Porto Alegre, v.14, n.2, jul./dez. 2011.
ISSN impresso 1516-084X ISSN digital 1982-1654

Ranking de Universidades Através da Visibilidade na Web de suas Siglas por Metabusca

Ranking of Universities Through their Acronyms Web Visibility Based on Metasearch

Resumo
A ciência de Webometrics visa realizar medições sobre a Web, obtendo dados quantitativos através de estudo e análise dos diferentes aspectos da Web. Numa das áreas de Webometrics está inserida a Web Visibility, que é a área de estudo das medidas de visibilidade na Web. Este trabalho propõe uma fórmula para o cálculo de visibilidade em um domínio homogêneo, baseando-se na visão proporcionada pelos motores de busca da Web. Para tal é utilizado um método de fusão de rankings, o MDPREF, que gera um ranking sobre o qual é calculada a precisão, principal parâmetro da fórmula proposta. O trabalho apresenta experimentos com rankings, justificando alguns dos parâmetros do método utilizado para o cálculo de Web Visibility. Ao final, as siglas de algumas universidades brasileiras são utilizadas como estudo de caso e é apresentado um ranking parcial de universidades do Brasil.
Palavras-chave: Visibilidade Web. Metabusca. Universidades Brasileiras. Siglas. MDPref.

Abstract
The Webometrics science aims to take measurements on the Web, obtaining quantitative data through study and analysis of various aspects of the Web. In the Webometrics field is embedded Web Visibility, which is the area of study of measures of visibility on the Web. This paper proposes a formula for the calculation of visibility on a homogeneous domain, based on the vision provided by Web search engines. For such, a ranking fusion method is used, called MDPREF, which generates a ranking on which is calculated the precision, the main parameter of the proposed formula. The paper also presents experiments with rankings, justifying some of the parameters of the method used to calculate Web Visibility. Finally, the abbreviations of some universities are used as a case study and a partial ranking of Web Visibility of universities in Brazil is showed.
Keywords: Web Visibility. Metasearch. Brazilian Universities. Acronyms. MDPref.

KLINGER, Augusto; PALAZZO M. de Oliveira, José. Ranking de Universidades Através da Visibilidade na Web de suas Siglas por Metabusca. *Informática na Educação*: teoria & prática, Porto Alegre, v. 14, n. 2, p. 113-127, jul./dez. 2011.

Augusto Klinger
José Palazzo Moreira de Oliveira
Universidade Federal do Rio Grande do Sul

1 Introdução

O enorme volume de dados disponíveis na *Web* pode revelar muito mais informações além do próprio conteúdo. Medições podem ser feitas aplicando técnicas de informetria e bibliometria, por exemplo, revelando dados estatísticos a respeito de popularidade, agrupamentos de *sites* e distribuição da informação.

Como existem inúmeras páginas *Web*, o ponto dominante de acesso são motores de busca. Mas a visão que eles oferecem é restrita, pois além de não cobrirem toda a rede, por fins de eficiência somente os *sites* melhor ranqueados são exibidos. O *ranking* é gerado de acordo com as heurísticas e técnicas próprias de cada motor de busca, resultando numa visão da *Web* de acordo com seus olhos.

Web Visibility é a medida da visibilidade de um dado *site* na *Web*. A visibilidade pode ser calculada de acordo com a popularidade da página, dada pelo número de *links* que levam àquele *site*, ou número de acessos (AALTOJÄRVI *et al*., 2008), ou posição no *ranking*

Porto Alegre, v.14, n.2, jul./dez. 2011.
ISSN impresso 1516-084X ISSN digital 1982-1654

gerado por um motor de busca, que emprega geralmente diversos critérios, podendo estes serem interessantes ou não, dependendo da avaliação que deseja-se fazer.

O critério de número de *links* é o mais empregado nos algoritmos que efetuam as classificações das páginas da *Web* nos motores de busca, como o mundialmente conhecido *PageRank* (PAGE *et al*., 1999), sendo a forma mais difundida na literatura e na Internet de se medir visibilidade na *Web* (AGUILLO; ORTEGA; GRANADINO, 2006, AGUILLO *et al*., 2006, CYBERMETRICS LAB, 2011). Não há trabalhos difundidos que empreguem diretamente resultados de motores de busca para o cálculo de *Web Visibility*, embora sejam pontos de partida para qualquer pesquisa na Internet.

A área de estudo na qual a *Web Visibility* está inserida é denominada de *Webometrics*. Segundo Björneborn e Ingwersen (2004), *Webometrics* é o estudo dos aspectos quantitativos de construção e uso da informação na *Web*, estruturas e tecnologias vistas sobre os pontos de vista bibliométricos e informétricos. Ou seja, a ciência de *Webometrics* tenta obter informação através de medições sobre os diversos aspectos da *Web*.

Medir a visibilidade na *Web* é importante em vários aspectos, sobretudo para avaliar o nível de publicidade ou medir o impacto de uma marca ou produto na rede. Como exemplo, pode-se medir a visibilidade de siglas de universidades, sendo este o domínio utilizado como estudo de caso neste trabalho por se tratar de um domínio homogêneo, onde cada componente é bem determinado (universidades) e podem-se fazer comparações.

2 Trabalhos Relacionados

Tratando-se de *ranking* de universidades brasileiras, o indicador de qualidade do Ministério da Educação (MEC) é o Índice Geral de Cursos (IGC). O índice leva em conta variáveis como corpo docente, infra-estrutura e organização didático-pedagógico, além é claro do Enade, que avalia o conhecimento dos alunos. A visibilidade *Web* ainda não é utilizada pelo órgão, mas é considerada por outros institutos de pesquisa (*Cybermetrics Lab, QS Quacquarelli Symonds Limited, ShanghaiRanking Consultancy*). A universidade de maior IGC atualmente é a Unifesp.

