UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO

AUGUSTO KLINGER

Visibilidade Web de Universidades

Dissertação apresentada como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação

Prof. Dr. José Palazzo Moreira de Oliveira
Orientador

Porto Alegre, setembro de 2011.

CIP – CATALOGAÇÃO NA PUBLICAÇÃO

Klinger, Augusto

Visibilidade Web de Universidades / Augusto Klinger. – 2011.

54 f.:il.

Orientador: José Palazzo Moreira de Oliveira; Co-orientador: José Valdeni de Lima.

Dissertação (mestrado) – Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Computação. Porto Alegre, BR – RS, 2011.

1.Visibilidade Web. 2.Ranking de Universidades I. Oliveira, José Palazzo Moreira de. II. Lima, José Valdeni de. III. Visibilidade Web de Universidades.

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
Reitor: Prof. Carlos Alexandre Netto
Vice-Reitor: Prof. Rui Vicente Oppermann
Pró-Reitor de Pós-Graduação: Prof. Aldo Bolten Lucion
Diretor do Instituto de Informática: Prof. Flávio Rech Wagner
Coordenador do PPGC: Prof. Álvaro Freitas Moreira
Bibliotecária-Chefe do Instituto de Informática: Beatriz Regina Bastos Haro

AGRADECIMENTOS

Meu pai, Hélio Klinger, é o principal responsável por este trabalho ter sido feito e por eu ter trilhado o caminho dos estudos. Agradeço ao meu pai por ter me ajudado a optar por este caminho, me encorajando e valorizando os estudos e o conhecimento. Aos meus pais por todo apoio financeiro e preocupação.

Agradeço a minha noiva Jordana, sempre me inspirando, me motivando, me apoiando e me dando muitas alegrias. Não obstante, é a razão de eu seguir em frente, buscando um crescimento contínuo para que nossas vidas sejam uma só, com paz, segurança, tranqüilidade e amor.

Agradeço aos meus amigos pelos momentos de descontração. Agradeço a dois amigos: Valdeni e Palazzo, por todas as idéias, conversas e discussões produtivas, e, principalmente, além de toda assistência dada ao longo do mestrado, agradeço por terem aberto as portas para eu prosseguir os estudos em um país de primeiro mundo. São verdadeiros mentores.

E finalmente, agradeço a universidade e ao Instituto de Informática, por toda infraestrutura e condições para que os alunos realizem suas pesquisas.

SUMÁRIO

LISTA DE ABREVIATURAS E SIGLAS **6**
LISTA DE FIGURAS **8**
LISTA DE TABELAS **9**
RESUMO **10**
ABSTRACT **11**
1 INTRODUÇÃO **12**
1.1 Objetivos **13**
1.2 Organização do Texto **14**
2 CONCEITOS **15**
2.1 Busca **15**
2.1.1 Metabusca 16
2.1.2 Refinamento 17
2.2 Ranking **17**
2.2.1 Fusão de Rankings 18
2.2.1.1 Borda Count (BC) 19
2.2.2 Avaliação de Rankings 20
2.2.2.1 Precisão e Recall 21
2.3 Visibilidade Web **21**
3 TRABALHOS RELACIONADOS **23**
3.1 Rankings de Universidades **23**
3.1.1 *Academic Ranking of World Universities* (ARWU) 23
3.1.2 *Webometrics Ranking of World Universities* (WR) 24
3.1.3 *Times Higher Education World University Ranking* (THE) 25
3.1.4 *QS World University Ranking* 25
3.1.5 *Performance Ranking of Scientific Papers for World Universities* 26
3.1.6 Análise e Comparação 26
3.2 Estado da Arte **27**
4 CÁLCULO DE VISIBILIDADE WEB **29**
4.1 Parâmetros **30**
4.1.1 Tamanho de Rankings VS. Precisão dos Motores de Busca 30
4.2 Protótipo **32**
4.2.1 Visão Geral 32
4.2.2 Interface do Usuário 32
4.2.3 Motores de Busca 34
4.2.4 Núcleo 35
4.2.4.1 Submit 35
4.2.4.2 SubmitRunner 35
4.2.4.3 Searcher 35
4.2.4.4 Clusterizador 35

4.2.5 Fusão 36
4.2.6 Tecnologias 37
5 EXPERIMENTOS NO DOMÍNIO DAS UNIVERSIDADES 38
5.1 Ranking de Universidades 38
5.1.1 Discussão e Problemas Encontrados 40
5.2 Expansão de Consulta 41
5.2.1 Novo Ranking 42
6 CONCLUSÕES 45
6.1 Trabalhos Futuros 45
REFERÊNCIAS 46
ANEXO A <UNIVERSIDADES BRASILEIRAS NO WR> 49
ANEXO B <VALORES DOS DOMÍNIOS DAS UNIVERSIDADES> 50
APÊNCIDE <HISTÓRICO DE DESENVOLVIMENTO> 53

LISTA DE ABREVIATURAS E SIGLAS

ARWU	Academic Ranking of World Universities
BC	Borda Count
BR	Brasil
CSS	Cascading Style Sheets
FGV	Fundação Getulio Vargas
HEEACT	Higher Education Evaluation & Accreditation Council of Taiwan
HTML	HyperText Markup Language
HTTP	HyperText Transfer Protocol
IR	Information Retrieval
JCR	Journal Citation Report
JVM	Java Virtual Machine
PUC-Rio	Pontifícia Universidade Católica do Rio de Janeiro
PUCRS	Pontifícia Universidade Católica do Rio Grande do Sul
PUC-SP	Pontifícia Universidade Católica de São Paulo
QS	Quacquarelli Symonds
SCI	Science Citation Index
SCIE	Science Citation Index-Expanded
SEO	Search Engine Optimization
SSCI	Social Sciences Citation Index
THE	Times Higher Education World University Ranking
UEL	Universidade Estadual de Londrina
UEM	Universidade Estadual de Maringá
UERJ	Universidade Estadual do Rio de Janeiro
UFBA	Universidade Federal da Bahia
UFC	Universidade Federal do Ceará
UFF	Universidade Federal Fluminense
UFG	Universidade Federal de Goiás

UFMG	Universidade Federal de Minas Gerais
UFPA	Universidade Federal do Pará
UFPB	Universidade Federal da Paraíba
UFPE	Universidade Federal de Pernambuco
UFPR	Universidade Federal do Paraná
UFRGS	Universidade Federal do Rio Grande do Sul
UFRJ	Universidade Federal do Rio de Janeiro
UFRN	Universidade Federal do Rio Grande do Norte
UFSC	Universidade Federal de Santa Catarina
UFSCAR	Universidade Federal de São Carlos
UFSM	Universidade Federal de Santa Maria
UFU	Universidade Federal de Uberlândia
UFV	Universidade Federal de Viçosa
UNB	Universidade de Brasília
UNESP	Universidade Estadual Paulista
UNICAMP	Universidade Estadual de Campinas
UNIFESP	Universidade Federal de São Paulo
UNISINOS	Universidade do Vale do Rio dos Sinos
URL	Uniform Resource Locator
USP	Universidade de São Paulo
VW	Visibilidade Web
WR	Webometrics Ranking

LISTA DE FIGURAS

Figura 4.1: Precisão do *Top-N*em 12 Buscadores para a Sigla UFRGS 31
Figura 4.2: Esquema de Funcionamento do Protótipo..... 32
Figura 4.3: Interface de Metabusca..... 33
Figura 4.4: Interface de Calculadora de Visibilidade Web 34
Figura 4.5: Formato de uma Consulta..... 34
Figura 4.6: Composição da Matriz 36
Figura 5.1: Precisão do *Top-N* em 12 Buscadores para a Sigla UFC 40
Figura 5.2: Precisão do *Top-N* em 12 Buscadores para a Sigla UFC + “Universidade” 41
Figura 5.3: Gráfico das Siglas que Perderam Pontuação 43
Figura 5.4: Gráfico das Siglas que Mais Ganharam Pontuação 44

LISTA DE TABELAS

Tabela 2.1: Pontuação para o BC 19
Tabela 3.1: Indicadores do ARWU..... 24
Tabela 3.2: Indicadores do *Webometrics Ranking* 24
Tabela 3.3: Indicadores do Ranking THE..... 25
Tabela 3.4: Indicadores do QSRanking 26
Tabela 3.5: Indicadores do HEEACT'sRanking 26
Tabela 4.1: Tecnologias Envolvidas na Implementação do Protótipo..... 37
Tabela 5.1: Ranking de Visibilidade Web por Metabusca..... 39
Tabela 5.2: Ranking de Visibilidade Web com Expansão de Consulta..... 42
Tabela A.1: Ranking Brasileiro do WR..... 50
Tabela B.1: Ranking de Universidades pelos Valores de seus Domínios..... 52

RESUMO

Visibilidade Web é a área de estudo das medidas de visibilidade de organizações, termos ou documentos na web, sendo um dos campos de pesquisa da ciência de Webmetria. Neste trabalho, uma fórmula para o cálculo de visibilidade na web é proposta, baseando-se na visão proporcionada por diferentes motores de busca. O resultado final é um indicador de Visibilidade Web para universidades.

São apresentados e analisados rankings de universidades do mundo todo encontrados na web, tais como o ARWU, o WR e o THE. Tais rankings trabalham com indicadores compostos, mas abordam a questão da visibilidade na rede de maneiras diferentes. Também são discutidos alguns trabalhos recentes no campo de Visibilidade Web.

O indicador elaborado baseia-se na classificação do *site* oficial da universidade nos buscadores como meio de pontuar a visibilidade, valendo-se de metabusca. Um método de fusão de rankings é utilizado para a pontuação. Siglas de universidades brasileiras foram submetidas ao cálculo de visibilidade e classificadas, mostrando uma aplicação da fórmula no mundo real. Dois rankings foram construídos, um deles apontando falhas no método proposto e o segundo contornando o problema encontrado, apresentando uma classificação justa, coerente com a realidade e com outros rankings de universidades brasileiras.

A principal contribuição do trabalho é a definição de um indicador de visibilidade na web que mostra como são classificadas as instituições através dos principais pontos de entrada da rede: os motores de busca.

Palavras-Chave: Visibilidade Web, Ranking de Universidades.

ABSTRACT

Web Visibility is the area of study of visibility measures of organizations, terms or documents on the web, one of the fields of the Webmetrics science. In this work, a formula for the calculation of visibility on the web is proposed, based on the vision provided by different search engines. The result is an indicator of Web Visibility for universities.

In related works section are presented and analyzed rankings of universities around the world from the web, such as ARWU, WR and THE. Such rankings work with composite indicators, but address the issue of visibility on the network in different ways. Also are covered some resent works in the field of Web Visibility.

The indicator developed is based on the classification of the universities' official website in search engines as a means of scoring the visibility, using a metasearch engine. A method of rankings fusion is used for the scoring process. Acronyms of Brazilian universities were subjected to the calculation of visibility and classified, showing an application of the formula in the real world. Two rankings were built, one of them pointing out flaws in the proposed method and another solving the problem encountered with a fair rating, consistent with reality and with other rankings of universities.

The main contribution of this work is a Web Visibility indicator that shows how the institutions are classified by the main entry points of the network: the search engines.

Keywords: Web Visibility, Universities Ranking.

1 INTRODUÇÃO

A Internet pode ser considerada como um meio digital para armazenar, publicar e recuperar informações. A facilidade e liberdade em publicar informações na Internet permitiram, por um lado, o surgimento do maior meio digital de armazenamento de informações da humanidade, e por outro lado, o aumento da complexidade em recuperar tais informações. Também, o grande volume e a falta de padronização das informações armazenadas dificultam a recuperação das mesmas.

O aumento do acesso à Internet juntamente com o avanço da tecnologia de recuperação de informações permitiram colocar a disposição dos bilhões de usuários da rede as referidas informações armazenadas.

Este grande volume de informações, resultado do armazenamento em milhões de *websites* (o número de *websites* em dezembro de 2010 já contabilizava 255 milhões de acordo com o *site* Pigdom[1]), referenciado na bibliografia como "repositório de dados", "repositório de informações", "meio digital" ou simplesmente web, é o alvo principal de nossa pesquisa.

Na busca de facilitar a tarefa de recuperação de informações têm surgido técnicas de medição tais como webmetria, informetria e bibliometria. Estas técnicas não levam em conta apenas o conteúdo procurado, elas abrangem parâmetros tais como medidas estatísticas, popularidade, agrupamentos de *sites* e distribuição deste conteúdo.

Webmetria é a técnica que estuda os parâmetros quantitativos de medidas relativas à informação armazenada na web. Exemplos de parâmetros são: o porquê de seu armazenamento, o porquê de sua recuperação, as estruturas de armazenamento, dados bibliométricos (BJÖRNEBORN; INGWERSEN, 2004) entre outros.

Uma medida (métrica) que é calculada nesta dissertação é a "Visibilidade Web" (VW), considerada uma subárea da Webmetria. Esta medida consiste do estudo de que fatores e como eles podem ser quantificados para a obtenção desta visibilidade. A visibilidade de um *site* ou página na rede pode ser determinante no momento de recuperar a informação. Ou seja, quanto mais fácil for recuperar uma informação da Internet maior será sua visibilidade. Fatores que promovem a VW de um *site* são, no contexto dos motores de busca, a similaridade com a consulta e a importância do site, medida pelo *PageRank* por exemplo (GORI, 2005). Os aspectos de qualidade são desprezados neste trabalho, assim o valor de VW calculado não distingue a publicidade positiva da negativa.

Medir a VW é importante em muitas áreas de conhecimento, sobretudo para avaliar o nível de publicidade ou medir o impacto de uma marca, produto ou instituição na rede.

---

[1]http://royal.pingdom.com/2011/01/12/internet-2010-in-numbers/

Existem diferentes formas de medir a visibilidade da informação: popularidade do *site* onde a informação está armazenada, número de acessos, número de *links* que levam ao *site*, posição que ocupa na lista ordenada resultante de um motor de busca entre outras.

Como a grande maioria das recuperações das informações na web passa por um motor de busca surgiram empresas especializadas para melhorar a posição de sites relacionados com a informação recuperada. Serviços de otimização de *sites* (SEO) são exemplos deste tipo de empresa. Estes serviços consistem em um conjunto de estratégias com o objetivo de potencializar e melhorar a posição de um *site* nas listas resultantes dos motores de busca. As empresas investem nestas estratégias para melhorar a visibilidade de seus *sites* na Internet, pois cerca de 62% das pessoas não vê além da primeira página de resultados dos motores de busca segundo Bacchin (2008). A medida de VW pode ser usada também para verificar se as estratégias adotadas pelas empresas são eficientes, ou seja, se elas estão melhorando a posição dentro das listas classificadas resultantes dos motores de busca existentes no mercado.

A elaboração de uma medida de visibilidade conforme os buscadores da web se apresenta como uma interessante alternativa às formas de cálculo de Visibilidade Web existentes. As medições são pertinentes não somente para a área de SEO, mas também para a elaboração de listas em determinados domínios (ex. universidades) e avaliações de popularidade de propósito geral na web. Sendo os motores de busca os pontos dominantes de acesso às páginas da web, a metabusca se apresenta como um recurso qualificado para medir visibilidade na rede, resultando em um dado quantitativo que representa o quão acessível e bem visto é o *site* de acordo com os motores de busca da web. Quanto mais motores de busca forem utilizados para medir a visibilidade, maior tende a ser a diversidade de critérios e abrangência, gerando um valor de VW mais fidedigno.

