RECUPERANDO INFORMAÇÕES DA WEB COM BASE NA ANÁLISE DE PREFERÊNCIA

WEB INFORMATION RETRIEVAL BASED ON PREFERENCE ANALYSIS

Elmário Gomes Dutra Jr. *
José Valdeni de Lima **

R e s u m o

Organizar um conjunto de objetos conforme sua relevância a determinados critérios tem se tornado uma tarefa dispendiosa, do ponto de vista computacional, em virtude da grande quantidade de informações disponível (*web*). Combinar *rankings* (consultas, ordenamentos) previamente feitos ajuda a reduzir tais custos. Este processo, no entanto, tem encontrado várias limitações na sua aplicação, tais como: diferentes estruturas de acesso, *rankings* heterogêneos (completos e parciais), distinção de importância dos *rankings* (automáticos e manuais), entre outros. Propõe-se, neste artigo, um modelo de Fusão de *rankings* baseado na Análise de Preferência, que visa a suprir as limitações citadas sem impactar na performance do *ranking* resultante e que possa ser aplicado na recuperação de informações na *web* (metabuscas). Os resultados dos experimentos possibilitaram validar o modelo de fusão de *rankings*, bem como sua aplicabilidade na recuperação de informações na *web*.

P a l a v r a s - c h a v e : Fusão de dados. Fusão de *rankings*. Recuperação de Informação.

* Mestre em Ciência da Computação pela UFRGS e licenciado em Matemática e Física pela FACOS. É professor no curso de Redes de Computadores na Faculdade SENAC, Coordenador de Extensão e Pós-graduação na Faculdade SENAI, além de atuar como professor nos cursos de Redes de Computadores e Sistemas de Informação na mesma instituição.
elmariodutra@hotmail.com

** Graduado em Processamento de Dados pela Universidade Federal do Ceará, Mestre em Ciências da Computação pela Universidade Federal do Rio Grande do Sul e doutor em Informática pela Université Joseph Fourier (antiga Université Scientifique Et Medicale) - Grenoble I. Atualmente é professor Associado IV da Universidade Federal do Rio Grande do Sul.
valdeni@inf.ufrgs.br

A b s t r a c t

Ordering a set of objects according to their relevance to certain criteria has become an expensive task to the computational point of view due to the large amount of information available (web). Combine rankings, which were previously made, helps to reduce such costs. However, this process has found several limitations in its application, such as different access structures, heterogeneous rankings (complete and partial), a distinction of rankings importance (automatic and manual), among others. We propose in this paper a model of Fusion rankings based on the Preference Analysis which aims to overcome those limitations without impacting the performance of the resulting ranking. The model can be used in web information retrieval applications (meta-search). The experiments results made possible to validate the model of ranking fusionas well as its applicability in web information retrieval.

K e y w o r d s : Data Fusion. Ranking Fusion. Information Retrieval.

1 Introdução

Historicamente, tem se constatado a necessidade humana de classificar objetos, sejam eles documentos, livros, músicas ou pessoas, entre outros. A diferença é que, atualmente, a quantidade de informações para se gerenciar é muito maior que a do século passado. Esse aumento na quantidade de informações armazenadas é fruto da disseminação da *internet,* que facilitou a publicação e o acesso às mesmas.

Da mesma maneira que a rede mundial facilitou a propagação da informação, sua recuperação tornou-se um problema, não só pelo grande volume de dados, mas pela grande quantidade de fontes com o mesmo conteúdo (sobreposição). Felizmente, ferramentas para encontrar a informação desejada foram construídas e hoje se tornaram um recurso indispensável aos usuários conectados à *internet*. Um exemplo clássico é o Google.

Mas a simples criação de um mecanismo de busca não foi o suficiente para resolver o problema de encontrar uma informação específica em um universo tão vasto e complexo. Fatores como a sobreposição de informações (com diferentes graus de precisão) e quantidade de critérios a serem avaliados têm tornado essa tarefa dispendiosa e alvo de muita pesquisa acadêmica. A principal motivação

dessas pesquisas é o desenvolvimento de técnicas que possam minimizar o custo computacional das consultas, bem como melhorar a qualidade de seus resultados (DWORK et al., 2001).

