransporte*. Na operação de teletransporte, o algoritmo pula da página atual para outra qualquer no grafo, simulando um ato de um usuário digitar diretamente algum endereço na barra de URL do *browser*. Nessa operação, a probabilidade é a mesma para todas as páginas, ou seja, se há N páginas no grafo G cada página possuirá a probabilidade de $1/N$ de ser a nova página atual. A operação de teletransporte pode acontecer em outros momentos, simulando que um usuário se entediou e decidiu ir para outra página qualquer. A operação de teletransporte é determinada pelo *fator de damping*, cujo valor típico é 0.15 [Baeza-Yates and Ribeiro-Neto, 2011]. O fator de damping representa a probabilidade de um usuário continuar navegando ou decidir ir para outra página. O resultado final do PageRank de cada página varia de 0 a 1 e é obtido por diversas interações que são realizadas até obter a convergência.

A fórmula para calcular o PageRank de uma página v_a é a seguinte:

$$PR(v_a) = \frac{c}{N} + (1 - c) \times \sum_{q \in i(v_a)} \frac{PR(q)}{Out(q)} \tag{2.3}$$

onde c é o fator de damping, $i(v_a)$ é o conjunto de páginas que apontam para v_a (inlinks de v_a), $Out(q)$ é a quantidade de páginas apontadas por q (quantidade de outlinks de q), e N é a quantidade de páginas da coleção. É importante observar que o valor obtido pela função de ranking para cada página é normalizado pela sua quantidade de links [Baeza-Yates and Ribeiro-Neto, 2011].

Considerando o exemplo da Figura 2.2 e utilizando o fator de damping $c = 0.5$, ou seja, o usuário possui 50% de probabilidade de continuar navegando pelo site e 50% de probabilidade de ir para outra página, é possível obter o valor de PageRank de cada nó por meio de um sistema linear. Dessa forma, podemos obter o valor de PageRank de cada nó por meio dos seguintes cálculos:

$$PR(v_1) = \frac{0.5}{3} + 0.5 \times (PR(v_2) + PR(v_3))$$

$$PR(v_2) = \frac{0.5}{3} + 0.5 \times \frac{PR(v_1)}{2}$$

$$PR(v_3) = \frac{0.5}{3} + 0.5 \times \frac{PR(v_1)}{2}$$

$$PR(v_2) + PR(v_3) = \frac{0.5}{3} + 0.5 \times \frac{PR(v_1)}{2} + \frac{0.5}{3} + 0.5 \times \frac{PR(v_1)}{2} = \frac{1}{3} + \frac{PR(v_1)}{2}$$

Substituindo o valor de $PR(v_2) + PR(v_3)$ no cálculo de $PR(v_1)$ temos:

$$PR(v_1) = \frac{0.5}{3} + 0.5 \times (\frac{1}{3} + \frac{PR(v_1)}{2}) = \frac{1}{3} + \frac{PR(v_1)}{4} = \frac{4 + 3 \times PR(v_1)}{12}$$

$$12 \times PR(v_1) = 4 + 3 \times PR(v_1)$$

$$9 \times PR(v_1) = 4$$

$$PR(v_1) = \frac{4}{9}$$

Os valores de PageRank dos nós v_2 e v_3 podem ser obtidos substituindo o valor de $PR(v_1)$ em $PR(v_2)$ e $PR(v_3)$:

$$PR(v_2) = \frac{0.5}{3} + 0.5 \times \frac{4}{18} = \frac{1}{6} + \frac{4}{36} = \frac{5}{18}$$

$$PR(v_3) = \frac{0.5}{3} + 0.5 \times \frac{4}{18} = \frac{1}{6} + \frac{4}{36} = \frac{5}{18}$$

Assim, o valor de PageRank do nó v_1 possui o valor $4/9$, sendo o que possui maior probabilidade de ser acessado, e o do nó v_2 e o nó v_3 possuem o valor $5/18$, portanto com menor probabilidade.

2.2.3 Variantes dos métodos clássicos

Métricas muito simples podem não ser confiáveis o suficiente e podem deixar a máquina de busca suscetível a spams, como discutido na Seção 2.1.1. Todos os métodos clássicos de análise de links consideram que a Web é um grafo em que todas as arestas são importantes[2]. Como é relativamente fácil para spammers criarem padrões de links que fazem com que alguns nós se reforcem mutuamente, várias estratégias foram propostas para dificultar a inclusão de nós ou para relativizar a importância das arestas. Alguns desses trabalhos [Berlt et al., 2010; Bharat et al., 2001; Bray, 1996; Xue et al., 2005] estudaram adaptações das métricas baseadas em *hosts* (e.g.: esporte.uol.com.br) e/ou *domínios* (e.g.: uol.com.br) das páginas. Os autores desses trabalhos demonstraram que, em geral, métricas que consideram descarte de alguns links ou fazem agrupamentos de páginas são mais robustas do que aquelas que consideram todos eles. Essas variações das métricas são interessantes porque eliminam links com ruídos, como links navegacionais e links informacionais redundantes, bem como dificultam a criação de padrões de links que aumentam artificialmente a importância de uma certa página. Por exemplo, na abordagem em que apenas são considerados os links entre páginas, se todas as páginas de um site apontarem para uma página v_1, todas elas estarão contribuindo para aumentar a posição de v_1 no ranking, o que não necessariamente quer dizer que v_1 é uma página relevante. Ao contrário, variações de hosts e domínios desconsideram links entre páginas do mesmo site, reduzindo os impactos causados pelo ruído. Nos parágrafos seguintes, será descrito um conjunto de algoritmos de análise de links para incorporar estratégias de descarte baseados no trabalho de Berlt et al. [2010].

Os métodos propostos por Berlt et al. [2010] tratam a Web como um hipergrafo ao invés de tratá-la com um grafo. O hipergrafo direcionado $\mathcal{H} = (\mathcal{V}, \mathcal{E})$, usado por Berlt et al. [2010], consiste de um conjunto de vértices $\mathcal{V}$ e um conjunto de hiperarcos $\mathcal{E}$, onde $\mathcal{E} \subseteq 2^{\mathcal{V}} \times \mathcal{V}$, $\epsilon = (G, v) \in \mathcal{E}$, $v \notin G$, e $G \subset \mathcal{V}$ é um conjunto

[2]É importante observar que o método PageRank se difere dos métodos Indegree e Outdegree por não tratar todas as arestas da mesma forma. Ainda que sejam aplicados pesos às arestas, todas elas são incluídas no cálculo.

de vértices. Dessa maneira, um hiperarco de um grupo G sempre aponta para um único vértice não pertencente a G. A Web continua sendo modelada de forma que cada página seja um vértice do grafo, porém o conjunto de páginas é particionado em blocos de páginas não-sobrepostas, onde as páginas são agrupadas de acordo com um critério de afinidade. Uma partição do bloco $\mathcal{B}$ no hipergrafo aponta para v por meio do hiperarco $\epsilon = (\mathcal{B}, v)$ se, e somente se, houver ao menos uma página de $\mathcal{B}$ que possui um link para a página Web v e $v \notin \mathcal{B}$. Assim, nesse modelo, hiperarcos representam o fato de uma página dentro de um bloco apontar para uma página de fora do bloco. Berlt et al. [2010] consideram três critérios de partição:

- Partições baseadas em páginas: o bloco é composto por uma única página Web. O grafo resultante corresponde precisamente à representação tradicional do grafo direcionado da Web;
- Partições baseadas em hosts: todas as páginas do bloco pertencem ao mesmo host na Web[3];
- Partições baseadas em domínios: todas as páginas do bloco pertencem ao mesmo domínio na Web[4].

Usando a representação de hiperarco na Web, Berlt et al. [2010] propuseram versões baseadas em hosts e domínios para o PageRank e para o Indegree. As variações do Indegree foram incluídas entre as métricas analisadas nesta pesquisa, uma vez que foram as que obtiveram o melhor desempenho nos experimentos de Berlt et al. [2010]. Neste trabalho, iremos nos referir ao HyperIndegree considerando partições baseadas em domínios e hosts como *HInD* e *HInH*, onde *HInD* se refere a "HyperIndegree baseado em domínio", e *HInH* se refere a "HyperIndegree baseado em host". Por meio de diversos experimentos em uma amostra da Web brasileira, Berlt et al. [2010] mostraram que seus métodos foram melhores que métodos tradicionais, como PageRank e Indegree, com a vantagem de serem menos suscetíveis a spam.

Os autores também implementaram variantes do PageRank e Indegree que não consideram links internos ao grupo de hosts e domínios. Neste trabalho

[3]Para obter o nome do host, a URL é primeiramente pré-processada, removendo os prefixos iniciais "http://" e "www.". Após isso, o nome do host é definido como a sequencia de caracteres que começa no início da URL restante e termina na posição anterior à primeira barra. Por exemplo, na URL "http://noticias.uol.com.br/politica", o nome do host será "noticias.uol.com.br".

[4]Para encontrar o nome do domínio, o nome do host será dividido em termos de acordo com a localização dos pontos (".."). Assim, para o termo "noticias.uol.com.br", teremos "noticias", "uol", "com" e "br". A última parte obtida possuirá o identificador do país e a penúltima o tipo de serviço prestado (em alguns casos, essas partes poderão estar vazias). A antepenúltima parte conterá o nome do servidor. Para o exemplo dado, todas as partes existem sendo que o identificador do país é "br", o serviço prestado é caracterizado por "com" e o nome do servidor é "uol". A união dessas três parcelas formará o domínio, sendo, portanto, "uol.com.br".

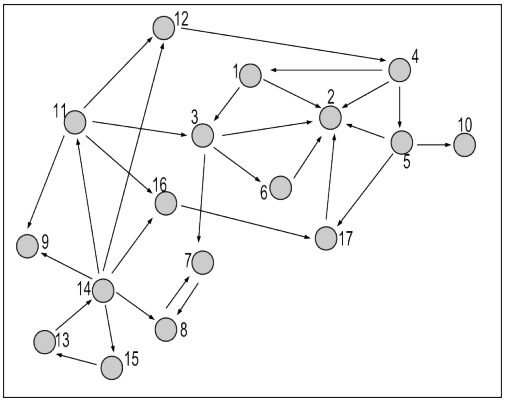

Figura 2.3: *Exemplo de uma representação de páginas Web em grafo baseado em páginas.*

também usaremos as implementações utilizadas por Berlt et al. [2010], contudo incluiremos versões para Outdegree. Iremos nos referir a esses métodos como PRH, InH, e $OutH$, nos quais os links internos aos hosts são descartados; e PRD, $OutD$ e InD, nos quais os links internos aos domínios são descartados. A Tabela 2.2 sumariza os métodos de análise de links que foram implementados considerando o hypergrafo de páginas $\mathcal{H}_p$, o hypergrafo de hosts $\mathcal{H}Hiper_h$, e o hipergrafo de domínios $\mathcal{H}Hiper_d$.

Método	Descrição
In(v)	quantidade de inlinks para v em $\mathcal{H}_p$
Out(v)	quantidade de outlinks para v em $\mathcal{H}_p$
PR(v)	pagerank em $\mathcal{H}_p$
InH(v)	quantidade de inlinks de i para v em $\mathcal{H}_p$, $\forall_i\ host(i) \neq host(v)$
OutH(v)	quantidade de outlinks de o para v em $\mathcal{H}_p$, $\forall_o\ host(o) \neq host(v)$
PRH(v)	PR(v), onde o link ℓ é visitado se $host(\ell) \neq host(v)$
InD(v)	quantidade de inlinks de i para v em $\mathcal{H}_p$, $\forall_i\ domain(i) \neq domain(v)$
OutD(v)	quantidade de outlinks de o para v em $\mathcal{H}_p$, $\forall_i\ domain(o) \neq domain(v)$
PRD(v)	PR(v), onde o link ℓ é visitado se $domain(\ell) \neq domain(v)$
HInH(v)	quantidade de inlinks para v em $\mathcal{H}Hiper_h$
HInD(v)	quantidade de inlinks para v em $\mathcal{H}Hiper_d$

Tabela 2.2: *Métodos de Análise de Links avaliados neste trabalho.*

Para ilustrar a aplicação dessas métricas, utilizaremos o grafo dado como exemplo na Figura 2.3. Para cada página nesse grafo, a Tabela 2.3 indica seus hosts e a Tabela 2.4, seus domínios.

O grafo da Figura 2.3 é obtido considerando todos os links do hypergrafo de páginas $\mathcal{H}_p$, sendo este o grafo utilizado pelas métricas clássicas, In(v), Out(v) e PR(v). Por meio dele e das informações de host e domínio das Tabelas 2.3 e 2.4 podemos obter os grafos das Figuras 2.4 e Figura 2.5. Ao longo deste trabalho, o grafo da Figura 2.3 será chamado de $\mathcal{H}_p$, o grafo da Figura 2.4 de $\mathcal{H}_h$ e o grafo da Figura 2.5 de $\mathcal{H}_d$. O grafo $\mathcal{H}_h$ é obtido retirando todos os links entre

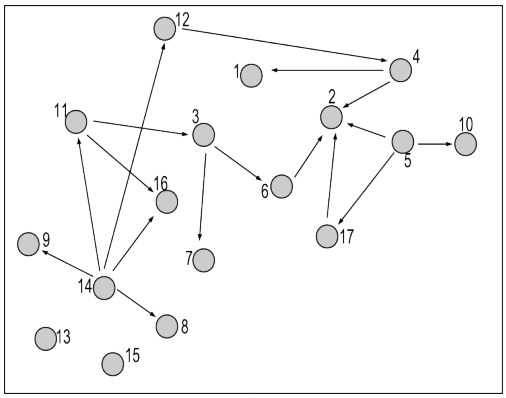

Figura 2.4: *Exemplo de uma representação de páginas Web em grafo baseado em hosts.*

Host H	**Páginas pertencentes ao host** H
h_1	v_1, v_2, v_3
h_2	v_4, v_5, v_6
h_3	v_7, v_8
h_4	v_9, v_{10}, v_{11}
h_5	v_{13}, v_{14}, v_{15}
h_6	v_{12}, v_{16}, v_{17}

Tabela 2.3: *Mapeamento de hosts para o exemplo da Figura 2.4*

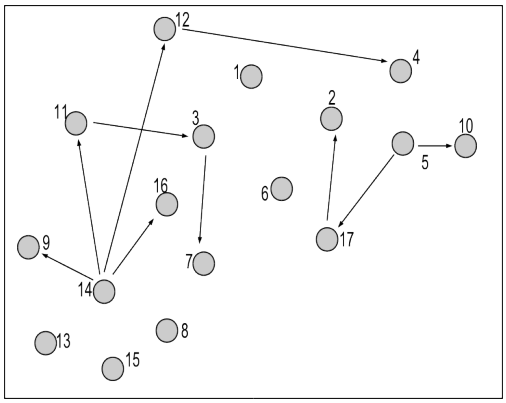

Figura 2.5: *Exemplo de uma representação de páginas Web em grafo baseado em domínios.*

Domínio D	**Páginas pertencentes ao domínio** D
d_1	v_1, v_2, v_3, v_4, v_5, v_6
d_2	v_7, v_8, v_{13}, v_{14}, v_{15}
d_3	v_9, v_{10}, v_{11}, v_{12}, v_{16}, v_{17}

Tabela 2.4: *Mapeamento de domínios para o exemplo da Figura 2.5*

páginas pertencentes ao mesmo host. A aplicação das métricas clássicas sobre o grafo $\mathcal{H}_h$ resulta nas variações InH(v), OutH(v) e PRH(v). Da mesma forma, ao eliminar links entre páginas de mesmo domínio, obtemos o grafo $\mathcal{H}_d$ e as variações originadas ao aplicarmos as métricas clássicas sobre esse grafo são InD(v), OutD(v) e PRD(v).