Em se tratando de rankings em domínios homogêneos, se destaca o domínio das universidades. Existem diversos rankings classificando instituições de ensino superior do mundo inteiro na web, como pode ser visto no trabalho do conselho de pesquisas Nórdicas NordForsk (2011). A análise da presença das universidades por meios de indicadores Cybermétricos se mostra como uma ferramenta importante para fins de avaliações e comparações, sendo cada vez mais relevante. Uma boa colocação dentro de um ranking pode atrair mais pesquisadores de alto nível, investimentos e alunos para a universidade.

As universidades têm se conscientizado da importância de suas presenças na web. Uma maneira de maximizar a visibilidade de uma instituição é manter um repositório digital que represente a produção científica da instituição (SWAN; CARR, 2008). E com o aumento da presença das universidades na web, torna-se importante o estudo das formas de medir a VW das instituições.

1.1 Objetivos

O objetivo principal desta dissertação é elaborar um novo indicador de VW para universidades, levando em conta as diferentes percepções da web que diferentes motores de busca possuem. Como objetivo secundário está a classificação de universidades conforme o indicador estudado, com o auxílio de um protótipo.

Para atingir os objetivos gerais foi feita uma divisão em objetivos mais específicos:

- Estudo sobre os conceitos envolvidos e trabalhos realizados no campo de VW;
- Pesquisa a respeito dos rankings de universidades existentes;
- Análise de parâmetros pertinentes a um cálculo de visibilidade;
- Elaboração de um indicador com um método de pontuação diferente dos estudados;
- Implementação de um protótipo;
- Classificação de universidades brasileiras segundo o indicador proposto;
- Análise do ranking construído e discussão sobre problemas encontrados;
- Aplicação de refinamento para atingir um melhor nível de pontuação;
- Elaboração de um novo ranking com as universidades e comparações com outros existentes;
- Análise do trabalho realizado e de possíveis direções a serem seguidas no futuro.

1.2 Organização do Texto

O trabalho está dividido em capítulos conforme:

- Capítulo 2, **Conceitos**, introduz os conceitos básicos pesquisados sobre assuntos pertinentes ao trabalho: Rankings, como unificá-los e extrair dados deles; Busca, conceitos de busca e metabusca na web; VW, visão da área de estudos.
- Capítulo 3, **Trabalhos Relacionados**, apresenta uma série de rankings de universidades encontrados na web, exibindo seus indicadores e analisando a abordagem da visibilidade. Também são discutidos trabalhos sobre VW no geral.
- Capítulo 4, **Cálculo de Visibilidade Web**, mostra como opera o cálculo efetuado pelo indicador de visibilidade elaborado, expondo os parâmetros e alguns detalhes do protótipo.
- Capítulo 5, **Experimentos no Domínio das Universidades**, apresenta dois rankings de universidades brasileiras, um deles expondo falhas no método e o outro as corrigindo. Os rankings são analisados e comparados.
- Capítulo 6, **Conclusões**, analisa os trabalhos realizados, propondo algumas melhorias e trabalhos futuros.

2 CONCEITOS

Para atingir os objetivos desta dissertação foram pesquisados os seguintes temas: Busca, Ranking e VW. Os resultados foram dispostos da seguinte forma:

- Busca – visão sobre o funcionamento de mecanismos de busca e metabusca, suas capacidades e características.
- Ranking – principais conceitos e visão geral sobre alguns modelos de fusão existentes, com atenção ao funcionamento do *Borda Count*; algumas medidas de avaliação para rankings.
- Visibilidade Web – visão geral sobre a área de pesquisa.

2.1 Busca

Uma busca se inicia a partir de um alvo. Um alvo pode ser uma lista de termos, uma imagem, um som, um dado estruturado, ou do que for constituída a base onde será efetuada a busca. O processo consiste basicamente em verificar cada elemento de um universo com o objetivo de encontrar aqueles elementos que mais se assemelham ao alvo da busca.

Motor de busca é um software projetado para encontrar e recuperar informações armazenadas em um sistema computacional, geralmente, a partir de palavras-chave entradas pelo usuário. Os motores de busca podem operar sobre redes corporativas ou computadores pessoais, mas a principal utilização deles é na web.

Com a popularização da Internet houve a necessidade de sistemas que extraíssem informações da rede de forma prática. Assim surgiram os buscadores da web. A idéia é simples: dado o termo alvo da busca, ou termos, é retornada ao usuário uma lista de páginas web que combinam com o alvo desejado.

Os motores de busca armazenam informações sobre um grande número de páginas, recuperadas através de *spiders*, que são uma espécie de robôs que seguem pelos *links* entre as páginas. O conteúdo de cada página é analisado e indexado para pesquisas futuras. Quando é feita uma busca, o sistema procura no índice e fornece uma lista das páginas que melhor combinam com o critério, trazendo também normalmente um breve resumo contendo o título e partes do texto.

Pode haver milhões de páginas incluindo uma palavra ou frase em particular, dentre as quais umas podem ser mais relevantes que outras. A maioria dos sistemas de busca cria listas visando colocar os “melhores resultados” entre os primeiros. Como cada motor de busca decide quais são as páginas que mais combinam com o alvo da busca e em qual ordem serão exibidas, os resultados podem variar muito de um sistema para outro.

2.1.1 Metabusca

Metabusca significa fazer uma busca sobre outra(s) busca(s). Essencialmente, o termo expressa o processo de consultar ao mesmo tempo diversos motores de busca, operando uma busca em dois (ou mais) níveis. No primeiro momento é realizada a busca pelo termo alvo da consulta nos diversos motores de busca e posteriormente, no segundo momento, o mesmo termo é buscado nos resultados.

Uma ferramenta de metabusca é basicamente um agregador de buscadores, fazendo uso das bases de dados dos motores de busca a que submete suas consultas. Quando se realiza uma metabusca na web, a ferramenta envia o pedido para diversos motores de busca simultaneamente como o Google, Yahoo, Bing, e outros, retornando o resultado deles em uma mesma tela. Alguns sistemas de metabusca ainda possuem recursos a mais, como unificar os resultados e alterar o posicionamento deles. Estas ferramentas de pesquisa, conforme Blattman (1999) podem ser utilizadas satisfatoriamente quando: procura-se por algum tópico obscuro; as buscas simples não retornam muitos *sites*; deseja-se recuperar o maior número de documentos possíveis.

A tarefa de um sistema de metabusca pode ser dividida em dois subproblemas. Um deles é como realizar a consulta a cada motor de busca. O outro consiste em como recuperar a informação retornada. Cada buscador precisa ter suas características estudadas individualmente para ser utilizado eficazmente.

Mas descobrir as particularidades de cada motor de busca pode ser uma tarefa trabalhosa, e não há garantias de que as interfaces não sofrerão mudanças com o tempo. Isso torna a manutenção de um sistema de metabusca bastante difícil. O ideal é verificar automaticamente as mudanças ao passo em que se manifestam em cada mecanismo de busca, mantendo o sistema sempre atualizado e funcional.

Para obter os resultados de motores de busca diversos, segundo Tous e Delgado (2002), pode ser explorado o fato de que eles utilizam, na maioria dos casos, os navegadores como interface, com requisições HTTP e resultados em páginas HTML, o que deixa a porta aberta para outras aplicações, como sistemas de metabusca, dispararem consultas.

O cenário perfeito para a aplicação de metabusca seria um universo no qual se tem uma ordem completa de todos os elementos em todas as listas. Porém isso não é possível na web, pois cada motor de busca tem uma cobertura diferente da mesma (DWORK et al., 2001). É improvável que os motores de busca sejam capazes de classificar toda a coleção de páginas da web dado seu constante crescimento.

Do ponto de vista do usuário praticamente não há diferença entre um *site* de busca tradicional e um de metabusca. Ambos utilizam a mesma interface padrão, na qual os elementos principais são uma caixa de texto, onde o usuário entra com o termo alvo da consulta, e o botão de <buscar>. Os resultados são exibidos, assim como nos motores de busca regulares, em forma de lista ordenada. Opções avançadas podem permitir a escolha dos motores de busca a serem utilizados, como ocorre no metabuscador web iBoogie[1].Outros sistemas de metabusca atualmente em funcionamento na web são IxQuick[2], Metacrawler[1] e DogPile[2].

---

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

[2] http://www.ixquick.com

2.1.2 Refinamento

Refinamento é uma técnica para melhorar a precisão do resultado de uma busca, ou metabusca, de acordo com a necessidade de informação. Neste trabalho são exploradas e identificadas três maneiras de fazer o refinamento de uma consulta:

1. Elaborar um filtro para excluir resultados não pertinentes;
2. Excluir manualmente resultados não pertinentes;
3. Melhorar a formulação da consulta (objeto buscado, condições, ajuste do alvo).

A primeira maneira de fazer refinamento pode ser através da elaboração de um filtro. Este filtro pode ser executado em tempo de busca, visando ignorar e não recuperar resultados não pertinentes. Isto pode ser feito através da indicação de termos indesejáveis nos resultados da busca, ou, mais sofisticadamente, através de uma ontologia que identifique a área de foco da consulta e despreze as demais áreas.

Na segunda maneira, a exclusão manual dos resultados não pertinentes é viável somente se o resultado é composto de uma lista pequena, em torno de uma dezena. A partir dos selecionados o sistema pode refazer a consulta os desconsiderando ou simplesmente excluí-los dos resultados.

Finalmente, a terceira maneira de fazer refinamento, que é melhorar a formulação da consulta, depende muito do motor de busca utilizado. Acrescentar palavras que aumentem o poder de descrição do objeto buscado é uma forma de refinamento muito utilizada. Esta forma pode ser usada no caso de um usuário que esteja buscando uma dada organização cujo nome é idêntico ao de outra organização de outro segmento, a adição do segmento junto às condições fará com que os resultados sejam mais coerentes.

Outra forma de refinar uma busca será através da Web Semântica, se bem difundida futuramente. Diretórios web, que são classificações por conteúdo das páginas web em distintas áreas e subáreas, tal como é encontrado no Yahoo, por exemplo, também são uma maneira de identificar a intenção de busca, porém exigem certo esforço ao navegar pelos níveis da hierarquia de diretórios.

2.2 Ranking

Um ranking é um conjunto ordenado de elementos, conforme sua relevância. Os elementos são organizados de maneira crescente de forma que, para cada dois itens, o primeiro é maior, menor ou igual ao segundo em grau de importância (DUTRA, 2008).

Em computação, rankings são muito utilizados na área de Recuperação de Informação (IR) na qual os algoritmos têm por objetivo buscar a informação desejada e, usualmente, classificá-la de acordo com determinados critérios de relevância. Os principais exemplos de aplicação na área são motores de busca que operam em grandes bases de dados ou na web. Neste contexto, um ranking pode ser visto como uma lista ordenada de textos, cada um deles com um *link* para uma página onde se encontra mais detalhes da informação recuperada, onde o primeiro da lista é mais relevante do que o

---

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

[2] http://www.dogpile.com

segundo e a materialização desta lista obedece esta ordem no dispositivo de apresentação da referida lista.

De acordo com Dutra (2008), existem duas classificações para rankings: completos e parciais. Um ranking completo possui todos os elementos de um universo. Um ranking parcial não possui todos os elementos de um universo, apresentando uma classificação com somente alguns dos elementos.

Os rankings gerados pelos motores de busca da web pertencem à classificação de rankings parciais, pois além dos algoritmos de indexação serem incapazes de cobrir todo o conteúdo da rede, os usuários estão interessados somente nos *k* primeiros elementos (*top-k*) de maneira geral ou simplesmente tendem a desconsiderar os resultados mais distantes dos primeiros colocados (GOULART, 2008).

Não incluir todos os elementos do universo na elaboração de rankings é uma necessidade em se tratando de web, mas processar somente os elementos do topo não implica em resultados ruins, como visto no artigo de Bast et al. (2006), no qual são apresentadas técnicas para o processamento de *top-k* elementos com o objetivo de otimizar a performance sem comprometer a recuperação da informação.

2.2.1 Fusão de Rankings

Fusão de rankings é o problema de computar um ranking consensual, dados diversos rankings contendo elementos classificados por diferentes juízes (DWORK et al., 2001). Um juiz é quem determina a ordem dos elementos de um dado universo, podendo ser um especialista humano ou um motor de busca, por exemplo. Em suma, o objetivo é combinar várias listas de objetos de forma robusta para produzir um único ranking que descreva todos os critérios envolvidos da melhor forma.

Existe uma diversidade de métodos para fusão de rankings, cada um com suas aplicabilidades, custos e limitações. Os métodos utilizam as informações dos rankings individuais, como o ordinal associado a um elemento (cada elemento associado a uma posição), uma nota dada a cada elemento (cada elemento associado a um valor) ou o próprio conteúdo dos elementos, com o objetivo de encontrar o melhor ranking resultante baseado em alguma métrica. Dentre os modelos propostos estão os de combinação linear, os probabilísticos e os de votação.

Os modelos de combinação linear fazem uso de um operador de agregação, que pode ser, por exemplo, a média das posições dos elementos, para unir duas listas. Adicionalmente, é possível utilizar coeficientes de peso para favorecer um dos rankings. Os modelos de combinação linear destacados são os que utilizam grafos de Cayley (HSU; PALUMBO, 2004) e o modelo que permite a inclusão de listas parciais e coeficientes de peso (DUTRA JÚNIOR; LIMA, 2006).

Já os modelos probabilísticos de fusão de rankings fazem inferências sobre os elementos envolvidos com base na teoria das probabilidades para efetuar a fusão de múltiplas listas. Esses modelos permitem analisar o posicionamento dos elementos e formar uma nova classificação considerando aspectos como a probabilidade de relevância de um elemento em relação a uma consulta ou a probabilidade da posição do elemento ser a melhor possível. Pode-se destacar dentre os métodos probabilísticos utilizados o Teorema de Bayes (ASLAM; MONTAGUE, 2001) e Cadeias de Markov (DWORK et al., 2001).

Os modelos de votação são baseados em métodos eleitorais. O processo de votação combina as preferências de diversos juízes, proporcionando algoritmos eficazes para a fusão de rankings. O método Borda Count (BC) é utilizado em uma série de aplicações, como no modelo de fusão por atribuição de pontos *Borda-fuse* (ASLAM; MONTAGUE, 2001).

Como os modelos de combinação linear exigem listas completas e operam com somente duas, e os modelos probabilísticos possuem grande complexidade, para este trabalho será dirigido o estudo do BC por se adequar melhor as necessidades, uma vez que as listas envolvidas não serão de fato fundidas em primeira instância, e é de interesse para a pesquisa somente a pontuação dos elementos, fornecida de forma prática pelo método.

2.2.1.1Borda Count (BC)

Borda Count é um método de votação classicamente utilizado em processos eleitorais, com eficiência e méritos justificados (BLACK, 1976). De fato, o método se mostra superior a outros métodos de votação de acordo com Saari (1985), que apresenta provas matemáticas e exemplos da vantagem do método em relação aos demais. A origem de suas idéias é desconhecida ao passo que estas foram aplicadas independentemente diversas vezes ao longo da história. Mas em 1770, o método foi nomeado e concebido pelo francês Jean-Charles de Borda, sendo utilizado ainda atualmente por organização privadas, competições e algumas entidades estatais como o parlamento de Nauru.