A reutilização de consultas previamente montadas tem se mostrado eficaz na redução do custo computacional do retorno de uma consulta (DUTRA JÚNIOR, LIMA, 2006; DWORK et al., 2001; FAGIN et al., 2004; RENDA, STRACCIA, 2003). Muitos estudos arguem que a fusão de *rankings*[1] tem grande potencial para combinar efetivamente várias fontes (FARAH; VANDERPOOTEN, 2007) e tem se mostrado útil e poderosa em várias aplicações, como metabuscas, procura de similaridade, classificações, banco de dados e recuperação de informação.

[1] A forma de referir-se a este problema varia de acordo com alguns autores. São sinônimos de fusão de *rankings* os termos agregação de *ranks* e fusão de dados. Neste trabalho, será convencionado o termo fusão de *rankings*.

A fusão de r*ankings* é o processo no qual unem-se as informações de diversas listas de objetos com o objetivo de gerar uma única lista que melhor represente todas as outras. Em outras palavras, a ideia é conseguir extrair de um conjunto de *rankings* uma preferência global ou consensual. Esse processo é vantajoso, pois possibilita que se filtre uma grande quantidade de objetos, priorizando os mais relevantes e permitindo o uso de *rankings* tanto automáticos como manuais (criados por especialistas).

Neste sentido, vários estudos foram desenvolvidos, e a maioria das abordagens propostas não contempla todos os aspectos inerentes ao problema, solucionando-o parcialmente. A grande limitação, porém, é em relação à atribuição de pesos distintos para cada uma das listas envolvidas e a possibilidade de elementos ocuparem a mesma posição (empatados).

Com base no exposto, percebe-se que a fusão de *rankings* tem grande potencial para ser aplicada na recuperação de informações na *web*, porém faz-se necessária a criação de um modelo que seja capaz de: (i) permitir a utilização de mais de dois *rankings* por vez; (ii) possibilitar o uso de listas completas e parciais, sem a necessidade de adequá-las; (iii) utilizar *rankings* estritamente posicionais ou baseados em *scores*; (iv) permitir a inserção de pesos distintos para cada um dos *rankings* envolvidos; e, (v) possibilitar que a mesma posição seja ocupada por mais de um elemento (empate).

Este artigo tem por objetivo, portanto, apresentar um modelo de fusão de *rankings* que supra as deficiências dos modelos existentes atualmente, preservando a performance dos *rankings* resultantes, e que possa ser aplicado no contexto da recuperação de informações na *web* (metabuscas).

O trabalho está estruturado da seguinte maneira: na próxima seção é apresentado um levantamento bibliográfico trazendo os conceitos e definições sobre o tema; a seção seguinte é responsável por apresentar o modelo de fusão de *rankings* proposto; a seção 4 traz a validação do modelo; na sequência apresenta-se a utilização do modelo em uma aplicação de metabuscas; a seção 6 reserva-se aos trabalhos relacionados; e, por fim, são apresentadas a conclusão e trabalhos futuros, na seção 7.

2 Referencial teórico

2.1 Rankings

Considere um conjunto de objetos ou universo $U = \{O_1, O_2, \ldots, O_m\}$, onde estes objetos podem representar qualquer coisa: palavras, pessoas, bebidas, documentos, entre outros. Um *ranking,* ou lista ordenada *r* em relação a U é um subconjunto ordenado $S \subseteq U$, em que $r = [x_1 \geq x_2 \geq \ldots \geq x_d]$ com $x_i \in S$ e $\geq$ é alguma relação de ordem em *S*.

Define-se ranking completo ou lista completa aquela lista que contém todos os objetos do universo U ($|r| = |U|$ e $S = U$). Já um *ranking* parcial é aquela lista que não contém todos os objetos de um universo U ($|r| < |U|$ e $S \subseteq U$). Denota-se que |r| e |U| representa o número de elementos do *ranking r* e do universo *U*, respectivamente.

Há várias situações em que a utilização de *rankings* parciais é preferida em detrimento dos *rankings* completos. Como exemplo, pode-se citar a necessidade de algum usuário ou sistema estar interessado nos *k* primeiros elementos de um *ranking (top k list)* (FAGIN, KUMAR, SIVAKUMAR, 2003); neste caso, os demais elementos são ignorados, restando apenas um subconjunto do universo.

Para cada elemento de um *ranking* pode-se associar um valor, que pode ser a posição deste elemento (lista baseada em *rank*) ou um valor de significância (lista baseadas em *score*). Define-se função de *ranking* a função *f* que, para dado elemento de uma lista r retorna a posição ou *score* do mesmo. De maneira análoga, a função inversa de *ranking*, f^{-1}, retorna o elemento de um *ranking* dado a sua posição ou *score*.

