Visibilidade Web Baseada em Metabusca

Augusto Klinger
PPGC
UFRGS - Instituto de Informática
Av. Bento Gonçalves, 9500
55 51 96874581
aklinger@inf.ufrgs.br

José Valdeni de Lima
PGIE-CINTED
UFRGS - Instituto de Informática
Av. Bento Gonçalves, 9500
55 51 33087019
valdeni@inf.ufrgs.br

José Palazzo Moreira de Oliveira
PPGC
UFRGS - Instituto de Informática
Av. Bento Gonçalves, 9500
55 51 33087019
palazzo@inf.ufrgs.br

ABSTRACT
Webometrics is the science which studies the various aspects of the Web in order to measure them. 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, 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. Finally, the abbreviations of some universities are used as a case study and are presented a partial ranking of Web Visibility of universities in Brazil.

RESUMO
Webometrics é a ciência que estuda os vários aspectos da Web com o objetivo de medi-los. Nela 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, 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, parâmetro principal da fórmula proposta. Ao final, as siglas de algumas universidades são utilizadas como estudo de caso e é apresentado um ranking parcial de universidades do Brasil.

Categories and Subject Descriptors
H.3.5 [**Information Storage And Retrieval**]: On-line Information Services – *Web-based services.*

General Terms
Measurement

Keywords
Web Visibility, metabusca, universidades brasileiras, fusão de rankings, siglas, MDPref.

1. INTRODUÇÃO
O grande 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 [7].

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 [1], ou posição no ranking gerado por um motor de busca, que emprega geralmente diversos critérios, podendo estes serem interessantes ou não, dependendo da avaliação que deseja-se fazer.

A área de estudo na qual a *Web Visibility* está inserida é denominada de *Webometrics*. Segundo Björneborn e Ingwersen [5], *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.

Um laboratório da Espanha elaborou o *Webometrics Ranking of World Universities* [12], que é um ranking que tenta englobar todas as universidades do mundo. No seu calculo de rank, a visibilidade representa 50% do valor agregado ao site da universidade, sendo que essa visibilidade foi obtida através do Yahoo Search, levando em conta o numero 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 [7], que são conjuntos de sites que se apontam entre si com o objetivo de ficarem mais bem ranqueados, e com isso, mais visíveis.

Estudos mostram que o numero 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* [2]. Levando em conta as visões proporcionadas por diferentes motores de busca e o crescente numero 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. 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 [7][10], 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 [8]. 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.

2. FUSÃO DE RANKINGS

Fusão de rankings é o problema de computar um ranking consensual, dados diversos rankings individuais contendo elementos classificados por diferentes juízes [13]. 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 [7][10].

2.1 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.

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.

2.2 Fusão de Rankings com o MDPref

O modelo de fusão baseado na Análise de Preferência foi proposto por Dutra [7]. 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 1: abaixo[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.

[1] Agregado

[2] Gerada pela ferramenta disponível em www.inf.ufrgs.br/~aklinger/mdpref

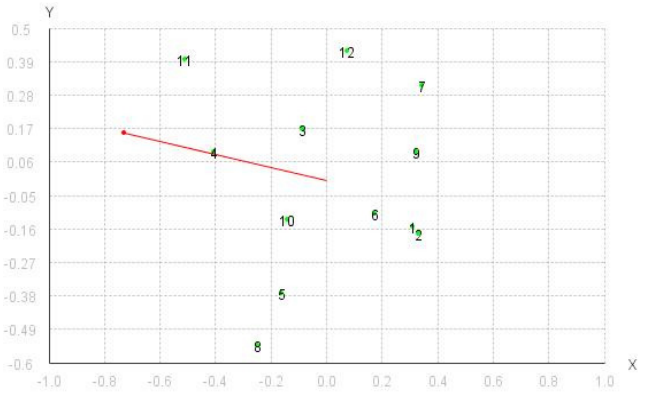

Figura 1: Gráfico resultante de fusão baseada no MDPREF.

3. AVALIAÇÃO DE RANKINGS

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

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 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 [11].

A relevância é julgada de acordo com a necessidade de informação, e não em relação à consulta, sendo subjetiva. 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, diferentes implementações podem discordar a respeito da relevância de um documento e de especialistas humanos.

3.1 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 numero **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 numero de relevantes retornados. Precisão e *recall* são definidos da seguinte maneira [3]:

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 numero total de documentos relevantes.