Essencialmente, BC é um método de eleição no qual cada eleitor define seus candidatos em ordem de preferência. O vencedor é determinado através de pontos dados a cada candidato de acordo com a posição em que se encontram na lista de preferência de cada eleitor. Aquele candidato com maior pontuação, ao final da contagem, é o vencedor. Como o método por vezes elege os candidatos mais aceitáveis ao invés dos preferidos pela maioria, é descrito freqüentemente como um sistema eleitoral consensual.

Para determinar a pontuação correspondente a cada colocação, é preciso conhecer o número de candidatos. Dessa forma, para *n* candidatos, o primeiro receberia *n* pontos, o segundo *n-1*, e assim por diante. Alternativamente, pode-se atribuir um ponto para o primeiro colocado, 1/2 para o segundo, 1/3 para o terceiro, e assim sucessivamente, dando maior ênfase a primeira colocação. A Tabela 2.1 abaixo ilustra os dois métodos de pontuação, para *n*=5.

Tabela 2.1: Pontuação para o BC

Classificação	*Formula A*	*Pontuação A*	*Formula B*	*Pontuação B*
1º	n	5 pontos	1/1	1.00
2º	(n-1)	4 pontos	1/2	0.50
3º	(n-2)	3 pontos	1/3	0.33
4º	(n-3)	2 pontos	1/4	0.25
5º	(n-4)	1 ponto	1/5	0.20

O método fracionado de pontuação é amplamente favorável para candidatos com muitos primeiros lugares, reduzindo o impacto de candidatos escolhidos aleatoriamente simplesmente para completarem as listas. Outra solução para o problema de candidatos

escolhidos aleatoriamente nas ultimas colocações é permitir rankings parciais, assim os eleitores não precisam elaborar uma lista completa de preferências.

Com a possibilidade de truncar o número de votos, a pontuação pode ser distribuída, para *n* candidatos, de duas maneiras:

1. O primeiro colocado recebe *n* pontos, o segundo *n-1*, conforme o método tradicional, até o ultimo colocado da lista parcial. Todos os candidatos que não figuram no ranking recebem a pontuação mínima, ou seja, um, ou zero caso o primeiro tenha recebido *n-1* pontos.
2. Para uma lista com *m* candidatos, o primeiro recebe *m* pontos, o segundo *m-1*, o terceiro *m-2*, até o ultimo da lista. Os candidatos que não figuram na lista não recebem pontuação.

O problema da segunda abordagem é que, para um *n* muito grande, aqueles eleitores com poucas preferências não influenciam muito no resultado final. Uma lista com somente um candidato resultará em um valor de *m* igual a um, o que influencia os eleitores a elaborarem rankings maiores em função de favorecerem mais seus candidatos preferenciais. Porém, no caso de todos os eleitores serem obrigados a votar em um mesmo número de candidatos, o problema é contornado.

Tratando-se de fusão de rankings, o método baseado no BC é aplicado para problemas computacionais como a metabusca (ASLAM; MONTAGUE, 2001). Por utilizar diretamente a colocação dos elementos de um ranking como meio de pontuação, o método satisfaz as necessidades de um cálculo de visibilidade baseado em rankings de motores de busca.

2.2.2 Avaliação de Rankings

A avaliação da qualidade de um ranking é dada por uma medida de precisão, que se baseia na relevância de cada elemento recuperado em relação à consulta realizada. O processo consiste em efetuar uma ou mais consultas a determinada base de dados e avaliar os documentos obtidos em resposta para mensurar a eficiência dos algoritmos de busca empregados.

Como em muitos casos, nas buscas efetuadas por motores de busca, nem todos os elementos recuperados em resposta são relevantes. Além disso, pode acontecer de muitos itens relevantes não participarem do ranking retornado.

A idéia de relevância é tratada como binária para o método padrão de avaliação, ou seja, com base em uma consulta proposta, os documentos são classificados como relevantes ou irrelevantes, não existindo categorias de muito relevante, razoavelmente relevante ou pouco relevante. Essa classificação binária é tratada como o padrão de ouro (*gold standard*) do julgamento de relevância (MANNING; RAGHAVAN;SCHÜTZE, 2008).

A relevância é julgada de acordo com a necessidade de informação, e não em relação à consulta. Isso significa que um documento pode ser relevante mesmo não contendo todas as palavras da consulta, e também que documentos não relevantes podem conter todas as palavras da consulta. A abordagem, apesar de ser padrão, é bastante criticada, pois a relevância é subjetiva. O que para um juiz é um aspecto que caracteriza a relevância de um documento, para outro juiz pode não caracterizar. Mesmo fazendo o julgamento de forma automatizada, diferentes implementações podem discordar a respeito da relevância de um documento e de especialistas humanos.

Precisão e *Recall* são as medidas mais utilizadas em avaliação de algoritmos e sistemas de recuperação de informações. O termo em inglês *recall*, em IR, é freqüentemente traduzido para o termo em português "abrangência", mas como não há um acordo sobre a tradução, neste trabalho é utilizado o termo em inglês. A próxima subseção descreve as duas técnicas de avaliação, provenientes da área de IR.

2.2.2.1Precisão e Recall

Considere uma coleção documentos *C* contendo um número *R* de documentos relevantes. Certa consulta é aplicada a um sistema de recuperação qualquer sobre a coleção *C* e retorna um conjunto de *A* documentos, sendo *RA* o número de relevantes retornados. As medidas de precisão e *recall* são definidas da seguinte maneira (BAEZA-YATES; RIBEIRO-NETO, 1999):

- Precisão é a quantidade de documentos retornados que são relevantes

$$\text{Precisão} = \frac{RA}{A}$$

- *Recall* é a quantidade de documentos relevantes retornados do total de relevantes

$$Recall = \frac{RA}{R}$$

Usualmente, é apresentado ao usuário do sistema de IR um ranking e, na medida em que o ranking vai sendo analisado a partir do topo, *recall* e a precisão variam. A cada nível de *recall*, tem-se uma precisão associada.

O problema com essas medidas, definidas da maneira acima, é que precisam que todo o conjunto de *A* documentos retornados seja examinado. E ainda que se conheça toda a coleção *C*, a fim de saber quais são todos os documentos relevantes contidos nela. Em coleções muito grandes isso é impraticável, como é o caso da web, onde é impossível descobrir o total de documentos relevantes e, portanto, medir o *recall.* Mesmo a precisão demanda esforço para ser computada na maioria dos casos, pois o total de elementos recuperados é geralmente grande.

Uma maneira interessante de medir precisão na web é considerar somente os dez primeiros documentos, já que raramente os usuários olham além do décimo *site* recuperado, isto é, além da primeira página de um motor de busca. A medida é chamada de *Precision at 10*, e tem a vantagem de não ser necessário conhecer o número total de documentos relevantes.

Mais genericamente, essa medida de precisão que considera somente os *k* primeiros resultados é chamada de *Precision at k* (MANNING; RAGHAVAN;SCHÜTZE, 2008). Sendo *Rt* o número de relevantes no *top-k*, a *Precision at k* é definida conforme segue:

$$Precision\ at\ k = \frac{Rt}{k}$$

2.3 Visibilidade Web

Em 1997 foi criado o termo Webmetria (também referenciado por Webometria), por Almind e Ingwersen, caracterizando a área de estudos dos aspectos quantitativos da

web. Nela está inserida a VW, que tem por objetivo medir o quão visível é dada página ou documento na rede. O termo Cybermetria é comumente utilizado como sinônimo de Webmetria, porém abrange os estudos e análises quantitativas das classes e meios de informações dentro de toda a Internet, ao passo que a Webmetria se restringe a web (JOURIS, 2011).

Uma boa visibilidade para uma página web é caracterizada por um *site* que recebe muitos acessos. A quantidade de acessos pode ser proveniente de, por exemplo, o *site* estar bem colocado em motores de busca, ou possuir grande quantidade de *links* que levem a ele, ou apresentar informações de qualidade reconhecida, ou por alguma estratégia de *marketing* eficiente, enfim, qualquer critério que facilite com que os usuários cheguem ao *site*.

A maneira mais comum de medir visibilidade na web, e mais difundida na literatura (AGUILLO et al., 2006; AGUILLO; GRANADINO; ORTEGA, 2006), é utilizar a quantidade de *links* que apontam para as páginas como indicador, inclusive utilizada no WR para a elaboração de rankings de universidades. O critério de número de *links* é geralmente empregado nos algoritmos de classificação de *sites* dos motores de busca, como o mundialmente conhecido *PageRank* (PAGE et al., 1999).

A classificação do *PageRank* baseia-se na estrutura de *links* para determinar o valor de uma página web, interpretando um *link* da página A para a página B como um voto de A para B. Mas não é somente o número de votos que é computado, o algoritmo leva em conta também a importância da página eleitora. Uma página de alta importância tem maior probabilidade de receber acessos que uma página de baixa importância, portanto seu voto pesa mais no cálculo do *PageRank*.

A quantidade de acessos a um *site* é um indicativo de popularidade, portanto também pode ser usado como uma medida de visibilidade na web. Apesar de simples, poucos adotam a medida diretamente tal como Aaltojärvi (2008) em sua análise do mecanismo *Google Scholar*.

Os trabalhos de Espadas (2008) e de Gori (2005) destacam pontos importantes da VW. O primeiro trata do problema de que os motores de busca não tornam visíveis grandes partes da web, propondo um método para indexação de *sites*. O segundo explica algumas heurísticas adotadas por motores de busca e expõe suas fraquezas ao permitirem a construção de comunidades artificiais de *sites* que se apontam entre si com o objetivo de melhorarem suas classificações. Uma possível solução para o problema está na Web Semântica, aponta o autor.

Mais trabalhos na área de VW são apresentados no final da seção de trabalhos relacionados, voltando-se para as medidas realizadas na área acadêmica.

3 TRABALHOS RELACIONADOS

Muitos trabalhos visam classificar universidades com o objetivo de evidenciar as melhores ou simplesmente posicionar as instituições segundo algum critério, tais como tamanho, área de atuação, produção científica, investimentos externos, e mais recentemente, visibilidade na web.

Este capítulo apresenta os mais populares rankings de universidades encontrados na Internet, fazendo ao final um comparativo entre eles com relação as suas classificações para as universidades brasileiras, e também avaliando a questão da VW. Um breve estudo sobre trabalhos que utilizam VW também é apresentado.

3.1 Rankings de Universidades

Surgiram alguns rankings de universidades na Internet durante a última década, tais como o *Academic Ranking of World Universities* da universidade chinesa *Shanghai Jiao Tong* (desde 2003), o *Webometrics Ranking of World Universities* do laboratório espanhol *Cybermetrics Lab* (desde 2004), o *Times Higher Education World University Ranking* (desde 2004), o *QS World University Ranking* da *Quacquarelli Symonds Limited* (desde 2004), e o *Performance Ranking of Scientific Papers for World Universities* da *Higher Education Evaluation and Accreditation Council of Taiwan*, HEEACT (desde 2007).

Todos esses rankings tratam as universidades como um todo, não levando em conta diferentes áreas de pesquisa isoladamente. Os indicadores são compostos, agregando diversos critérios. O objetivo desses rankings é posicionar uma universidade em particular e informar o quão melhor, ou pior, ela é em relação às outras. Boas colocações são interpretadas normalmente como um indicativo de que a instituição pertence às melhores do mundo.

A seguir são apresentadas as características e critérios dos rankings mundiais de universidades encontrados na web.

3.1.1 *Academic Ranking of World Universities* (ARWU)

O chinês ARWU teve a sua primeira versão em 2003. São incluídas somente universidades que possuem Prêmios Nobel, *Fields Medalists*, pesquisadores altamente citados ou artigos publicados em *Nature and Science*. Universidades com um número significativo de documentos indexados pela *Science Citation Index-Expanded* (SCIE) e *Social Citation Index* (SCI) também são incluídas, somando por volta de 1000 universidades classificadas, das quais as quinhentas melhores são publicadas na web.

Os indicadores quantitativos utilizados para a classificação do ARWU são apresentados na tabela 3.1. O maior diferencial da pontuação é que 30% dela é baseada em prêmios.

Tabela 3.1: Indicadores do ARWU

Critério	*Indicador*	*Peso*
Qualidade da Educação	Número de alunos que ganharam um Prêmio Nobel ou *Fields Medal* desde 1901	10%
Qualidade Docente	Número de pesquisadores que ganharam um Prêmio Nobel emfísica, química, medicina ou economia e/ou *Fields Medal* em matemática desde 1911	20%
	Número de pesquisadores altamente citados nas áreas de ciências da vida, medicina, física, engenharia e ciências sociais	20%
Pesquisa	Número de artigos publicados em *Nature and Science* durante os últimos 5 anos	20%
	Número de artigos listados no SCIE e SSCI no ultimo ano	20%
Tamanho da instituição	A soma ponderada dos 5 indicadores acima dividida pelo número de pesquisadores de tempo integral	10%

3.1.2 *Webometrics Ranking of World Universities* (WR)

O *Webometrics Ranking*, elaborado pelo laboratório espanhol *Cybermetrics Lab*, tenta englobar todas as universidades do mundo em sua classificação, somando por volta de 20.000 instituições. Desde 2004, seu ranking é divulgado duas vezes por ano. Ele pode ser acessado de diversas formas: por país, por continente, e por determinadas regiões, além do ranking global.

O atributo fundamental para a avaliação de uma universidade é o seu domínio institucional. Somente universidades com um domínio web independente são consideradas. Os indicadores utilizados pelo laboratório são todos extraídos diretamente da web, conforme a tabela 3.2 abaixo. O objetivo é mostrar a presença das instituições na Internet e motivar a publicação das produções cientificas na web de modo a refletir suas atividades.

Tabela 3.2: Indicadores do *Webometrics Ranking*

Critério	*Indicador*	*Peso*
Tamanho	Número de páginas recuperadas por 3 motores de busca da web: *Bing*, *Google* e *Yahoo*	20%
Visibilidade	Número de *links* externos únicos (*inlinks*) recebidos pelo *site*, de acordo com o *Yahoo Site Explorer*	50%
Documentos Eletrônicos	Quantidade de documentos em Adobe Acrobat (.pdf), Adobe PostScript (.ps), Microsoft Word (.doc) e Microsoft Powerpoint (.ppt)	15%
Scholar	Número combinado de publicações incluídas no *Google Scholar* entre 2007 e 2011 e *Scimago* SIR entre 2005 e 2009	15%

No seu cálculo de pontuação, a VW representa 50% do valor agregado à universidade, sendo que essa visibilidade é obtida através do *Yahoo Search*, levando em conta o número total de *links* externos únicos que o *site* de cada instituição recebe

(*inlinks*). Publicações dos últimos quatro anos disponibilizadas na web representam 15% da pontuação. Documentos eletrônicos, outros 15%.

Páginas web recuperadas nos três principais motores de busca da atualidade pesam 20%. A inclusão do número total de páginas, segundo os autores, é baseada no reconhecimento de um novo mercado global para informações acadêmicas, sendo a web a plataforma adequada para a internacionalização das instituições.