Para melhor entendimento, seja o exemplo da página p_2 e as diferentes variantes do método Indegree. Ao observar o grafo $\mathcal{H}_p$, vemos que a página p_2 recebe links das páginas p_1, p_3, p_4, p_5, p_6 e p_{17} e, portanto, o resultado obtido para In(p_2) é 6. Como as páginas p_1 e p_3 possuem o mesmo host que p_2, durante a extração do grafo $\mathcal{H}_h$, esses links são eliminados, fazendo com que o valor de InH(p_2) seja 4. Ao utilizar o grafo $\mathcal{H}_d$, o mesmo processo de eliminação ocorre para os links restantes com exceção daquele feito pela página p_{17}, pertencente ao domínio d_3, resultando em InD(p_2)=1.

Para a obtenção de HInH(v) e HInD(v), é importante lembrar que, nesses casos, as páginas são separadas em grupos de hosts, sendo que cada grupo aponta uma única vez para páginas que não pertencem a ele. Assim, HInH(v) é a quantidade de grupos de hosts que apontam para a página v e HInD(v) é a quantidade de grupos de domínios que apontam para v. Ainda utilizando a página p_2 como exemplo, obtemos HInH(v_2)=2, pois as páginas v_4, v_5 e v_6 pertencem ao grupo de páginas do host h_2 e a página v_{17} ao grupo de páginas do host h_{17}. Similarmente, o resultado de HInD(v_2) é 1, pois há somente uma página (v_{17}) que não possui domínio d_1 apontando para v_2, fazendo esta parte do grupo de domínios d_3.

Conforme é possível observar, a quantidade de processamento necessária para obter os resultados das métricas varia conforme a quantidade de informações existente no grafo analisado e, portanto, a complexidade dessa tarefa varia também. Além da vantagem de diminuir ruídos, o processamento em grafos menores tende a ser mais rápido, visto que há uma quantidade menor de dados para serem analisados, e o espaço físico ocupado por eles também diminui. Dessa forma, dependendo da eficácia das variações das métricas de análise de links quando comparadas às métricas clássicas, o seu uso pode tornar-se vantajoso.

2.3 Métricas de correlação de rankings

Métricas de correlação de rankings são utilizadas para averiguar o quão diferentes são os resultados de duas funções de ranking conhecidas. Neste trabalho, usaremos métricas de comparação de ranking para verificar a correlação entre as métricas de análise de links e os fatores qualidade e importância atribuídos aos artigos da Wikipédia, bem como a popularidade dos artigos. Para tanto, a ordenação produzida por cada método de ranking será comparada com a orde-

nação dos artigos considerando esses fatores, conforme estabelecida pela Wikipédia. A métrica de correlação que iremos aplicar é a Kendall τ [Kendall, 1948] apud. [Baeza-Yates and Ribeiro-Neto, 2011] p.146.

2.3.1 Propriedades das métricas de correlação de ranking

As métricas de correlação de ranking comparam as ordens relativas das posições que os documentos (neste trabalho, artigos da Wikipédia) obtiveram na função de ranking aplicadas a eles. Assim, para os resultados de um ranking R_x e de um ranking R_y, podemos representar a correlação existente entre eles por $C(R_x, R_y)$. A correlação entre duas métricas possui as seguintes propriedades [Baeza-Yates and Ribeiro-Neto, 2011]:

- $-1 \leq C(R_x, R_y) \leq 1$;
- se $C(R_x, R_y) = 1$, a concordância entre os dois métodos de ranking é total, ou seja, os rankings comparados são iguais;
- se $C(R_x, R_y) = -1$, a disconcordância entre os dois métodos de ranking é total, ou seja, os rankings comparados são opostos um ao outro.

2.3.2 Coeficiente de Kendall τ

O coeficiente de Kendall τ considera que dois rankings podem variar em uma mesma direção e de maneira similar, ou seja, leva em consideração o deslocamento das posições das páginas e não a posição delas individualmente. Dessa forma, o método Kendall τ utiliza os conceitos de pares concordantes e pares discordantes. Estes conceitos estão ligados a diferenças de posições de dois documentos. A diferença de posição de dois documentos v_a e v_b nos rankings pode ser obtida da seguinte forma:

$pos_{x,a} - pos_{x,b}$ para a diferença das posições de v_a e v_b no ranking R_x
$pos_{y,a} - pos_{y,b}$ para a diferença das posições de v_a e v_b no ranking R_y

Pares concordantes são pares de documentos cuja diferença de suas posições em um ranking R_x possui o mesmo sinal que a diferença de suas posições em um ranking R_y. Pares discordantes são pares de documentos cuja diferença de suas posições em um ranking R_x possui o sinal diferente do sinal da diferença de suas posições em um ranking R_y.

A Tabela 2.5 será utilizada para exemplificar o funcionamento do método Kendall τ. Vamos considerar que a posição de um documento é representada por $pos_{x,a}$, onde x é o método de ranking R_x aplicado e a representa o documento v_a. Assim, a posição do artigo v_6 obtida pelo método R_1 será representada por

Documento a	$pos_{1,a}$	$pos_{2,a}$
d_{12}	1	2
d_5	2	1
d_{34}	2	2
d_{67}	3	3
d_{28}	4	4
d_{19}	4	3

Tabela 2.5: *Representação das posições de documentos obtidas pelos métodos R_1 e R_2*

$pos_{1,6}$. Chamaremos de $\mathcal{V}$ o conjunto com todos os documentos e a quantidade de documentos total desse conjunto de $\mathcal{N}$.

Tomando como exemplo os rankings obtidos por R_1 e R_2 apresentados na Tabela 2.5, para o par (d_{12}, d_{19}) o valor da diferença de suas posições no ranking R_1 é -3 e no ranking R_2 é -1, possuindo em ambos o sinal negativo e sendo, portanto, um par concordante. Já o par (d_{12}, d_5) é uma par discordante, pois a diferença das posições deles em R_1 é -1 e em R_2 é +1, ou seja, os sinais obtidos pelos dois rankings se diferem.

Uma forma simples de medir a correlação de duas métricas é contar os números de pares concordantes e discordantes e verificar a diferença. A quantidade de pares possíveis em um conjunto de páginas V com N documentos é N $\times$ (N-1). Considerando que $disc(R_x, R_y)$ seja a quantidade de pares discordantes entre os rankings R_x e R_y, podemos concluir que a quantidade de pares concordantes é N $\times$ (N-1) - $disc(R_x, R_y)$. A probabilidade de R_x ser igual a R_y, $P(R_x = R_y)$, é proporcional à quantidade de pares concordantes existentes, enquanto a probabilidade de R_x ser diferente de R_y, $P(R_x \neq R_y)$ é proporcional à quantidade de pares discordantes existentes. Assim, é possível inferir que:

$$P(R_x = R_y) = \frac{N \times (N-1) - disc(R_x, R_y)}{N \times (N-1)}$$

$$P(R_x \neq R_y) = \frac{disc(R_x, R_y)}{N \times (N-1)}$$

O coeficiente de Kendall τ é obtido por meio da diferença da probabilidade $P(R_x = R_y)$ pela probabilidade $P(R_x \neq R_y)$. Dessa forma, a fórmula para obter o coeficiente de Kendall τ é:

$$Kendall(R_x, R_y) = 1 - \frac{2 \times disc(R_x, R_y)}{N \times (N-1)} \tag{2.4}$$

Para o exemplo da Tabela 2.5, vamos ordenar todos os documentos de acordo com a sua ordem nos rankings R_1 e R_2. Assim, chamaremos de $\vec{R_1}$ o vetor composto pelos documentos segundo sua ordem no ranking R_1 e de $\vec{R_2}$ o vetor

composto pelos documentos segundo sua ordem no ranking R_2:

$$\vec{R_1} = \{d_{12}, d_5, d_{34}, d_{67}, d_{28}, d_{19}\}$$

$$\vec{R_2} = \{d_5, d_{12}, d_{34}, d_{19}, d_{67}, d_{28}\}$$

Considerando que desejamos avaliar apenas os 4 primeiros documentos de cada função de ranking, para o ranking $\vec{R_1}$ os pares existentes são:

$$(d_{12}, d_5), (d_{12}, d_{34}), (d_{12}, d_{67}),$$
$$(d_5, d_{34}), (d_5, d_{67}), (d_{34}, d_{67})$$

Dos pares existentes em $\vec{R_1}$, podemos dizer que os pares (d_{12}, d_{67}), (d_5, d_{67}) e (d_{34}, d_{67}) são concordantes e os pares (d_{12}, d_5), (d_{12}, d_{34}) e (d_5, d_{34}) são discordantes.

Os pares existentes em $\vec{R_2}$ são:

$$(d_5, d_{12}), (d_5, d_{34}), (d_5, d_{19})$$
$$(d_{12}, d_{34}), (d_{12}, d_{19}), (d_{34}, d_{19})$$

Dos pares existentes em $\vec{R_2}$, podemos dizer que os pares (d_5, d_{19}), (d_{12}, d_{19}) e (d_{34}, d_{19}) são concordantes e os pares (d_5, d_{12}), (d_5, d_{34}) e (d_{12}, d_{34}) são discordantes.

Somando o total de pares discordantes existentes em $\vec{R_1}$ e $\vec{R_2}$ temos 7 pares. Aplicando a fórmula do coeficiente de Kendall τ ao exemplo, temos:

$$Kendall(R_1, R_2) = 1 - \frac{2 \times 6}{4 \times (4-1)} = 1 - \frac{12}{12} = 1 - 1 = 0$$

Podemos notar que, avaliando os 4 primeiros documentos dos rankings R_1 e R_2, para o exemplo da Tabela 2.5 o valor do coeficiente de Kendall τ é 0, o que quer dizer que as duas funções de ranking são independentes entre si, não sendo rankings completamente concordantes nem discordantes um em relação ao outro.

2.4 Qualidade na Wikipédia

A Wikipédia é um enciclopédia digital disponível na Web- em mais de 280 idiomas e com mais de 700.000 artigos em português. A principal diferença da Wikipédia para as enciclopédias impressas, além de ela ser acessada pela Web, é que nela qualquer pessoa é capaz de modificar o conteúdo, não havendo necessariamente um controle de autoria. Os autores dos artigos da Wikipédia trabalham como uma grande comunidade virtual e seu crescimento acontece pelo trabalho mútuo e colaborativo de pessoas com interesses em um mesmo tema. As pessoas que participam na construção dos artigos da Wikipédia são chamadas de Wikipedistas[5]. É permitido que qualquer pessoa edite um artigo, sendo

[5] http://pt.wikipedia.org/wiki/Wikipéia:Wikipedistas

que atualizações inadequadas podem ser retiradas por outros Wikipedistas, ou a versão anterior à modificação pode ser restaurada. Com a finalidade de auxiliar neste processo, os editores trabalham seguindo regras. Para preservar um bom convívio entre eles são disponibilizadas normas de conduta[6] e políticas[7] que se baseiam nos princípios de civismo, respeito e boa comunicação.

O controle de qualidade dos artigos é executado por meio das políticas desenvolvidas e pelo monitoramento dos próprios editores. Para um maior controle, editores podem monitorar os artigos modificados recentemente, acompanhar os artigos que são de seu interesse, marcar artigos que apresentam algum problema e participar de discussões. Os melhores artigos são chamados de artigos destacados e são apresentados na página principal da Wikipédia. Existem 6 níveis de qualidade dos artigos[8] (descrições dos níveis serão apresentadas na Seção 2.4.3). O tipo de artigo e seu nível de qualidade ficam disponíveis aos leitores para que estes tenham maior informação sobre o artigo que está sendo lido. A qualidade atribuída a um artigo é realizada por meio de um processo democrático, o qual é descrito na Seção 2.4.3. Existem artigos aos quais não foram atribuídas notas de qualidade; estes casos são descobertos e avaliados por programas denominados robôs. O nível de qualidade de um artigo permite que se tenha conhecimento sobre quais artigos precisam ser aperfeiçoados e quais artigos aparentam mais indícios de serem fontes confiáveis de informação.

Alguns artigos possuem caráter informativo[9] para explicar para a comunidade sobre o funcionamento da Wikipédia. Os artigos informativos podem conter informações sobre a Wikipédia, sobre a comunidade, auxiliar na edição de artigos e servir de ajuda no modo geral.

Os dados da Wikipédia estão disponíveis integralmente ao público, fornecendo um ambiente rico de informações e, por causa disso, muitas pesquisas são realizadas com a Wikipédia. A Wikipédia é apenas um dos projetos da Wikimedia Fundation[10], organização sem fins lucrativos que visa incentivar a produção de conteúdo e disponibilizá-lo na forma de *wikis*. Alguns outros projetos são Wikicionário[11], Wikinotícias[12], Wikisources[13] e Wikiversidade[14].

A seguir são detalhados aspectos da Wikipédia que são importantes para esta pesquisa.

[6] http://pt.wikipedia.org/wiki/Wikipédia:Normas_de_conduta
[7] http://pt.wikipedia.org/wiki/Wikipédia:Política_de_ediç~ao
[8] http://pt.wikipedia.org/wiki/Predefiniç~ao:Escala_de_avaliaç~ao
[9] http://pt.wikipedia.org/wiki/Wikipédia:Artigo_informativo
[10] http://wikimediafoundation.org/wiki/Home
[11] http://pt.wiktionary.org/
[12] http://pt.wikinews.org/
[13] http://pt.wikisource.org/
[14] http://pt.wikiversity.org/

2.4.1 Páginas de discussões

Páginas de discussões[15] são páginas especiais nas quais se discute sobre os artigos aos quais elas se referem. Qualquer usuário registrado na Wikipédia pode participar de uma discussão. As mensagens devem possuir títulos sucintos, claros e objetivos e as mensagens não devem possuir caráter ofensivo entre os participantes, devendo servir como uma forma de contribuição para a melhoria de um artigo.

Figura 2.6: *Informações sobre o artigo "Futebol"*

Para artigos enciclopédicos (não informativos sobre a própria Wikipédia), na página de discussão são apresentadas informações relevantes sobre o artigo como a sua qualidade (a ser explicado na Seção 2.4.3), categorias às quais ele se refere (Wikiprojetos), notas de importância para as categorias (a ser explicado na Seção 2.4.4) e até informações sobre o mesmo artigo em outros idiomas. A Figura 2.6 apresenta o topo da página de discussão do artigo "Futebol".

2.4.2 Pilares, Políticas e Recomendações para edição

As edições realizadas devem obrigatoriamente seguir certos princípios, os cinco pilares[16] da Wikipédia, que moldam suas características. Os cinco pilares são:

[15] http://pt.wikipedia.org/wiki/Wikipédia:Página_de_discuss˜ao
[16] http://pt.wikipedia.org/wiki/Wikipédia:Cinco_pilares

1. Enciclopedismo: características de enciclopédias especializadas e almanaques. Deve conter toda a informação necessária sobre seus artigos de forma clara, sendo essas informações verificáveis e relevantes para seus artigos. Os artigos devem possuir ligações com outros artigos que possuam conteúdos relacionados ao seus;

2. Neutralidade de ponto de vista: autores devem ser imparciais, informações devem ser precisas e contextualizadas, verbetes devem ser justificados com fontes reputadas e todos os pontos de vista devem ser representados proporcionalmente;

3. Licença livre: qualquer pessoa possui licença para editar;

4. Convivência comunitária: respeito às normas de conduta;

5. Liberalidade nas regras: autores possuem liberdade para serem criativos desde que os textos sejam bons.

Para tornar estes princípios mais claros e evitar conflitos, foram criadas políticas e recomendações[17]. As políticas descrevem padrões a serem seguidos e as recomendações indicam boas práticas, de forma que as ações possam ser padronizadas e colaborações mais fáceis. Alguns editores (ou grupos de editores) podem propor ensaios, os quais podem ganhar aprovação comunitária e virar políticas.