O *QS World University Rankings* (QS QUACQUARELLI SYMONDS LIMITED, 2011) classifica universidades de todo o mundo, inclusive separando-as por áreas. O *ranking* leva em conta o tamanho dos corpos discente e docente, abrangência de áreas de estudo, e intensidade de pesquisa medida através dos documentos recuperáveis pelo *Scopus*, que é uma base de dados (disponível em versão na *Web*) de resumos e citações da literatura de produções científicas, contendo cerca de 18.000 títulos.

O *Academic Ranking of World Universities* (ARWU) (SHANGHAIRANKING CONSULTANCY, 2011) também promove uma classificação por áreas de pesquisa e engloba universidades de todo o mundo. Os critérios envolvem número de publicações, citações e prêmios recebidos pelos pesquisadores, sendo que as fontes de dados estão presentes na *Web*.

Um laboratório da Espanha elaborou o *Webometrics Ranking of World Universities* (CYBERMETRICS LAB, 2011), que é um *ranking* que tenta englobar todas as universidades do mundo. No seu cálculo de *rank*, a visibilidade *Web* representa 50% do valor agregado ao *site* da universidade, sendo que essa visibilidade foi obtida através do *Yahoo Search*, levando em conta o número total de *links* externos únicos que cada *site* recebe (*inlinks*). A universidade brasileira melhor colocada é a USP, na 122ª colocação do *ranking* geral

mundial. Ainda de acordo com os resultados do laboratório espanhol, as três universidades nacionais de maior visibilidade são a USP, UNICAMP e UFRJ.

No geral, nota-se que o cálculo de visibilidade na *Web* é bastante simples, e há espaço para novos métodos que englobem mais características. A abrangência utilizada no cálculo é um fator determinante, pois quanto maior a cobertura de *sites*, mais precisa será a medida de visibilidade. Outra questão importante é fugir das bolhas de visibilidade (GORI; WITTEN, 2005) que são conjuntos de *sites* que se apontam entre si com o objetivo de ficarem mais bem ranqueados, e com isso, mais visíveis.

3 Visibilidade por Metabusca

Estudos mostram que o número de páginas na Web sobre domínio das universidades brasileiras tem crescido exponencialmente, assim como tem crescido também a visibilidade dessas universidades em toda a rede, medida através de *inlinks* (AGUILLO; ORTEGA; GRANADINO, 2006). Levando em conta as visões proporcionadas por diferentes motores de busca e o crescente número de páginas *Web* relacionadas a universidades brasileiras, uma medida de visibilidade diferente pode ser obtida através da consulta e análise dos *rankings* produzidos por motores de busca, utilizando-se a sigla da universidade como entrada.

Utilizar diversos motores de busca é uma maneira de aumentar o *recall*. O processo de consultar diversos motores de busca ao mesmo tempo é conhecido como metabusca. Um sistema de metabusca permite ao usuário pesquisar simultaneamente em vários motores de busca simples, tornando a pesquisa mais vasta e mais eficiente. Metabuscadores geralmente não possuem nenhum tipo de base de dados, baseando-se puramente nos dados de outros mecanismos de busca (BLATTMANN; FACHIN; RADOS, 1999).

Como retorno tem-se os vários *rankings*, sendo necessário unificá-los. Existem diversos métodos para realizar fusão de *rankings*, em particular o método baseado na análise de preferência (MDPREF), utilizado em trabalhos anteriores (DUTRA, 2008, KLINGER, 2009), se mostrou uma excelente alternativa para o problema.

O cenário perfeito para a aplicação de metabusca seria aquele no qual se tem uma ordem completa de todos os elementos do universo em todos os *rankings*. Porém isso não é possível, pois cada motor de busca tem uma cobertura diferente da *Web* (DWORK *et al*., 2001). É improvável que os motores de busca sejam capazes de rankear toda a coleção de páginas da *Web*, que cresce a uma taxa bastante rápida.

A metabusca se apresenta, então, como uma técnica alternativa no sentido de aumentar a abrangência e obter-se uma visibilidade mais fidedigna. Uma vez que cada motor de busca emprega técnicas diferentes, o resultado é, teoricamente, uma visão mais ampla da *Web*.

Este trabalho visa definir uma nova forma de medir a visibilidade na *Web*, utilizando a metabusca com fusão de *rankings*. O estudo de caso das siglas de universidades mostra um cenário atual para a aplicação de uma fórmula de *Web Visibility*, com o propósito de ranquear e também mostrar como são vistas no mundo da *Web* as universidades brasileiras.

3.1 Fusão de Rankings

Um *ranking* é um conjunto ordenado de elementos. Ou seja, os elementos são organizados de forma crescente, de acordo com sua relevância, sendo que o primeiro tem o maior grau de importância e o ultimo o menor grau.

Porto Alegre, v.14, n.2, jul./dez. 2011.
ISSN impresso 1516-084X ISSN digital 1982-1654

Em computação, *rankings* são muito utilizados na área de Recuperação de Informação (IR) na qual os algoritmos visam buscar a informação desejada e, usualmente, classificá-la de acordo com determinados critérios de relevância. Motores de busca que operam em grandes bases de dados ou na *Web* são os principais exemplos de aplicação na área.

Existem duas classificações para *rankings*: completos e parciais. Um *ranking* completo possui todos os elementos de um universo. Um *ranking* parcial possui apenas parte dos elementos de um universo. A segunda classificação é onde se enquadram os *rankings* gerados pelos motores de busca da *Web*, pois além de não serem capazes de indexar todo o conteúdo da rede, os usuários estão interessados somente nos *k* primeiros elementos (*top-k*). Desconsiderar 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 conforme Bast *et al*. (2006), que apresenta técnicas para o processamento de *top-k* elementos objetivando otimizar a performance.