3.1.3 *Times Higher Education World University Ranking* (THE)

O ranking THE era até 2010 um ranking conjunto com o QS. Atualmente independente, apresenta um *top-200* mundial e seis rankings por áreas contendo as 50 melhores universidades de cada área.

Universidades que não possuem cursos de graduação, que não publicam ao menos 50 artigos por ano ou que são focadas em uma única área, são excluídas da classificação. O ranking utiliza 13 indicadores de desempenho visando capturar todos os aspectos das atividades acadêmicas, divididos em cinco categorias apresentadas na tabela 3.3 abaixo.

Tabela 3.3: Indicadores do ranking THE

Critério	*Indicador*	*Peso*
Ensino	Cinco indicadores, sendo o principal uma pesquisa sobre a reputação do ensino (realizada pela Thomson Reuters)	30%
Pesquisa	65% baseado em uma pesquisa de reputação, o restante da categoria é determinado pela renda da universidade com pesquisa e medidas simples de volume de pesquisa contra o número de efetivos. Para esta última medida, THE conta o número de artigos publicados em periódicos acadêmicos indexados pela *Thomson Reuters* por membro	30%
Citações	O impacto das citações é normalizado para refletir as variações do volume de citações entre as diferentes áreas de pesquisa	32,5%
Investimentos	Renda proveniente da indústria como indicador de inovação	2,5%
Internacionalização	Estrangeiros no corpo docente e discente	5%

É interessante notar que 34,5% da pontuação total é atribuída por pesquisas de reputação (total de cerca de 13.000 respostas, segundo o *site* do THE), somados 15% pelo critério de ensino e 19,4% por pesquisa.

3.1.4 *QS World University Ranking*

O ranking da *Quacquarelli Symonds*, assim como o THE, em parte se baseia em pesquisas de reputação. Ele permite visualizar as classificações em categorias diferentes de tamanho, foco e intensidade de pesquisa. Isso possibilita que se evite comparar universidades de diferentes tamanhos e tipos umas com as outras.

Como se pode observar na tabela 3.4, 20% da pontuação é baseada em dados bibliométricos. Mas o indicador principal, representando metade da pontuação total, é baseado nos resultados de uma pesquisa de reputação na qual 5.000 empregados e cerca

de 15.000 acadêmicos e administradores de universidades do mundo todo participam com suas opiniões acerca das melhores universidades.

Tabela 3.4: Indicadores do *QS Ranking*

Critério	*Indicador*	*Peso*
Reputação	Em uma pesquisa de reputação, acadêmicos e empregadores são convidados a ranquear até 30 universidades no seu domínio de atuação (excluindo sua própria instituição)	50%
Citações	Número de citações por pessoa, baseado em dados do Scopus (período de 5 anos), juntamente com dados sobre o número de docentes	20%
Ensino	Relação faculdade/aluno	20%
Globalização	Percentagem de contratados internacionais	5%
	Percentagem de alunos internacionais	5%

3.1.5 *Performance Ranking of Scientific Papers for World Universities*

O ranking da *Higher Education Evaluation & Accreditation Council of Taiwan* (HEEACT) foi elaborado pela primeira vez em 2007. Baseando-se em números de artigos publicados em revistas e números de citações, são selecionadas as maiores universidades do mundo, totalizando 820 instituições no universo coberto pelo HEEACT. A classificação envolve métodos bibliométricos para analisar e classificar as produções científicas das 500 melhores universidades do mundo, levando em conta diversos indicadores, apresentados na tabela 3.5. As informações utilizadas provêm da *Thomson Reuters*' SCI, SSCI, e *Journal Citation Report* (JCR).

Tabela 3.5: Indicadores do *HEEACT's Ranking*

Critério	*Indicador*	*Peso*
Produção	Número de artigos nos últimos 11 anos	10%
	Número de artigos no ano atual	10%
Impacto	Número de citações nos últimos 11 anos	10%
	Número de citações nos últimos 2 anos	10%
	Número médio de citações nos últimos 11 anos	10%
Excelência	H-Index dos 2 últimos anos	20%
	Número de artigos altamente citados nos últimos 11 anos	15%
	Número de artigos em revistas de grande impacto no ano atual	15%

3.1.6 Análise e Comparação

Todos os rankings apresentados utilizam de alguma maneira dados recuperados da web. Quatro dos cinco apresentam no mínimo um indicador baseado em citações, e três utilizam no mínimo um indicador baseado em publicações. Cada ranking possui um diferencial, seja por considerar pesquisas de reputação, seja por valorizar premiações ou se basear em motores de busca.

Os cinco rankings apresentados possuem diferentes critérios de classificação, porém possuem um ponto em comum: visam classificar universidades do mundo todo. Mas como eles vêem as universidades do Brasil?

No ranking THE nenhuma universidade da América Latina aparece em seu *top-200*. No ranking da QS figuram algumas universidades nacionais, a melhor colocada é a USP, na 253ª posição, seguida pela UNICAMP, na 292ª colocação, e UFRJ na 381ª.

USP e UNICAMP são também primeira e segunda colocadas brasileiras, respectivamente, no ARWU. No intervalo da 301ª à 400ª posição se encontram UNESP, UFRJ e UFMG. Também figuram no ranking UFRGS e UNIFESP.

No ranking da HEEACT aparecem sete universidades brasileiras, as mesmas presentes no ARWU. As três melhores colocadas nacionais são a USP (74º lugar geral), UFRJ (341º) e UNICAMP (345º).

O WR utiliza informações provenientes de mecanismos de busca, como número de páginas recuperadas, número de *links* e quantidade de documentos em repositórios. Portanto, merece atenção ao basear suas classificações em recursos da web, tal como a proposta deste trabalho. A universidade brasileira melhor colocada segundo o WR é a USP, com uma notória 43ª colocação mundial. A UFRGS é a segunda do Brasil na versão mais recente do ranking, ocupando a 150ª posição geral. Na seqüência estão, respectivamente, UNICAMP, UFRJ e UFSC. O WR também é o ranking com maior número de universidades brasileiras, são 584 no total.

De acordo com os rankings estudados, a USP é a melhor universidade do Brasil, sendo de consenso comum sua primeira colocação em todos os rankings quando analisadas somente universidades brasileiras. UNICAMP e UFRJ também figuram entre as melhores em todos os rankings.

3.2 Estado da Arte

Aguillo (2006) analisou a presença de universidades brasileiras na web. Segundo o autor, os países em desenvolvimento da América Latina estão fazendo esforços para publicar eletronicamente os resultados de suas pesquisas e estudos. O tamanho de seus domínios web tem crescido, assim como sua visibilidade na web. Os resultados apontam uma presença cada vez mais forte das universidades do Brasil na web, mas ainda longe dos países desenvolvidos. Os dados do estudo foram obtidos de oito motores de busca. Os indicadores foram o número de páginas (tamanho), número de *links* externos (visibilidade) e número de visitas (popularidade). A USP liderou os rankings conforme os três indicadores. UNICAMP foi a segunda em tamanho e visibilidade, UFRJ a segunda em popularidade.

No trabalho reportado pela NordForsk (2011) foram comparados diversos rankings internacionais de universidades avaliando o posicionamento das universidades nórdicas. É feita uma crítica aos indicadores compostos utilizados pelos rankings da web e é sugerido que indicadores bibliométricos avançados provêem a informação necessária melhor que rankings para a avaliação das instituições. Também é salientada a importância da caracterização de diferentes perfis de pesquisa para comparações mais justas. Métodos de normalização propostos possibilitam a comparação de instituições de diferentes perfis.

Uma análise acerca de como são representados no buscador *Google Scholar* dezesseis departamentos de sociologia Nórdicos (AALTOJÄRVI, 2008) revelou que

VW, citações e posicionamento se reforçam mutuamente. Os fatores são influenciados pelo gênero do corpo docente, tipo e idade da publicação. Ao final, conclui-se que nenhum método individual para medir produção científica é objetivo. O número de acessos foi considerado como indicador de VW.

Na dissertação de Jouris (2011) é apresentado um modelo para o processo de avaliação de conferências a partir de dados obtidos na web. O modelo pode ser utilizado como um indicador de qualidade para Conferências Científicas. Os indicadores webométricos utilizados foram: citações, a partir do Google Acadêmico; visibilidade, medida por *inlinks*; *pagerank* do *site* da conferência; H-index da conferência. Um novo ranking de conferências baseado na VW demonstra uma nova maneira de medir a qualidade das conferências.

Indicadores de VW foram examinados para avaliar até que ponto a colaboração de publicações em ciências e tecnologia é visível na web (AGUILLO; KRETSCHMER, 2004). O estudo revelou que cerca de 80% da bibliografia com múltiplos autores é visível através de motores de busca. Já os estudos de Kretschmer (2007) mostram que as estruturas de *hyperlinks* não refletem as estruturas de colaboração dos dados bibliográficos, porém indicadores de VW são diferentes de *hyperlinks* e podem ser usados com sucesso como indicadores web de colaboração.

Barjak (2008) reporta os resultados de um estudo sobre a relação entre a contagem de *inlinks* e a real relevância dos *sites* de cerca de 400 grupos de pesquisa europeus. A análise confirmou que o tamanho dos grupos de pesquisa e suas presenças na web são importantes para atrair *links*, ao passo que a produção científica em si não. A interpretação dos dados provenientes de motores de busca precisa ser mais bem estudada antes de tirarem-se conclusões sobre sua utilidade como indicadores, conclui.

Porém dados de motores de busca são amplamente utilizados em diversos trabalhos e com diferentes finalidades. Uma rápida pesquisa na web revela as muitas aplicações, além dos indicadores para universidades apresentados, tais como: avaliadores de domínio (DNSCOOP, 2011; CUBESTAT, 2011), que fornecem um valor em moeda para o *site* alvo; árvores de palavras, que se baseiam nas consultas dos usuários de motores de busca com o propósito de identificar o que as pessoas dos diferentes sexos pensam (VIEGAS, 2011); rankings de *sites* populares na web (ALEXA, 2011).

Um trabalho prévio utilizou noções de relevância e precisão de rankings aliados a metabusca com fusão de rankings para elaborar um cálculo de VW (KLINGER; LIMA; OLIVEIRA, 2011). A fórmula proposta pode ser aplicada a domínios homogêneos, como o caso de estudo apresentado das universidades brasileiras. Os resultados servem como indicativo de que os motores de busca provêem dados interessantes para a avaliação de visibilidade e estudos futuros apontam para a aplicação do cálculo de maneira genérica, ou seja, a qualquer coisa existente na web a qual se queira medir a visibilidade na rede. A pesquisa serviu de base para a presente dissertação.

4 CÁLCULO DE VISIBILIDADE WEB

Como visto no capítulo anterior, a visibilidade de uma instituição na web pode ser medida de diversas formas. A forma mais comum é a contagem de *links* externos que levam ao *site* da instituição, como utilizado pelo WR. Outra maneira de medir a presença na rede pode ser representada pela quantidade de páginas recuperadas em motores de busca, ou de artigos indexados. A maioria dos rankings de universidades da Internet utiliza, em grande parte, indicadores bibliométricos, especialmente citações. A hipótese é que se pode construir um novo indicador de VW para universidades e validá-lo através de experimentos e comparações com rankings da web, especialmente com o WR por ser o mais completo.

A contribuição deste trabalho é uma nova maneira de calcular visibilidade na web, um indicador baseado em dados provenientes de motores de busca. Ao realizar uma pesquisa em algum buscador, os usuários tendem a olhar somente os primeiros resultados. Ou seja, um *site* bem colocado em um motor de busca possui, em outras palavras, uma boa visibilidade naquele *site* de busca. É essa a idéia a ser explorada: a colocação da página em diversos motores de busca, ao invés do número de páginas sobre o domínio da instituição, número de documentos, citações ou *links*.

Considerando a página web oficial de uma instituição como marco principal de sua presença na rede, é esperado que ao consultar um motor de busca usando como argumento de busca o nome da instituição, ou sua sigla, a página oficial esteja entre as primeiras colocadas. Isso significaria que o *site* da instituição em questão, e conseqüentemente o seu próprio nome, possuem uma boa visibilidade de acordo com o motor de busca utilizado. Boas colocações estão sempre atreladas a uma boa VW.

Para uma instituição alvo a qual se pretende medir a visibilidade na web, o cálculo de visibilidade por metabusca na web baseia-se nas posições do *site* oficial da instituição em *m* rankings, procedendo da seguinte maneira:

1. Busca pela sigla (ou nome) da instituição em *m* motores de busca;
2. Busca pela página oficial dentro de cada um dos *m* rankings;
3. Distribuição da pontuação de acordo com posição em cada ranking;
4. Valor de visibilidade resultante da soma dos pontos.

A consulta é primeiramente enviada aos *m* motores de busca e somente os dez primeiros colocados de cada ranking são considerados. Dentre estes são buscadas as colocações do *site* oficial da instituição. Para cada colocação é atribuída uma pontuação, de acordo com o método *Borda Count*, assim para a primeira colocação são atribuídos dez pontos, nove para a segunda, oito para a terceira, até a décima colocação que recebe somente um ponto. A soma desses pontos resulta no valor de visibilidade, o qual poderá atingir o máximo de dez vezes *m*.

É utilizado o sistema de pontuação tradicional do BC, pois os motores de busca exibem seus resultados em forma de lista, apresentado o primeiro colocado praticamente tão visível (aos olhos do usuário) quanto o segundo colocado, e assim sucessivamente. Então não é desejável dar uma ênfase maior à primeira colocação, como sugere o método de pontuação alternativo. Experimentos anteriores (KLINGER, 2011) efetuaram fusão de rankings antes do momento do cálculo da VW, mas o novo método não necessita de tal complexidade, objetivando um cálculo simples e claro.

São utilizados somente os dez primeiros dos rankings com base em experimentos realizados, descritos nas próximas seções.

Um exemplo do cálculo do indicador de visibilidade, dada a sigla UFRGS:

1. A sigla UFRGS é aplicada como argumento de busca no sistema de metabusca o qual repassa a sigla para três motores de busca: x, y, z;
2. O *site* www.inf.ufrgs.br é encontrado em primeiro lugar no motor de busca x, em segundo no motor de busca y e em oitavo no motor de busca z;
3. A pontuação é dada de acordo com as colocações em cada motor de busca, x: dez pontos; y: nove pontos; z: três pontos;
4. Somando os pontos, visibilidade igual a vinte e dois.

4.1 Parâmetros

Os principais parâmetros do indicador de visibilidade proposto para universidades são dois:

- Sigla: a sigla da universidade a ser avaliada (ex.: ufrgs);
- *Site* oficial: a página web oficial da instituição (ex.: www.ufrgs.br);

Além destes, também são necessários para o cálculo a determinação dos motores de busca envolvidos e o tamanho de rankings a ser considerado.

No quesito motores de busca é interessante avaliar a região em que os mesmos atuam. O Google, por exemplo, direciona as buscas de forma a priorizar as páginas da região em que o usuário está acessando. Outros, como o próprio Google, além de Yahoo e Alta Vista, possuem versões brasileiras, que priorizam os *sites* do Brasil, e versões mundiais, não priorizando os resultados do país. Ainda, nem todos os buscadores da web são de propósito geral, alguns são focados, além da questão de região, em determinados assuntos, tais como medicina ou puramente notícias, livros, blogs, etc.

