# Combinação de Métodos para Pesquisa de Informação

João Ferreira
ISEL
jferreira@deetc.isel.ipl.pt

Alberto Rodrigues da Silva
INESC-ID, IST
alberto.silva@acm.org

José Delgado
Instituto Superior Técnico
Jose.Delgado@tagus.ist.utl.pt

*Sumário: Pretende-se investigar diversos caminhos para combinar métodos de pesquisa por forma a melhorar o desempenho dos sistemas, oferecendo uma nova perspectiva da investigação dos sistemas de pesquisa, à descoberta da melhor estratégia, propondo um método de combinação baseado na combinação de três modelos: Textual, ligações e de classificação.*

**Palavras chave**: Combinações, Pesquisa Informação

## 1 Introdução

Descobrir o melhor caminho para encontrar a informação desejada, especialmente na Web é difícil senão mesmo impossível. Como ainda não foi encontrada uma teoria unificada de pesquisa, pode investigar-se a escolha do melhor de cada método e estudar a melhor estratégia para as combinar.

Outro grande desafio é a optimização dos parâmetros das combinações onde diferentes combinações resolvem problemas diferentes. Como [1] referiram, "deve haver uma ordem de combinação óptima, com os pesos dos termos óptimos para uma dada situação na pesquisa de informação". Por exemplo, ao combinar várias fontes de evidências para uma colecção simples de documentos podem obter-se resultados diferentes dos obtidos combinando resultados da pesquisa em colecções múltiplas ou combinando múltiplos paradigmas de pesquisa. Assim, há vários tipos de combinações de métodos para lidar com diferentes situações de pesquisa.

## 2 Combinação de Resultados

A combinação faz-se de acordo com a Figura 1, ao nível da:

- Combinação de colecções, usado essencialmente para sistemas distribuídos de recuperação, em que se faz a análise das combinações de matéria-prima dos sistemas de informação (i.e., colecções de documentos). Na Figura 1, faz-se referência a três métodos principais de combinação, informação mais detalhada encontra-se na página pessoal autor;
- Combinação de classificadores, elaborando-se uma síntese dos principais métodos de catalogação de documentos;
- Meta pesquisa, onde se aborda o tema da combinação de resultados de diferentes sistemas de pesquisa na Web;
- Combinação de resultados de representações de uma colecção;

- Perguntas; combina diferentes formulações de perguntas as quais podem resultar em diferentes expressões de necessidades de informação, bem como a escolha de representativos e do tipo de perguntas (e.g. booleana, linguagem natural);
- Documentos; combina diferentes representações de documentos, as quais podem resultar, na escolha das fontes (e.g. título, termos, sumários, texto total) e o método (e.g. *stopwords*, radicalizar, peso termos, escolha termos) ou a indexação;
- Pesos dos termos indexados; combina diferentes esquemas de pesos para os termos;

– Combinação de resultados de diferentes processos de comparação, ao seja por meio de algoritmos apropriados (formulas) combinam-se resultados obtidos por processos diferentes:

- Métodos, combinam diferentes métodos de pesquisa (e.g. probabilístico, espaço vectorial, seguimento das ligações) estimando a relevância de documentos em diferentes caminhos com ênfase nas diferentes características de documentos e perguntas;
- Retroacção, combinação de diferentes métodos de retroacção.