Fusão de *rankings* é o problema de computar um *ranking* consensual, dados diversos r*ankings* individuais contendo elementos classificados por diferentes juízes (RENDA; STRACCIA, 2002). Um juíz é quem determina a ordem dos elementos de um dado universo, podendo ser um especialista humano ou um motor de busca, por exemplo.

Existem vários métodos para se realizar a fusão de *rankings*. Os métodos utilizam as informações dos *rankings* individuais, como o ordinal associado a um elemento, uma nota dada a cada elemento ou o próprio conteúdo. Neste trabalho, é dada uma atenção especial ao MDPref, escolhido para a fusão com base em estudos prévios (DUTRA, 2008, KLINGER, 2009).

3.1 MDPref

MDPref é um método de mapeamento perceptual, técnica proveniente na área de *Marketing,* onde o estudo das preferências de grupos de consumidores em relação a determinados produtos é de extremo interesse. O mapeamento perceptual permite a análise conjunta dos atributos de um produto, podendo gerar um gráfico com o posicionamento, em um espaço comum, dos produtos e consumidores com suas respectivas preferências. Em suma, métodos de mapeamento perceptual permitem determinar a preferência de um grupo de indivíduos em relação a um conjunto de elementos (DUNN-RANKIN, 2004).

O MDPref é baseado no modelo desenvolvido por J. D. Carrol e J. J. Chang em 1973, que faz uso do teorema de decomposição de Eckart-Young (SVD) ou da análise de componente principal (PCA), executado sobre os dados de preferência dos consumidores para cada produto. Graficamente, cada juíz ou grupo de juízes é representado como um vetor, o qual indica a sua direção de preferência, os estímulos são representados como pontos. Cada ponto de estímulo é projetado sobre cada vetor, revelando a sua preferência. A Figura 1 exemplifica a visualização.

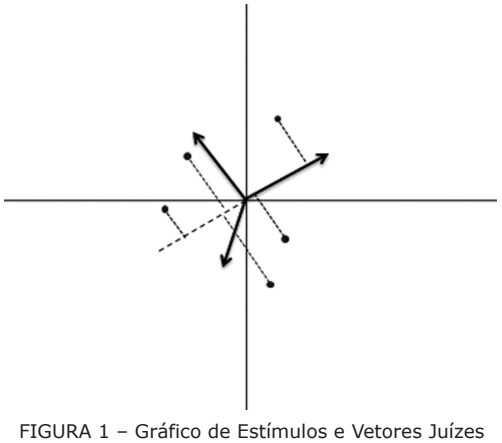

FIGURA 1 – Gráfico de Estímulos e Vetores Juízes
FONTE: Autor

3.2 Fusão de Rankings com MDpref

O modelo de fusão baseado na Análise de Preferência foi proposto por Dutra (2008). Os dados de entrada são uma matriz com os *rankings* de cada juíz e uma matriz de pesos. A matriz dos *rankings* é colocada na forma de um *cluster*[1] de *rankings*, de tamanho *p* por *n*, sendo *p* a dimensão dos objetos avaliados e *n* a dimensão dos juízes.

Iniciando o processo, primeiramente é construída uma matriz tridimensional D (matriz de *scores* primários) a partir da matriz de *rankings* (*cluster* de *rankings*). Em uma das dimensões de D estão os juízes, e as outras duas representam a comparação entre pares de elementos. Cada par *jk* de elementos é avaliado, para o juíz *i*, da seguinte forma:

D[i][j][k] = 1, se o elemento *j* foi avaliado melhor que *k*
D[i][j][k] = -1, se o elemento *j* foi avaliado pior que *k*
D[i][j][k] = 0, se o elemento *j* foi avaliado igual *k* ou não foi avaliado.

A partir da matriz D e da matriz de pesos (opcional) é formada uma matriz bidimensional S (matriz de *scores* secundários), que define a diferença entre a preferência dos elementos *j* sobre *k* para cada juíz. S é de tamanho *n* por *p*. Cada elemento seu é preenchido com o somatório dos resultados da diferença dos valores nas posições *jk* e *kj*. O valor somado é multiplicado pela raíz quadrada do peso do *ranking* do juíz *i* e colocado em S na posição *ij*, correspondente ao juíz e objeto avaliado respectivamente.

Através da decomposição matricial SVD de S, são obtidas três novas matrizes: U, L e A. L é uma matriz diagonal, contendo os autovalores. U e A são matrizes que contém os autovetores, sendo suas colunas ortogonais. As três matrizes são ordenadas de acordo com a magnitude dos autovalores e utilizadas para obter as matrizes solução X e Y.

Apenas as duas componentes mais significativas de U, L e A são utilizadas para formar X e Y, procedendo-se da seguinte forma:

X = UL
Y = A

O vetor de preferência é calculado com base na matriz X, e normalizado. Cada uma das duas componentes do vetor é dada pelo somatório das linhas da matriz X em cada uma das duas colunas.

Finalmente, a matriz Y é projetada sobre o vetor de preferência, resultando no *ranking* consensual. Note que Y contém a posição no espaço para cada elemento.

A visualização gráfica pode ser vista na FIGURA 2 - Gráfico Resultante de Fusão Baseada no MDPREF[2], gerada por um conjunto de dados aleatórios. Os pontos são todos os elementos avaliados, contidos em Y. O vetor de preferência é representado como uma linha com um ponto no final. Os algarismos são os rótulos de identificação de cada elemento. Quanto mais próximo do vetor está o elemento, melhor é sua colocação no *ranking*.

3.3 Avaliação de Rankings

Como serão utilizados *rankings* provenientes de metabuscas na *Web* para o cálculo da visibilidade, é necessário um estudo de como avaliar tais *rankings*.