Portanto, os motores de busca envolvidos no cálculo de visibilidade devem atender a necessidade básica de retornar resultados coerentes com o alvo da busca. Na aplicação em questão, foram selecionados quatorze motores de busca de propósito geral que recuperam os *sites* das universidades brasileiras, as quais constituem o caso de estudo do indicador de visibilidade proposto. São eles: versões brasileiras do Alta Vista, Ask, Google e Yahoo; versões globais do Alexa, All the Web, Alta Vista, Ask, Aol, Exalead, Google, Icerocket, Lycos e Yahoo.

Quanto ao tamanho dos rankings, foi determinado pela avaliação de precisão dos motores de busca conforme a subseção a seguir.

4.1.1 Tamanho de Rankings VS. Precisão dos Motores de Busca

Foram realizados experimentos com motores de busca individualmente para analisar os seus comportamentos em buscas por siglas de universidades. O objetivo principal

destes experimentos foi determinar o tamanho dos rankings a serem utilizados para o cálculo de visibilidade e observar o conteúdo recuperado de acordo com a consulta submetida, permitindo a análise de precisão.

Em doze motores de busca foram realizadas consultas com as siglas de algumas universidades brasileiras, dentre elas a UFRGS. Analisou-se a relevância dos *sites* retornados em rankings de diferentes tamanhos. Foram considerados *sites* relevantes todos àqueles que pertencem à própria universidade, ou que fazem referência a ela.

Utilizou-se, de cada motor de busca, rankings de tamanho $n = \{10, 20, 30, 40, 50\}$, sendo que os participantes de cada ranking são os n primeiros colocados. A precisão foi calculada através da divisão do número de *sites* relevantes recuperados no *top-n* pelo total de *sites* no ranking (n). Ou seja, foi utilizada a medida *precision at k*, onde $k = n$.

O gráfico para a sigla UFRGS (Figura 4.1:) mostra os valores de precisão para cada tamanho de ranking. Pode-se ver uma tendência aos rankings ficarem mais poluídos, isto é, contendo menos *sites* relevantes conforme mais resultados são considerados. Nota-se que, para $n = 10$, o menor valor de precisão para a sigla UFRGS foi 0.8, enquanto que, para $n = 50$, o buscador Lycos chegou a cair para 0.5 sua precisão. De modo geral, a diversidade de valores de precisão é claramente maior aumentando-se o valor de n.

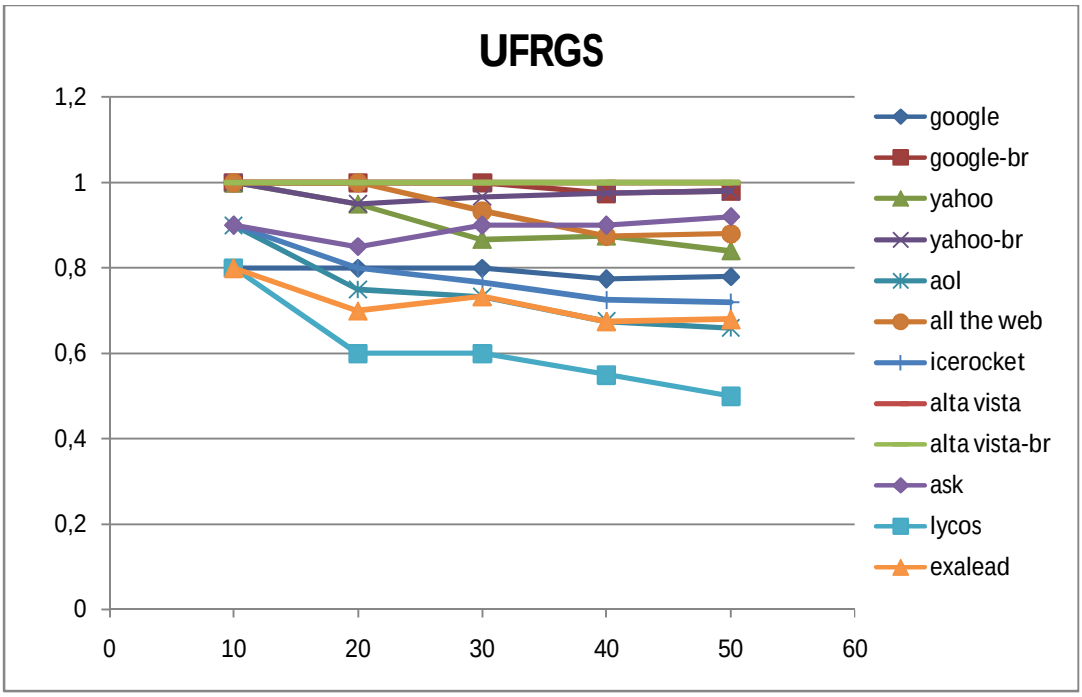

Figura 4.1: Precisão do *top-n* em 12 buscadores para a sigla UFRGS

Como ilustrado no gráfico acima e observado com os experimentos, quanto maior o ranking, pior a precisão, porque mais *sites* de menor relevância para a consulta são incluídos. Baseado nesse fato determinou-se a utilização do *top-10* somente de cada motor de busca.

4.2 Protótipo

Um protótipo foi elaborado para fins experimentais, possibilitando a elaboração eficiente de rankings de instituições e o cálculo automatizado do novo indicador de visibilidade auxiliado por metabusca.

O sistema está subdividido em quatro módulos: Interface do Usuário, Motores de Busca, Núcleo e Fusão. A linguagem de programação utilizada foi Java.

4.2.1 Visão Geral

A figura 4.2 ilustra, de forma geral, o funcionamento do protótipo. Para o cálculo do indicador de VW, o usuário entra com a sigla da universidade e o *site* oficial na interface de calculadora de visibilidade. A sigla é repassada ao Núcleo, onde são efetuadas todas as consultas aos diversos motores de busca e é montada uma matriz, entregue ao módulo de Fusão. Com o auxílio do endereço do *site* oficial da universidade informado no *applet*, é calculado o valor de visibilidade, em cima da matriz, de acordo com o BC. Então, o valor é finalmente exibido no *applet* da calculadora de visibilidade.

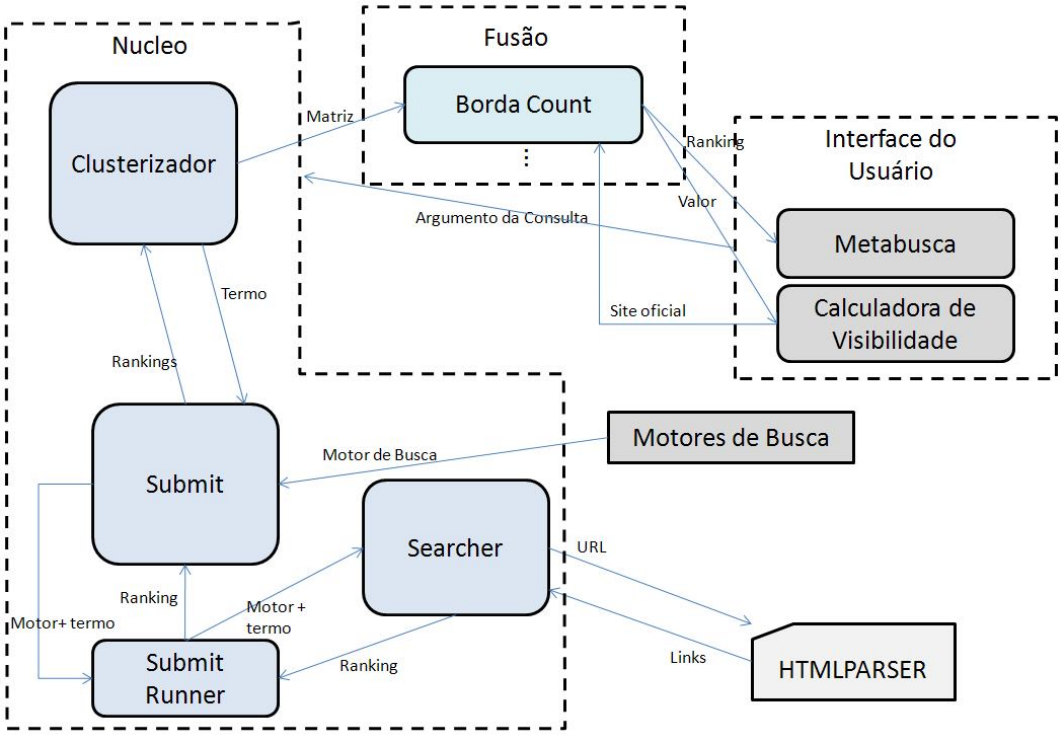

Figura 4.2: Esquema de funcionamento e organização do protótipo

A seguir são descritos os módulos e suas classes internas.

4.2.2 Interface do Usuário

O módulo contém as classes de entrada e saída de dados. Dois *applets* se encarregam da comunicação com os algoritmos principais do Núcleo e exibição dos resultados.

A escolha de *applets* como interface se deu por conta de sua fácil publicação na web e não necessidade de servidor, uma vez que executam diretamente na máquina do

cliente. Um *applet* pode ser utilizado através de qualquer navegador com suporte a Java, a partir de qualquer máquina que possuir uma JVM instalada.

Uma das interfaces do sistema é um metabuscador. Ele pode ser utilizado para observar o comportamento dos motores de busca de acordo com diferentes termos alvo. Não é de importância para o cálculo de visibilidade, mas pode servir como, além de curiosidade, no caso das universidades, para visualizar o *site* oficial da instituição em um ranking geral com, inclusive, a exibição de suas colocações em cada motor de busca. A figura 4.3 apresenta esta interface de metabusca.

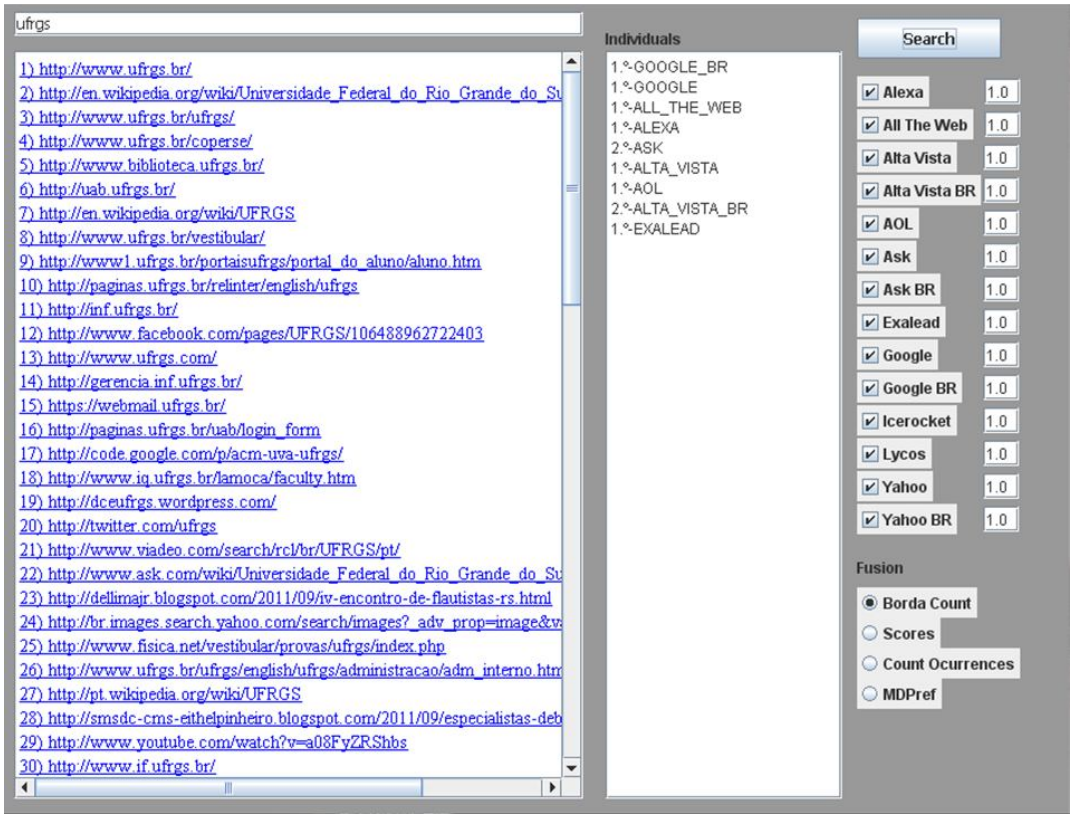

Figura 4.3: Interface de metabusca

A interface segue os padrões adotados pela maioria dos sistemas de busca, com uma caixa de texto para o usuário digitar o termo da pesquisa e um botão de busca. Abaixo são exibidos os resultados. Ao lado são exibidos os motores de busca e colocações em que foi encontrado o *link* que estiver selecionado pelo cursor do mouse (no momento em que foi extraída a figura 4.3, o mouse estava posicionado no primeiro *link* do ranking). À direita, podem-se selecionar os motores de busca desejados e atribuir pesos diferentes a eles, possibilitando um estudo futuro sobre um diferente balanceamento de pesos. Os botões de rádio, no canto direito inferior, permitem escolher outros métodos de fusão de rankings que não o *Borda Count*, como o MDPref, utilizado em outro trabalho (KLINGER, 2011), e métodos que poderão ser utilizados futuramente.

Mas a interface principal é a calculadora de VW (figura 4.4). É ela que, a partir dos parâmetros do usuário, atribui um valor de visibilidade à instituição desejada. Em <Objeto> é colocado o alvo da busca, ou seja, uma sigla de universidade. Em <Site> é

colocado o endereço do *site* oficial da universidade. Após clicar no botão <Calcular Visibilidade>, é exibido o valor correspondente abaixo e o total de pontos que poderia atingir de acordo com o número de motores de busca utilizado. No exemplo da figura, o *site* da UFRGS obteve uma pontuação de visibilidade de 116 de um máximo de 140, que corresponderia à pontuação máxima em todos os quatorze motores de busca utilizados.

Figura 4.4: Interface de calculadora de VW

Os *applets* podem ser acessados através do *site* do autor[1], disponibilizados a qualquer pessoa a título de curiosidade.

4.2.3 Motores de Busca

Todas as informações referentes aos motores de busca utilizados para o cálculo de visibilidade estão contidas neste módulo. Para cada motor de busca é armazenada a sua forma de consulta (*query*), seu separador (caractere usado para separar palavras no lugar de um espaço em branco) e a classe CSS dos resultados da consulta. Não é realizada nenhuma ação dentro deste pacote. Todos os motores de busca são descritos por um *enum*[2].

O formato de cada consulta, bem como o separador, são importantes para o sistema de metabusca repassar o argumento da consulta para o motor de busca em questão. A consulta montada será usada para obter os resultados, através do pacote *htmlparser*[3], em métodos do módulo Núcelo do sistema. A figura 4.5, a seguir, identifica os elementos de uma consulta base.