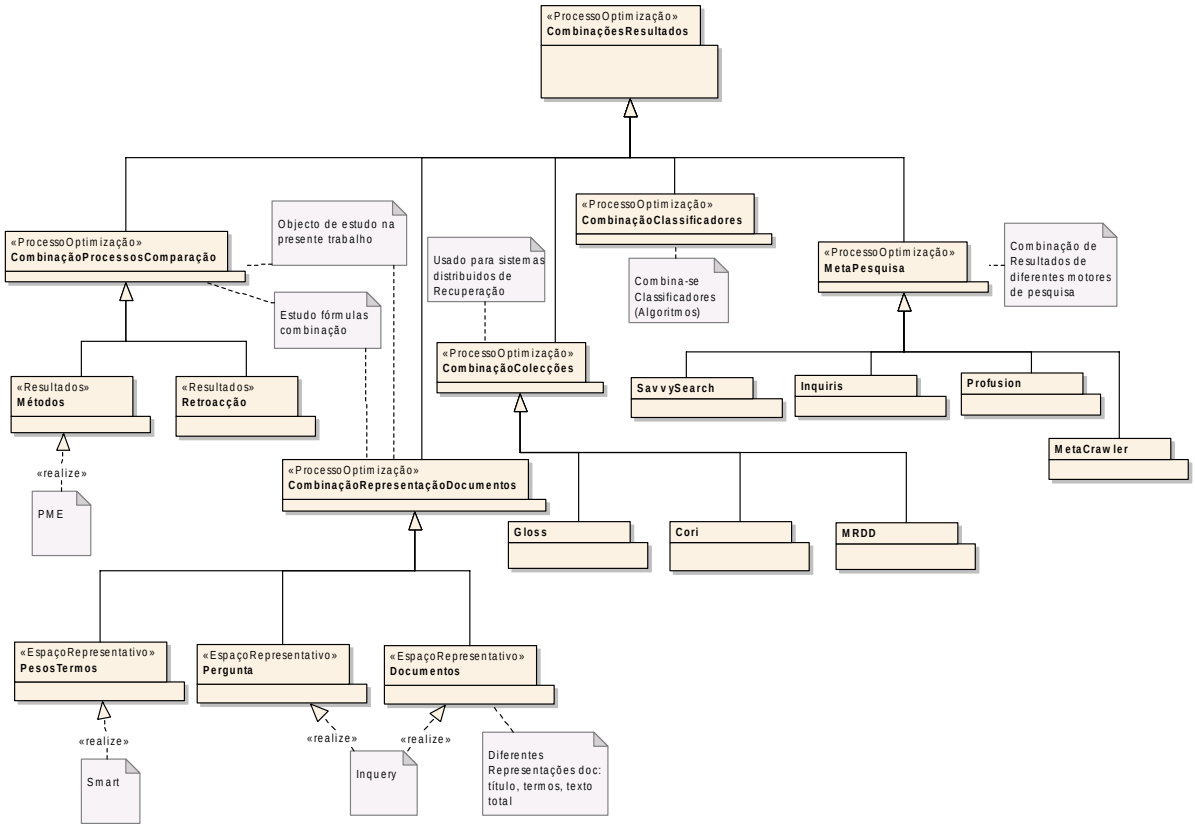


Figura 1: Principais formas de combinações.

Projectos e experiências realizadas, bem como uma descrição dos processos enunciados não abordados na presente dissertação, podem ser encontrados na página pessoal do autor, salientando-se uma predominância das combinações de diferentes métodos de pesquisa.

# 3 Fórmulas de Combinações

O problema da combinação de resultados obtidos dos métodos individuais é determinar quais são os parâmetros e os métodos mais importantes, sendo um problema actual da recuperação de informação [2,3,4,5,6]. Como combinar ou integrar as diferentes componentes é a questão central desta fase da investigação. Os caminhos mais usuais resumem-se a aplicar a combinação no momento da pesquisa (i.e. componentes combinados são integrados para produzir um único conjunto de resultados) ou após a pesquisa (i.e. múltiplos conjuntos de resultados são produzidos pela combinação de métodos aplicados em paralelo após a pesquisa). Na presente dissertação, é aplicada a combinação de métodos após a pesquisa usando duas das fórmulas de combinação mais comuns:

– Combinação de semelhanças (Fox e Shaw 1994 1995; Lee 1996 1997);

– Somas pesadas [4,10,3];

– Ambas as fórmulas calculam uma medida de combinação linear das componentes que medem as semelhanças das perguntas e dos documentos, numa escala ordenada.