1 Agregado.

2 Gerada pela ferramenta disponível em: <www.inf.ufrgs.br/~aklinger/mdpref>

Porto Alegre, v.14, n.2, jul./dez. 2011.
ISSN impresso 1516-084X ISSN digital 1982-1654

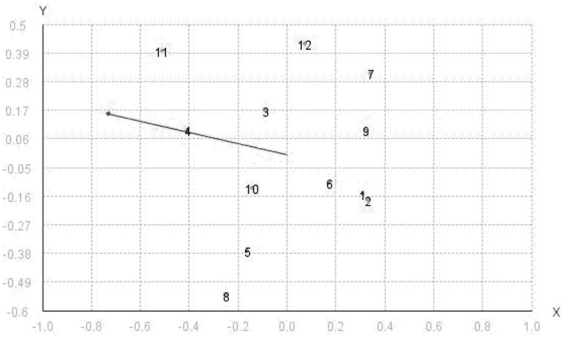

FIGURA 2 – Gráfico Resultante de Fusão Baseada no MDPREF
FONTE: Ferramenta de Fusão em www.inf.ufrgs.br/~aklinger/mdpref

Para avaliar a qualidade de um *ranking* é utilizada uma medida de precisão, que toma por base o quão relevante é cada elemento retornado para a consulta realizada. No caso de motores de busca, nem todos os elementos recuperados em resposta a uma consulta são relevantes. Além disso, pode acontecer de muitos itens relevantes não participarem do *ranking* retornado.

A ideia 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, sendo subjetiva. 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 por sua subjetividade. O que para um juiz é um aspecto que caracteriza a relevância de um documento, para outro juiz pode não caracterizar. Mesmo fazendo os julgamentos automaticamente por meios de programas de computador, diferentes implementações podem discordar a respeito da relevância de um documento e de especialistas humanos.

3.4 Precisão e *Recall*

Precisão e *Recall* são as medidas mais utilizadas em avaliação de algoritmos e sistemas de recuperação de informações.

Considere uma coleção documentos *C* contendo um número *R* de documentos relevantes. Uma 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. Precisão e *recall* são definidos da seguinte maneira (BAEZA-YATES; RIBEIRO-NETO, 1999):

Precisão é a quantidade de documentos retornados que são relevantes:

RA/A

Recall é a quantidade de documentos relevantes retornados do total de relevantes:

RA/R

O problema com essas medidas, definidas da maneira acima, é que precisam que todo o conjunto de *A* documentos retornados seja examinado, além de conhecer 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. A medida é chamada de *Precision at 10*, e tem a vantagem de não ser necessário saber o número total de documentos relevantes.

Mais genericamente, essa medida de precisão considerando 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:

Rt/k

4 Experimentos

4.1 Motores de Busca

O primeiro experimento realizado foi com alguns motores de busca isolados, sem fazer a fusão de *rankings*. O objetivo desse experimento foi determinar o tamanho dos *rankings* a serem utilizados para a fusão e observar o conteúdo recuperado de acordo com a consulta submetida.

Em doze motores de busca foram realizadas consultas com as siglas de duas universidades: UFC e UFRGS. Analisou-se a relevância dos *sites* retornados em *rankings* de diferentes tamanhos. Foram considerados *sites* relevantes todos àqueles que pertencem a 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*). Os gráficos gerados são correspondentes as Figuras 3 e 4, e mostram os valores de precisão para cada tamanho de *ranking*.

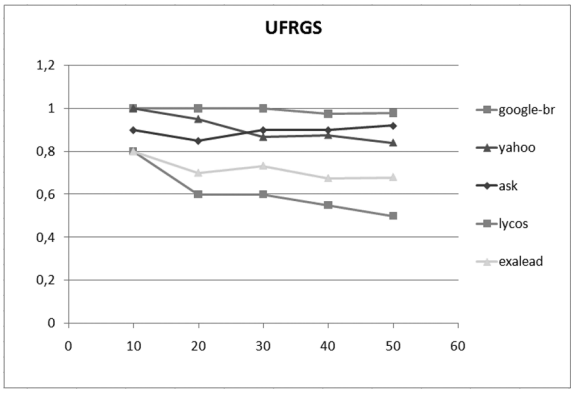

FIGURA 3 – Precisão dos *top n* Para a Sigla UFRGS em 12 Motores de Busca
FONTE: Medição a partir dos motores de busca

A FIGURA 3 - Precisão dos top n Para a Sigla UFRGS em 12 Motores de Busca ilustra os resultados obtidos para a sigla UFRGS. Apenas cinco dos doze motores de busca foram ilustrados para maior clareza. Pode-se ver uma tendência aos *rankings* ficarem mais poluídos, ou seja, com menos *sites* relevantes, conforme mais resultados são considerados. As linhas de tendência dos buscadores que não constam no gráfico seguem o mesmo padrão dos presentes.

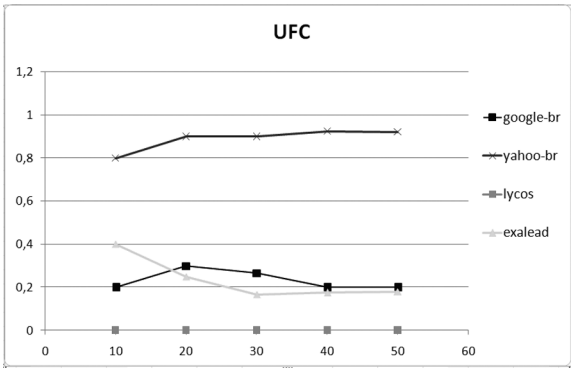

FIGURA 4 – Precisão dos *top n* Para a Sigla UFC em 12 Motores de Busca
FONTE: Medição a partir dos motores de busca

Porto Alegre, v.14, n.2, jul./dez. 2011.
ISSN impresso 1516-084X ISSN digital 1982-1654

A FIGURA 4 - Precisão dos top n Para a Sigla UFC em 12 Motores de Busca mostra que, para a sigla UFC, a precisão dos motores de busca é bem inferior ao caso da UFRGS. Isso se deve ao fato de existir outra instituição, mais famosa mundialmente, que partilha da mesma sigla: a liga de vale-tudo *Ultimate Fighting Championship.* Nota-se bastante divergência entre as linhas, o *Yahoo-Br* retornou um bom número de *sites* relevantes, por exemplo, e o *Lycos* nenhum em todos os tamanhos de *rankings* analisados.