2.4.3 Qualidade dos artigos

Os artigos são classificados em seis níveis de qualidade que são impróprio[18], spam[19] e mínimo[20](nível 1); esboço[21](nível 2, 3 ou 4); bom[22](nível 5); e destaque[23](nível 6), sendo os artigos de destaque os melhores artigos da Wikipédia.

A atribuição da qualidade dos artigos é importante para que o leitor esteja informado sobre a qualidade do texto que ele está lendo, e para que editores saibam quais artigos precisam de atenção.

A Wikipédia utiliza três processos para avaliar artigos. Cada processo de avaliação visa atingir um público. O primeiro deles realiza avaliação pela comunidade, na qual, para níveis de um a quatro, qualquer editor pode mudar a qualidade de um artigo desde que o artigo esteja de acordo com o nível a ser

[17] http://pt.wikipedia.org/wiki/Wikipédia:Políticas_e_recomendaç˜oes
[18] http://pt.wikipedia.org/wiki/Wikipédia:Impróprio
[19] http://pt.wikipedia.org/wiki/Wikipédia:Spam
[20] http://pt.wikipedia.org/wiki/Wikipédia:Mínimo
[21] http://pt.wikipedia.org/wiki/Wikipédia:Esboço
[22] http://pt.wikipedia.org/wiki/Wikipédia:Artigos_bons
[23] http://pt.wikipedia.org/wiki/Wikipédia:Destaque

designado a ele. O segundo processo é automático, realizado por meio de robôs. Este tipo de avaliação foi criada visto que muitos artigos não possuíam notas de qualidade e é apenas implantada para avaliar os artigos fornecendo níveis de 1 a 4, conforme as regras estabelecidas pela Wikipédia. Essa avaliação, por não ser muito precisa (principalmente em relação aos níveis 3 e 4), é descartada quando um usuário fornece uma avaliação. O terceiro processo é mais recente, no qual foi desenvolvido uma ferramenta que permite que leitores e editores avaliem os artigos com uma escala de 1 a 5.

Artigos impróprios são artigos com material inadequado por ser ofensivo, sem sentido ou não ter caráter enciclopédico. Quando um artigo é marcado como impróprio, ele é automaticamente colocado na sessão de eliminação rápida, por isso, quando apenas parte do artigo é considerada imprópria essa parte deve ser apagada e, se a versão anterior for uma versão estável, é ela que deverá permanecer.

Spams são artigos que visam promover um produto ou um site por meio de propagandas em artigos (utilização de linguagem de anúncios) ou links externos (redirecionamento a um Website comercial). Quando isso acontece, o artigo pode ser reescrito sob um ponto de vista neutro ou pode ser automaticamente deletados, colocados na sessão de eliminação rápida ou anunciados para serem eliminados. O conteúdo também pode ser retirado se violar direitos autorais.

São considerados mínimos os artigos muito pequenos, normalmente com poucas linhas (ou uma frase), mas que contêm informações aproveitáveis e por isso não devem ser descartados. Esses artigos devem ser desenvolvidos em esboços ou em artigos de tamanho maior. Estes tipos de artigos são qualificados como nível 1.

Esboço são artigos curtos que não estão completos ou que possuem pouca informação. Normalmente possuem dois parágrafos. Quando marcado como esboço, o artigo fica dependendo de outras pessoas que saibam melhor sobre o assunto para melhorá-lo e expandi-lo. A ideia é que esboços sejam modificados continuamente pelos vários editores até se tornarem artigos completos. Deve haver um cuidado para não confundi-lo com artigos mínimos. A grande diferença é que esboços possuem mais dados e se referem a um pesquisa maior. Os esboços podem possuir o nível de qualidade 2, sendo que esboços muito limitados recebem nível 1.

Artigos maiores que um esboço, mas pouco desenvolvidos, recebem nível 3 de qualidade. Artigos desenvolvidos mas que não possuem qualidade para ser um artigo bom ou não conseguiram se eleger como um artigo bom possuem nível 4. Artigos que não possuem verificabilidade (seus dados não são comprovados com fontes externas) estão entre os níveis 1 e 3.

Artigos bons (nível 5) são aqueles que foram definidos como de boa qualidade (qualidade acima da média) e que, durante o processo de validação, não conseguiram atingir o nível de destaque. Esses artigos possuem conteúdo satisfatório e são bem escritos. Eles devem estar escritos de maneira clara e correta (redação), cobrir todos os principais pontos sem perder o foco (completo), fazer referências a fontes confiáveis (verificabilidade), conter todos os verbetes relacionados a ele (links internos), ser estáveis (estabilidade), seguir o livro de estilos da Wikipédia (Wikificação), possuir imagens (imagens) e seu tamanho deve estar de acordo com o assunto abordado (layout). A Figura 2.6 apresenta um exemplo de um artigo definido como "artigo bom".

Os artigos de destaque (nível 6) são artigos que são eleitos pela comunidade para participar da página principal devido a sua alta qualidade. Para um artigo se tornar destaque, ele deve estar de acordo com os padrões impostos pela Wikipédia (redação, estabilidade, imagens, Wikificação, completo, verificabilidade, layout e links internos). Caso ele esteja, um editor pode apresentar a candidatura do artigo. Podem votar apenas as pessoas que são cadastradas e estas devem ter, no mínimo, 300 edições válidas e sua primeira contribuição válida deve ter sido há, no mínimo, 90 dias antes do início da votação. Os eleitores devem conhecer todos os critérios para um artigo ser considerado um destaque. Os editores podem marcar na votação com uma estrela dourada, caso considerem que o artigo deva ser um artigo de destaque, ou com uma estrela cinza, caso acreditem que o artigo deva ser apenas considerado bom. Quando eles julgam que o artigo não está apto a ser bom ou de destaque, eles devem marcar com um "X". Todos os votos devem ser justificados para que os autores dos artigos possam mudar possíveis problemas reportados. Os eleitores também podem anular o seu voto dado anteriormente. A votação possui um período máximo de 30 dias. Para artigo bom, ela pode ser encerrada após 15 dias se o artigo possuir 5 votos para artigo bom e nenhum para artigo normal. Caso o artigo chegue ao prazo de 30 dias sem ser definido como bom, será verificado se ele possui no mínimo 7 votos favoráveis com 75% de votos a seu favor. O artigo é considerado de destaque quando possui 7 votos para artigo de destaque e 75% dos votos a seu favor. Se o artigo não for aceito como destacado, ele pode, depois de um mês do fim do prazo de votação, apresentar nova candidatura. A nova candidatura é aceita apenas se houver modificações significativas no conteúdo do artigo. Um conjunto composto por artigos de destaque ou bons que estejam relacionados a um mesmo tema é chamado de "tópico de destaque".

O status de artigo de destaque pode ser retirado no processo de revalidação caso o artigo não esteja cumprindo os critérios vigentes. O processo é semelhante ao da eleição para avaliá-lo como destaque (apresentação de candidatura

e votação justificada pelos membros por um período de 30 dias). Nesse processo, é possível votar para o artigo permanecer como destaque, remover seu status (tornar-se um artigo normal) ou um mudá-lo para um artigo bom. Um artigo continuará sendo destacado se, após 30 dias, houver um consenso para que ele permaneça. Caso haja um consenso para o artigo tornar-se bom, ele será modificado para bom. Se houver algum voto indicando que ele deve ser um artigo normal, ele passará a possuir a nota 4 de qualidade.

A Tabela 2.6 apresenta os 6 níveis de qualidade existentes para um artigo da Wikpédia e quais os critérios para que um artigo seja classificado em uma desses níveis.

Nível de qualidade	Critério de Classificação
Para avaliar	Reservada aos artigos cuja qualidade ainda não foi avaliada e a artigos antigos que não foram incorporados no projeto
1	Reservada aos artigos esboçados não referenciados e aos artigos mínimos
2	Reservada aos artigos extensos não referenciados e aos esboços referenciados
3	Reservada aos artigos pouco desenvolvidos
4	Reservada aos artigos desenvolvidos
5	Reservada exclusivamente aos artigos bons
6	Reservada exclusivamente aos artigos destacados

Tabela 2.6: *[Wikipedia, 2011] Níveis de qualidade estabelecidos pela Wikipédia e seus critérios de classificação*

Não há evidências de que tais estratégias para a avaliação da qualidade dos artigos da Wikipédia realmente promovem a produção de artigos de qualidade. Enquanto não conhecemos nenhum trabalho que tenha avaliado o processo em si, vários trabalhos têm estudado a Wikipédia para averiguar se o conteúdo de seus artigos possuem qualidade.

Em 2005, por exemplo, foi apresentado um estudo comparando a qualidade dos artigos de ciências da Enciclopédia Britânica com os da Wikipédia [Giles, 2005], no qual foram encontrados erros em ambas as enciclopédias. Este estudo mostrou que a diferença de acurácia da Enciclopédia Britânica para a Wikipédia não era grande, estando a Wikipédia próxima a uma enciclopédia impressa. Reavley et al. [2011] procuraram avaliar a qualidade de conteúdos relacionados a tópicos de doenças mentais de 14 sites frequentemente acessados (entre eles a Wikipédia), juntamente com a Enciclopédia Britânica e um livro de psiquiatria. O objetivo do trabalho era verificar a qualidade das contribuições de especialistas na Wikipédia. Os conteúdos de cada fonte foram avaliados por

6 especialistas nas áreas de acordo com 5 critérios (acurácia, cobertura, referenciamento, legibilidade e atualidade). Os resultados obtidos para cada fonte variaram significantemente de acordo com o tópico analisado. Contudo, com exceção de legibilidade, a Wikipédia foi a melhor conceituada em todos os tópicos, mostrando que, para os artigos médicos sobre doenças mentais avaliados, ela era uma fonte de qualidade e confiável. Ainda sobre artigos da área médica, Heilman et al. [2011] discutem os pontos fortes e fracos da Wikipédia como fonte de informações de saúde e a comparam com outras wikis médicas, mostrando que a Wikipédia pode ser uma importante fonte de distribuição de conhecimento. Eles também divulgam o Wikiprojeto Medicina criado em 2004 para coordenar a edição de artigos em inglês relacionados à medicina, e convidam a comunidade médica a melhorar os conteúdos de saúde da Wikipédia.

2.4.4 Wikiprojetos e Notas de Importância

Os Wikiprojetos[24] reúnem esforços de diversos editores com a finalidade de coordenar a escrita de artigos da Wikipédia referentes a um tema ou a um conjunto de temas. Nos Wikiprojetos os autores trabalham em equipe por meio de ações como discussões e criação de listas de tarefas a serem feitas. O seu principal objetivo é possibilitar uma dinâmica sistemática para a produção de excelentes artigos.

É permitido que um artigo faça parte de mais de um Wikiprojeto, já que um mesmo artigo pode estar relacionado a diversos temas. Isto traz duas grandes vantagens. Primeiro, é vantajoso no sentido de que sempre as pessoas que fazem parte de um Wikiprojeto podem monitorar as pessoas do outro Wikiprojeto e vice-versa, diminuindo as chances de vandalismos. A outra vantagem é que editores que fazem parte de um Wikiprojeto podem pedir a colaboração de outras pessoas de outros Wikiprojetos para tornarem os textos mais completos e auxiliar de diversas formas, como com verificação gramatical, fornecendo materiais e até mesmo dando conselhos. Um Wikiprojeto também pode estar dentro de um Wikiprojeto mais abrangente.

Para cada Wikiprojeto do qual um artigo faz parte, ele possui uma nota de importância, sendo que o valor mínimo é 1 e o máximo é 4. Na Figura 2.6 é possível ver a quais Wikiprojetos o artigo “Futebol” pertence (Futebol e Wikipédia Offline) e a importância deste artigo para cada um desses Wikiprojetos (nota máxima em ambos). É necessário observar que a nota de importância do artigo para cada Wikiprojeto que ele pertence difere-se da nota de qualidade do artigo, que é geral e independe de Wikiprojetos.

Artigos que não foram avaliados segundo a sua importância para um Wiki-

[24] http://pt.wikipedia.org/wiki/Wikipédia:WikiProjeto

projeto recebem como nota uma interrogação (?). A nota 1 é atribuída aos artigos com pouca importância para o projeto, por se focar especificamente em uma determinada área do conhecimento. Artigos com nota 2 são aqueles que possuem importância relativa para o projeto, cobrindo uma área de conhecimento específica, mas sobre assuntos relativamente conhecidos. Normalmente artigos com nota 2 são artigos que são sub-tópicos sem muita importância de um artigo mais importante. A nota 3 é atribuída a artigos importantes que cobrem a área de conhecimento geral a qual o projeto abrange em diversos aspectos significantes, e normalmente trata de sub-tópicos mais importantes da área. Por fim, os artigos mais importantes para o projeto, com nota 4, são aqueles que tem fundamental importância para um projeto, por conter informações de base sobre o tema do projeto. No exemplo da Figura 2.6, o artigo "Futebol" é considerado um artigo de maior importância para o Wikiprojeto Futebol, pois contém informações que formam a base do tema do projeto. O Wikiprojeto Wikipédia Offline é um projeto com o objetivo de reunir os melhores artigos para compor uma versão *offline* da Wikipédia para pessoas que não possuem uma conexão com a Internet. Podemos notar que este Wikiprojeto considera o artigo "Futebol" muito importante para uma versão *offline*.

2.5 Considerações finais

Neste capítulo, foram apresentadas informações necessárias para o desenvolvimento da pesquisa. A análise de links fornece uma compreensão sobre as relações e as propriedades existentes na estrutura de grafos baseada em links da Web. As métricas de análise de links utilizam tais relações e propriedades para avaliar a relevância de uma página. Visto que desejamos verificar a correlação existente entre o resultado de métricas de análise de links e fatores como qualidade, importância e popularidade de páginas da Wikipédia, foram selecionadas 11 métricas para realizar o estudo (In, Out, PR, InH, OutH, PRH, InD, OutD, PRD, HInH, HInD), as quais foram explicadas neste capítulo. O coeficiente de Kendall τ foi o responsável por obter essas correlações. Também foi apresentado o processo de atribuição de notas de qualidade e importância na Wikipédia, importante para a compreensão de particularidades da Wikipédia e de como os níveis para cada artigo são obtidos.

Capítulo

3

Trabalhos relacionados

Neste capítulo, são apresentados trabalhos publicados na literatura que se relacionam com o trabalho realizado. Em particular, são apresentados (a) trabalhos que propõem métricas de análise de links com foco em questões como qualidade ou uso da Wikipédia; (b) trabalhos que tentam estimar a qualidade de artigos na Wikipédia de forma automática e que, para tanto, usam análise de links; e (c) outros trabalhos que discutem temas relacionados indiretamente ao nosso, como análise comparativa da estrutura de links da Wikipédia.

3.1 Métricas de Análise de Links

Alguns trabalhos estudam métricas de análise de links relacionadas com a Wikipédia, ou discutem o conceito de qualidade no contexto de análise de links.