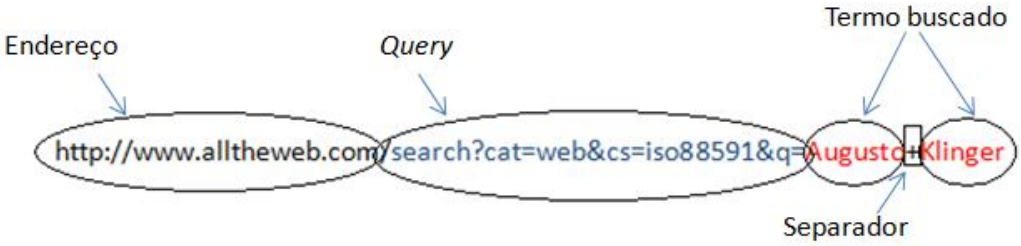

Figura 4.5: Formato de uma consulta

---

[1] http://www.inf.ufrgs.br/~aklinger

[2] Tipo enum é um tipo de dado cujos campos consistem de um conjunto fixo de constantes

[3] Biblioteca Java de *parsing*, disponível em: http://htmlparser.sourceforge.net/

A classe CSS serve para identificação dos elementos do ranking retornado pelo motor de busca. O pacote *htmlparser*, quando invocado dentro do Núcleo, irá ler e obter os dez primeiros *links* que estiverem com esta classe dentro da página retornada pela consulta. Dessa maneira, somente motores de busca que possuem uma classe CSS única para identificação dos resultados são utilizados no protótipo, padronizando o processo de metabusca.

4.2.4 Núcleo

As classes do núcleo disparam as consultas aos motores de busca, lêem os resultados e preparam os dados. Ao todo são quatro classes, descritas nesta subseção.

4.2.4.1Submit

Classe que recebe o argumento da consulta como parâmetro e passa para todos os motores de busca. Cada motor de busca é invocado dentro de uma *thread*[1] diferente, que é implementada pela classe *SubmitRunner* (4.2.4.2). Após todos os motores de busca terem devolvido seus resultados, eles são armazenados em uma lista de rankings, que será usada pelo Clusterizador (4.2.4.4).

4.2.4.2SubmitRunner

Classe que estende a classe *Thread* do Java. Basicamente, ao disparar uma instância de *SubmitRunner*, é instanciado um objeto *Searcher* (4.2.4.3) e invocado o método que efetua de fato a busca. O ranking resultante é devolvido pelo *Searcher* e adicionado a uma lista parcial dos rankings já obtidos.

Sumariamente, o método *run()*, principal de uma *Java Thread*, invoca uma consulta a um motor de busca, verifica se retornou resultados e então os armazena. O nome do motor de busca é armazenado, neste momento, em outra estrutura paralela, tornando assim possível a identificação do buscador que retornou dado ranking.

4.2.4.3Searcher

Contém o método que realiza as buscas, o qual recebe dois parâmetros de entrada: o motor de busca e o argumento da consulta. Primeiramente é montada a consulta, conforme apresentada na figura 4.5. De acordo com o motor de busca recebido como parâmetro, os espaços em branco do argumento da consulta, quando presentes, são substituídos pelo caractere separador, e o argumento é então concatenado a *query* e ao endereço, formando uma URL.

A URL formada é passada como parâmetro para o objeto *Parser* do pacote *htmlparser*, que extrai todos os *links* das páginas. Em cada *link* é filtrado o atributo *class*e são adicionados ao ranking somente os *links* que possuem a classe CSS de resultado da consulta, identificada previamente. Assim são evitados os demais *links* de patrocínios, propagandas, navegacionais e outras coisas que não são pertinentes ao ranking.

4.2.4.4Clusterizador

Classe responsável pela preparação dos dados, dispondo os diversos rankings em uma estrutura matricial. O Clusterizador recebe o termo alvo da busca e repassa para sua instância de Submit (4.2.4.1), a qual irá devolver os rankings de todos os motores de

---

[1] Forma de um processo dividir a si mesmo em duas mais tarefas que podem ser executadas concorrentemente

busca. A seguir é feita uma lista com todas as páginas web encontradas. Para *links* diferentes, mas que levam a uma mesma página, como por exemplo www.site.com/index, www.site.com/index.htm e www.site.com, é armazenada somente uma entrada na lista.

A partir da lista de *sites* é criada uma matriz. Cada *site* da lista corresponde a uma linha da matriz e cada motor de busca é representado por uma coluna. Em cada célula da matriz é armazenada a colocação em que o *site* representado pela linha obteve no motor de busca correspondente a coluna. Quando um *site* não figura no ranking de um motor de busca a célula correspondente recebe o valor zero, conforme ilustrado na figura 4.6. A matriz é utilizada pelo módulo de fusão, descrito a seguir.

Lista

Sites	Buscador X	Buscador Y	Buscador Z	Buscador W
www.ufrgs.br	1	2	1	1
neaufrgs.wordpress.com	3	0	2	0
queroentrarnaufrgs.blogspot.com	2	3	3	2
www.poa.ifrs.edu.br	7	4	4	3
pt.wikipedia.org/wiki/Universidade_F...	6	1	5	4
noticias.terra.com.br/vestibular/interna	5	5	6	10
twitter.com/ufrgs	8	6	7	5
www.ppgbm.com.br	4	7	8	6
www.universitario.com.br	0	8	9	9
bcufrgs.blogspot.com	9	9	10	7
dceufrgs.wordpress.com	10	10	0	8

Motores de busca

Colocações do site no ranking

Figura 4.6: Composição da matriz

O nome da classe não faz referência ao termo utilizado na área de mineração de dados. O Clusterizador em questão, em suma, agrupa rankings.

4.2.5 Fusão

Módulo responsável por métodos de fusão de rankings. Essencialmente, possui uma implementação do método BC. A partir da matriz descrita anteriormente, os *sites* recebem uma pontuação e são ordenados conforme ela para exibição do ranking no *applet* de metabusca. Para cada primeira colocação o *site* recebe dez pontos, nove pontos para segunda, oito para a terceira, até a décima colocação que soma somente um ponto.

Para o cálculo de VW é utilizado outro método, no qual são somados somente os pontos do *site* oficial da universidade em questão e é retornado, ao invés de uma lista, um valor correspondente a visibilidade. Somente é considerada uma colocação por ranking, assim se, por exemplo, em um motor de busca aparecer o endereço www.ufrgs.br/ufrgs em primeiro lugar e aparecer o endereço www.ufrgs.br em uma posição inferior, é computada somente a melhor colocação, nesse caso a do endereço www.ufrgs.br/ufrgs. Ou seja, é levada em conta a página web mais bem colocada sob o domínio da instituição. Assim, não é somente a linha da matriz correspondente ao *site* oficial da universidade que é necessária. A cada coluna da matriz (motor de busca) é buscada a página sob o domínio da instituição de melhor colocação, quando houver.

Alternativamente, a classe acomoda a possibilidade de receber por parâmetro diferentes pesos para cada motor de busca. Porém não foi estudada nenhuma

distribuição de pesos e tal parâmetro foi ignorado. Todas as pontuações são uniformes, mas seriam multiplicadas pelo peso correspondente ao motor de busca caso fosse desejado.

Também outros métodos de fusão de rankings podem ser incluídos no módulo de fusão, desde que operem com a matriz montada pelo Clusterizador.

4.2.6 Tecnologias

O sistema foi construído com o auxílio das tecnologias apresentadas na tabela 4.1. Foram utilizados dois pacotes de terceiros, encontrados na web. Um deles para efetuar operações matriciais eficientemente, facilitando o trabalho com rankings, e o outro para realizar o *parsing*[1] das páginas de resultados dos motores de busca, recolhendo os *sites* dos rankings gerados.

Tabela 4.1: Tecnologias envolvidas na implementação do protótipo

Categoria	*Tecnologia*	*URL*
Ambiente de Desenvolvimento	NetBeans 6.7	http://www.netbeans.org/
Linguagem de Programação	Java	http://java.sun.com
Biblioteca Matricial	JAMA	http://math.nist.gov/javanumerics/jama/
Biblioteca de Parsing	HTML Parser	http://htmlparser.sourceforge.net/

[1] Análise de palavras; processo de analisar uma seqüência de entrada para determinar sua estrutura gramatical.

5 EXPERIMENTOS NO DOMÍNIO DAS UNIVERSIDADES

As universidades constituem um domínio homogêneo, onde cada instituição tem uma sigla e uma página na web que não varia muito de certo padrão, caracterizando, assim, um caso de estudo ideal para a elaboração de rankings baseados na visão de seus *sites* oficiais pelos motores de busca.

Foram vistos alguns rankings da web que classificam universidades e observou-se que, exceto pelo WR, as universidades brasileiras não possuem significativas participações. Nota-se que há espaço para novos rankings e formas de classificar universidades através de suas presenças e dados disponíveis na web. A seguir, é apresentada a aplicação do indicador proposto por este trabalho ao domínio das universidades brasileiras.

5.1 Ranking de Universidades

Trinta universidades do Brasil foram submetidas ao cálculo do indicador de visibilidade baseado em metabusca na web. As universidades participantes do ranking foram escolhidas com base no ranking nacional do WR para fins comparativos. A sigla de cada universidade foi utilizada como parâmetro de consulta no protótipo elaborado, junto com os respectivos *sites* oficiais previamente identificados, revelando a visibilidade da sigla vinculada à universidade na web. A tabela 5.1 contém o ranking das trinta universidades amostradas no experimento, sendo 140 o valor de visibilidade máximo, correspondendo ao primeiro lugar nos quatorze motores de busca empregados no cálculo. Não foi utilizado critério de desempate, universidades com mesma pontuação pertencem, na prática, à mesma colocação.

No ranking da tabela 5.1 observa-se a PUC-SP como líder e única com pontuação máxima. Logo após está a PUC-Rio, mostrando que ambas as siglas da Pontifícia Católica possuem um ótimo poder de descriminação referente às suas instituições e uma ótima visibilidade aos olhos dos motores de busca. A UNIFESP completa o *top-3*, seguido de cinco universidades empatadas na quarta colocação, dentre elas a UFRGS, e mais duas empatadas na nona posição. No ranking do WR, constante no anexo A, PUC-SP é a trigésima, enquanto que PUC-Rio está apenas na 13ª colocação. USP lidera o ranking, seguida pela UFRGS.

Tabela 5.1: Ranking de VW por metabusca

Colocação	*Sigla (Universidade)*	*Pontuação*
1º	PUC-SP	140
2º	PUC-Rio	138
3º	UNIFESP	137
4º	UFRGS	135
5º	PUCRS	135
6º	UFRN	135
7º	UFSM	135
8º	UERJ	135
9º	UFRJ	134
10º	UFSCAR	134
11º	UNICAMP	133
12º	UFPE	131
13º	UFPB	131
14º	UNISINOS	131
15º	UFPR	130
16º	UFSC	129
17º	UFMG	127
18º	FGV	126
19º	UFPA	124
20º	UFBA	123
21º	UNESP	122
22º	UFG	112
23º	UFU	110
24º	UFV	83
25º	UFF	82
26º	UEM	77
27º	UEL	66
28º	USP	63
29º	UNB	60
30º	UFC	48

Considerando o *top-10* do ranking elaborado, e também o *top-10* dos rankings em anexo (A, WR; B, valor do domínio), somente duas universidades, UFRGS e UFRJ, encontram-se entre as dez nos três rankings. Uma das universidades do *top-10* do ranking elaborado, a PUC-Rio, figura também no *top-10* por valor do domínio da instituição (anexo B).

Com relação ao *top-10* dos dois rankings anexados, oito universidades são comuns, havendo apenas divergências de colocações. Isto serve de indicativo de falhas na forma de cálculo do indicador apresentado. Ainda mais levando em conta que USP e

UNICAMP, bem colocadas nos dois rankings e reconhecidamente duas das principais universidades brasileiras, não atingiram boas colocações.

5.1.1 Discussão e Problemas Encontrados

Um grande obstáculo em se realizar consultas utilizando somente uma sigla, ou um nome de organização, é que pode haver outras organizações de outros segmentos compartilhando a mesma identificação. Ou seja, falta semântica para determinar o domínio em que se está medindo a visibilidade.

O problema dos homônimos ocorre quando outra organização possui a mesma sigla de uma universidade. Assim, *sites* não relacionados à universidade competem juntamente no ranking por colocações, favorecendo amplamente aquelas universidades cuja sigla está somente ligada a ela. Tal problema prejudicou diversas instituições no ranking apresentado anteriormente.

Um exemplo é a UFC, ultima colocada na tabela 5.1. A Universidade Federal do Ceará utiliza a mesma sigla da liga de vale-tudo *Ultimate Fighting Championship*, que além de compartilhar a sigla, é uma organização muito mais famosa mundialmente. O gráfico de precisão (figura 5.1) ilustra o problema da UFC.

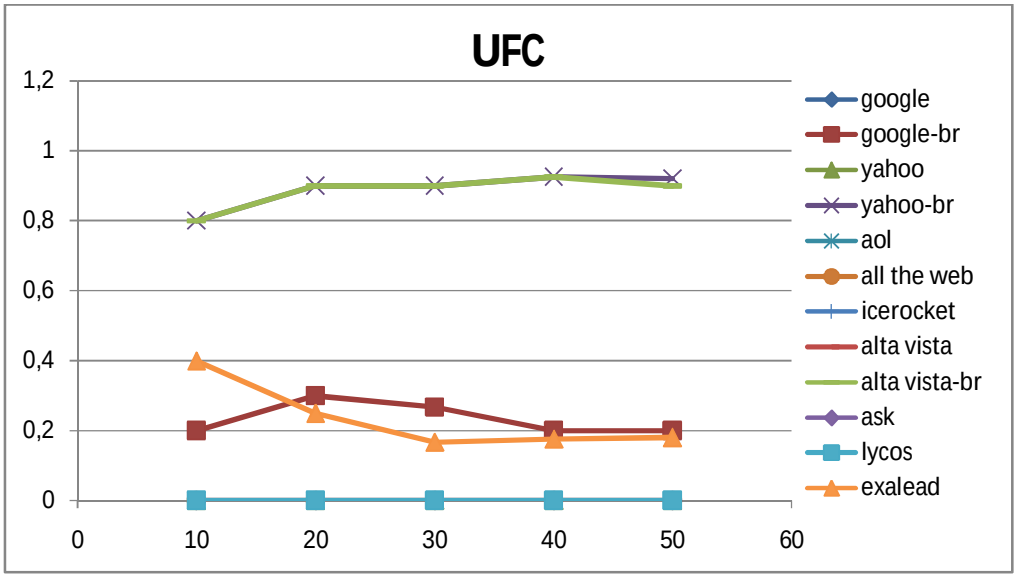

Figura 5.1: Precisão do *top-n* em 12 buscadores para a sigla UFC

Dos doze motores de busca dos quais foram tiradas as medidas de precisão, oito mantiveram valor zero ao longo da variação de *n* de dez até o tamanho cinqüenta. Mesmo considerando cinqüenta *sites*, nenhum foi relevante em relação à Universidade Federal do Ceará. Somente Yahoo-Br e Alta Vista-Br retornaram resultados favoráveis, sendo que Google-Br e Exalead ainda conseguiram recuperar alguns *sites* pertinentes a universidade.

Para contornar o problema exposto, faz-se necessário um método de refinamento que restrinja o domínio dos experimentos, tornando a pontuação de visibilidade mais justa entre as siglas.