Ainda conforme o experimento com a sigla da universidade do Ceará, no correspondente gráfico foram omitidos sete motores de busca cujo comportamento foi o mesmo do *Lycos*, retornando zero relevantes em todos os casos. O *Alta Vista-Br*, omitido por clareza da FIGURA 4 ☐ Precisão dos top n Para a Sigla UFC em 12 Motores de Busca teve um bom desempenho, apresentando uma tendência praticamente igual à do *Yahoo-Br*.

O caso da UFC expõe um problema: nem todos os motores de busca sabem que estamos procurando a Universidade Federal do Ceará. Para contornar este problema foi realizado o experimento descrito a seguir, empregando um artifício de Recuperação da Informação.

4.2 Expansão de Consulta

Uma técnica bastante popular na área de RI é a expansão de consultas. Basicamente consiste em acrescentar algumas palavras a consulta a fim de discriminar melhor o que se deseja recuperar.

Como o que se quer recuperar, no contexto deste trabalho, são *sites* referentes a universidades, a palavra *universidade* foi incluída ao lado da sigla nas consultas. As mesmas duas consultas do experimento anterior foram submetidas, agora expandidas, aos mesmos doze motores de busca do experimento anterior. Obteve-se os resultados ilustrados pelos gráficos das Figuras 5 e 6 (não contém todos os buscadores para melhor visualização).

É interessante comparar a FIGURA 3 - Precisão dos top n Para a Sigla UFRGS em 12 Motores de Busca, do experimento anterior, com a FIGURA 5 ☐ Precisão dos top n Para a Sigla UFRGS + Universidade, ambas referentes a sigla UFRGS. Observa-se que para uma sigla de boa expressividade o uso da palavra *universidade* junto à consulta não influi muito nos resultados. Isso porque a sigla UFRGS não é ambígua, e a Universidade Federal do Rio Grande do Sul é a entidade de maior relevância que detém a sigla. Para alguns motores de busca, como o *Google-br*, o uso da palavra *universidade* aumentou a precisão dos *rankings* de tamanhos maiores, e para outros casos, como o do *Lycos* e do *Exalead*, a palavra *universidade* aumentou sutilmente o nível de poluição nos seus *rankings*.

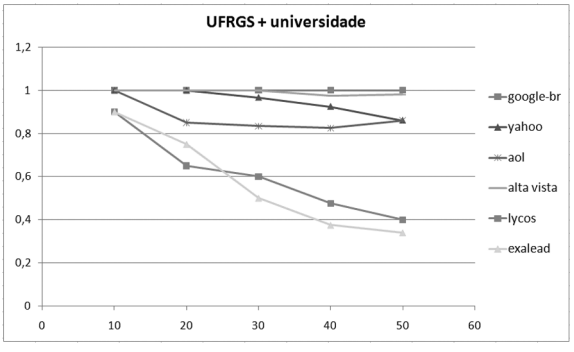

FIGURA 5 – Precisão dos *top n* Para a Sigla UFRGS + Universidade
FONTE: Medição a partir dos motores de busca

Já para a sigla UFC, pertencente a Universidade Federal do Ceará, pode-se ver no gráfico da FIGURA 6 - Precisão dos top n para a sigla UFC + universidade que o uso da palavra u*niversidade* melhora a precisão em todos os motores de busca, sendo o intuito de recuperar *sites* relevantes a Universidade do Ceará,

em comparação ao gráfico da FIGURA 4 - Precisão dos top n Para a Sigla UFC em 12 Motores de Busca. Nenhum buscador retornou zero relevantes dessa vez.

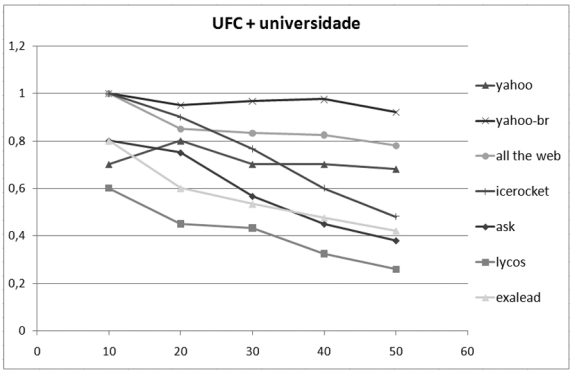

FIGURA 6 – Precisão dos top n para a sigla UFC + universidade
FONTE: Medição a partir dos motores de busca

Pode-se notar através da análise dos gráficos que, quanto maior o *ranking*, no geral, pior a precisão, pois mais *sites* de menor relevância para a consulta são incluídos. Baseado nesses experimentos determinou-se que utilizar somente o *top 10* de cada motor de busca para a fusão de *rankings* é o ideal, pois se observa claramente nas Figuras 3 a 6 que a precisão é mais alta considerando apenas os dez primeiros resultados, e os primeiros colocados tendem a ser os mais relevantes (BAST *et al*., 2006).

4.3 Motores de Busca Brasileiros e Globais

Outra interessante questão a se considerar em um cálculo de visibilidade na *Web* através de motores de busca é a restrição geográfica dos mesmos.

Motores de busca regionais, ou seja, que priorizam *sites* de uma determinada região, tendem a produzir *rankings* com uma melhor precisão para as buscas intencionadas em recuperar resultados da região específica a qual o motor de busca se concentra. Por outro lado, motores de busca que não se concentram em uma determinada região, buscando resultados ao redor de toda a *Web*, tendem a produzir *rankings* mais poluídos, ou de entidades mais propagadas mundialmente.