A relação entre métricas de análise de links e qualidade de páginas foi alvo do estudo de Amento et al. [2000]. Os autores foram motivados pela observação de que, embora seja fácil encontrar páginas relevantes para um determinado tópico, nem sempre elas possuem qualidade. Assim, eles resolveram verificar esta relação usando especialistas humanos. Em particular, 5 sites foram avaliados (15 páginas por site), conforme sua qualidade, por 19 especialistas. Cada site abordava um tópico distinto. Métricas de análise de links (Indegree, Outdegree, pontuação de autoridade da métrica HITS[1], pontuação de hub da métrica HITS

[1]A métrica HITS Kleinberg [1999] é uma métrica que se diferente dos métodos estudados neste trabalho, por considerar em seus cálculos os termos por computar duas pontuações. A primeira pontuação reflete a ideia de autoridade no assunto. Uma página é considerada um autoridade quando ele possui alta reputação indicando ser uma fonte de informação confiável. Espera-se que páginas de autoridade tenham conteúdos relevantes. A segunda pontuação considera que algumas páginas funcionam melhor como *hubs*, ou seja, um

e PageRank) e de termos foram aplicadas sobre essas páginas. Um problema encontrado foi a natureza subjetiva do conceito de qualidade que faz com que cada especialista julgue as páginas de acordo com aspectos tão diversos quanto a sua organização, quantidade de informação e particularidades. Como consequência, é necessária a opinião de muitos especialistas para se alcançar consenso, o que não foi possível nesse estudo. O pequeno número de especialistas, páginas e tópicos, somados às diferenças de opinião observadas, dificultaram a obtenção de resultados estatisticamente significativos. Contudo, os resultados obtidos sugeriram que (a) métricas de análise de links ranqueiam melhor páginas de qualidade mais alta e (b) a métrica Indegree é tão eficaz para esta tarefa quanto outras mais complexas, como PageRank e HITS. Os autores concluíram seu estudo observando que um trabalho de maior escala era necessário, em particular, um que tire proveito de um número maior de domínios, de tópicos e de especialistas. Sem dúvida, dos trabalhos que analisamos, este é o mais relacionado ao nosso. De fato, ao usar a Wikipédia, tivemos a oportunidade de confirmar os resultados de Amento et al. [2000], pois possuímos mais artigos revisados, cobrindo tópicos em um número muito maior de domínios. Desta forma, obtivemos resultados estatisticamente significativos. Além disso, diferente de Amento et al. [2000], verificamos o impacto de outros fatores como popularidade e importância, que também puderam ser obtidos de forma independente na Wikipédia. Devemos observar que ao usar a Wikipédia, temos a oportunidade de contar com um mecanismo de avaliação robusto, baseado em princípios criteriosamente definidos, como apresentado na Seção 2.4.

Gleich et al. [2010] estudaram o impacto do parâmetro de teletransporte (coeficiente de damping) da métrica PageRank e utilizaram como coleção de estudo a Wikipédia. O coeficiente de damping é responsável por determinar quando um usuário fará um salto aleatório de uma página para outra, desconsiderando se a página nova é ou não apontada pela página atual. O valor obtido pela métrica PageRank varia de acordo com o seu coeficiente de damping. Para avaliar o seu impacto, os autores testaram diferentes valores e, para cada valor, calcularam o PageRank das páginas na Wikipédia. Os resultados foram então comparados com *logs* de browsers. Os autores observaram que os rankings obtidos foram diferentes do esperado, levantando uma discussão sobre a diferença entre o grafo de links da Wikipédia e a Web. Os autores afirmaram saber sobre diferenças existentes entre a estrutura da Web e da Wikipédia, porém as desconsideram em seus experimentos. A principal diferença desse trabalho para o nosso é que,

centro de divulgação de informações. Páginas de hub concentram links para diversas páginas de autoridade, não sendo páginas de autoridade em si, mas coleções de páginas que alguém com interesse em determinado tópico se dispôs a reunir.

nesse trabalho, a métrica PageRank é aplicada dentro da Wikipédia e restrita aos links internos desta. Uma contribuição desse trabalho para o trabalho que realizamos foi a discussão sobre as possíveis diferenças entre a Wikipédia e a Web, uma vez que essa discussão nos permite saber o quanto é possível estender conclusões obtidas na Wikipédia para a Web em geral.

Berlt et al. [2010] sugeriram que a Web fosse representada como um hipergrafo de forma a permitir que grupos de sites pudessem apontar para páginas isoladas. Dessa forma, novas estratégias de análise de links foram propostas. Em particular, as métricas PageRank e Indegree foram modificadas de acordo com a representação proposta e comparadas com as versões originais que modelam a Web como um grafo. Os autores também adaptaram essas métricas, de forma que as páginas fossem agrupadas por host e domínio. Essas variações foram denominadas PageRankHost, IndegreeHost, PageRankDomain e IndegreeDomain. Os autores concluíram que todos os métodos apresentados tiveram desempenho melhor ou igual ao PageRank e ao Indegree tradicional com a vantagem adicional de serem menos suscetíveis a spam. Ao contrário deste trabalho, esse estudo foi focado na proposição e na avaliação de novas métricas de análise de links. Ele é relacionado ao nosso estudo por apresentar uma série de adaptações aos métodos tradicionais que também utilizamos. A razão para escolher esses métodos é o fato de os autores reforçarem a ideia que links originados independentemente e de diferentes fontes representam votos mais confiáveis de qualidade da página apontada. Outro aspecto em comum entre esse trabalho e o que realizamos é que ambos usam uma coleção derivada do domínio *br* como amostra da Web. Além dos métodos propostos por Berlt et al. [2010], usamos o método clássico Outdegree e aplicamos sobre eles as variações sugeridas para Indegree e PageRank.

3.2 Estimativa automática de qualidade de conteúdo na Wikipédia

Entre os vários trabalhos que estudam o problema de estimativa automática de qualidade na Wikipédia (e.g. Blumenstock [2008]; Liu and Ram [2009]; Stvilia et al. [2005]; Wöhner and Peters [2009]), alguns sugerem o uso de informação de análise de links para esta tarefa (e.g. Dalip et al. [2011]; Dondio and Barrett [2007]; Rassbach et al. [2008]). Em todos esses casos, apenas links internos da Wikipédia são usados. Além disso, os autores não estão interessados em analisar métricas de links fora do contexto de previsão de qualidade na Wikipédia.

Dondio et al. [2006] sugeriram uma metodologia para estimar a qualidade e a credibilidade de artigos da Wikipédia. Os autores notaram que peculiaridades da Wikipédia, como a alta taxa de atualização e a independência de qualidade

entre as versões, dificultam a aplicação da maioria dos métodos propostos anteriormente. Assim, para avaliar um artigo, eles listaram várias características que deveriam ser utilizadas para construir sistemas de ranking correspondendo a diferentes aspectos de qualidade (estabilidade, controlabilidade e qualidade de conteúdo). Os rankings foram obtidos por meio da combinação de indicadores, como histórico de revisão, conteúdo, estrutura e links, combinando-os em um único valor. Entre as evidências consideradas, os autores usaram o Indegree e o Outdegree com links internos. Como os autores não estudaram o impacto de cada evidência isoladamente, não foi possível determinar qual a contribuição das métricas de link na estimativa de qualidade por meio de seu trabalho.

Rassbach et al. [2008] aplicaram técnicas de aprendizagem de máquina para estimar automaticamente a qualidade dos artigos da Wikipédia. Em seu estudo, os autores propuseram uma série de características indicativas de qualidade e verificaram quais delas são mais eficientes para a tarefa. Para tanto, foram utilizados artigos da Wikipédia em inglês. Da mesma forma que em português, esses artigos são classificados em até 6 níveis, sendo eles (em ordem decrescente de nível): “Featured", “A", “Good", “B", “Started"e “Stub". Pelo fato de artigos “Stub"não possuírem muitos dos elementos adequados para uma fonte confiável, eles foram retirados da coleção. No trabalho, foram processados 168.183 artigos no total. O método utilizado foi um modelo de Entropia Máxima, técnica de aprendizado de máquina para classificação. Os artigos foram divididos em duas coleções, uma de treino e outra de teste. A coleção de treino, utilizada pelo classificador, possuía 650 artigos de cada nível. Os artigos restantes foram utilizados na coleção de teste. Baseado na coleção de teste, o modelo de Entropia Máxima ajustava, iterativamente, pesos para cada característica de acordo com a sua relevância ao inferir a nota de qualidade de um artigo da Wikipédia. Foram analisadas 50 características, divididas nas seguintes categorias: 1) medidas de tamanho, 2) conteúdo textual, 3) características específicas da rede, e 4) métricas de legibilidade. As características específicas da rede consideravam o número de imagens e links internos da Wikipédia. O modelo apresentado obteve acurácia de quase 75%. Como trabalhos futuros, os autores discutiram o uso de outras características, como a categoria a qual a página se refere e a qualidade das imagens dos artigos (em termos de utilidade para a compreensão do artigo). Eles também sugeriram que o PageRank, implementado internamente, poderia ser um bom indicador de qualidade dos artigos e que, de forma mais abrangente, uma implementação utilizando links da Web, externos à Wikipédia, também poderia. Neste trabalho, verificamos analisar a ideia sugerida por Rassbach et al. [2008], estendendo-a para outras métricas de análise de links além da métrica PageRank.

Da mesma forma que Rassbach et al. [2008], Dalip et al. [2011] aplicaram técnicas de aprendizado de máquina para o problema de previsão de qualidade na Wikipédia. Seu trabalho, contudo, teve ênfase no estudo de indicadores de qualidade e no impacto desses indicadores na previsão. Os indicadores estudados correspondiam a características extraídas do texto do artigo (tamanho, estrutura, estilo e legibilidade), das revisões e do grafo de links interno (características de rede). As características de rede estudadas foram os valores de PageRank, Indegree e Outdegree dos artigos, contagem de links, contagem de traduções, reciprocidade dos links, coeficiente de cluster e associatividade. Foram avaliados 874 artigos, considerando um grafo de 3.165.998 nós (artigos e páginas que redirecionam para artigos) e 86.077.675 links entre os nós. Dos experimentos realizados, os autores concluíram que os indicadores mais importantes para a atribuição da qualidade são os mais fáceis de extrair, ou seja, características textuais relacionadas ao tamanho, estrutura e estilo. Em particular, os indicadores mais complexos, como os de análise de links, não contribuem significantemente para a tarefa. Ao usar a melhor combinação dos indicadores propostos, os autores obtiveram os menores erros de previsão, superando os métodos propostos anteriormente por Rassbach et al. [2008] e Dondio et al. [2006].

Em todos os casos apresentados, os links usados foram restritos à própria Wikipédia. No trabalho que realizamos, ao contrário, foram utilizadas diversas métricas de análise de links com uma amostra da Web como um todo e, portanto, não restrita ao grafo interno da Wikipédia. A vantagem de usar citações externas é que, ao fazer isso, utilizamos links independentes que não são usados com natureza enciclopédica, eliminando a padronização de links imposta pela Wikipédia. Além disso, os links externos proporcionam uma ampla vizinhança de páginas, necessárias para algumas métricas, como o PageRank.

3.3 Outros trabalhos relevantes

Considerando que links na Wikipédia podem ser mais semelhantes a citações em documentos bibliográficos que links em páginas Web, torna-se importante verificar em que medida a Wikipédia se assemelha à Web. Finalmente, é importante conhecer quais fatores influenciam no desempenho de métricas de análise de links no contexto da Wikipédia. Nesta seção, apresentamos trabalhos que abordam estes temas.

Smith [2004] realizou um estudo no qual são verificadas as diferenças entre links e citações no escopo de pesquisas científicas. Para isso, foram obtidos sites de institutos de pesquisa, com os quais Smith [2004] identificou tipos de páginas existentes e os motivos de autores de páginas fazerem links entre as páginas.

Como resultado, o autor elencou 8 motivos principais: download de arquivos, publicidade, informações sobre áreas geográficas, páginas relacionadas, informação sobre o criador da página, reconhecimento do suporte, citação formal de um estudo e informações gerais sobre o link. Em seu trabalho, Smith [2004] concluiu que citações e ligações mostraram-se semelhantes quando o motivo para fazer o link referia-se a pesquisas, como busca de informações, e-journals, trabalhos técnicos, reportagens, artigos de conferências e fontes de programas, o que totalizava apenas 20% da Web em 2004, ano em que a pesquisa foi realizada.

A diferença entre a estrutura da Web e a Wikipédia também foi estudada por Kamps and Koolen [2009]. Nesse trabalho, foram comparados os comportamentos das páginas do domínio *gov*, coletadas da base TREC Web, com as páginas da Wikipédia. Também foram investigadas as evidências de links a fim de melhorar buscas nas duas estruturas. Segundo Kamps and Koolen [2009], a Wikipédia diverge da estrutura da Web am alguns pontos: (a) a densidade dos seus links é mais alta que os da Web; (b) tanto inlinks quanto outlinks são igualmente úteis para determinar importância na Wikipédia, enquanto na Web, inlinks são mais úteis; e (c) na Wikipédia, ao ser utilizadas evidências de links para fazer busca, evidências de links externos falham e é necessário considerar o contexto. É importante observar que nesse trabalho, a Wikipédia é comparada com uma coleção restrita a páginas do domínio *gov*, enquanto em nosso trabalho utilizamos um conjunto de páginas mais representativo da Web. Em nossa pesquisa, realizamos uma comparação da estrutura de links da Wikipédia com a nossa amostra da Web, similar à realizada pelos autores em Kamps and Koolen [2009]. Por meio desta comparação, pudemos estabelecer em que medida nossas conclusões para Wikipédia podiam ser estendidas para a Web.

Yamada et al. [2006] analisaram a estrutura da Wikipédia por meio de links. Os autores mostraram que a estrutura de links da Wikipédia torna-se mais densa conforme aumenta o número de artigos. Segundo os autores, o comportamento dos links entre os artigos varia de acordo com a categoria ao qual eles pertencem, com alguns nós (artigos) que são mais interconectados do que os outros (normalmente artigos que interconectam muitos artigos, como artigos de Categoria, os quais apontam para todos os outros artigos relacionados à sua categoria). Por causa disso, nesse trabalho foram utilizadas as métricas de análise de links (PageRank, HITS e Indegree) como medidas de centralidade, tendo elas sido aplicadas nas redes de categorias. As páginas que obtiveram melhores posições no PageRank foram as páginas de conceitos fundamentais da Wikipédia (artigos informativos), como as páginas de “Categorias”, “Fundamentos”, “Wikiportais”, “Culturas” e “Humanidade”, mostrando que o PageRank é capaz de mostrar a estutura básica de categorias da Wikipédia, o que foi muito dife-

rente do resultado apresentado pelo HITS, que trouxe como melhores posições páginas relacionadas à músicas, “Álbum por artista”, “Álbuns americanos”, “Álbuns canadenses”, “Álbuns de Rock Alternativos” e “Álbuns britânicos”. Todas as 50 primeiras páginas retornadas pelo HITS estavam fortemente relacionadas à primeira página (“Álbum por artista”). Esse trabalho reforça a ideia de que métricas de análise de links podem ter comportamentos distintos de acordo com os tópicos dos documentos analisados.

3.4 Considerações finais

Neste capítulo, foram apresentados trabalhos que se relacionam ao nosso. Dentre esses trabalhos, foram discutidos artigos que focam em estimativa automática da qualidade de artigos da Wikipédia, explorando diversas características, entre elas, os links internos da Wikipédia. Diferentemente destes trabalhos, em nossa pesquisa, fizemos uso de links internos e externos. Dado que nosso objetivo era verificar métricas que fossem capazes de inferir os níveis da qualidade da Wikipédia, além de importância e popularidade, a pesquisa diferenciou-se desses trabalhos também por não possuir enfoque em encontrar melhores formas de estimar a qualidade dos artigos. Também foram apresentados trabalhos que aplicaram análise de links na Wikipédia ou que fizeram uso da Análise de Links para discutir conceitos de qualidade de páginas. Esses trabalhos foram importantes por fornecer embasamento teórico para a nossa pesquisa e pudemos comprovar alguns de seus resultados. Por fim, foram apresentados trabalhos que compararam a estrutura de links da Wikipédia com a da Web. Esses trabalhos são importantes, uma vez que decidimos aplicar na Wikipédia técnicas de Análise de Links que são comumente utilizadas na Web. Essa revisão bibliográfica nos mostrou que muitos aspectos da nossa proposta não foram investigados anteriormente, o que nos permitiu concluir que a pesquisa realizada pode fornecer contribuições para esse tema de pesquisa.

Capítulo

4

Coleções e Fatores utilizados

Neste capítulo, são apresentadas as coleções usadas neste estudo. Em particular, essas coleções são descritas e comparadas. Além disso, são apresentadas caracterizações detalhadas das informações de interesse que são objetos de estudo deste trabalho: qualidade, importância e popularidade.