Dados dois *rankings,* r_1 e r_2, a medida de similaridade $sim(r_1, r_2)$ expressa o quão próximos esses rankings estão e é calculada com base em alguma métrica de distância (FAGIN, KUMAR, SIVAKUMAR, 2003) (Kendall-tau, Spearman

footrule) entre as duas listas. O valor normalizado deve estar no intervalo [0,1], sendo que valores próximos de 1 indicam um grau maior de similaridade. Além da medida de similaridade, outras duas métricas são bastante importantes: ruído e qualidade de um *ranking*.

Dado um conjunto de *rankings* e formando clusters (grupos) conforme a sua similaridade (*rankings* com alta similaridade juntos), define-se ruído (*noise*) como uma medida de perturbação média destes clusters. Em outras palavras, o ruído mede a divergência média das posições de um elemento nos clusters formados.

A medida de qualidade (*missinformation*) de um conjunto de *rankings* ou cluster permite mensurar a assimetria entre estes *rankings*; ou seja, a razão média entre as larguras dos clusters e suas distâncias. De acordo com Adali, Magdon-Ismail e Marshall (2007), para determinar a qualidade de um cluster é necessário conhecer a largura do cluster e a distância entre os mesmos.

2.2 Fusão dos dados

A fusão de dados é o problema de combinar várias listas ordenadas de uma forma robusta para produzir um único *ranking* de objetos, cuja aplicação tem se mostrado bastante usual e poderosa para uma série de aplicações como metabuscas, procura de similaridade (FAGIN et. al., 2004), banco de dados (DAS et al., 2006), recuperação de informações (VOGT, COTTREL, 1999), *data cleaning* (GUHA et al., 2004), entre outros. O objetivo principal da fusão de dados é, portanto, encontrar uma alternativa[2] que descreva, da melhor forma, todos os critérios envolvidos. Essa alternativa é uma espécie de consenso entre os julgamentos (*rankings*) apresentados por cada um dos juízes[3].

Os esforços de pesquisas sobre o tema fusão de dados gerou uma grande quantidade de propostas em torno deste assunto, e cada uma delas procura atacar um ou mais problemas referentes à fusão de dados.

[2] Uma alternativa, neste caso, é a organização de determinados objetos de um universo com base na avaliação de algum sistema.

[3] O termo juiz é aplicado de forma a generalizar qualquer sistema capaz de organizar objetos com base em algum critério.

2.3 Análise de preferência

A necessidade de determinar a preferência de um grupo de indivíduos[4] em relação a um conjunto de elementos não está presente apenas na área da Ciência da Computação; outras áreas de pesquisa também abordaram este problema, como é o caso do *Marketing*, da Matemática Estatística, da Psicologia, entre outros.

A principal contribuição neste sentido vem da área de *Marketing, em que* a necessidade de determinar a preferência de um grupo de consumidores em re-

[4] Termo trazido da área de *Marketing*, mas pode ser aplicado ao contexto deste trabalho, como sistemas ou juiz.

lação a um ou mais produtos é de extrema relevância. Para tanto, são utilizados, entre outros, métodos de mapeamento perceptual (*perceptual mapping*), os quais permitem a análise conjunta dos atributos[5] de um determinado produto, possibilitando a produção de gráficos que mostram tanto o posicionamento do produto quanto do consumidor e sua respectiva preferência em um espaço comum.

[5] São características de um determinado objeto. Em *Marketing* podem-se definir vários atributos a um produto, por exemplo, como cor, tamanho, entre outros.

O mapeamento perceptual recorre a técnicas como *Multidimensional preference analysis* (MDPREF) e *preference mapping* (PREFMAP), que ajudam a visualizar a estrutura competitiva de mercados através da percepção dos consumidores, com base na avaliação de vários atributos. Esta técnica permite ao pesquisador realizar uma série de inferências sobre o conjunto de dados (preferências dos consumidores), as quais não são possíveis apenas pela observação dos mesmos. Dentre essas inferências, podem-se citar a visualização de grupos de preferências de consumidores, grupos de similaridade de produtos, entre outros.

As técnicas para gerar mapas perceptuais fazem uso de métodos estatísticos mais complexos, conhecidos como *multivariate data analysis*, os quais podem ser definidos, de forma bastante geral, como métodos estatísticos que analisam simultaneamente múltiplas medidas em cada objeto a ser investigado (HAIR et al., 1998). Em outras palavras, *multivariate analysis* é um conjunto de procedimentos que envolvem a observação e análise simultânea de mais de duas variáveis estatísticas.