Realizou-se um experimento com a fusão de *rankings*, separando os motores de busca e fundindo-os com o MDPREF em duas categorias: brasileiros e mundiais, conforme:

Brasileiros: versões brasileiras do *Alta Vista, Ask, Google* e *Yahoo*

Mundiais: *All the Web, Alta Vista, Ask, Aol, Exalead, Google, Icerocket, Lycos* e *Yahoo.*

Com o intuito de verificar a melhor precisão em motores de busca regionais, dezessete siglas de universidades brasileiras foram avaliadas com quatro consultas diferentes, sendo duas em cada grupo de motores de busca em suas versões normal e expandida. O gráfico de precisão de oito das siglas é exibido na Figura 7.

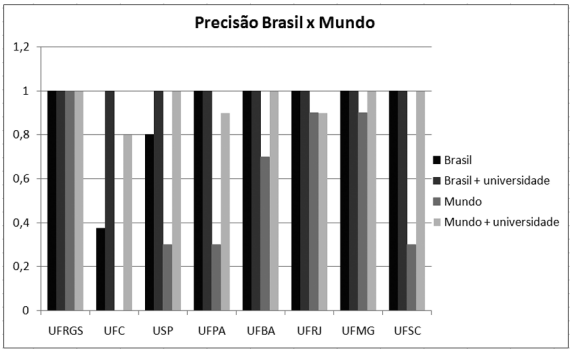

FIGURA 7 – *Precision at 10* do MDPREF Para as Siglas das Universidades
FONTE: Medição a partir dos motores de busca

Analisando a Erro: Origem da referência não encontrada 7, observa-se que a precisão do *ranking* gerado pelo MDPREF para siglas de universidades do Brasil é, de fato, maior quando se fundem *rankings* de motores de

Porto Alegre, v.14, n.2, jul./dez. 2011.
ISSN impresso 1516-084X ISSN digital 1982-1654

busca direcionados ao Brasil somente (preto). Em alguns casos, quando a sigla é bastante expressiva, como o da UFGRS, o resultado é o mesmo, mas em todos os casos, conforme o gráfico, a precisão é maior com o uso de motores de busca regionais. Ainda pode-se observar que, confirmando os dados anteriores, o uso da palavra *universidade* nas consultas junto à sigla de fato melhora a precisão do *ranking*. Nenhuma sigla obteve maior precisão com o uso de motores de busca globais.

Outro dado interessante, analisando o caso da universidade UFC cuja sigla pertence também a uma entidade mais famosa mundialmente, é que os três motores de busca brasileiros retornaram sete *sites* (buscando somente pela sigla UFC), sendo que dois eram relevantes, enquanto que os dez outros motores de busca globais retornaram vinte e um sites e nenhum relevante entre eles.

4.4 Considerações

Os experimentos demonstraram alguns parâmetros importantes para o cálculo de visibilidade *Web* baseado em fusão de *rankings*:

- Quanto maiores os *rankings*, pior a precisão;
- expansão de consulta aumenta o poder de descrição do alvo que se pretende recuperar;
- motores de busca locais geram *rankings* de maior precisão para alvos de sua região.

5 Fórmula de *Web Visibility*

Com base na distinção entre os resultados gerados por motores de busca regionais e globais, criou-se uma fórmula para o cálculo de visibilidade na *Web* de universidades do Brasil. Não somente universidades, a fórmula pode ser usada para calcular a *Web Visibility* de qualquer marca, instituição, pessoa ou produto.

Como os experimentos demonstram que os dez primeiros colocados de cada motor de busca geram *rankings* de maior precisão, somente estes são utilizados para a fusão. Para a avaliação, também são considerados somente os dez primeiros colocados do *ranking* proveniente da fusão (*Precision at 10*).

Sendo *Pb* a precisão do *ranking* resultante da fusão dos motores de busca regionais e *Pm* a precisão do *ranking* resultante da fusão dos motores de busca mundiais, a fórmula proposta para o cálculo de *Web Visibility* (*WV*) é como segue:

$$WV = 1 - (Pb - Pm)$$

Note que o valor estará no intervalo [0, 1]. O caso de a precisão *Pm* ser maior do que a *Pb* resultaria em um valor fora desse intervalo, porém não houveram situações em que isso ocorreu durante os experimentos. Como a fórmula está voltada para coisas regionais, a tendência é que $Pb \geq Pm$.

Os motores de busca regionais não precisam ser restritos ao Brasil. Pode-se querer calcular a visibilidade de algo mais restrito a determinado estado ou outro país, por exemplo. Porém, neste trabalho voltado a classificar universidades brasileiras, foram utilizados os motores de busca direcionados ao Brasil para gerar a precisão *Pb*.

5.1 *Ranking d*as Universidades

Vinte e duas universidades foram submetidas ao cálculo de *Web Visibility*. Os dois grupos de motores de busca usados são compostos da mesma maneira que apresentados na subseção Motores de Busca Brasileiros e Globais.

Foram feitos testes com a consulta pela si-

gla e com o uso da palavra *universidade* junto da sigla. A relevância dos *sites* contidos nos *rankings* do MDPref foi julgada por um humano através da análise direta do conteúdo das páginas. Foram considerados válidos todos os *sites* internos da instituição e *sites* que referenciam ou citam de fato a instituição buscada.

Levando em conta a consulta expandida para cada sigla de universidade, dentre as testadas, temos o *ranking* da Tabela 1, a seguir. O critério de desempate utilizado foi o índice de visibilidade da versão sem a expansão de consulta.