5.2 Expansão de Consulta

Uma técnica bastante popular na área de IR é a expansão de consultas. Basicamente consiste em acrescentar algumas palavras ao argumento de consulta a fim de discriminar melhor o que se deseja recuperar. É uma maneira fácil e eficiente de adicionar alguma semântica ao processo, sem a necessidade de implementação de filtros ou análises de relevância por especialistas.

Como o indicador de VW baseia-se em *sites* referentes a universidades, a palavra "universidade" (sem as aspas) foi incluída ao lado da sigla nas consultas. O novo gráfico de precisão (figura 5.2) para a sigla UFC demonstra uma melhora significativa às precisões medidas no gráfico anterior. Observa-se que, desta vez, nenhum motor de busca zerou a precisão em qualquer tamanho de ranking e, ainda, oito deles obtiveram valor máximo no *top-10*. Analisando o ranking resultante da fusão com o BC, o *site* da UFC pulou da vigésima primeira para a primeira posição.

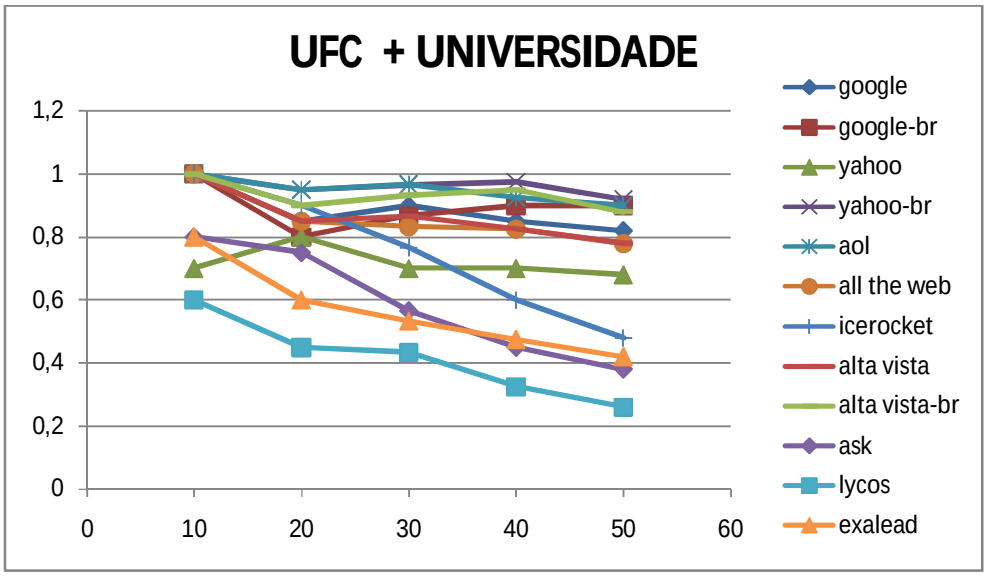

Figura 5.2: Precisão do *top-n* em 12 buscadores para a sigla UFC + "universidade"

Claro que se houvesse outra organização também relacionada a universidades partilhando da mesma sigla o problema dos homônimos persistiria em certo grau. Neste caso a solução poderia estar no uso de uma ontologia, porém a abordagem seria mais complexa.

Tratando-se da aplicação às siglas de universidades brasileiras do indicador elaborado, a técnica de expansão de consultas resolveu o problema dos homônimos, adicionando a semântica necessária.

Com o auxílio do protótipo de metabusca foi observado que acrescentar a palavra "universidade" não implica em posicionar o *site* oficial da instituição na primeira colocação sempre. Muitas vezes *sites* que se referem à instituição estão melhores posicionados pelos critérios de certos motores de busca. A expansão de consulta simplesmente torna mais justa a classificação das instituições, reduzindo o número de resultados irrelevantes.

5.2.1 Novo Ranking

A tabela 5.2 apresenta o novo ranking de universidades segundo o indicador proposto, agora sem o problema dos homônimos. Foram utilizadas as mesmas siglas da versão anterior. Novamente, universidades com mesma pontuação pertencem, na prática, a mesma colocação da melhor colocada de mesma pontuação, desde que não foram empregados critérios de desempate.

Tabela 5.2: Ranking de VW com expansão de consulta

Colocação	*Sigla (Universidade)*	*Pontuação*
1º	UNICAMP	140
2º	UEM	140
3º	UFV	140
4º	USP	136
5º	UFRGS	136
6º	UNB	136
7º	PUC-Rio	136
8º	UFPE	136
9º	UFF	136
10º	UFRN	136
11º	PUC-SP	136
12º	UFPB	136
13º	UNESP	135
14º	UFSC	135
15º	UFC	135
16º	UNIFESP	135
17º	UFMG	132
18º	UFSCAR	132
19º	UFG	132
20º	UNISINOS	131
21º	UFU	130
22º	UFPR	129
23º	PUCRS	129
24º	UFBA	128
25º	UFPA	126
26º	UFSM	120
27º	UERJ	118
28º	FGV	117
29º	UFRJ	112
30º	UEL	90

Na primeira colocação houve um empate entre três universidades, as quais, inclusive, não estavam entre as dez melhores classificadas anteriormente, agora com

pontuação máxima. UNICAMP, UEM e UFV obtiveram uma pontuação de 133, 77 e 83 no experimento anterior, respectivamente, obtendo pontuação de 140 na versão com consulta expandida. Logo após, nove universidades ficaram empatadas com 136 pontos, dentre elas PUC-SP e PUC-Rio, mostrando que a técnica aumentou o nível de ruído nos rankings gerados pelos motores de busca, que já possuíam uma ótima precisão para estas siglas, mas não comprometeu, já que continuam bem colocadas.

Ainda na tabela 5.2, salienta-se a boa colocação da USP, que subiu de 63 para 136 pontos, entrando no *top-10*. A UFC, prejudicada anteriormente, subiu para a décima terceira colocação na prática, obtendo 135 pontos em contraste com os 48 obtidos anteriormente.

De modo geral, a palavra "universidade" junto da sigla elevou a pontuação das universidades amostradas. Na nova versão do ranking, a universidade com menos pontos ficou com 90, em contraste com 48 no primeiro experimento. O total de pontuação de todas as trinta universidades, utilizando a consulta somente pela sigla foi somado 3466 pontos, e utilizando a consulta expandida a soma chegou a 3910 pontos. Algumas universidades foram prejudicadas pela expansão de consulta, como a UFRJ que caiu vinte posições. Este é um caso em que a sigla já é bastante expressiva e que o uso da palavra universidade só aumenta o nível de poluição nos rankings. A solução está em uma maneira mais elaborada de adicionar semântica ao cálculo de visibilidade. Mas, essencialmente, o novo ranking apresenta uma classificação mais justa para as siglas de universidades brasileiras, contornando o problema dos homônimos.

O gráfico da figura 5.3 mostra as nove universidades que tiveram suas pontuações reduzidas ao aplicar a expansão de consulta. A UFRJ perdeu 22 pontos, UERJ 17 e UFSM 15. As demais siglas constantes no gráfico não tiveram um prejuízo tão significativo.

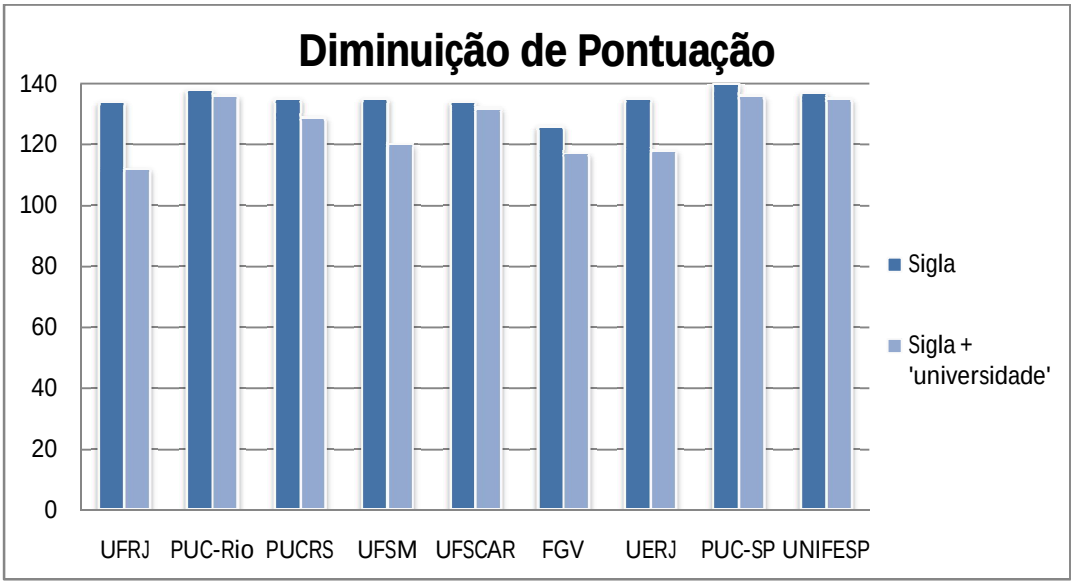

Figura 5.3: Gráfico das siglas que perderam pontuação

Em contraste, o gráfico da figura 5.4 apresenta as universidades que mais se beneficiaram com a técnica de expansão de consulta. A UFC ganhou 87 pontos, UNB e USP ganharam 76 e 73, respectivamente. Apesar de algumas perdas, a distribuição da pontuação se tornou muito mais homogênea.

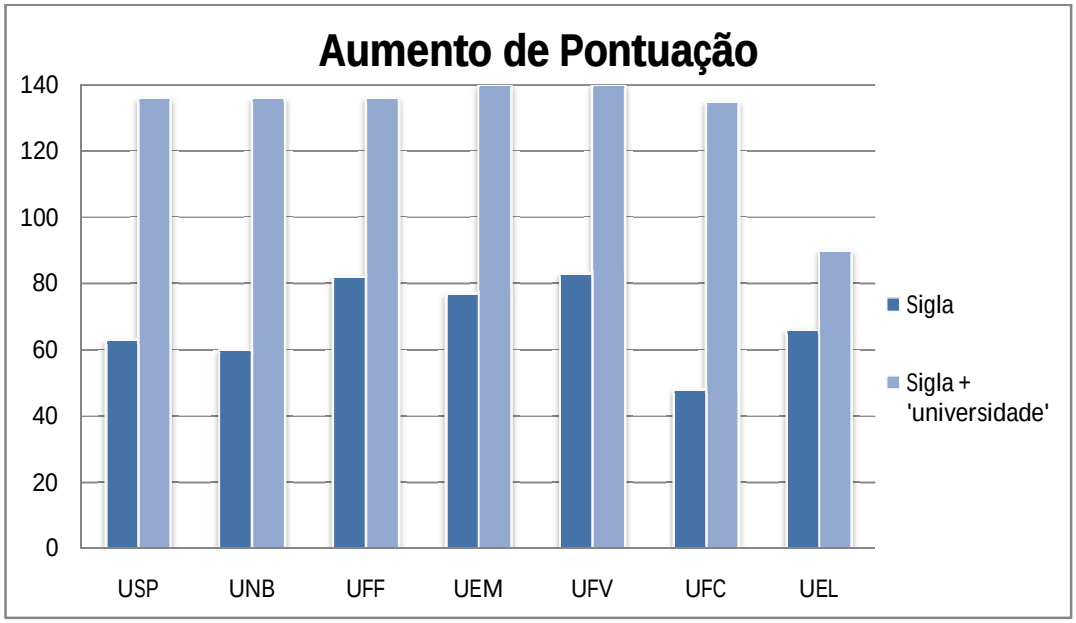

Figura 5.4: Gráfico das siglas que mais ganharam pontuação

Em comparação com os rankings anexados, os *top-3,* tanto do WR como do ranking por valor de domínio, estão contidos dentre os dois mais altos níveis de pontuação obtidos segundo o indicador de VW baseado em metabusca. Quatro universidades são comuns entre os três *top-10*.

Houve muitos empates, decorrente da distribuição mais uniforme de pontos. Ocorreram quatorze níveis de pontuação para trinta universidades. Considerando cada nível como somente uma colocação no ranking, sete universidades seriam comuns entre os *top-10* dos três rankings: USP, UFRGS, UNICAMP, UFSC, UFMG, UNB e UFPR. Sendo renomadas estas universidades do Brasil, o indicador de VW apresenta coerência.

Uma interessante forma de visualizar os resultados, dada a quantidade de empates, seria através da criação de categorias de níveis de VW, agregando em cada categoria universidades de mesma pontuação, ou até mesmo de pontuações próximas. Com um conjunto de siglas maior tal visualização deverá ser mais clara que o formato tradicional de ranking.

6 CONCLUSÕES

Foram apresentados diversos rankings de universidades que trabalham com dados recolhidos da web. A maioria dos rankings, a exceção do WR, não cobre grande parte das universidades brasileiras. Estes rankings são elaborados com base em indicadores compostos, os quais são bem variados, englobando critérios como número de páginas web recuperadas em motores de busca, ou artigos, ou documentos, mas essencialmente utilizam indicadores bibliométricos.

Um estudo sobre como é empregado o cálculo de VW na literatura revelou que a maioria dos trabalhos analisa as estruturas de *hyperlinks*, por exemplo, efetuando a contagem de *links* externos para determinar um valor de visibilidade.

Foi apresentada uma forma de cálculo de visibilidade na web através dos motores de busca, levando em conta as colocações da página web vinculada ao argumento da busca nos diversos rankings. Eliminando o problema dos homônimos, o cálculo apresentado resulta na principal contribuição deste trabalho: um indicador de VW para universidades, que mostra quão bem percebida é determinada entidade pelos buscadores da web.

Através de um estudo de caso, visando classificar universidades do Brasil segundo o indicador elaborado, observou-se a aplicabilidade em um cenário atual que é alvo de muitas pesquisas. Aplicando metabusca nas siglas das universidades, foram elaborados dois rankings: um mostrando a visibilidade na web das páginas das instituições quando feita uma busca somente com a sigla, revelando falhas na falta de semântica da abordagem, e outro, utilizando-se de uma técnica de expansão de consultas para descrever melhor o domínio, aumentando a pontuação das universidades abordadas no experimento e tornando a classificação mais justa.

Desde que a classificação gerada pelo indicador foi coerente com outros rankings de universidades, é válida sua usabilidade como uma forma de medir a VW de uma universidade. A presença na web das instituições é importante para atrair pesquisadores, investimentos e parcerias, então medir suas presenças é importante e, cada vez mais, essencial para o crescimento das universidades.

6.1 Trabalhos Futuros

A pesquisa permitiu a identificação de muitos trabalhos futuros e direções a serem seguidas, as principais são:

- Considerar uma distribuição de diferentes pesos para cada motor de busca no momento de pontuar a visibilidade, seguindo algum critério a ser estudado;

- Elaborar uma classificação englobando mais universidades, analisando o comportamento inclusive para universidades de fora do Brasil. Comparar o ranking gerado com outros existentes;
- Estudar outros parâmetros que possam ser extraídos para o cálculo de visibilidade na web, construindo um indicador composto para universidades;
- Analisar a aplicabilidade do indicador proposto em outros domínios. Demonstrar se o cálculo pode ser aplicado de maneira genérica;
- Incluir critérios de desempate nos rankings;
- Criar outras maneiras, que não através de rankings, de visualizar os resultados;
- Adicionar mais semântica, considerando o uso de ontologias para uma melhor descrição do domínio;
- Extrair dados temporais e espaciais com a metabusca, objetivando descobrir quando e onde a visibilidade da universidade foi melhor ou pior, e quais fatores a influenciam.