3 Modelo de fusão de rankings baseado em análise de preferência

Esta seção descreve a proposta para o modelo de fusão de dados baseado na análise de preferência, no qual o objetivo principal é gerar um único *ranking* dos objetos de um universo (documentos, pesquisadores, entre outros), a partir da união das informações obtidas de diversas listas ordenadas.

A análise de preferência permite, por meio de um conjunto de técnicas, que se descreva tanto gráfica como analiticamente o julgamento de diversos juízes em relação a um conjunto de objetos. Tal descrição é possível, pois o método reduz a dimensionalidade dos dados e faz uma separação das configurações dos juízes e dos objetos.

O modelo aqui apresentado é bastante versátil, uma vez que possibilita:

- a fusão de dados envolvendo um número maior que dois *rankings*;
- o uso de listas baseadas em *ranks* ou *scores*;
- a inserção da ideia de empate em cada *ranking*;

- a utilização de listas parciais e completas;
- a definição de pesos diferenciados para cada um dos *rankings* envolvidos;
- a mensuração da distância entre dois elementos após a fusão.

3.1 Dados de entrada

Neste modelo são definidos como dados de entrada os *rankings* gerados por um juiz ou sistema (*matriz de dados*) e o coeficiente de peso de cada um deles (*matriz de pesos*).

3.1.1 Matriz de Dados

A matriz de dados será formada a partir do conjunto de rankings com que se deseja realizar a fusão. Estes rankings, por sua vez, são provenientes de avaliações realizadas por diversos julgadores.

Assim, seja r_1, r_2,..., r_N, N *rankings*, e U um universo de p elementos, formado pela união dos elementos destes rankings, a matriz de dados F_{Nxp}, com linhas representando a avaliação dos juízes e colunas representando os elementos, é dada por:

$$F[j,i] = \begin{cases} f_{r_i}(j), \text{se } r_i \text{ é baseado em scores} \\ f_{r_i^*}(j), \text{se } r_i \text{ é baseado em } r\text{anks} \end{cases} \quad (1)$$

Caso os *rankings* envolvidos não sejam baseados em *scores*, os mesmos devem ser ajustados de maneira a serem entendidos como *scores*. Desta forma o *ranking* ajustado r^* é dado por:

$$r_i^* = \left[max - \frac{f_{r_i}(z) - 1}{|r_i| - 1}(max - min) \right] \; z \in r_i, \; i = 1,2,...,N \quad (2)$$

3.1.2 Matriz de pesos

A matriz de pesos é opcional, pois pode haver ou não a necessidade de privilegiar um ou outro *ranking*. É definida por uma matriz W_{1xN} , onde cada elemento da matriz armazena os coeficientes de peso w_i relativos a cada um dos *rankings* envolvidos.

Os valores dos *elementos* w_i devem estar no intervalo [0,1], sendo que $\sum_i w_i = 1$, pois estes valores representam percentuais de preferência.

3.2 Análise de preferência

A análise de preferência dos julgamentos está embasada no modelo MDPREF. O objetivo desta análise é conseguir diminuir a dimensionalidade dos dados e gerar duas matrizes contendo as configurações dos objetos e dos juízes.

A base do modelo MDPREF está na definição de uma matriz de concordância *S,* a qual pode ser escrita, pelo uso de PCA (*Principal Component Analysis*), como combinação de outras duas matrizes *X* e *Y,* as quais contêm as configurações dos juízes e dos objetos, respectivamente.

Para construir a matriz *S*, são definidas, inicialmente, matrizes D^i, para $i = 1,2,...,N$, contendo a comparação entre os pares de elementos *j* e *k*, referente ao juiz *i*, onde d^i_{jk} é dado por:

$$d^i_{jk} = \begin{cases} 1, \text{se o juiz } i \text{ avaliou } j > k \\ -1, \text{se o juiz } i \text{ avaliou } j < k \\ 0, \text{se o juiz } i \text{ avaliou } j = k \text{ ou não respondeu} \end{cases} \quad (3)$$

A matriz de concordância *S* será formada pela diferença entre as preferências de *j* sobre *k* para um juiz *i* e, cada elemento desta matriz é dado por:

$$s_{ij} = \sqrt{w_i} \sum_{j \neq k} \left(d^i_{jk} - d^i_{kj} \right) \quad (4)$$

Para expressar a matriz de concordância como uma combinação de outras duas, a mesma é decomposta na forma $S = ULA'$, pelo método SVD (*Singular Value Decomposition*), onde:

- U e A são matrizes cujas colunas são ortogonais $(U'U = I \text{ e } A'A = I)$ e contém os autovetores de SS′ e S′S, respectivamente.
- L é uma matriz diagonal contendo autovalores.