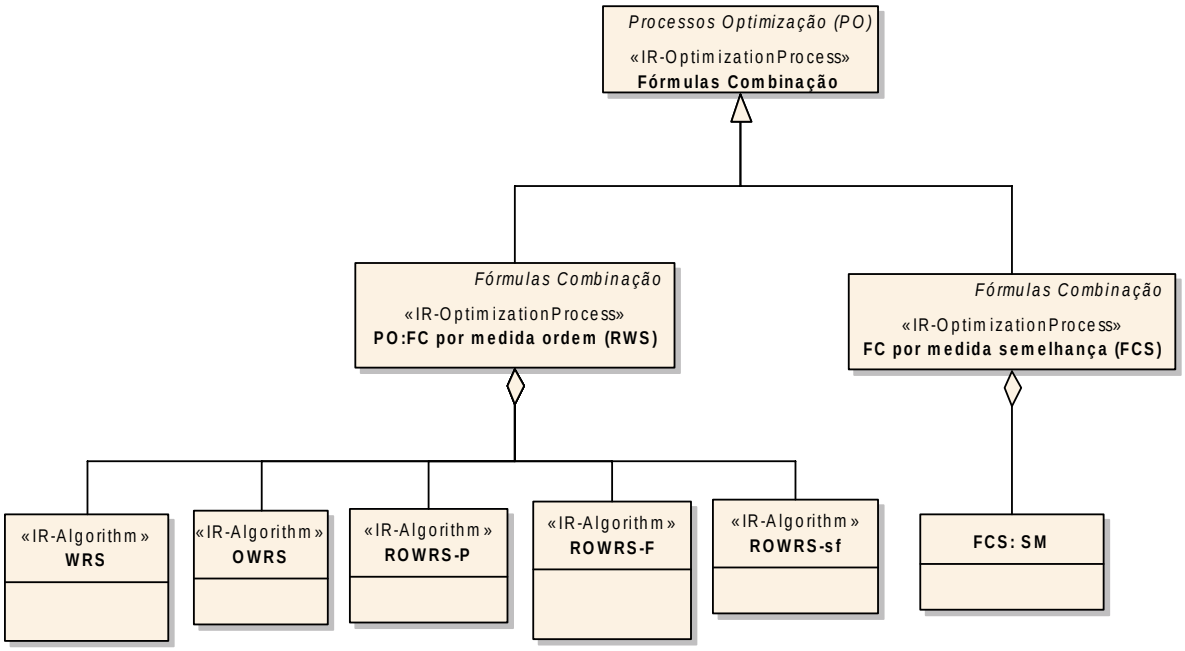


Figura 2: Principais fórmulas de Combinação

## 3.1 União de Semelhanças

O conceito da união de semelhanças (*Similarity Merge*, SM), nas fórmulas combinadas, foi introduzido inicialmente por Fox e Shaw [7] e refinado por Lee [2,8], calculando a medida combinada de um documento pela soma das medidas normalizadas estimuladas pela sobreposição da pesquisa. Quando combinada com um grande número de conjuntos resultantes, a sobreposição torna-se relevante com medidas normalizadas. Contudo, um método com mais variações de

sistemas pode dominar o processo de combinações pois tem tendência para apresentar uma medida mais elevada.

A sobreposição é normalizada pelo número de sistemas num determinado método. A Fórmula F1 descreve a forma de combinação usada para ordenar documentos pesquisados por sistemas diferentes:

$$FS = (\sum NS_i) * \frac{olp}{m(i)} \qquad \text{(F1)}$$

$FS$ = medida de combinação de um determinado documento; $NS_i$ = medida normalizada do documento pelo sistema $i$; $olp$ = número de sistemas que pesquisaram um determinado documento; $m(i)$ = número de métodos a que o sistema $i$ pertence.

A medida normalizada do documento $NS_i$ é calculada pela fórmula min-máx de Lee [2,8] sendo que $S_i$ é a medida de pesquisa de um determinado documento e $S_{máx}$ e $S_{min}$ são as medidas máxima e mínima dos documentos no sistema $i$:

$$NS_i = (S_i - S_{min}) / (S_{máx} - S_{min}) \qquad \text{(F2)}$$

Esta fórmula (SM) é de simples implementação não requerendo dados de treino ou qualquer refinamento, sendo de baixo custo computacional e dando ênfase à sobreposição. Por outro lado, esta fórmula (SM) não leva em consideração a diferença dos vários componentes combinados nem distingue a sobreposição de diferentes sistemas.

### 3.2 Soma ordenada de pesos

Quando os componentes dos sistemas combinados são distintos uns dos outros, a normalização das medidas dos documentos entre sistemas pode não compensar as diferenças nas ordens dos documentos apresentados. Este é o caso da combinação de métodos de sistemas de pesquisa textual, de ligações e de classificação, cujas medidas de semelhança documento / pergunta são calculadas de forma diferente:

- Sistemas vectoriais, medem a semelhança entre perguntas e documentos;
- Sistemas probabilísticos medem a probabilidade de relevância;
- Sistemas HITS representam as autoridades das ligações de um documento em relação ao assunto da pergunta;
- Sistemas de classificação medem a probabilidade do documento pertencer à mesma categoria da pergunta.

Neste cenário, é útil combinar as ordens dos documentos em vez de combinar as medidas.

Para compensar as diferenças entre a combinação das componentes dos sistemas surge a fórmula Soma das Ordens Pesadas (*Weighted Rank Sum* (WRS)), que usa medidas baseadas em ordens (i.e.

1/ordem) no lugar das medidas dos documentos na fórmula F3:

$$FS = \sum (w_i * RS_i) \quad \text{(F3)}$$

$w_i$ = peso do sistema *i*; $RS_i$ = medida de ordem do documento pelo sistema *i*.

Apesar de a Fórmula WRS tentar pesar as contribuições individuais dos componentes da combinação na pesquisa dando ênfase à sua força relativa, não explicita a diferença entre sobreposição ou não de instâncias, (sumário das medidas das componentes dos sistemas de combinação implicitamente recompensam a sobreposição). Por outras palavras, a contribuição absoluta do documento pesquisado por um sistema permanece a mesma independentemente de ser ou não pesquisado por outro sistema. O que a fórmula WRS despreza é a possibilidade de a contribuição de um documento poder ser diferente tendo em conta a sobreposição de partições (i.e. documentos pesquisados por um ou dois sistemas apenas, etc.).