4.1 Coleções obtidas

Para a realização deste trabalho, foram usadas duas coleções. A primeira, chamada de WikiPt neste trabalho, foi a Wikipédia, usada como fonte de artigos para os quais informações de qualidade, de importância e de popularidade puderam ser obtidas. A segunda foi uma amostra representativa da Web brasileira, referenciada neste trabalho como WBR10. Como a Wikipédia possui links externos à Wikipédia, ela fornece uma estrutura de links em torno da Wikipédia que possibilita o uso de métricas complexas como Pagerank, além de métricas simples baseadas em informação externa à Wikipédia, como métricas derivadas do Outdegree.

A WBR10 é um base coletada por pesquisadores do projeto INWEB[1], composta por documentos do domínio *br*. Essa coleção foi obtida durante o final do mês de Setembro e início do mês de Outubro de 2010, por um período aproximado de 7 dias. Em sua forma original, ela contém 125 milhões de documentos, com 6,8 bilhões de links entre eles, o que correspondia à maior amostra da Web brasileira disponível para estudo em 2012.

[1] http://www.inWeb.org.br

	Quantidade de páginas	Quantidade de artigos/páginas extraídos(as)
WBR10	126.985.671	123.668.772 (Wbr10)
WikiPt	2.363.341	611.811 (Wpt10)
Total	129.349.012	124.300.583

Tabela 4.1: *Quantidade de documentos existentes nas coleções obtidas (WBR10 e WikiPt) e quantidade de documentos extraídos (Wbr10 e Wpt10).*

A coleção WBR10 consiste de vários arquivos, em formato proprietário, usados para armazenar informações como o título da páginas, suas URLs, outlinks, textos de âncoras dos outlinks e conteúdo textual das páginas, entre outras. Dessa coleção, foram extraídas a sua estrutura de links e alguns metadados de interesse, o que resultou em 750 GB de dados. A manipulação eficiente desses dados foi um dos desafios deste trabalho.

No caso da coleção WikiPt, foi utilizada uma cópia completa da Wikipédia. A Wikipédia disponibiliza para o público cópias livres de todo o seu conteúdo[2] em vários idiomas. Como as páginas da coleção WBR10 correspondem a uma coleta de 2010, foi obtida uma cópia da Wikipédia do mesmo período. Em particular, a cópia adquirida foi a do dia 26 de setembro de 2010, que possui 2.363.341 páginas[3] e 20.311.293 revisões de artigos. Nessa coleção é possível encontrar informações referentes a artigos, ao seu histórico de edição e a discussões relacionadas a ele. Dentre as informações dos artigos, estão o seu conteúdo textual, seu título, Wikiprojetos aos quais ele pertence, suas notas de importância para os Wikiprojetos, sua qualidade, páginas internas e externas que o artigo referencia, entre outras. Podemos notar que, com exceção da popularidade do artigo, todos os elementos fundamentais para a execução da investigação reportada nessa dissertação puderam ser extraídos diretamente dessa base.

Assim como na coleção WBR10, a manipulação dos dados da WikiPt não foi trivial, uma vez que não foi possível obter documentação que orientasse a extração das informações de qualidades, de importâncias e URLs dos artigos. Para a extração dos dados recorremos a softwares disponíveis na Internet como MWDumpper[4] e Wikipédia Extractor[5], além de desenvolvermos parsers próprios.

Para a realização de estudos com as coleções WBR10 e WikiPt, foi necessário combiná-las visto que a WBR10 não inclui páginas da Wikipédia, uma vez que sua coleta foi restrita ao domínio *br* e a Wikipédia pertence ao domínio *org*. Da coleção WBR10, 123.668.772 páginas foram selecionadas, visto que foram

[2] http://en.wikipedia.org/wiki/Wikipedia:Database_download

[3] É importante informar que, quando nos referimos à Wikipédia neste trabalho, o termo "páginas"se refere a todos os tipos de páginas existentes na Wikipédia, como páginas de discussões, páginas informativas e, inclusive, páginas de artigos

[4] http://www.mediawiki.org/wiki/Manual:MWDumper

[5] http://medialab.di.unipi.it/wiki/Wikipedia_Extractor

	Quantidade de links para Wbr10	Quantidade de links para Wpt10	Total de links realizados
Wbr10	4.471.078.526	48.629	4.471.127.155
Wpt10	49.184	22.108.902	22.158.086

Tabela 4.2: *Quantidade de links existentes nas coleções extraídas (Wbr10 e Wpt10).*

		Min	Max	Média	Mediana	Desvio Padrão
Wbr10	Indegree	0	6.594.549	36,14	1	1.205,32
	Outdegree	0	11.973	36,14	19	81,73
	Tamanho das páginas	0	10.428.864	4.270,26	2.055	39.204,40
Wpt10	Indegree	0	113.330	36,23	3	503,29
	Outdegree	0	2.488	36,19	17	72,09
	Tamanho dos artigos	0	5.760.990	53.699,80	30.043	98.484,50

Tabela 4.3: *Estatísticas de links das coleções Wbr10 e Wpt10.*

identificadas réplicas de páginas dentro da coleção[6]. Da coleção WikiPt foram selecionadas somente páginas caracterizadas como artigos, o que resultou em 611.811. Chamaremos a coleção extraída da WBR10 de Wbr10 e a coleção extraída da WikiPt de Wpt10. A quantidade de páginas coletadas e extraídas são sumarizadas na Tabela 4.1. Caracterizações e comparações das coleções extraídas serão apresentadas a seguir.

4.1.1 Análise Comparativa das Estruturas de Links

A estrutura dos links relacionados a todas as páginas da Wbr10 e da Wpt10 compõem o grafo de páginas $\mathcal{H}_p$. É importante notar que links externos às duas coleções não fazem parte do grafo.

Das 124.300.583 de páginas que compõem $\mathcal{H}_p$, 99,5% são documentos da Wbr10, o que corresponde a uma rica quantidade de informações disponíveis para a aplicação de métricas que necessitam de informações externas à Wpt10, como a métrica PageRank. Os 123.688.772 documentos da Wbr10 que compõem $\mathcal{H}_p$ fazem 4.471.127.155 de links em $\mathcal{H}_p$, sendo que somente 48.629 links são para páginas na coleção Wpt10. Os artigos da Wpt10, por sua vez, realizam 22.158.086 de links, dos quais 22.108.902 são para artigos da própria Wpt10. Essas informações sobre quantidade de links existentes nas duas coleções podem ser vistas na Tabela 4.2.

Na Tabela 4.3 são apresentadas estatísticas relacionadas com inlinks (Indegree), outlinks (Outdegree) e o tamanho das páginas (em caracteres) de ambas as coleções.

Como apresentado na Tabela 4.3, as páginas da Wbr10 têm um tamanho

[6]Foram consideradas réplicas as páginas que, após a normalização, possuíam a mesma URL. Para normalizar as URLs, utilizou-se a RFC-3986 (cf. http://tools.ietf.org/html/rfc3986). Entre os tratamentos realizados, citamos a eliminação de partículas como “http”, “www” e “\..\”, provenientes de diretórios, e a substituição de caracteres em maiúsculo por minúsculo.

médio de cerca de 4.270 caracteres (mediana 2.063), e os artigos da Wpt10, mais longos, possuem em média 53.709 caracteres (mediana 30.043). Esse fato se deve, em grande parte, às regras da Wikipédia que exigem que os artigos devem fornecer informação completa sobre os seus tópicos. Como esperado, valores máximos de Outdegree e Indegree são maiores na Wbr10, uma vez que a mesma possui uma ordem de grandeza maior que a Wpt10. Ambas as coleções têm um número médio de aproximadamente 36 inlinks e outlinks por documento. Contudo, a mediana de inlinks na Wpt10 (3) é três vezes maior que na Wbr10a (1), o que sugere que a Wpt10 é mais densamente ligada.

A maior densidade de links na Wikipédia é confirmada por várias outras caracterizações estatísticas reportadas na literatura, por exemplo, no trabalho de Kamps and Koolen [2009]. A alta densidade está associada a dois motivos principais. Primeiro, as diretrizes da Wikipédia definem claramente quando e para onde devem ser feitas ligações, incentivando que elas sejam para dentro da própria Wikipédia (o que fornece maiores informações sobre o assunto tratado) ou para páginas externas que comprovem a veracidade do artigo (Enciclopedismo, ver Seção 2.4.2). Além disso, robôs constantemente verificam links que faltam, de forma que estes são inseridos automaticamente em algum momento. Como resultado da grande densidade de links, ao contrário de uma coleção Web comum, o grafo da Wpt10 consiste em único componente, ou seja, um grande conjunto de páginas conectadas.

A Figura 4.1 mostra a distribuição de tamanhos das páginas da Wbr10 e dos artigos da Wikipédia, além dos seus valores de Indegree e Outdegree. Em particular, as Figuras 4.1(c)–(f) apresentam o número de inlinks e outlinks das coleções. Mais precisamente, foram contados o número de páginas únicas que apontam para uma certa página ou que são apontadas por aquela página.

Nas Figuras 4.1(c)–(f) é possível observar que todas as distribuições seguem uma lei de potência. Na Wbr10, essa distribuição é observada com mais clareza para valores de Indegree que para valores de Outdegree. Na Wpt10, as distribuições apresentam-se mais suaves. A diferença entre as distribuições de inlinks e outlinks é menor na Wpt10 que na Wbr10, o que sugere que outlinks se comportam como inlinks na Wikipédia, algo também destacado por Kamps and Koolen [2009]. Isso se deve provavelmente à natureza semântica dos links na Wikipédia onde, se um link de p_1 para p_2 significa que p_2 é relevante para p_1, espera-se que p_1 também seja relevante para p_2. Considerando as Figuras 4.1(a)–(b), que apresentam as distribuições de tamanho dos documentos nas duas coleções, é possível observar que, ao contrário do que ocorre com inlinks e outlinks, nenhuma das distribuições parece seguir uma lei de potência.

A Tabela 4.4 apresenta as correlações entres as três medidas avaliadas nessa

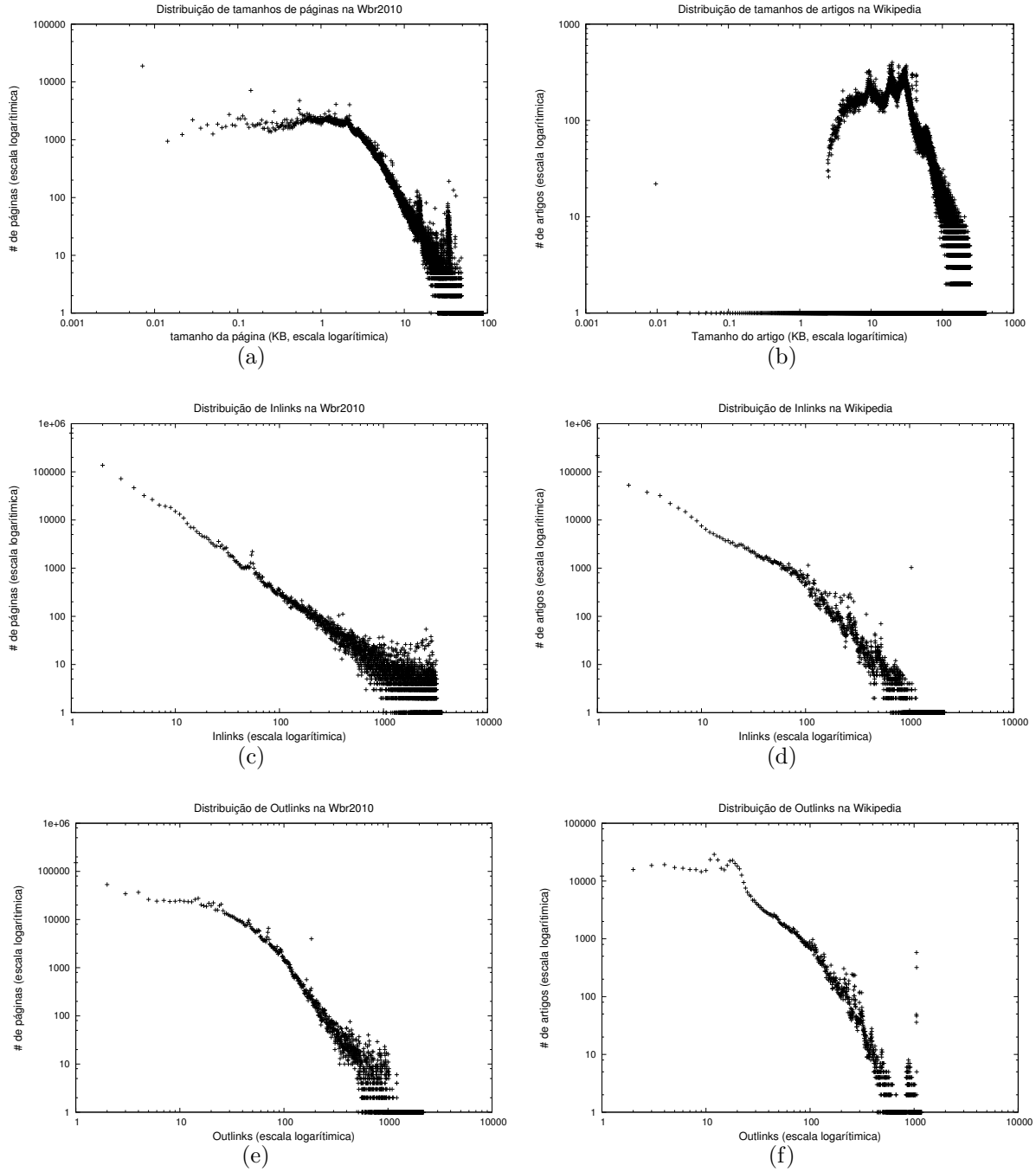

Figura 4.1: *Distribuição de tamanhos de páginas para Wbr10 (a); Distribuição de tamanhos de artigos para Wpt10 (b); Distribuição de inlinks para Wbr10 (c); Distribuição de inlinks para Wpt10 (d); Distribuição de outlinks para Wbr10 (e); Distribuição de outlinks para Wpt10 (f).*

	Indegree x Outdegree	Indegree x Tamanho dos documentos	Outdegree x Tamanho dos documentos
Wbr10	0.310	-0.005	-0.004
Wpt10	0.468	0.2555	0.444

Tabela 4.4: *Correlações entre valores de Indegree, Outdegree e tamanho das páginas/artigos das coleções Wbr10 e Wpt10.*

	Quantidade existente	Quantidade de documentos conectados	Total de links realizados
$\mathcal{H}_p$	123.688.772	123.688.772	4.471.078.526
$\mathcal{H}_h$	444.503	74.898.838	616.869.548
$\mathcal{H}_d$	325.638	64.044.188	264.169.559

Tabela 4.5: *Dados referentes aos grafos utilizados nos experimento $\mathcal{H}_p$, $\mathcal{H}_h$, $\mathcal{H}_d$. A obtenção dos grafos considera documentos das coleções Wbr10 e Wpt10).*

análise para as coleções Wbr10 e Wpt10. As correlações foram obtidas aplicando a métrica de correlação Kendall τ em 20.000 documentos selecionados aleatoriamente de cada coleção. Observando a Tabela 4.4, é possível comprovar que na Wpt10 o comportamento dos inlinks e dos outlinks são semelhantes. Além disso, é possível observar uma relação clara entre a quantidade de outlinks dos artigos da Wikipédia com os tamanhos dos artigos, o que não ocorre na Wbr10. Essa correlação do valor de Outdegree com o tamanho dos documentos dos artigos na Wikipédia pode ser explicada pelo fato de que as regras da Wikipédia obrigam que artigos criem links internos para outros artigos relacionados aos seus. Esses links são realizados ao longo do conteúdo do artigo. Assim, quanto maior o conteúdo, maior é a necessidade de criar links para outros artigos.