Por fim, são tomados os dois autovetores mais significativos de *U* e *A*, de acordo com a importância dos autovalores, e as matrizes *X* e *Y* são dadas por:

$$X = U_2 L_2 \quad (5)$$

$$Y = A_2 \quad (6)$$

Neste ponto, a análise gráfica das informações é permitida, plotando-se as matrizes *X* e *Y,* as quais mostrarão os elementos posicionados em um espaço bidimensional através de pontos e cada juiz será representado por um vetor dirigido.

3.3 Ranking consensual

Com a decomposição realizada anteriormente, a matriz Y carrega as preferências dos juízes de forma vetorial; assim, o vetor de preferência ideal é obtido pela soma dos vetores de Y, ou seja, pela contribuição de todos os julgamentos.

Seja P o vetor de preferência consensual normalizado, então:

$$P = \left[\sum_{i=1}^{N} x_i \quad \sum_{i=1}^{N} y_i \right] \Big/ \left\| \left[\sum_{i=1}^{N} x_i \quad \sum_{i=1}^{N} y_i \right] \right\| \tag{7}$$

O *ranking* consensual é gerado a partir da projeção das configurações dos objetos sobre o vetor P, no qual o objeto mais preferido é aquele cuja projeção é maior. Desta forma, define-se r_c o *ranking* proveniente da fusão das demais listas como $r_c = XP$.

4 Validação do modelo de fusão de rankings

Nesta seção serão apresentados os experimentos realizados com a intenção de validar o modelo apresentado e seus resultados.

4.1 Conjunto de dados

Os experimentos foram realizados com base em dois conjuntos de dados, provenientes de duas aplicações: uma de busca de competências (RECH, 2007), e outra de descoberta de qualificação de pesquisadores (HANNEL, 2008), as quais utilizam uma metodologia própria para determinar um *ranking* resultante.

Os conjuntos de dados referem-se a indicadores extraídos do CV-Lattes de 12 pesquisadores doutores da área da Ciência da Computação. O primeiro conjunto foi extraído do trabalho de Rech (2007) e apresenta 21 indicadores quantitativos bibliográficos (CJ1b) e 23 do currículo (CJ1c). Já o segundo (CJ2) foi extraído do trabalho de Hannel (2008), o qual possui 23 indicadores semelhantes aos de CJ1.

4.2 Ensaios

Os experimentos foram realizados em duas etapas. A primeira destina-se à aplicação do método sobre um conjunto de dados; já a segunda etapa preocupa-se com a determinação dos valores de qualidade e ruído das fusões.

4.2.1 Etapa 1: Aplicação do Modelo

Esta etapa dos experimentos foi dividida em duas baterias, reproduzindo os experimentos originais, mas com a aplicação do método de fusão proposto. Após o processo de fusão, os *rankings* obtidos foram comparados com os apresentados originalmente.

A primeira bateria consistiu em aplicar o modelo sobre o conjunto de dados CJ1 e, foi dividida em três momentos: no primeiro momento, gerou-se um *ranking* a partir dos dados do conjunto CJ1b; no segundo momento, o processo foi realizado com base em CJ1c; e, no terceiro momento, o *ranking* geral de pesquisadores foi obtido através da união dos dados de CJ1b e CJ1c, ou seja, sobre todos os indicadores. Na segunda bateria, um *ranking* de pesquisadores foi obtido a partir das informações de CJ2.

4.2.2 Etapa 2: Cálculo da Qualidade e Ruído

Nesta etapa foram calculados os valores de qualidade e ruído: (i) para cada uma das fusões da Etapa 1 e (ii) para os resultados originais, tendo como objetivo compará-los. Para se efetuar o cálculo, foi definido um conjunto de *rankings* composto pelas listas de entrada e pelo *ranking* resultante. Desta forma, para cada fusão realizada pelo método proposto, dois valores de qualidade e ruído foram determinados: um utilizando o *ranking* da fusão, e outro utilizando o *ranking* da proposta original.

Os experimentos foram realizados em duas baterias, reproduzindo os experimentos originais, mas com a aplicação do método de fusão proposto. Após o processo de fusão, os *rankings* obtidos foram comparados com os apresentados originalmente.