Mais genericamente, essa medida de precisão considerando somente os **k** primeiros resultados é chamada de *Precision at k* [11]. Sendo **Rt** o numero 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 numero de sites relevantes recuperados no *top* ***n*** pelo total de sites no ranking (**n**). Os gráficos gerados são correspondentes as Figuras 2 e 3, e mostram os valores de precisão para cada tamanho de ranking.

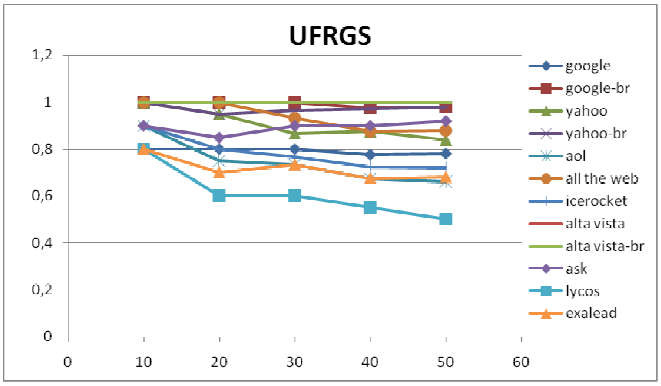

Figura 2: Precisão dos *top n* para a sigla UFRGS em 12 motores de busca.

A Figura 2: ilustra os resultados obtidos para a sigla UFRGS. Pode-se ver uma tendência aos rankings ficarem mais poluídos, ou seja, com menos sites relevantes, conforme mais resultados são considerados.

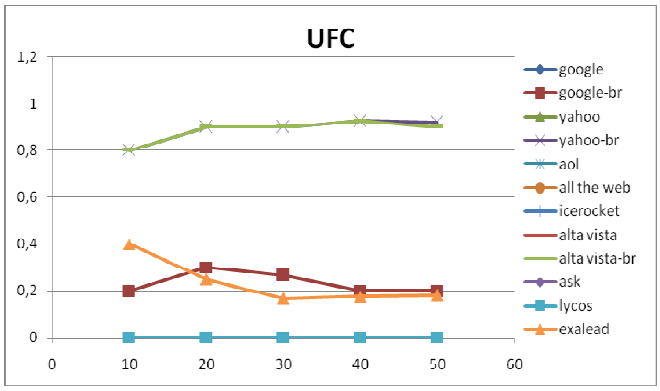

Figura 3: Precisão dos top n para a sigla UFC em 12 motores de busca.

A Figura 3: 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 numero de sites relevantes, por exemplo, e o Lycos nenhum em todos os tamanhos de rankings analisados.

O caso da UFC expõe um problema: nem todos os motores de busca sabem que estamos procurando a Universidade Federal do Ceara. 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" (sem as aspas) 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 4 e 5.

É interessante comparar a Figura 2:, do experimento anterior, com a Figura 4:, 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, a palavra "universidade" aumentou sutilmente o nível de poluição nos seus rankings.

Já para a sigla UFC, pertencente a Universidade Federal do Ceará, pode-se ver no gráfico da Figura 5: que o uso da palavra "universidade" 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 3:.

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 2 a 5 que a precisão é mais alta considerando apenas os dez primeiros resultados.

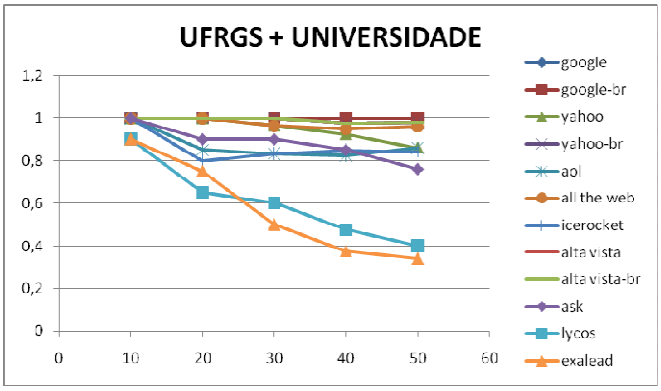

Figura 4: Precisão dos top n para a sigla UFRGS + universidade.