TABELA 1 – *Top*-10 Universidades Amostradas

Posição	*Universidade*	*Visib. expansão*	*Visib.*
1	UFRGS	1	1
2	UNESP	1	1
3	UFPR	1	1
4	UFMG	1	0,9
5	UNICAMP	1	0,8
6	UFBA	1	0,7
7	PUC-RIO	1	0,7
8	USP	1	0,5
9	UFAM	1	0,4
10	FURG	1	0,3

FONTE: Dados da pesquisa

Em primeiro lugar há um empate técnico entre as siglas de UFRGS, UNESP e UFPR, todas com índice máximo de visibilidade nas duas modalidades de consulta. Também ficaram empatadas UFBA e PUC-Rio na sexta colocação.

Comparado com o *ranking* brasileiro do *Webometrics Ranking of World Universities* (CYBERMETRICS LAB, 2011), que leva em conta ainda outros três aspectos além da visibilidade na *Web*, o resultado é bastante diferente. A UFRGS aparece na terceira posição do *ranking* do *Webometrics*, onde a USP, oitava colocada no experimento, está em primeiro. Levando em conta somente a visibilidade no *Webometrics* e as universidades amostradas nos experimentos, a UFRGS ficaria em quinto, atrás de USP, UNICAMP, UFRJ e PUC-Rio (na versão adaptada do *Webometrics*). UNESP e UFPR, que também dividem a primeira colocação no *ranking* deste trabalho, ocupariam a sexta e a décima segunda colocações, respectivamente, no *Webometrics* segundo visibilidade.

A UFC, utilizada com destaque em experimentos anteriores, ficou em vigésimo lugar, à frente ainda de UFG e UFS.

Para fins comparativos, dentre as 22 universidades submetidas ao cálculo de visibilidade estão as 10 melhores colocadas do *Webometrics.* Destas, apenas UFRJ, UFSC e UNB não entraram no *top*-10 apresentado.

Vale ressaltar que o critério de desempate utilizado influiu diretamente no resultado da tabela 1. Caso tivesse sido utilizada uma média entre os dois valores de visibilidade UFES e UFRJ entrariam no *top-10,* pois ficaram com 0,9 em ambos valores. As universidades subiriam das décima quinta e décima sexta colocações para quinto e sexto lugares.

O restante do *ranking* elaborado consta na Tabela 2. Do total de siglas testadas, a maioria obteve pontuação máxima no cálculo de visibilidade com expansão da consulta, havendo muitos empates segundo esta coluna. A visibilidade, quando da busca somente pela sigla, apresentou resultados mais diversos, propiciando um bom critério de ordenação para desempates.

Caso a classificação tivesse sido feita de acordo com a coluna da visibilidade da sigla sem expansão, o resultado seria diferente, porém as primeiras colocadas ainda seriam as mesmas. Como a palavra *universidade* descreve bem o domínio, reduzindo a influência de múltiplas organizações compartilhando a mesma sigla, a classificação em primeira or-

Porto Alegre, v.14, n.2, jul./dez. 2011.
ISSN impresso 1516-084X ISSN digital 1982-1654

TABELA 2 – *Ranking* a Partir da 11ª Colocação

Posição	*Universidade*	*Visib. expansão*	*Visib.*
11	UFAC	1	0,3
12	UFSC	1	0,3
13	UFMS	1	0,3
14	UNB	1	0,2
15	UFES	0,9	0,9
16	UFRJ	0,9	0,9
17	UFPA	0,9	0,3
18	UFV	0,9	0,2
19	UNIR	0,8	0,8
20	UFC	0,8	0,625
21	UFG	0,8	0,2
22	UFS	0,8	0,1

FONTE: Dados da pesquisa

dem de acordo com a versão com expansão de consulta foi escolhida.

6 Conclusões

Foi apresentada uma maneira de calcular a visibilidade na *Web* a partir da fusão dos *rankings* gerados por motores de busca. O método de fusão utilizado permite determinar a preferência consensual de um grupo de juízes (*rankings*) através de análise multivariada e decomposição matricial (SVD). A precisão calculada sobre o *ranking* resultante serviu de parâmetro para a fórmula de *Web Visibility* proposta.

Experimentos demonstraram que a precisão tende a cair conforme se analisa mais resultados de um *ranking* e que a expansão da consulta melhora os resultados para o caso das siglas de universidades.

O uso da palavra *universidade*, junto à sigla na consulta, resulta em uma maior expressividade. Isto é, discrimina o que se está de fato buscando. Os experimentos comprovaram que o número de sites relevantes recuperados aumenta nos casos em que a sigla possui pouca representatividade, e praticamente não se altera quando a sigla já possui um bom poder de expressão. Ou seja, tem-se um resultado mais justo para as siglas de universidades que coincidentemente pertencem também a outras entidades.

Também foi mostrado que existe uma diferença nos *rankings* produzidos por motores de busca brasileiros e motores de busca mundiais, e que essa diferença pode ser explorada na elaboração de uma fórmula de visibilidade *Web*.

Um *ranking* parcial com algumas universidades do Brasil foi elaborado de acordo com a fórmula de *Web Visibility* estudada mostrando a visão dos motores de busca da *Web* para as siglas dessas universidades, definindo assim uma forma nova de medir visibilidade na *Web*.

É importante salientar que essa forma de cálculo serve para qualquer conteúdo de domínio homogêneo, não somente universidades. Estudos futuros devem mostrar essa aplicabilidade.

6.1 Trabalhos Futuros

A pesquisa realizada permitiu a identificação de futuros trabalhos:

- Englobar mais universidades, principalmente aquelas que ficam bem colocadas em outros *rankings* similares, para produzir um *ranking* mais significativo e próximo de completo;
- comparar o *ranking* de *Web Visibility* com mais *rankings* de universidades;
- analisar o comportamento da fórmula proposta para consultas mais genéricas, como por exemplo, marcas registradas com a intuição de medir a divulgação;
- aprimorar a formula de visibilidade com novos parâmetros (estão sendo

estudadas novas formas de medir visibilidade através de motores de busca, os avanços podem ser conferidos na *web*[3]);

- propor uma forma automática de avaliar a relevância de *sites* em um *ranking*;
- estudar a influência de pesos diferentes para cada motor de busca, e como distribuí-los;
- utilizar métodos mais simples para fusão de *rankings* e cálculo de visibilidade;
- buscar os últimos resultados dos motores de busca da *Web* para identificar *sites* pouco visíveis, fazendo uma fusão de *rankings* invertidos com o objetivo de visualizar os resultados mais improváveis.