4.3 Resultados

4.3.1 Etapa 1

A Tabela 1 apresenta a comparação entre os *rankings* obtidos através do modelo de Análise de Preferência e os *rankings* obtidos por Rech (2007) e Hannel (2008), nas duas baterias de testes. Os valores entre parênteses, na tabela representam a distância normalizada entre o elemento e seu antecessor no *ranking*.

Tabela 1: Resultados dos experimentos da etapa 1.

Posição	1ª bateria						2ª bateria	
	CJ1b		CJ1c		CJ1		CJ2	
	Proposto	Original	Proposto	Original	Proposto	Original	Proposta	Original
1	P11	P11	P4	P4	P11	P11	P4	P11
2	P4 (0,37)	P4	P11 (0,09)	P11	P4 (0,17)	P4	P11(0,10)	P8
3	P8 (0,05)	P8	P10 (0,08)	P10	P8 (0,30)	P8	P8 (0,07)	P4
4	P3 (0,07)	P3	P5 (0,15)	P8	P3 (0,01)	P10	P3 (0,09)	P10
5	P12(0,04)	P10	P8 (0,10)	P5	P10(0,00)	P3	P5 (0,13)	P5
6	P5 (0,17)	P12	P3 (0,09)	P3	P5 (0,03)	P5	P9 (0,09)	P3
7	P10(	P1	P12 (0,16)	P9	P12(0,06)	P12	P6 (0,01)	P6
8	P6 (0,12)	P5	P6 (0,00)	P12	P6 (0,22)	P9	P12(0,03)	P9
9	P2 (0,05)	P7	P9 (0,1)	P6	P7 (0,08)	P1	P7 (0,22)	P12
10	P7 (0,00)	P6	P1 (0,05)	P7	P9 (0,02)	P7	P1 (0,04)	P7
11	P1 (0,06)	P2	P7 (0,06)	P1	P1 (0,04)	P6	P10(0,08)	P1
12	P9 (0,01)	P9	P2 (0,07)	P2	P2 (0,02)	P2	P2 (0,07)	P2

Fonte: os autores.

4.3.2 Etapa 2

Os valores de qualidade e ruído foram calculados para cada uma das fusões realizadas pelo modelo proposto. Os mesmos cálculos foram feitos utilizando os resultados dos trabalhos de Rech (2007) e Hannel (2008), conforme Tabela 2.

Tabela 2: Comparativo dos valores de qualidade e ruído dos experimentos.

		Proposto	Original
CJ1b	*Qualidade*	0,4667	0,4707
	Ruído	0,4337	0,4442
CJ1c	*Qualidade*	0,6242	0,6251
	Ruído	0,5247	0,5247
CJ1	*Qualidade*	0,5818	0,5827
	Ruído	0,1389	0,1392
CJ2	*Qualidade*	0,3984	0,4019
	Ruído	0,1237	0,1247

Fonte: os autores.

4.4 Discussão

Pode-se perceber, por meio dos resultados, que os *rankings* obtidos pela aplicação do modelo proposto não são idênticos aos *rankings* apresentados originalmente. Entretanto, há algumas semelhanças entre eles, principalmente nas fusões ocorridas na primeira bateria da Etapa 1. Tal inferência está expressa na Tabela 3, que apresenta o valor de similaridade entre eles.

Tabela 3: Similaridade entre os *rankings* proposto e original.

	Similaridade
CJ1b	0,88
CJ1c	0,94
CJ1	0,91
CJ2	0,84

Fonte: os autores.

Em relação a qualidade e ruído, percebe-se que houve uma redução, embora muito pequena, nos valores dos mesmos, perceptível, na maioria das vezes, apenas no terceiro dígito significativo. Como a diferença não é significativa, não se pode afirmar que o modelo apresentado é melhor ou preferível que os métodos utilizados originalmente.

Considerando que o ruído é uma medida de erro, pode-se estabelecer uma razão entre tal medida e a qualidade, com o objetivo de se mensurar o desempenho da fusão (ξ) baseando-se nestas métricas.

Define-se o desempenho da fusão da seguinte maneira:

$$\xi = \frac{ruido}{1 - qualidade} \tag{8}$$

A Tabela 4 apresenta os valores de ξ para cada uma das fusões tanto originais como proposta.

Tabela 4: Comparativo entre o desempenho das fusões.