4.1.2 Hosts e Domínios

A implementação das variantes das métricas PageRank, Outdegree e Indegree requer que dois grafos sejam implementados além de $\mathcal{H}_p$. Para as variações baseadas em hosts, o grafo $\mathcal{H}_h$ deve desconsiderar links entre páginas que possuam o mesmo host. Por sua vez, as variações baseadas em domínios utilizam um grafo $\mathcal{H}_d$, o qual desconsidera links entre páginas do mesmo domínio. Da mesma forma que é adquirido $\mathcal{H}_p$, esses grafos são obtidos considerando todas as páginas existentes na coleção Wbr10 e todos os artigos existentes na coleção Wpt10.

A Tabela 4.5 apresenta a quantidade de nós (documentos) que compõem os três grafos utilizados. É possível observar que, quando utilizamos grafos com eliminação de links para páginas com mesmos hosts ou domínios, essa quantidade diminui quase pela metade quando comparada com o grafo sem tratamentos ($\mathcal{H}_p$). A quantidade de links trabalhados nos grafos de hosts e domínios é menor também. O número de links trabalhados nos grafos derivados equivaleram a menos de 15% do total de links utilizados em $\mathcal{H}_p$. Como consequência, a utilização de grafos menores implica em processamento mais rápido dos dados e menor espaço físico para armazenamento.

4.2 Qualidade, importância e popularidade na Wikipédia

Inicialmente, o objetivo principal deste trabalho é analisar a correlação entre as métricas de análise de links com a qualidade dos conteúdos dos artigos da Wi-

	Q1	Q2	Q3	Q4	Q5	Q6	Total
Wpt10	78.822	6465	3218	265	153	30	88.961
Wpt10a	1.997	996	500	265	147	30	3.935

Tabela 4.6: *Distribuições de qualidade dos artigos da Wikipédia para as coleções Wpt10 e Wpt10a.*

kipédia. Uma vez que também é possível obter dados de importância junto aos Wikiprojetos aos quais os artigos pertencem, bem como valores de popularidade desses artigos, foi possível estender o estudo e verificar as correlações que podiam ser encontradas também para esses fatores. A forma como os dados foram obtidos e as estatísticas sobre eles são apresentadas a seguir.

4.2.1 Qualidade

Dos 611.811 artigos existentes na Wpt10, apenas 14,5% possuíam nota de qualidade. Foram consideradas notas de qualidade válidas qualquer valor atribuído de 1 a 6, ou seja, não foram designadas notas para artigos que apresentavam caracteres especiais, como “?", ou valor 0 como valor de qualidade.

Como mostra a Tabela 4.6, mais de 88,5% dos artigos com qualidade possuem qualidade 1, enquanto os artigos com a maior qualidade (6) equivalem a menos de 0.05% do total. Podemos inferir que quanto maior é a o grau de qualidade de uma página, menor é a sua quantidade de artigos na base, o que é esperado visto que o processo de avaliação se torna mais preciso e a exigência sobre o conteúdo do artigo é maior.

Dentre os artigos com qualidade anotada, optou-se por extrair uma amostra para a obtenção de rankings. Contudo, dada a distribuição de classes de qualidade desbalanceada da coleção original, uma amostra aleatória estratificada seria formada por mais de 88,5% de documentos de qualidade 1. Quaisquer rankings gerados com tal amostra seriam muito semelhantes a um ranking ideal, dado o grande número de empates. Assim, optou-se por uma amostra menor, porém mais diversificada, que facilitasse as análises. Por esse motivo, procuramos utilizar todos os artigos de qualidades 6, 5 e 4, dado o seu pequeno número, e amostras aleatórias dos artigos nas classes 1, 2 e 3, respeitando uma lei de potência mais suave que a observada na coleção original. Como resultado, a amostra usada consistiu de 3935 artigos, com distribuição de classes de qualidade apresentada na Tabela 4.6. Atribuímos o nome de Wpt10a a essa amostra.

Com a finalidade de contextualizar quais artigos compõem cada classe de qualidade, a Tabela 4.7 apresenta exemplos de títulos de artigos e suas notas de qualidade (*Qua*).

Qua	Título
1	Paisagem cultural de Sukur
1	Igreja Matriz de Vila de Frades
2	Pinhalzinho (Santa Catarina)
2	Fábrica de porcelanas Allach
3	Região metropolitana de Natal
3	Bahamas nos jogos olímpicos de verão de 2008
4	Monobloco (Banda Musical)
4	Segunda Guerra Mundial
5	Harvey Milk
5	Alemanha
6	Rio de Janeiro (cidade)
6	Xadrez

Tabela 4.7: *Exemplos de títulos de artigos por classe de qualidade.*

4.2.2 Importância

Diferente do que acontece com as notas de qualidade, as notas de importância dos artigos são fornecidas pelos seus Wikiprojetos (ver Seção 2.4.4). Dessa forma, um artigo pode ter mais do que uma nota de importância. Visto que os Wikiprojetos se referem a temas diferentes (eg. Astronomia, Brasil, Geografia, 2a Guerra Mundial, Rock, etc.), neste trabalho eles serão tratados como "categorias". A maioria das categorias estavam relacionadas com os temas "Música"e "Geografia". Foram encontradas 133 categorias na Wpt10, sendo que os artigos existentes em Wpt10a estão incluídos em 104 delas. Apenas artigos com nota de qualidade apresentaram qualquer nota de importância. O número de artigos com notas de importância foi 33.995, representando 5,5% dos artigos da nossa coleção Wpt10 e aproximadamente 38% dos artigos com qualidade. Dentre os artigos selecionados para a amostra Wpt10a, 2.304 artigos possuem nota de importância, o equivalente a cerca de 60% da amostra.

Devido à possibilidade de um artigo possuir mais do que uma nota de importância, consideramos Imp como sendo a média aritmética de todas as notas de importância atribuídas ao artigo em todos os Wikiprojetos. Optou-se por calcular Imp da maneira descrita pelo fato de haver somente uma nota de importância para maior parte dos artigos. Para os casos nos quais haviam mais de uma avaliação, apenas aproximadamente 50 artigos possuíam notas de importância diferentes. E destas, apenas em 2 casos, a diferença era maior que 1. Os experimentos realizados para avaliar importância levaram em consideração Imp. Na Tabela 4.8, são apresentadas as distribuições encontradas para Imp agrupadas de acordo com notas de qualidade. Pelo fato de Imp não apresentar somente valores discretos, os valores apresentados para Imp na Tabela 4.8 estão arredondados.

		Imp					
		-	1	2	3	4	total
Qua	1	1344	107	204	320	22	1997
	2	198	87	364	193	154	996
	3	36	62	208	90	104	500
	4	28	11	21	71	134	265
	5	23	10	21	45	48	147
	6	2	1	5	8	14	30
	total	1631	278	823	727	476	3935

Tabela 4.8: *Distribuição dos níveis de qualidade (Qua) e importância (Imp) em Wpt10a.*

IR	Título
1	Eremospatha
1	Canoagem nos jogos Pan-Americanos de 2007 -k-2500m masculino
2	Cordilheira Alta
2	Xadrez por computador
3	Anarquismo em Cuba
3	Jornalismo
4	Administração
4	Guerra Fria

Tabela 4.9: *Exemplos de títulos de artigos por classe de importância.*

A Tabela 4.8 mostra que, para a amostra do Wpt10a, a maioria dos artigos de boa qualidade (artigos 5 e 6) possui boa nota de importância (4). O oposto ocorre com artigos com qualidade 1, estando mais distribuídos pelas outras classes de importância e, em sua maioria, sem valor algum. Esse resultado está de acordo com o esperado, uma vez que se espera que páginas com conteúdo de qualidade sejam páginas que abordem assuntos importantes sobre as categorias as quais ela se refere.

A Tabela 4.9 apresenta alguns exemplos de artigos para cada classe de importância. Como é possível observar na Tabela 4.9, à medida em que o grau de importância aumenta, aumenta também a abrangência dos assuntos dos artigos.

4.2.3 Popularidade

A Wikipédia originalmente registrava o número de visualizações de uma página em sua base de dados. Contudo, para evitar um acesso ao banco de dados cada vez que um usuário solicitava uma página, este registro passou a ser feito como um serviço separado da base principal, que pode ser acessado através da ferramenta “Wikipedia article traffic statistics"[7]. Usando esse serviço, foi possível obter uma estimativa para a popularidade dos artigos da Wpt10a.

[7] http://stats.grok.se/

		Grupos de popularidade					
		G1	G2	G3	G4	G5	Total
Qua	1	321	60	658	678	280	1997
	2	322	332	50	90	202	996
	3	117	181	61	49	92	500
	4	40	105	50	21	49	265
	5	25	87	15	9	11	147
	6	3	14	5	4	4	30
	total	828	779	839	851	638	3935

Tabela 4.10: *Distribuição de qualidade (Qua) entre os grupos de popularidade selecionados.*

Especificamente, as popularidades foram coletadas como a média mensal das visitas aos artigos entre os meses de Janeiro a Setembro de 2010. Para os casos de artigos criados nesse período, foram considerados todos os seus meses de vida. Nenhum artigo tinha menos que dois meses de vida. Optamos por verificar a quantidade de visitas em mais de um mês para evitar vícios relacionados a aumentos repentinos de interesse causados por eventos incomuns ou sazonais (e.g. aniversário de uma cidade, lançamento de algum filme, eventos musicais envolvendo artistas famosos, etc.). Devido ao custo de obter essas informações, foram coletadas a quantidade de visitas apenas dos artigos da amostra Wpt10a. As popularidades obtidas (*Pop*) para os 3.935 artigos foram divididas em 5 grupos. Procuramos colocar uma quantidade balanceada de artigos em cada grupo, os quais serão descritos a seguir:

- G1: $0 \leq Pop <$ 300, totalizando 828 artigos;
- G2: $300 \leq Pop <$ 600, totalizando 779 artigos;
- G3: $600 \leq Pop <$ 20.000, totalizando 839 artigos;
- G4: $20.000 \leq Pop <$ 30.000, totalizando 851 artigos;
- G5: $Pop >$ 30.000, totalizando 628 artigos.

A Tabela 4.10 apresenta a distribuição de qualidade em cada um dos grupos de popularidade.

A Tabela 4.11 apresenta os 20 artigos com maior numero de visualizações durante os meses de Janeiro a Setembro de 2010, juntamente com os seus números de visitas e suas notas de qualidade e importância. É possível observar nessa tabela que dos 20 artigos mais populares, 6 possuem boas notas de qualidade (5 e 6). Sobre as 5 primeiras posições, os artigos Mitologia Grega e Portugal são atualmente (Janeiro/2013) artigos destacados, possuindo qualidade 6. O artigo Lady Gaga possui atualmente qualidade 5. O aumento das notas

	Título	*Pop*	*Qua*	*Imp*
1	Brasil	447.471,0	6	4
2	Portugal	151.029,0	3	4
3	Lady Gaga	136.119,0	3	3
4	Segunda Guerra Mundial	120.679,0	4	4
5	Mitologia Grega	114.301,0	5	4
6	Luiz Inácio Lula da Silva	104.099,0	3	4
7	Região Nordeste di Brasil	86.022,2	4	4
8	Europa	83.116,8	3	4
9	Revolução Francesa	81.614,3	5	4
10	Espanha	79.282,1	2	4
11	São Paulo (cidade)	76.175,9	6	4
12	Alemanha	74.957,6	5	4
13	Guerra Fria	69.598,0	3	4
14	França	69.204,1	3	4
15	Administração	68.927,1	2	4
16	Adolf Hitler	68.798,2	4	4
17	Itália	67.773,0	4	4
18	Futebol	64.216,4	3	4
19	Região Norte do Brasil	62.580,0	4	4
20	Japão	61.574,9	5	4

Tabela 4.11: *Títulos dos 20 artigos mais populares de Wpt10a e suas respectivas médias de visitas, notas de qualidade e notas de importância.*

de qualidade de quando a coleção foi obtida em relação a versão da Wikipédia disponível atualmente na Internet pode indicar que páginas mais visualizadas tendem a atrair a atenção dos editores para melhorar o conteúdo do artigo.

4.3 Considerações finais

A obtenção de informações a respeito de links internos e externos à Wikipédia, bem como acerca de avaliações de qualidade, de importância e de popularidade, objetos de estudo deste trabalho, é essencial para a realização deste trabalho. As informações de links puderam ser capturadas utilizando duas bases, uma contendo páginas não pertencentes à Wikipédia e outra contendo páginas internas à Wikipédia e fatores de interesse utilizados para realizar as comparações. As coleções adquiridas, os dados extraídos e a amostra utilizada nos experimentos foram detalhadas nesse capítulo. Também foram apresentadas estatísticas e caracterizações dos dados trabalhados. A compreensão dos dados é fundamental para o entendimentos e discussão dos resultados no Capítulo 5.

Capítulo

5

Experimentos realizados

Neste capítulo, apresentamos os experimentos realizados e seus resultados. As métricas de análise de links avaliadas foram as métricas clássicas Indegree, Outdegree e PageRank, e algumas de suas variações considerando hosts e domínios. Detalhes sobre essas métricas foram apresentados na Seção 2.2.3. Informações sobre as coleções e fatores utilizados para a realização dos experimentos foram expostas no Capítulo 4.

5.1 Metodologia

Inicialmente, foram implementados algoritmos para computar as onze métricas descritas na Tabela 2.2 (pag. 16), Indegree, Outdegree e Pagerank de páginas (In, Out e PR), Indegree, Outdegree e Pagerank de hosts (InH, OutH e PRH), Indegree, Outdegree e Pagerank de domínios (InD, OutD e PRD), Hiper-Indegree de hosts (HInH) e Hiper-Indegree de domínios (HInD). Em particular, no caso da métrica Pagerank e suas variantes, foi utilizada a implementação de Kyrola et al. [2012]. Essa implementação[1] foi escolhida por processar a métrica do PageRank em computador isolado durante um tempo satisfatório, considerando a grande quantidade de documentos trabalhados nessa pesquisa.

Uma vez implementadas as métricas, elas foram usadas para ranquear os artigos da coleção Wbr10a. As listas ordenadas de artigos foram então comparadas com as listas ordenadas de acordo com a avaliação humana de qualidade e importância, e a estimativa de popularidade baseada em visitas às páginas. Devemos observar que os valores dos métodos Indegree e Outdegree foram calcula-

[1] https://github.com/bmabey/graphchi/blob/master/example_apps/pagerank.cpp

dos para todas as páginas de Wbr10 e para todos os artigos de Wpt10, uma vez que os mesmos foram utilizados para fazer a comparação entre essas coleções descrita na Seção 4.1.1, e da, mesma forma, os valores da métrica PageRank foram calculados para todas os documentos de ambas as coleções, visto que, segundo essa métrica, a pontuação de um documento é influenciada pela pontuação dos documentos aos quais ele se conecta. Porém, para os experimentos, somente foram analisadas as pontuações obtidas para os artigos selecionados em Wpt10a.

Para analisar a correlação entre as listas ordenadas obtidas, foi computado o método Kendall τ (ver Seção 2.3.2). Para obter os valores das correlações, utilizamos a implementação disponibilizada pelo ambiente R [2]. Esse método foi escolhido por considerar empates nas posições dos rankings, uma vez que empates são muito comuns na ordenação dos fatores qualidade e importância. Para o fator qualidade, isso ocorre porque as notas de qualidade possuem número discretos de 1 a 6. Embora esse não seja o caso das notas de importância, visto que essas foram obtidas como média aritmética das notas atribuídas por diferentes Wikiprojetos (ver Seção 4.2.2), empates são quase igualmente frequentes. Finalmente, para o fator popularidade, empates são incomuns visto que a estimativa de popularidade foi obtida como a média das visitas mensais que o artigo obteve em um período de até nove meses (ver Seção 4.2.3).