6.2 Trabalho em Andamento

A visibilidade *Web* de uma universidade também pode ser medida através da posição de seu site oficial nos diversos *rankings* provenientes da metabusca.

O método que está sendo estudado consiste dos seguintes passos:

- Identificação da página *Web* oficial da universidade;
- busca pela sigla da universidade em *n* motores de busca;
- busca pela página oficial dentro de cada um dos *n rankings*;
- distribuição de uma pontuação de acordo com posição do *site* em cada *ranking*;
- valor de visibilidade resultante da soma dos pontos.

Um próximo trabalho divulgará os resultados e apresentará um novo *ranking* de visibilidade *Web* de siglas de universidades do Brasil.

Referências

AALTOJÄRVI, I. *et al*. Scientific Productivity, Web Visibility and Citation Patterns in Sixteen Nordic Sociology Departments. *Acta Sociologica*, Oslo, v. 51, no. 1, p. 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*, New York, v. 57, p. 1296-1302, 2006.

AGUILLO, I.F.; ORTEGA, J.L.; GRANADINO, B. Brazil Academic Webuniverse Revisited: A Cybermetric Analysis. In: INTERNATIONAL WORKSHOP ON WEBOMETRICS, INFORMETRICS AND SCIENTOMETRICS, 2006, Nancy; COLLNET MEETING, 7., Maio 10 -12, 2006, Nancy (França). *Conference Paper.*

BAEZA-YATES, R.; RIBEIRO-NETO, B. Retrieval Evaluation. In: BAEZA-YATES, R.; RIBEIRO-NETO, B. *Modern Information Retrieval*. New York: ACM Press, 1999. Cap. 3.

3 http://www.inf.ufrgs.br/~aklinger

Porto Alegre, v.14, n.2, jul./dez. 2011.
ISSN impresso 1516-084X ISSN digital 1982-1654

BAST, H. *et al*. IO-Top-k: Index-access Optimized Top-k Query Processing. In: *INTERNATIONAL CONFERENCE ON VERY LARGE DATA BASES* – VLDB '06, 2006, Seoul, Korea. *Proceedings*. [S.l.: s.n], 2006. p. 475-486

BJÖRNEBORN, L.; INGWERSEN, P. Toward a Basic Framework for Webometrics. *Journal of the American Society for Information Science and Technology*, New York, v. 55, p. 1216-1227, dec. 2004.

BLATTMANN, U.; FACHIN, G.R.B.; RADOS, G.J.V. Recuperar a Informação Eletrônica Pela Internet. *Revista da ACB*, Florianópolis, v. 4, n. 4, p. 9-27, 1999. Disponível em: <DOI=http://www.ced.ufsc.br/~ursula/papers/buscanet.html> Acesso em: 10 abr. 2011.

CYBERMETRICS LAB. *Ranking Web of World Universities*. Madrid: CSIC, 2011. Disponível em: <DOI=http://www.webometrics.info> Acesso em: 11 abr. 2011.

DUNN-RANKIN, P. *et al*. Mapping Individual Preference. In: DUNN-RANKIN, P. *et al*. *Scaling Methods*. 2. ed. Cidade: Lawrence Erlbaum, 2004. Cap 13.

DUTRA, E.G.J. *Um Modelo de Fusão de Rankings Baseado em Análise de Preferência*. 2008. Dissertação (Mestrado em Ciência da Computação) – Instituto de Informática, Universidade Federal do Rio Grande do Sul, 2008, Porto Alegre, BR-RS.

DWORK, C. *et al*. Rank Aggregation Methods for the Web. In: INTERNATIONAL WORLD WIDE WEB CONFERENCE - WWW10, 10., 2001, New York. *Proceedings*. S.l.: s.n.], 2001. P. 613-622.

GORI, M.; WITTEN, I. The Bubble of Web Visibility. *Communications of the ACM*. New York, v. 48, p. 115-117, 2005.

KLINGER, A. *O Modelo de Fusão de Rankings Baseado em Análise de Preferência Aplicado a Metabusca*. 2009. Projeto de Diplomação (Bacharelado em Ciência da Computação) – Instituto de Informática, Universidade Federal do Rio Grande do Sul, 2009, Porto Alegre, BR-RS.

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

PAGE, L. *et al*. *The PageRank Citation Classificação*: Bringing Order to the Web: Technical Report. [S.l.]:Stanford InfoLab, 1999.

QS QUACQUARELLI SYMONDS LIMITED. *QS Top Universities*. 2011. Disponível em: <DOI=http://www.topuniversities.com> Acesso em: 4 abr. 2011.

RENDA, M.E.; STRACCIA, U. *Metasearch*: Rank vs. Score Based Rank List Fusion Methods (without Training Data). Pisa, It.: Instituto di Elaborazione della Informazione, 2002.

SHANGHAIRANKING CONSULTANCY. *Academic Ranking of World Universities*. 2011. Disponível em: <DOI=http://www.arwu.org> Acesso em: 4 abr. 2011.

Recebido em maio de 2011.
Aprovado para publicação em julho de 2011.

Augusto Klinger
Universidade Federal do Rio Grande do Sul, Porto Alegre/RS, Brasil. E-mail: aklinger@inf.ufrgs.br

José Palazzo Moreira de Oliveira
Universidade Federal do Rio Grande do Sul, Porto Alegre/RS, Brasil. E-mail: palazzo@inf.ufrgs.br