	Proposto	Original
CJ1b	0,8132	0,8503
CJ1c	1,3962	1,3995
CJ1	0,3321	0,3335
CJ2	0,2056	0,2084

Fonte: os autores.

É possível perceber-se que, embora com diferenças bem pequenas, os resultados obtidos através do modelo proposto tiveram um melhor desempenho, indicando que a relação ruído *versus* qualidade foi mais favorável ao modelo apresentado.

5 Aplicação em metabusca

Com a intenção de aplicar o modelo de fusão de *rankings* apresentado, Klinger (2009) desenvolveu um protótipo para testar tal abordagem em um cenário de metabusca. Neste trabalho, foi desenvolvida uma interface em que se insere um ou mais termos de busca, e a aplicação busca os *rankings* em diversos motores de busca, tais como Google, AOL, Yahoo, Alta Vista, entre outros e, em seguida, é feita a fusão desses *rankings* e apresentado o ranking resultante.

Os testes realizados envolveram buscas com um único termo, termos compostos sem aspas e termos compostos com aspas nos metabuscadores Ixquick, Iboogie e Dogpile, e os resultados foram comparados com o modelo de fusão de *rankings* proposto neste artigo. Após obter os resultados, o autor utiliza uma métrica para comparar a divergência entre os *rankings* apresentados utilizando a distância Hamming.

Conforme Klinger (2009), o método MDPREF mostrou-se adequado para o cenário de metabuscas comparando-se com outros sistemas com a mesma finalidade e que este modelo, de fato, busca um consenso entre os juízes envolvidos (motores de busca).

> De acordo com o MDPREF, nem sempre o site que aparece em mais *rankings* ocupa uma posição melhor no ranking final. Através de testes percebe-se que, por exemplo, um site que aparece em seis motores de busca, mas tira três primeiros lugares fica melhor posicionado que um site que possa ter figurado entre todos os motores de busca em posições de pouco destaque. (KLINGER, 2009)

6 Trabalhos relacionados

Na intenção de prover um método geral para a fusão de dados, Dwork et al. (2001) propoem o uso de cadeias de Markov (MC) motivados por vários aspectos: (i) necessidade de tratamento de listas parciais, (ii) tratamento de comparações desiguais, (iii) melhoria de heurísticas para fusão de *rankings* e, (iv) eficiência computacional.

Este método foi aplicado nos contextos de metabuscas, redução de *spam* e *Word association,* e apresentou bons resultados, conforme os experimentos realizados pelos autores. Os mesmo ainda destacam que a proposta é de fácil implementação, não possui excessos computacionais e supera em desempenho os métodos

tradicionais. Entretanto, não contempla a utilização de listas baseadas em *scores* e não permite que se determinem pesos diferenciados para cada *ranking* envolvido, limitações estas que a proposta deste artigo contempla.

Dutra Júnior (2008) apresenta um modelo de fusão de dados que visa a possibilitar a utilização de *rankings* parciais e coeficientes de peso. O modelo proposto por ele baseia-se em um método linear de fusão de dados estritamente posicional; ou seja, a fusão se processa considerando apenas a posição dos elementos em cada lista. Para permitir que *rankings* parciais estejam envolvidos, é definido o completamento de *rankings* parciais (DUTRA JÚNIOR; LIMA, 2006) para ajustar as diferenças entre os elementos de cada lista. Em seus experimentos, foi possível demonstrar a validade do método.

Farah e Vanderpooten (2007) propõem um modelo de fusão de dados em que o *ranking* consensual é obtido com base em regras de decisão que identificam aspectos positivos e negativos para que um elemento seja posicionado melhor que outro. Para tanto são apresentados dois tipos de condições: (i) *condição de concordância,* que assegura que a maioria dos *rankings* concorda que o elemento *i* seja posicionado melhor que *j* e, (ii) c*ondição de discordância,* que assegura que nenhuma das listas envolvidas rejeita fortemente que *i* seja melhor que *j*.

O *ranking* consensual é obtido através de um processo de "purificação" em série, na qual são formados grupos de elementos com características de concordância e discordância semelhantes (necessita estabelecer um valor de *threshold* para a concordância máxima e discordância mínima). A união desses grupos representa a resultante da fusão. Embora com limitações em relação ao tipo de *rank* e contexto de aplicação, a abordagem mostrou-se superior ao método MC.