Considerando que $C(R_x, R_y)$ é a correlação existente entre os resultados de um ranking R_x e de um ranking R_y, adotamos 3 graus de correlação em nossa análise:

- Forte:0.7 < $C(R_x, R_y) \leq$ 1.0 (rankings diretamente correlacionados) ou -0.7 > $C(R_x, R_y) \geq$ -1.0 (rankings inversamente correlacionados);
- Moderada: 0.3 $\leq C(R_x, R_y) \leq$ 0.7 (rankings diretamente correlacionados) ou -0.3 $\geq C(R_x, R_y) \geq$ -0.7 (rankings inversamente correlacionados);
- Fraca: 0.0 $\leq C(R_x, R_y) <$ 0.3 (rankings diretamente correlacionados) ou 0.0 > $C(R_x, R_y) >$ -0.3 (rankings inversamente correlacionados).

Para analisar diferentes aspectos destas correlações, várias sub-amostras de Wpt10a foram criadas. Ao longo deste capítulo, elas serão referenciadas pelos seguintes nomes:

1. Wpt10a: amostra original da Wikipédia com 3935 artigos. Todos os artigos nesta amostra possuem informação de qualidade e popularidade. Nela, os fatores qualidade e popularidade são moderadamente correlacionados (Kendall τ = 0.567);

[2] http://www.r-project.org/

	In	Out	PR	InH	OutH	PRH	InD	OutD	PRD	HInH	HInD
Qualidade	0.444	**0.516**	0.468	0.451	**0.396**	0.450	0.452	**0.396**	0.451	0.461	0.463
Popularidade	**0.467**	0.406	**0.558**	**0.561**	0.356	**0.553**	**0.561**	0.356	**0.553**	**0.574**	**0.573**

Tabela 5.1: *Valores de Kendall τ para a amostra Wpt10a (artigos com e sem importância).*

2. Wpt10i: sub-amostra da Wpt10a, composta por 2304. Todos os artigos nesta amostra possuem informação de qualidade, popularidade e importância. Nela, todos os fatores são fracamente correlacionados (Kendall τ = 0.263 para popularidade e importância; 0.204 para qualidade e popularidade; e 0.132 para qualidade e importância);

3. Wpt1234x56: amostra Wpt10a com duas classes de qualidade. A primeira corresponde às classes 5 e 6 originais, e a segunda às classes 1 a 4. Os artigos nas classes 5 e 6 são aqueles que passaram por um rigoroso processo de avaliação, enquanto os que receberam notas entre 1 e 4 podem ter sido avaliados de forma isolada e mais superficial;

4. Wpt1x23456: amostra Wpt10a com duas classes de qualidade. A primeira corresponde às classes 2 a 6 originais, e a segunda à classe 1. Esta divisão foi feita para separar a classe 1, que corresponde à maioria dos artigos avaliados na Wikipédia. Esta avaliação pode ser compreendida como uma indicação explícita de que um documento ainda é um esboço ou representa apenas um conteúdo mínimo que precisa ser verificado;

5.2 Resultados e discussão

Nesta seção descrevemos os resultados obtidos, ou seja, os valores de correlação, de acordo com a métrica de correlação de rankings Kendall τ, entre as métricas de análise de links e os fatores qualidade, popularidade e importância.

5.2.1 Correlação entre Métricas e Fatores

A Tabela 5.1 apresenta os valores de Kendall τ obtidos quando correlacionadas as métricas de análise de links com os fatores qualidade e popularidade na Wpt10a, a amostra maior. Para cada métrica, está destacado em negrito os fatores de maior correlação.

Como podemos observar, em geral, as métricas apresentaram maior correlação com o fator popularidade que com o fator qualidade. Essa observação não foi válida apenas para Outdegree e suas variações. Contudo, não há correlações fortes em nenhum caso, sendo todas elas moderadas. A maior correlação com o fator qualidade foi obtida por Outdegree de páginas (Out).

O resultado obtido pelo Out é esperado uma vez que a Wikipédia recomenda que artigos façam referências a páginas externas que comprovem o conteúdo

dos artigos ou páginas internas que dêem ao leitor maior contexto. Essa recomendação é levada em consideração nas avaliação de qualidade dos artigos da Wikipédia. O resultado obtido por Out também pode ser relacionado a sua correlação com o tamanho do artigo (Kendall τ = 0.555). De acordo com a literatura, o tamanho do artigo é altamente correlacionado com a qualidade [Blumenstock, 2008]. Em particular, essa afirmação também foi observada na amostra Wpt10a, onde a correlação entre qualidade e tamanho do artigo apresentou Kendall τ de 0.682. Esse resultado é reforçado pela observação feita na Seção 4.1.1 e destacado por outros autores [Kamps and Koolen, 2009] de que outlinks se comportam como inlinks na Wikipédia. Assim, contrário ao observado na Web, na qual métricas baseadas em inlinks são dominantes, para o caso da Wikipédia, métricas baseadas em outlinks podem ser igualmente úteis.

Também podemos observar na Tabela 5.1 desempenhos muito similares entre as métricas de PageRank e Indegree. Esse resultado está claramente relacionado à forte correlação observada entre essas métricas (valores de Kendall τ entre 0.795 e 0.965). Vale ressaltar a coincidência entre os desempenhos obtidos por Indegree, Pagerank e suas variantes e a correlação entre eles. Esse é um resultado importante, pois o Indegree é uma métrica que demanda uma quantidade menor de recursos para ser calculada. Esse resultado também confirma observações feitas anteriormente para a Web [Amento et al., 2000; Berlt et al., 2010].

De maneira geral, observamos que variantes do Indegree baseadas em hosts e domínios foram mais correlacionadas à qualidade que o Indegree tradicional. Além disso, variantes baseadas em Hipergrafo, propostas por Berlt et al. [2010], obtiveram algumas das maiores correlações com o fator qualidade. Esses resultados reforçam as observações feitas para a Web pelos mesmos autores.

Apesar de a maior correlação com qualidade ter sido da métrica Outdegree, suas variantes foram as que tiveram menor correlação com qualidade. Esse resultado pode ser explicado pelo fato de a maioria dos links na Wikipédia apontarem para a própria Wikipédia e, portanto, haver pouca variedade de hosts e domínios. No caso dos experimentos realizados neste trabalho, esse problema é maior ainda uma vez que na coleção da Wikipédia usada poucos artigos externos estão na Wbr10, conforme apresentado na Tabela 4.2 (pag. 41). Isso ocorre porque muitos artigos da Wikipédia em português são traduções de suas versões em inglês. Assim, muitos links externos não fizeram parte dos nossos cálculos. Como resultado, ao utilizar os grafos de hosts e domínios, a maioria dos outlinks existentes foi descartada. O fato de os outlinks descartados para os grafos de hosts e domínios serem os mesmos influenciou para que os resultados das variações de Outdegree nesses casos fossem semelhantes.

	In	Out	PR	InH	OutH	PRH	InD	OutD	PRD	HInH	HInD
Popularidade	**0.376**	**0.347**	**0.529**	**0.603**	**0.379**	**0.589**	**0.603**	**0.379**	**0.589**	**0.616**	**0.615**
Importância	0.309	0.238	0.339	0.233	0.105	0.226	0.234	0.105	0.226	0.230	0.228
Qualidade	0.100	0.236	0.152	0.164	0.161	0.163	0.164	0.161	0.164	0.171	0.167

Tabela 5.2: *Valores de Kendall τ para a amostra Wpt10i (somente artigos com importância).*

Para o fator popularidade, podemos observar que as variantes de Outdegree também possuem as menores correlações, ainda como reflexo da pequena quantidade de links externos catalogados em nossas coleções. Assim como com o fator qualidade, os resultados para as métricas de Indegree se aproximaram dos resultados das métricas de PageRank. Contudo, visivelmente as variações de Indegree são mais correlacionadas com popularidade do que as variações de PageRank. Assim como ocorreu em qualidade, as variantes de Indegree que utilizam o conceito de hipergrafo (HInH, HInD) apresentaram bons resultados, sendo as mais correlacionadas ao fator popularidade.

Esse primeiro estudo não envolveu o fator importância, uma vez que apenas 2307 artigos da amostra Wpt10a foram rotulados quanto à importância. Para analisar esse fator, a amostra de artigos Wpt10i foi utilizada como detalhado a seguir. A Tabela 5.2 apresenta as correlações entre as métricas de links e os fatores importância, qualidade e popularidade na amostra Wpt10i.

Assim como observado na Tabela 5.1, as métricas se mostraram mais correlacionadas com o fator popularidade que com o fator qualidade. E embora elas tenham apresentado maiores correlações com importância do que com qualidade, essas correlações foram, em sua maioria, fracas. De fato, para todas as métricas, as maiores correlações foram com os fatores popularidade.

As menores correlações encontradas foram obtidas entre as métricas e o fator qualidade dos artigos. A diminuição brusca dos valores dessas correlações no segundo experimento, quando comparadas ao primeiro, resulta do uso de uma amostra menor com menos artigos de baixa qualidade. Como descrito anteriormente, a maioria dos artigos que não possuem nota de importância apresentam qualidade 1 (em particular, dos 1.633 artigos eliminados, mais de 65% possuem qualidade 1). A mudança brusca observada nas correlações indica que esses artigos têm características bem específicas em relação aos demais artigos, pois a sua eliminação afetou os resultados globais.

Apesar do valor de todas as correlações terem diminuído consideravelmente, o método Outdegree (Out) e as variações do Indegree utilizando hipergrafo (HInD e HInH) continuam sendo os mais correlacionados à qualidade. Para grafos que eliminam links para os mesmos hosts ou domínios, houve poucas variações entre as correlações das métricas de Indegree, Outdegree e PageRank.

Assim como aconteceu com o fator qualidade, os resultados das correlações

	In	Out	PR	InH	OutH	PRH	InD	OutD	PRD	HInH	HInD
Wpt10a	**0.444**	**0.516**	0.468	**0.451**	**0.396**	**0.450**	**0.452**	**0.396**	**0.451**	**0.461**	**0.463**
Wpt1x23456	**0.444**	0.476	**0.472**	0.441	0.350	0.440	0.441	0.350	0.439	0.452	0.451
Wpt1234x56	0.147	0.232	0.168	0.199	0.187	0.198	0.199	0.187	0.199	0.203	0.202

Tabela 5.3: *Valores de Kendall τ para a amostra Wpt10a (classes de qualidade original, de 1 a 6), Wpt1234x56 (duas classes: artigos avaliados de forma superficial e artigos avaliados de forma rigorosa) e Wpt1x23456 (duas classes: artigos avaliados como esboços/mínimos e artigos maiores).*

das métricas com popularidade pouco mudaram em relação à Tabela 5.1. As mais correlacionadas continuaram sendo as variações de Indegree (principalmente HInH e HInD), sendo que, nesses casos, os valores de PageRank foram sempre próximos, mas nunca superiores aos de Indegree. Quando consideramos o grafo completo de páginas, sem descarte de links, a métrica PageRank possui a maior correlação com popularidade, porém, ainda assim, sua correlação é inferior às das variações de Indegree.

Para o fator importância, os resultados foram surpreendentes. Enquanto todas as variações do Indegree foram as mais correlacionadas com popularidade, as métricas mais correlacionadas com importância foram as tradicionais (In, Out, PR), aplicadas diretamente sobre o grafo original de páginas. O melhor resultado foi o do PageRank, seguido pelo Indegree.

Na maioria dos casos analisados, os valores computados para as métricas de Indegree (In, InH, InD) foram similares aos computados para as métricas de PageRank (PR, PRH, PRD). Para cada fator, pudemos perceber uma leve diferença entre a correlação dessas métricas. Diferente do que ocorreu com o fator qualidade, para os fatores popularidade e importância tanto Out quanto as variações da métrica Outdegree (OutH e OutD) apresentaram correlação inferior às outras métricas.

5.2.2 Correlações considerando diferentes taxonomias de qualidade

Considerando o impacto que a remoção de uma classe de qualidade teve nos experimentos anteriores, em um novo experimento, verificamos como as correlações são afetadas quando diferentes taxonomias de qualidade são usadas.

A Tabela 5.3 apresenta os resultados obtidos considerando amostras com três diferentes taxonomias de qualidade, Wpt10a, Wpt1234x56 e Wpt1x23456. O primeiro caso corresponde à taxonomia original da Wpt10a com classes de qualidade de 1 a 6. No segundo caso, é utilizada uma taxonomia com duas classes: artigos avaliados de forma menos rigorosa e mais rigorosa. No terceiro caso, são utilizadas também duas classes: artigos avaliados como esboços ou mínimos e demais artigos.

Como pode ser observado na Tabela 5.3, há pouco impacto nas correlações quando a classe de qualidade 1 é separada de cada uma das outras (Wpt10a)

ou de todas elas em conjunto (Wpt1x23456). Isso sugere que grande parte da correlação observada reside na separação entre a classe de menor qualidade e as demais, o que reforça a observação feita na Tabela 5.2.

Ainda nessa tabela, é possível notar que a separação entre as classes de qualidade menos rigorosamente avaliadas das mais rigorosamente avaliadas (Wpt1234x56) provocou uma grande queda na correlação observada quando todas as classes estão separadas (Wpt10a), mostrando que as métricas avaliadas não são boas para distinguir artigos de qualidade alta dos demais. Esses resultados mostram que informações de links são pouco úteis para discernir qualidade na Wikipédia, exceto para o caso da classe 1.

5.3 Considerações Finais

Neste capítulo foram apresentados os experimentos realizados e os resultados obtidos, considerando 4 amostras: 1) Wpt10a, contendo todos os artigos e suas notas de qualidade e popularidade); 2) Wpt10i, contendo somente artigos que possuem nota de importância e suas notas de qualidade, importância e popularidade; 3)Wpt1x23456, contendo todos os artigos e suas notas de qualidade, as quais estão divididas em dois grupos, artigos de menor conteúdo e artigos de maior conteúdo; e 4)Wpt1234x56, contendo todos os artigos e suas notas de qualidade, as quais estão divididas em dois grupos, artigos mais rigorosamente avaliados e artigos menos rigorosamente avaliados.

A partir dos experimentos executados, foi possível observar que métricas de Análise de Links são mais correlacionadas com popularidade do que com qualidade e importância. Em diversos casos, métricas simples como Indegree (mais frequente) e Outdegree (apenas para qualidade dos artigos) apresentaram desempenhos comparáveis ao PageRank, métrica mais complexa. Para importância e popularidade, as variações de hipergrafo apresentaram resultados satisfatórios, podendo ser boas opções para medir esses fatores, dado que se necessita de pouco processamento para essas métricas. Detalhes das conclusões alcançadas são apresentados no Capítulo 6.

Capítulo

6

Conclusões

6.1 Caracterização da Pesquisa Realizada

Neste trabalho verificamos a "Hipótese da Qualidade", comumente usada em Análise de Links, no contexto de uma enciclopédia criada colaborativamente, a Wikipédia. Mais especificamente, foram observadas as correlações entre onze métricas de Análise de Links e três fatores usualmente associados a essas métricas, ou seja, qualidade, popularidade e importância. As métricas foram utilizadas para ordenar conjuntos de artigos. Os rankings gerados foram então comparados com rankings dos mesmos artigos, considerando os aspectos qualidade, popularidade e importância. Esta pesquisa se diferenciou das demais pesquisas relacionadas à qualidade dos artigos da Wikipédia pelo fato de não considerar somente links internos a ela, como também links externos os quais possibilitaram o uso de métricas baseadas em informação externa.

6.2 Contribuições

Entre os principais resultados deste trabalho podemos citar que, considerando a Wikipédia, métricas de Análise de Links são mais correlacionadas com popularidade do que com qualidade e importância. As correlações observadas variam de fracas a moderadas. Métricas baseadas em hosts e domínios apresentam desempenho similar, com exceção do Indegree. Métricas simples como as baseadas em Indegree e Outdegree são "competitivas" com o PageRank e artigos de baixa qualidade são determinantes para as correlações. Um resultado adicional deste trabalho foi uma comparação das estruturas de links de amostras da Wikipédia

e da Web brasileira.