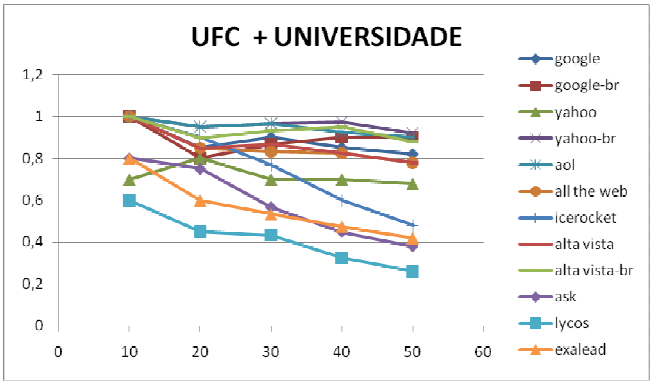

Figura 5: Precisão dos top n para a sigla UFC + universidade.

4.3 Motores de Busca Brasileiros VS. Globais

Outra interessante questão a se considerar em um calculo 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 é exibido na Figura 6:.

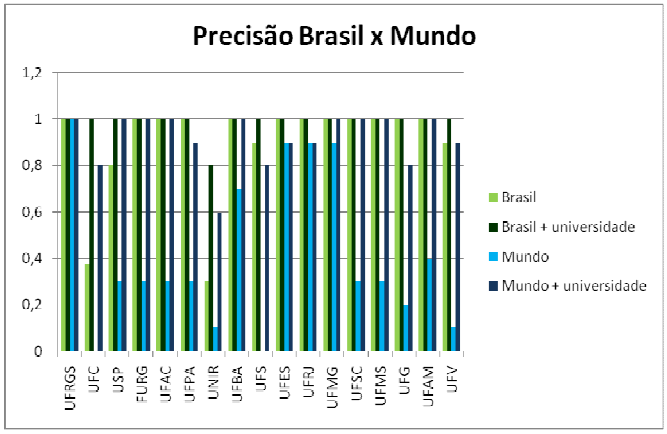

Figura 6: *Precision at 10* do MDPREF para as siglas das universidades.

Analisando a Figura 6:, 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 busca direcionados ao Brasil somente (verde claro). 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.

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 7 sites distintos (buscando somente pela sigla UFC), sendo que 2 eram relevantes, enquanto que os 10 outros motores de busca globais retornaram 21 sites e nenhum relevante entre eles.

4.4 Considerações

Os experimentos demonstraram alguns parâmetros importantes para o calculo 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 comente 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** ≥ **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 das Universidades

Dezessete 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 4.3.

Foram feitos testes com a consulta pela sigla 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.

A Figura 7: apresenta o gráfico gerado. Em azul tem-se a visibilidade da sigla da universidade, e em vermelho a visibilidade da sigla com a consulta expandida. Pode-se ver que a UFRGS foi a única universidade a apresentar um índice máximo de visibilidade na Web sem o uso da palavra "universidade", dentre as siglas testadas. Isso porque sua sigla já é bastante expressiva. O uso da consulta expandida também não alterou a visibilidade das universidades UNIR, UFRJ e UFES. Nove siglas, no total, alcançaram pontuação máxima na versão com expansão de consulta.

Já para algumas universidades, como UFAM, UFV, UFSC, UFAC, FURG, UFS e UFPA, o uso da palavra "universidade" foi essencial para discriminar a instituição buscada.

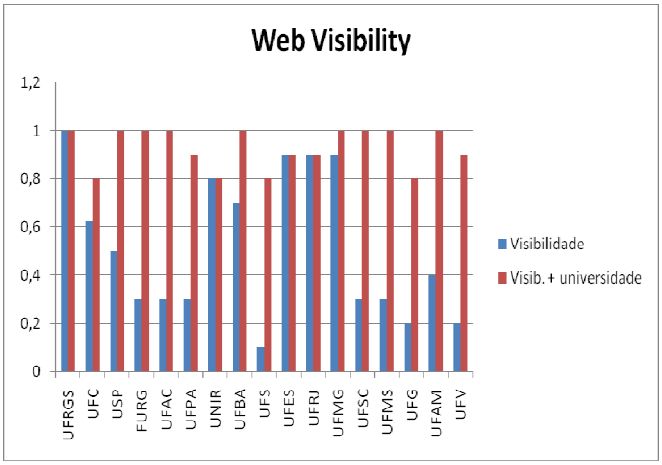

Figura 7: Visibilidade das siglas de universidades.

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.