A soma das medidas de ordem sobrepostas (*Overlap Weighted Rank Sum* (OWRS)) tenta suprir o problema anteriormente referido tendo em conta a sobreposição de partições.

$$FS = \sum (wik * RSi) \quad \text{(F4)}$$

$w_{ik}$ = peso do sistema i na sobreposição da partição *k;* $RS_i$ = medida de ordem do documento pelo sistema *i*.

A soma das medidas de ordem sobrepostas ordenadas (*Rank-Overlap Weighted Rank Sum* (ROWRS)) é uma variação da fórmula OWRS que considera não só a sobreposição de partições como também a ordem pela qual um documento é pesquisado. A Fórmula F5 descreve a fórmula ROWRS:

$$FS = \sum (w_{ikj} * RS_i) \quad \text{(F5)}$$

$w_{ikj}$ = peso do sistema *i* na sobreposição da partição *k* na ordem *j;* $RS_i$ = medida de ordem do documento pelo sistema *i*.

Em todas as Fórmulas F3, F4 e F5, os conjuntos de treino usados como dados de treino para determinar os pesos:

- $w_i$ (Fórmula F3) é determinada pela média da precisão geral (i.e. média dos valores de precisão média das perguntas de treino), que é uma simples medida que reflecte o desempenho geral sobre todos os documentos;
- $w_{ik}$ (Fórmula F4) precisão média global é multiplicada pela média da precisão sobreposta. Esta precisão média é calculada para cada partição sobreposta. Numa combinação de três sistemas, a precisão média é calculada para cada uma das quarto partições sobrepostas de cada sistema, Tabela 1. De outra forma, o conjunto de resultados de um sistema é dividido em partições sobrepostas (i.e. para o sistema A: documentos pesquisados pelo sistema A e B por sistema A e C por sistema A B e C) e a precisão média é calculada para cada partição de cada

sistema;

| | | |
|---|---|---|
| A: Método A apenas | B: Método B apenas | C: Método C apenas |
| AB: Método A e B | BC: Método B e C | AC: Método A e C |
| | ABC: Método A, B e C | |

Tabela 1: Sobreposição da partição de 3 sistemas.

– $w_{ikj}$ (Fórmula F5), é necessário estimar o desempenho numa dada ordem, e por isso a precisão média global não é adequada. Assim três medidas de ordem em cada ordem são usadas para calcular os pesos das três versões da fórmula ROWRS: (1) eficiência *(F);* (2) precisão (P); (3) sucesso/falhas (sf).

## 4 Conclusões

Elaborou-se uma sistematização dos conceitos que as combinações de métodos de pesquisa envolvem. Este trabalho esta envolvido num geral de criar uma linguagem específica para a pesquisa de informação, a qual permita a sistematização e conceitos e uniformização de notação.

Referências


[1] Belkin N. J. Cool. C. Croft W. B. e Callan J. P. (1993). The effect of multiple query representations on information retrieval system performance. Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval 339-346.

[2] Lee J. H. (1997). Analyses of multiple evidence combination. Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval 267-276.

[3] Modha D. e Spangler W. S. (2000). Clustering hypertext with applications to Web searching. Proceedings of the 11th ACM Hypertext Conference 143-152.

[4] Bartell B. T. Cottrell G. W. e Belew R. K. (1994). Automatic combination of multiple ranked retrieval systems. Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval.

[5] Beitzel et al. (2003) Recent Results on Fusion of Effective Retrieval Strategies in the Same Information Retrieval. Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval.

[6] Montague, M., & Aslam, J. (2002, November). Condorcet Fusion for Improved Retrieval. Paper presented at the 11th Annual ACM Conference on Information and Knowledge Management (CIKM-2002), Tyson's Corner, VA.

[7] Fox E. A. e Shaw J. A. (1994). Combination of multiple searches. In D. K. Harman (Ed.) The Second Text Rerieval Conference (TREC-2) (NIST Spec. Publ. 500-215 pp. 243-252). Washington DC: U.S. Government Printing Office.

[8] Lee J. H. (1996). *Combining multiple evidence from different relevance feedback methods* (Tech. Rep. No. IR-87). Amherst: University of Massachusetts Center for Intelligent Information Retrieval.

[9] Lee J. H. (1997). Analyses of multiple evidence combination. Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval 267-276.

[10] Larkey L. e Croft W. B. (1996). Combining Classifiers in Text Categorization. Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval 289-297.