Finalmente, a motivação para esta pesquisa foi sintetizada em cinco perguntas, apresentadas na Seção 1.4, para as quais obtivemos as seguintes respostas:

- *Como os fatores qualidade, importância e popularidade se relacionam entre si?* Em uma amostra, observou-se uma correlação moderada entre popularidade e qualidade. Em uma segunda amostra, observaram-se fracas correlações entre todos os fatores, com popularidade mais relacionada com importância que com qualidade. A diferença entre as amostras se deveu à remoção de um grande número de artigos de baixa qualidade.

- *Na Wikipédia, qual a relação entre esses fatores e métricas de Análise de Links?* As métricas estudadas estão mais correlacionadas com popularidade que com qualidade e importância e mais com importância que com qualidade. As correlações são moderadas para qualidade e popularidade e fracas para importância.

- *Considerando cada fator, quais as métricas mais adequadas para medi-los?* Conforme apresentado nas Tabelas 5.1 e 5.2, o fator qualidade é mais correlacionado com a métrica Outdegree (Kendal τ = 0,516), o fator popularidade com a métrica HInD (Kendal τ = 0,574) e o fator importância com a métrica Pagerank (Kendal τ = 0,339).

- *Em quais pontos os resultados obtidos durante a pesquisa divergem dos reportados na literatura?* As principais diferenças estão relacionadas com observações feitas para a Web e, portanto, esse pode ser o motivo das divergências. Considerando a literatura, observamos as seguintes diferenças em resultados: (a) métricas baseadas em hosts e domínios tiveram desempenho pior que métricas baseadas em páginas; (b) o Pagerank teve apenas um desempenho similar ao de outras métricas muito mais simples, embora seja importante ressaltar que não há tanto consenso na literatura sobre vantagens do Pagerank.

- *Em que medida, as conclusões obtidas para Wikipédia podem ser estendidas à Web em geral?* Com base na discussão apresentada no Capítulo 4, é possível dizer que há diferenças significativas na estrutura de links da Wikipédia e da Web. Entre algumas dessas diferenças podem ser citadas, em relação à Wikipédia, (a) a maior densidade de links, (b) a natureza distinta dos outlinks que parecem se assemelhar a inlinks, (c) a correlação entre outlinks e tamanho dos artigos e, provavelmente, (d) um ambiente com menor ocorrência de spam. Dadas tais diferenças, é arriscado estender observações feitas na Wikipédia para a Web, no contexto de análise de links. Ainda

assim, se considerarmos os casos onde as divergências não são grandes (por exemplo, a natureza dos inlinks) e resultados similares já foram observados para Web, pode-se afirmar que (a) métricas de menor custo, como Indegree, são tão eficazes quanto métricas complexas, como Pagerank; (b) todas estas métricas estão mais relacionadas com popularidade que com qualidade.

6.3 Dificuldades e Limitações

Uma limitação desta pesquisa foi o pequeno número de artigos anotados simultaneamente quanto à sua importância e qualidade na base da Wikipédia. Essa limitação pode em parte ser resolvida com a coleta de uma cópia mais recente da coleção, fornecendo acesso a um número maior de julgamentos de qualidade. Por outro lado, seria necessário obter também uma amostra mais recente da Web.

Uma segunda limitação da nossa abordagem reside no fato de estarmos limitados a páginas e artigos em língua portuguesa. Contudo, para trabalhar com a Wikipédia em língua inglesa e uma amostra da Web como um todo, os recursos necessários em hardware seriam consideravelmente maiores. Uma alternativa, no entanto, seria continuar este estudo sem incluir métricas complexas como o Pagerank.

Uma última dificuldade que obtivemos foi a ausência de documentação ou a desatualização da documentação disponível sobre as coleções e as suas ferramentas de apoio.

6.4 Trabalhos Futuros

Este é um estudo em andamento que oferece oportunidades de pesquisa tanto em curto quanto em longo prazo. Em particular, ainda em curto prazo, pretendemos verificar como as métricas estudadas se correlacionam com qualidade, importância e popularidade, quando combinadas. Em principio, pretendemos aplicar uma análise de regressão para cada fator de forma a determinar qual o peso que cada métrica teria em uma combinação linear. Em longo prazo, este estudo deve evoluir para a sugestão de novas métricas para a previsão dos fatores dados, usando técnicas mais sofisticadas de aprendizado de máquina.

Ainda em curto prazo, pretendemos estudar o impacto das categorias nas correlações e nos métodos baseados em tópicos, como o HITS [Kleinberg, 1999]. No primeiro caso, queremos entender se as observações feitas neste trabalho são igualmente válidas independente dos tópicos das páginas. No segundo caso, pretendemos verificar como fatores de interesse correlacionam com métodos que consideram a relatividade das noções de reputação e qualidade entre diferentes

tópicos. O HITS, por exemplo, considera que uma autoridade em um tópico não é necessariamente autoridade em outro tópico. Logo, a reputação dela varia conforme o grupo de tópicos em análise.

Outra possibilidade de continuação do trabalho realizado que pretendemos explorar em curto prazo é a utilização de outros elementos que compõem os artigos. Desejamos investigar se a quantidade de informações extra-textuais, como a utilização de sons ou de muitas imagens, pode ou não ser um indício de qualidade dos artigos.

Finalmente, em mais longo prazo, pretendemos estudar a possibilidade de criar métricas que tirem proveito de diferentes tipos de links, em lugar de tratá-los de forma homogênea. No caso da Wikipédia, em particular, tais métricas deveriam levar em conta a diferença entre links enciclopédicos e as citações a outras fontes, o que atualmente é representado de forma distinta na coleção da Wikipédia, facilitando essa diferenciação. Também pretendemos verificar se os resultados obtidos se aplicam para outros ambientes de wikis, como o Wikicionário, Wikinotícias e Wikiversidade.

Referências Bibliográficas

Amento, B., Terveen, L., and Hill, W. (2000). Does authority mean quality? predicting expert quality ratings of web documents. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '00, pages 296–303, New York, NY, USA. ACM. Available from: http://doi.acm.org/10.1145/345508.345603.

Baeza-Yates, R., Boldi, P., and Castillo, C. (2006). Generalizing pagerank: damping functions for link-based ranking algorithms. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '06, pages 308–315, New York, NY, USA. ACM. Available from: http://doi.acm.org/10.1145/1148170.1148225.

Baeza-Yates, R. and Ribeiro-Neto, B. (2011). *Modern Information Retrieval, the concepts and technology behind search*, volume Second Edition. Pearson Education Limited.Edinburgh Gate, Pearson Education Limited.Edinburgh Gate , Harlow, England.

Bendersky, M., Croft, W. B., and Diao, Y. (2011). Quality-biased ranking of web documents. In *Proceedings of the fourth ACM International Conference on Web Search and Data Mining*, WSDM '11, pages 95–104, New York, NY, USA. ACM. Available from: http://doi.acm.org/10.1145/1935826.1935849.

Berlt, K., de Moura, E. S., Carvalho, A., Cristo, M., Ziviani, N., and Couto, T. (2010). Modeling the web as a hypergraph to compute page reputation. *Information Systems Frontiers*, 35(5):530–543. Available from: http://dx.doi.org/10.1016/j.is.2009.02.005.

Bharat, K., Chang, B.-W., Henzinger, M. R., and Ruhl, M. (2001). Who links to whom: Mining linkage between web sites. In *Proceedings of the 2001 IEEE International Conference on Data Mining*, ICDM '01, pages 51–58, Washing-

ton, DC, USA. IEEE Computer Society. Available from: http://dl.acm.org/citation.cfm?id=645496.757722.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA.

Blumenstock, J. E. (2008). Size matters: word count as a measure of quality on wikipedia. In *Proceedings of the 17th International Conference on World Wide Web*, WWW '08, pages 1095–1096, New York, NY, USA. ACM. Available from: http://doi.acm.org/10.1145/1367497.1367673.

Borodin, A., Roberts, G. O., Rosenthal, J. S., and Tsaparas, P. (2005). Link analysis ranking: algorithms, theory, and experiments. *ACM Trans. Internet Technol.*, 5(1):231–297. Available from: http://doi.acm.org/10.1145/1052934.1052942.

Bray, T. (1996). Measuring the web. *Proceedings of the 5th International World Wide Web Conference on Computer Networks and ISDN Systems*, 28(7-11):993–1005.

Caverlee, J., Webb, S., and Liu, L. (2007). Spam-resilient web rankings via influence throttling. In *21th IEEE International Parallel and Distributed Processing Symposium/International Parallel Processing Symposium*, IPDPS '07, pages 1–10.

Cho, J. and Roy, S. (2004). Impact of search engines on page popularity. In *Proceedings of the 13th International Conference on World Wide Web*, WWW '04, pages 20–29, New York, NY, USA. ACM. Available from: http://doi.acm.org/10.1145/988672.988676.

Dalip, D. H., Gonçalves, M. A., Cristo, M., and Calado, P. (2011). Automatic assessment of document quality in web collaborative digital libraries. *Journal of Data and Information Quality*, 2(3):14:1–14:30. Available from: http://doi.acm.org/10.1145/2063504.2063507.

Dondio, P. and Barrett, S. (2007). Computational Trust in Web Content Quality: A Comparative Evalutation on the Wikipedia Project. *Informatica*, 31(2):151–160. Available from: http://www.informatica.si/.

Dondio, P., Barrett, S., and Weber, S. (2006). Calculating the trustworthiness of wikipedia articles using dante methodology. *IADIS International Conference on e-Society*, pages 354–359. Available from: http://www.iadis.org/es2006.

Giles, J. (2005). Internet encyclopaedias go head to head. *Nature*, 438(7070):900–901. Available from: http://dx.doi.org/10.1038/438900a.

Gleich, D. F., Constantine, P. G., Flaxman, A. D., and Gunawardana, A. (2010). Tracking the random surfer: empirically measured teleportation parameters in pagerank. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 381–390, New York, NY, USA. ACM. Available from: http://doi.acm.org/10.1145/1772690.1772730.

Heilman, M. J., Kemmann, E., Bonert, M., Chatterjee, A., Ragar, B., Beards, M. G., Iberri, J. D., Harvey, M., Thomas, B., Stomp, W., Martone, F. M., Lodge, J. D., Vondracek, A., de Wolff, F. J., Liber, C., Grover, C. S., Vickers, J. T., Meskó, B., and Laurent, R. M. (2011). Wikipedia: A key tool for global public health promotion. *J Med Internet Res*, 13(1):e14. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21282098.

Henzinger, M. R. (2000). Link analysis in web information retrieval. *IEEE Data(base) Engineering Bulletin*, 23:3–8. Available from: http://sites.computer.org/debull/A00SEP-CD.pdf.

Kamps, J. and Koolen, M. (2009). Is wikipedia link structure different? In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, WSDM '09, pages 232–241, New York, NY, USA. ACM. Available from: http://doi.acm.org/10.1145/1498759.1498831.

Kendall, M. (1948). *Rank correlation methods*. Griffin, London. Available from: http://gso.gbv.de/DB=2.1/CMD?ACT=SRCHA&SRT=YOP&IKT=1016&TRM=ppn+18489199X&sourceid=fbw_bibsonomy.

Kleinberg, J. and Lawrence, S. (2001). The structure of the web. *Science*, 294(5548):1849–1850.

Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of ACM*, 46(5):604–632. Available from: http://doi.acm.org/10.1145/324133.324140.

Kyrola, A., Blelloch, G., and Guestrin, C. (2012). GraphChi: Large-Scale Graph Computation on Just a PC. In *10th USENIX Symposium on Operating Systems Design and Implementation*, pages 31–46. Available from: https://code.google.com/p/graphchi/.

Lempel, R. and Moran, S. (2001). Salsa: the stochastic approach for link-structure analysis. *ACM Trans. Inf. Syst.*, 19(2):131–160. Available from: http://doi.acm.org/10.1145/382979.383041.

Lim, E.-P., Vuong, B.-Q., Lauw, H. W., and Sun, A. (2006). Measuring qualities of articles contributed by online communities. In *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, WI '06, pages 81–87, Washington, DC, USA. IEEE Computer Society. Available from: http://dx.doi.org/10.1109/WI.2006.115.

Liu, J. and Ram, S. (2009). Who does what: Collaboration patterns in the wikipedia and their impact on data quality. *19th Workshop on Information Technologies and Systems*, pages 175–180. Available from: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1565682.

Manning, C. D., Raghavan, P., and Schutze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press, New York. USA.

Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab. Previous number = SIDL-WP-1999-0120. Available from: http://ilpubs.stanford.edu:8090/422/.

Rassbach, L., Pincock, T., and Mingus, B. (2008). Exploring the feasibility of automatically rating online article quality. Available from: http://upload.wikimedia.org/wikipedia/wikimania2007/d/d3/RassbachPincockMingus07.pdf.

Reavley, N. J., Mackinnon, A. J., Morgan, A. J., Alvarez-Jimenez, M., Hetrick, S. E., Killackey, E., Nelson, B., Purcell, R., Yap, M. B., and Jorm, A. F. (2011). *Quality of information sources about mental disorders: a comparison of Wikipedia with centrally controlled Web and printed sources*, volume FirstView. Psychological medicine.

Smith, A. G. (2004). Web links as research indicators: analogues of citations? *Information Research*, 9(4). Available from: http://informationr.net/ir/9-4/paper188.html.

Stvilia, B., Twidale, M. B., Smith, L. C., and Gasser, L. (2005). Assessing information quality of a community-based encyclopedia. In *Proceedings of the International Conference on Information Quality - ICIQ 2005*, pages 442–454. Available from: http://mailer.fsu.edu/~{}bstvilia/papers/quantWiki.pdf.

Surdeanu, M., Ciaramita, M., and Zaragoza, H. (2011). Learning to rank answers to non-factoid questions from web collections. *Comput. Linguist.*, 37(2):351–383. Available from: http://dx.doi.org/10.1162/COLI_a_00051.

Suryanto, M. A., Lim, E. P., Sun, A., and Chiang, R. H. L. (2009). Quality-aware collaborative question answering: methods and evaluation. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, WSDM '09, pages 142–151, New York, NY, USA. ACM. Available from: http://doi.acm.org/10.1145/1498759.1498820.

Vadrevu, S. and Velipasaoglu, E. (2011). Identifying primary content from web pages and its application to web search ranking. In *Proceedings of the 20th International World Wide Web Conference*, WWW '11, pages 135–136, New York, NY, USA. ACM. Available from: http://doi.acm.org/10.1145/1963192.1963261.

Wang, S. and Iwaihara, M. (2011). Quality evaluation of wikipedia articles through edit history and editor groups. In *Proceedings of the 13th Asia-Pacific Web Conference on Web Technologies and Applications*, APWeb'11, pages 188–199, Berlin, Heidelberg. Springer-Verlag. Available from: http://dl.acm.org/citation.cfm?id=1996794.1996820.

Wikipedia (2011). Predefinição:escala de avaliação. Access date: 29 feb. 2012. Available from: http://pt.wikipedia.org/wiki/Predefiniç˜ao:Escala_de_avaliaç˜ao [cited 1 jun. 2011].

Wöhner, T. and Peters, R. (2009). Assessing the quality of wikipedia articles with lifecycle based metrics. In *Proceedings of the 5th International Symposium on Wikis and Open Collaboration*, WikiSym '09, pages 16:1–16:10, New York, NY, USA. ACM. Available from: http://doi.acm.org/10.1145/1641309.1641333.

Xue, G.-R., Yang, Q., Zeng, H.-J., Yu, Y., and Chen, Z. (2005). Exploiting the hierarchical structure for link analysis. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '05, pages 186–193, New York, NY, USA. ACM. Available from: http://doi.acm.org/10.1145/1076034.1076068.

Yamada, T., Saito, K., and Kazama, K. (2006). Network analysis to understand the structure of wikipedia. In *Symposium on Network Analysis in Natural Sciences and Engineering*, page 108.