Conclusões e trabalhos futuros

O presente artigo propôs um modelo de fusão de dados baseado na análise de preferência que visa a permitir: (i) a fusão dos *rankings* utilizando várias listas, (ii) a utilização de listas parciais e/ou completas, (iii) a definição de coeficientes de pesos distintos para cada uma das listas, (iv) a possibilidade de utilizar listas baseadas em *ranks* e/ou *scores*, (v) a visualização gráfica dos resultados em um plano, e (vi) a possibilidade de mensurar a distância entre um elemento e outro após a fusão. Também teve como objetivo utilizar este modelo em uma aplicação de recuperação de informações na *web*.

Considerando-se os experimentos realizados no modelo proposto, percebe-se que o mesmo é capaz de efetuar a fusão dos *rankings* suprindo as deficiências encontradas em outros modelos sem degradar a performance dos *rankings* resultantes, validando, assim, o modelo em questão. A aplicação da fusão de *rankings* com base na análise de preferência na recuperação de informações na *web* foi testada em um sistema metabuscador, e seus resultados puderam comprovar que o modelo permite sua utilização neste contexto.

Sugere-se como trabalhos futuros uma análise comparativa entre a performance dos *rankings* gerados pelo modelo proposto e um *ranking* especialista (gerado manualmente).

Referências

ADALI, S.; MAGDON-ISMAIL, M.; MARSHAL, B. A Classification Algorithm for Finding the Optimal Rank Aggregation Method. In: INTERNATIONAL SYMPOSIUM ON COMPUTER AND INFORMATION SCIENCES. 22, 2007. *Proceedings***...** Ankara: IEEE, 2007.p. 1-6.

DAS, G. et al. Ordering the Attributes of Query Results. In: ACM SIGMOD INTERNATIONAL CONFERENCE ON MANAGEMENT OF DATA, 25, 2006. *Proceedings...* New York: ACM, 2006. p. 395-406.

DUTRA JUNIOR, E. G. *Um modelo de fusão de rankings baseado na análise de preferência*. 2008. 74 f. Dissertação (Mestrado em Ciência da Computação) – Instituto de Informática, 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, 2006. p. 148-154.

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

FAGIN, R.; KUMAR, R.; SIVAKUMAR, D. Comparing top k lists. In: ANNUAL ACM-SIAM SYMPOSIUM ON DISCRETE ALGORITHMS, 14., 2003. *Proceedings***...** Philadelphia: Society for Industrial and Applied Mathematics, 2003. p. 28-36.

FAGIN, R. et al. Comparing and Aggregating Rankings with Ties. In: ACM SIGMOD-SIGACT-SIGART SYMPOSIUM ON PRINCIPLES OF DATABASE SYSTEMS, 33., 2004. *Proceedings***...** New York: ACM, 2004. p. 47-58.

FARAH, M; VANDERPOOTEN, D. An Outranking Approach for Rank Aggregation in Information Retrieval. In: ANNUAL INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 30., 2007. *Proceedings...* New York: ACM, 2007. p. 591-598.

GUHA, S. et al. Merging the Results of Approximate Match Operations. In: INTERNATIONAL CONFERENCE ON VERY LARGE DATA BASES, VLDB, 30., 2004. Proceedings... [S.l.:s.n.], 2004. p. 636-647.

HAIR, J. F. et al.. *Multivariate Data Analysis*. 5. ed. Porto Alegre: Bookman, 2005.

HANNEL, K. *Qualificação de pesquisadores por área da ciência da computação com base em uma ontologia de perfil*. 2008. 99 f. Dissertação (Mestrado em Ciência da Computação) – Instituto de Informática, UFRGS, Porto Alegre, 2008.

KLINGER, A. *O modelo de fusão de rankings baseado em análise de preferência aplicado a metabusca*. 2009. 40 f. Monografia (Graduação em Informática) – Instituto de Informática, UFRGS, Porto Alegre, 2009.

RECH, R. O. *Um Modelo de Pontuação na Busca de Competências Acadêmicas de Pesquisadores*. 2007. 92 f. Dissertação (Mestrado em Ciência da Computação) – Instituto de Informática, UFRGS, Porto Alegre, 2007.

RENDA, M. E.; STRACCIA, U. Web Metasearch: Rank vs. Score Based Rank Aggregation Methods. In: ACM SYMPOSIUM ON APPLIED COMPUTING, 2003. *Proceedings***...** New York: ACM, 2003. p. 841-846.

VOGT, C. C.; COTTRELL, G. W. Fusion Via a Linear Combination of Scores. *Information Retrieval,* v. 1, n. 3, p. 151-173, Oct. 1999.