REFERÊNCIAS

AALTOJÄRVI, I. et al. Scientific Productivity, Web Visibility and Citation Patterns in Sixteen Nordic Sociology Departments. **Acta Sociologica**, V. 51, No. 1, 5-22. 2008.

AGUILLO, I. F. et al. Scientific Research Activity and Communication Measured with Cybermetrics Indicators. **Journal of the American Society for Information Science and Technology**, V. 57, 1296–1302. 2006.

AGUILLO, I. F. and KRETSCHMER, H. Visibility of Collaboration on the Web. **Scientometrics**, V. 61, 405-426. 2004.

AGUILLO, I. F., GRANADINO, B. and ORTEGA, J. L. Brazil Academic Webuniverse Revisited: A Cybermetric Analysis. In: INTERNATIONAL WORKSHOP ON WEBOMETRICS, INFORMETRICS AND SCIENTOMETRICS& SEVENTH COLLNET MEETIN,2006. **Proceedings…** Nancy, France.2006.

ALEXA. The Web Information Company, Top Sites. Disponível em: <http://www.alexa.com/topsites>. Acesso em: out. 2011.

ARWU, Academic Ranking of World Universities. Disponível em: <http://www.arwu.org>. Acesso em: jul. 2011.

ASLAM, J. A. and MONTAGUE, M. Models for Metasearch. In: ANNUALINTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH ANDDEVELOPMENT IN INFORMATION RETRIEVAL, 24., 2001. **Proceedings…** New York: ACM, 276-284. 2001.

BACCHIN, T. Aumentando a Visibilidade na Web. As Aplicações dos Webstandard. **Os Padrões Web e seus Impactos no Futuro do Governo Eletrônico**. Brasília, out. 2008. Disponível em: <http://www.w3c.br/palestras/eventopadroesegov-out-2008/Apresentacao_Thiago-Bacchin_W3C-BSB.pdf>. Acesso em: jul. 2011.

BAEZA-YATES, R., RIBEIRO-NETO, B. **Modern Information Retrieval**. Cap3: Retrieval Evaluation. ACM Press, New York. 1999.

BARJAK, F. and THELWALL, M. A Statistical Analysis of the Web Presences of European Life Sciences Research Teams. **Journal of the American Society for Information Science and Technology**, V. 59, 628-643. 2008.

BAST, H. et al. IO-Top-k: Index-access Optimized Top-k Query Processing. In: INTERNATIONAL CONFERENCE ON VERY LARGE DATABASES, 32., 2006, VLDB '06. **Proceedings…**Seoul, Korea. 2006.

BJÖRNEBORN, L., INGWERSEN, P. Toward a Basic Framework for Webometrics. **Journal of the American Society for Information Science and Technology**. V. 55 (Dec. 2004), 1216-1227. 2004.

BLACK, D. Partial Justification of the Borda Count. **Public Choice**, V. 28, no.1, 1-15. 1976.

BLATTMANN, U.; FACHIN, G. R. B.; RADOS, G. J; V. Recuperar a Informação Eletrônica pela Internet. **Revista da ACB**, Florianópolis, V.4, no. 4, 9-27. 1999.

CUBESTAT. Cubestat, The Free Website Value Calculator. 2008. Disponível em: <http://www.cubestat.com>. Acesso em: out. 2011.

DNSCOOP. Domain and Site Value Tool. 2009. Disponível em: <http://www.dnscoop.com>. Acesso em: out. 2011.

DUTRA, E. G. J. **Um Modelo de Fusão de Rankings Baseado em Análise de Preferência.** 2008. 74 f. Dissertação (Mestrado) – Ciência da Computação, UFRGS, Porto Alegre. 2008.

DUTRA JÚNIOR, E. G.; LIMA, J. V. Supplement of partial ranks to the data fusion. In: BRAZILIAN SYMPOSIUM ON MULTIMEDIA AND THE WEB, WEBMEDIA, 12., 2006, Natal, Rio Grande do Norte. **Proceedings…** New York: ACM, 148-154. 2006.

DWORK, C.et. al. Rank Aggregation Methods for the Web. In: INTERNATIONAL CONFERENCE ON WORLD WIDE WEB, 10., WWW10. **Proceedings…**ACM, New York, NY, USA, 613-622. 2001.

ESPADAS, J., CALERO, C. and PIATTINI, M. Web Site Visibility Evaluation. **Journal of the American Society for Information Science and Technology**, V. 59, 1727-1742. 2008.

GORI, M. and WITTEN, I. The Bubble of Web Visibility. **Commun.**.. ACM, V. 48, 115-117. 2005.

GOULART, R. 2008. **Os Mecanismos de Busca e suas Implicações em Comunicação e Marketing**. Líbero, Brasil, V. 11, No. 21, 119-132. 2008.

HEEACT, Higher Education Evaluation & Accreditation Council of Taiwan. **Performance Ranking of Scientific Papers for World Universities**. Disponível em: <http://ranking.heeact.edu.tw>. Acesso em: ago. 2011.

HSU, D. F. and PALUMBO, A. A Study of Data Fusion in Cayley Graphs G(S{n},P{n}).In: INTERNATIONAL SYMPOSIUM ON PARALLEL ARCHITECTURES,ALGORITHMS AND NETWORKS, 7., 2004. **Proceedings…** [S.l.:s.n.], 2004.

JOURIS, A. **Avaliação da Qualidade de Conferências Científicas Através da Visibilidade na Web**. 2011, 76 p. Dissertação (Mestrado em Ciência da Computação), Instituto de Informática, UFRGS, Porto Alegre. 2011.

KLINGER, A., LIMA, J. V., OLIVEIRA, J. P. M. Visibilidade Web Baseada em Metabusca. In: VI CONGRESSO IBERO-AMERICANO DE TELEMÁTICA. CITA2011. Anais… Gramado. **Cadernos de Informática**, V. 6, No. 1, 71-77. 2011.

KRETSCHMER, H., KRETSCHMER, U., KRETSCHMER, T. Reflection of Co-Authorship Networks in the Web: Web Hyperlinks Versus Web Visibility Rates. **Scientometrics**, V. 70, 519-540. 2007.

MANNING, C. D., RAGHAVAN, P., SCHÜTZE, H. **Introduction to Information Retrieval**. Cap 8: Evaluation in Information Retrieval. Cambridge University Press. 2008.

NORDFORSK. **Comparing Research at Nordic Universities using Bibliometric Indicators.** NORIA-net. 2011. Disponível em: <http://www.nordforsk.org/files/rapp.bib.2011.pub_21.5.11>. Acesso em: ago. 2011.

PAGE, L et al. **The Page Rank Citation Ranking**: Bringing Order to the Web. Technical Report. Stanford Info Lab. 1999.

QS, Quacquarelli Symonds. **World University Rankings.** Disponível em: <http://www.topuniversities.com>. Acesso em: ago. 2011.

SAARI, D. G. **The Optimal Ranking Method is the Borda Count**. Discussion Papers, no. 638.Northwestern University.1985. Disponível em: <http://econpapers.repec.org/paper/nwucmsems/638.htm>. Acesso em: jul. 2011.

SWAN, A. and CARR, L. Institutions, their Repositories and the Web. **Serials Review**, V. 34, 31-35. 2008.

THE, Times Higher Education. **World University Rankings**. Disponível em: <http://www.timeshighereducation.co.uk>. Acesso em: ago. 2011.

TOUS, R. and DELGADO, J. Advanced Meta-Search of News In The Web. In: INTERNATIONAL ICCC/IFIP CONFERENCE ON ELECTRONIC PUBLISHING, 6., 2002. **Proceedings…**Czech Republic, VWF Berlim. 2002.

VIEGAS, F. B. Word Tree. Disponível em: <http://fernandaviegas.com/wordtree.html>. Acesso em: ago. 2011.

WR, Webometrics Ranking of World Universities. Disponível em: <http://www.webometrics.info>. Acesso em: jun. 2011.

ANEXO A <UNIVERSIDADES BRASILEIRAS NO WR>

Tabela A.1: Ranking Brasileiro do WR

Colocação	*Sigla (Universidade)*
1º	USP
2º	UFRGS
3º	UNICAMP
4º	UFRJ
5º	UFSC
6º	UNESP
7º	UFMG
8º	UNB
9º	UFPR
10º	UFBA
11º	UFF
12º	UFPE
13º	PUC-Rio
14º	UFRN
15º	PUCRS
16º	UFV
17º	FGV
18º	UEM
19º	UFU
20º	UFC
21º	UERJ
22º	UFSCAR
23º	UFSM
24º	UFG
25º	UFPA
26º	UFPB
27º	UEL
28º	UNISINOS
29º	UNIFESP
30º	PUC-SP

Fonte: http://www.webometrics.info/rank_by_country.asp?country=br

ANEXO B <VALORES DOS DOMÍNIOS DAS UNIVERSIDADES BRASILEIRAS>

A seguir é apresentado um ranking ordenado pelo valor, em moeda, do domínio web de cada universidade. Os valores foram calculados pelo site Cubestat. Não foram estudas as métricas abordadas pelo site, servindo o ranking a seguir apenas como comparativo.

Os dados foram extraídos em abril de 2011.

Tabela B.1: Ranking de universidades pelos valores de seus domínios.

Colocação	*Sigla (Universidade)*	*Valor ($)*
1º	USP	370370,61
2º	UNB	218826,99
3º	UNICAMP	160124,04
4º	UFSC	135714,3
5º	UFRGS	129120,21
6º	UFRJ	123007,92
7º	UFMG	98525,91
8º	UFPR	95037,24
9º	FGV	86310,09
10º	PUC-Rio	71560,44
11º	UFF	69168,96
12º	UFBA	61440,45
13º	UFG	60643,29
14º	UFC	55630,38
15º	UEL	53133,78
16º	UERJ	51824,16
17º	UFPE	46837,53
18º	PUCRS	45444,69
19º	PUC-SP	44446,05
20º	UFPA	41747,97
21º	UEM	40140,51
22º	UFRN	36562,05
23º	UNISINOS	35013,72
24º	UFPB	34133,34
25º	UFV	32567,49
26º	UFSCAR	31984,95
27º	UFU	30460,71
28º	UFSM	29527,77
29º	UNESP	23570,97
30º	UNIFESP	0

Fonte: http://www.cubestat.com

APÊNDICE <HISTÓRICO DE DESENVOLVIMENTO>

O protótipo apresentado nesta dissertação utilizado como calculadora do indicador de VW possui um interessante histórico de desenvolvimento.

Originalmente, foi concebido como uma ferramenta de metabusca para validar a aplicação de um método de fusão de rankings baseado na análise de preferência, no ano de 2009. O método de fusão em questão, MDPref (DUTRA, 2008), foi elaborado durante o período da graduação, quando o autor deste presente trabalho colaborou com o idealizador do MDPref sob a tutela do mesmo orientador.

O pacote de fusão do MDPref foi desenvolvimento em JAVA, o que incentivou a utilização da mesma tecnologia no protótipo de metabusca. Inicialmente, foram feitos diversos testes com pacotes de *parsing* e APIs necessárias para se trabalhar com informações de motores de busca. A disponibilidade de recursos possibilitou o avanço do trabalho e o crescimento de um metabuscador que operava com dezessete motores de busca.

O aplicativo de metabusca tomou forma através de um *applet*, disponibilizado ao publico na Web. Os excelentes resultados recuperados com a ferramenta renderam um interessante Trabalho de Diplomação, o qual foi premiado com o Prêmio Terra em 2009/2.

Com a qualidade do trabalho realizado na graduação evidenciou-se a idéia de explorar o metabuscador durante o mestrado. Como o numero de motores de busca abrangidos era grande, especulou-se que, ao se efetuar uma busca por alguma marca no protótipo, o ranking retornado reflete de alguma maneira a visibilidade da marca consultada. Para o estudo da hipótese foi escolhido o domínio das universidades brasileiras, mais precisamente suas siglas.

Então, primeiramente foi realizada a extração de parâmetros de precisão a partir de consultas pelas siglas das universidades. Assim, foi estabelecida uma relação entre a precisão do ranking gerado pela metabusca e a visibilidade na web da universidade. Para contornar o evidente problema dos homônimos, logo se experimentou a expansão da consulta e foram obtidos resultados satisfatórios.

O método utilizado pelo metabuscador para unir os resultados dos motores de busca, neste ponto, era ainda o MDPref, e foi com ele que o primeiro trabalho foi publicado no CITA'2011, explorando a diferença da precisão de motores de busca locais e motores de busca globais.

Os resultados, porém, eram variáveis e o processo era lento. Cada sigla de universidade precisava ser consultada separadamente no protótipo de metabusca e a relevância dos *sites* recuperados era julgada manualmente. Além do mais, o método de fusão não evidenciava claramente o porquê certos *sites* estavam mais bem posicionados

que outros, além de demandar esforço computacional para calcular a decomposição matricial (SVD) e dificultar a depuração da pontuação real dos elementos do ranking final.

Diversos métodos de fusão de rankings haviam sido estudados previamente, surgindo a idéia de aplicar um método simples e direto de pontuação, que utilizasse diretamente a colocação dos sites dentro de cada motor de busca como meio de pontuação.

Foram, então, implementado métodos alternativos ao MDPref na ferramenta de metabusca: um método que atribui notas para cada faixa de colocações; um método que conta as ocorrências de cada *site* dentro dos *top-10* de cada motor de busca; e o método BC. Através destes métodos, a colocação dos sites dentro do ranking final ficou muito mais clara.

Mas o processo de cálculo de VW ainda estava trabalhoso. Para eliminar a necessidade do julgamento de relevância, optou-se por levar em conta somente a página oficial da organização. Assim, uma calculadora de VW podia ser facilmente elaborada, necessitando apenas de dois parâmetros de entrada: a sigla, e o endereço oficial.

O protótipo da calculadora de VW foi então implementado, utilizando o BC como meio de pontuação por possuir ampla literatura favorável ao método. A aplicação do cálculo às siglas de universidades retornou resultados interessantes e possibilitou uma fácil coleta de dados e a escrita da dissertação, evidenciando que havia sido criado um novo indicador de VW para universidades.

A idéia inicial era que a calculadora de VW funcionasse para qualquer marca, organização ou entidade presente na web, porém não foram realizados trabalhos nesse sentido. A título de curiosidade pode-se utilizar o protótipo para medir a visibilidade de qualquer termo, mas esforços futuros talvez tratem de comprovar a qualidade do valor obtido.

Quanto ao numero de motores de busca cobertos pela ferramenta de metabusca foi bastante variável ao longo dos anos. As interfaces de consulta mudam de maneira imprevisível, afetando a inclusão ou exclusão dos motores de busca. Quando há uma mudança em algum motor de busca são necessárias alterações no código da ferramenta. A manutenção constante é uma dificuldade em se manter um sistema de metabusca operante ao longo do tempo e foram necessárias inúmeras intervenções neste sentido desde a concepção e publicação na Web do protótipo.

A interface atual do metabuscador permite a escolha dos motores de busca, escolha do método de fusão e distribuição de pesos.