Do 6° ao 9° lugar há um empate técnico, assim como na 10ª colocação. Comparado com o ranking brasileiro do *Webometrics Ranking of World Universities* [12], que leva em conta ainda outros três aspectos além da visibilidade na Web, o resultado é bastante diferente. A UFRGS aparece na 6ª posição do ranking do *Webometrics*, onde a USP está em primeiro. Levando em conta somente a visibilidade e as universidades amostradas nos experimentos, a UFRGS ficaria em 4°, atrás de USP, UFRJ e UFSC (na versão adaptada do *Webometrics*). A UFMG, segunda colocada de acordo com a fórmula proposta neste trabalho, ocuparia a 4ª colocação segundo o site.

Tabela 1. Top-10 universidades amostradas

Colocação	*Universidade*	*Visib. expansão*	*Visib.*
1	UFRGS	1	1
2	UFMG	1	0,9
3	UFBA	1	0,7
4	USP	1	0,5
5	UFAM	1	0,4
6	FURG	1	0,3
7	UFAC	1	0,3
8	UFMS	1	0,3
9	UFSC	1	0,3
10	UFES (ou UFRJ)	0,9	0,9

Para fins comparativos, cinco universidades do *top-10* brasileiro do *Webometrics* não foram submetidas aos experimentos, sendo uma delas a UNICAMP, que figura na segunda posição do ranking. Experimentos futuros deverão conter tais universidades a fim de comparar os resultados de forma mais satisfatória, possivelmente tomando por base não somente as 10 melhores classificadas universidades segundo o *Webometrics*, mas um conjunto maior a fim de verificar mais aprofundadamente a diferença entre os métodos de medição

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 numero 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.

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.

É importante salientar que essa forma de calculo serve para qualquer conteúdo, não somente universidades.

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 outros 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 calculo de visibilidade.

7. REFERÊNCIAS

[1] Aaltojärvi, I., Arminen, I., Auranen, O., Pasanen, H-M. 2008. Scientific Productivity, Web Visibility and Citation Patterns in Sixteen Nordic Sociology Departments. *Acta Sociologica*, vol. 51, no. 1, P. 5-22.

[2] Aguillo, I. F., Ortega, J. L., Granadino, B. 2006. Brazil Academic Webuniverse Revisited: A Cybermetric Analysis. *Proceedings... International Workshop on Webometrics, Informetrics and Scientometrics & Seventh COLLNET Meeting*, França. 2006.

[3] Baeza-Yates, R., Ribeiro-Neto, B. 1999. *Modern Information Retrieval.* Cap3: Retrieval Evaluation. ACM Press, New York.

[4] Bast, H. et al. 2006. IO-Top-k: Index-access Optimized Top-k Query Processing. *Proceedings of the 32nd International Conference on Very Large Data Bases*. VLDB '06. Seoul, Korea.

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

[5] Björneborn, L., Ingwersen, P. 2004. Toward a Basic Framework for Webometrics. *Journal of the American Society for Information Science and Technology***.** Vol. 55 (Dec. 2004), 1216–1227.

[6] Dunn-Rankin, P. et al. 2004. *Scaling Methods***.** 2ª ed. Lawrence Erlbaum, 221p. Cap 13: Mapping Individual Preference.

[7] Dutra, E. G. J. 2008. *Um Modelo de Fusão de Rankings Baseado em Análise de Preferência***.** 74 p. Dissertação (Mestrado em Ciência da Computação) – Instituto de Informática, UFRGS, Porto Alegre.

[8] Dwork, C et. al. 2001. *Rank Aggregation Methods for the Web*. WWW10, May 1-5, 2001.

[9] Gori, M., Witten, I. 2005. The Bubble of Web Visibility. *Communications of the ACM.* Vol. 48, 115-117.

[10] Klinger, A. 2009. *O Modelo de Fusão de Rankings Baseado em Análise de Preferência aplicado a Metabusca*. 40 p. Projeto de Diplomação (Bacharelado em Ciência da Computação) – Instituto de Informática, UFRGS, Porto Alegre.

[11] Manning, C. D., Raghavan, P., Schütze, H. 2008. *Introduction to Information Retrieval.* Cap 8: Evaluation in Information Retrieval. Cambridge University Press, 2008.

[12] Cybermetrics Lab. 2010. *Ranking Web of World Universities*. CSIC, Spain. DOI=http://www.webometrics.info.

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