Combinações de Sistemas de Pesquisa de Informação

João Ferreira
Instituto Superior de Engenharia de Lisboa
jferreira@deetc.isel.ipl.pt

Alberto Rodrigues da Silva
INESC-ID, Instituto Superior Técnico
alberto.silva@acm.org

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

SUMÁRIO

O presente trabalho explora o problema da pesquisa de informação na Web, explorando a combinação de resultados dos principais sistemas de pesquisa (textual, seguimento de ligações e classificação) de forma a melhorar os resultados da pesquisa de informação (i.e., aumento da precisão e cobertura). São analisados resultados das diferentes fórmulas de combinação tendo como objectivos a melhoria de resultados.

PALAVRAS CHAVE

Pesquisa de Informação, Classificação, Combinações

1 Introdução e Objectivos

Os avanços tecnológicos permitem uma maior facilidade na produção e difusão de informação. Esta realidade conduz muitas vezes a situações de excesso de informação, incapacitando o utilizador de obter a informação desejada, a qual se torna cada vez mais indispensável e crítica. Por estas e outras razões, torna-se relevante o estudo e desenvolvimento de mecanismos que dada uma necessidade conduzam à obtenção da informação desejada no mais curto espaço de tempo. Os principais métodos de pesquisa são: (1) o textual, (2) seguimento de ligações e (3) de classificação.

A natureza do ambiente de pesquisa na Web é tal que as aproximações de pesquisas baseadas em fontes simples de evidências, apresentam fraquezas que degradam o desempenho da pesquisa em certas situações. Por exemplo, as aproximações baseadas no conteúdo textual têm dificuldade em lidar com a diversidade de vocabulário e qualidade de documentos da Web, enquanto que as aproximações baseadas nas ligações sofrem do ruído ou das ligações incompletas existentes na topologia das ligações.

Como combinar ou integrar componentes das combinações é a questão central da investigação feita neste domínio. 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). No presente artigo é aplicada a combinação de métodos após a pesquisa usando duas das fórmulas de combinação mais usadas: combinação de semelhanças (Fox e Shaw 1993 1994; Lee 1996 1997) e somas pesadas (Bartell et al. 1994; Larkey e Croft 1996; Modha e Spangler 2000; Thompson 1990).

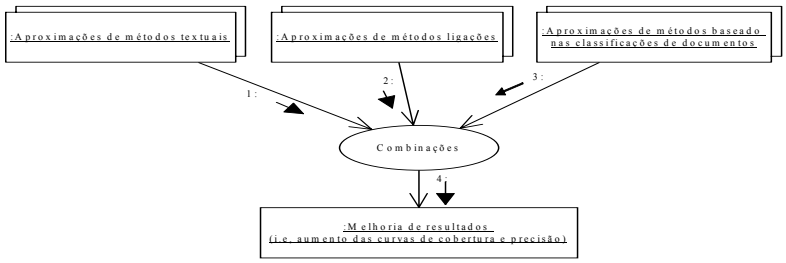

Figura 1: Combinação de diferentes métodos de pesquisa.

In *Proceedings of Ibero-Americana WWW/Internet 2004 (IADIS) Conference (CIAWI2004)*.

2 Fórmulas de combinações

A **fórmula de combinação de semelhanças** multiplica a soma das medidas individuais do documento pelo número de componentes combinadas que pesquisaram o documento (i.e. sobreposição) baseado no pressuposto de que os documentos com maior sobreposição tendem a ser mais relevantes. A **fórmula das somas pesadas** adiciona os pesos dos componentes com as respectivas contribuições, os quais são estimados com base nos dados de treino. Ambas as fórmulas calculam uma medida linear de combinação linear das componentes de combinação as quais tendem a medir as semelhanças das perguntas e dos documentos, numa escala ordenada.

As fórmulas de combinação investigadas nesta dissertação são baseadas nestas duas fórmulas de combinação conforme a pesquisa descrita nas secções seguintes.

2.1 União de Semelhanças

A união de semelhanças (*Similarity Merge* - SM) nas fórmulas combinadas foi introduzida inicialmente por Fox e Shaw (1993;1994) e refinadas por Lee (1996;1997), 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 dá importância à sobreposição 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.

Existe o problema de combinar um largo número de sistemas com contribuições desiguais (36 VSM 6 HITS 120 DC 24 TM sistemas) 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:

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

onde: 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 (1996 1997) onde S_i é a medida de pesquisa de um determinado documento e $S_{máx}$ e S_{min} são as medidas máxima e mínima do documento 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 dá ênfase à sobreposição. Por outro lado, esta fórmula (SM) não leva em consideração a diferença dos vários componentes combinados nem distingue a sobreposição de sistemas VSM e HITS com sistemas HITS e *Web Directory-based* (WD).

A fórmula da soma dos pesos (WS) tenta compensar a fórmula SM pela contribuição pesada das componentes de acordo com a sua força relativa estimando a relevância de uma documento a uma pergunta pela soma estimadas dos pesos individuais. O conjunto de pesos óptimos pode ser determinado, seleccionando o melhor conjunto de pesos de 0.1 a 0.9 em passos de 0.1 (Larkey e Croft 1996) por métodos de optimização numérica (Bartell et al. 1994; Modha e Spangler 2000) ou baseados no desempenho individual das componentes combinadas (Thompson 1990).

Não há na literatura uma indicação clara de quanto se pode ganhar pelo emprego de diferentes fórmulas de combinações. Ambos os métodos têm as suas fraquezas especialmente quando aplicadas no contexto da pesquisa e informação na Web. A segunda fórmula de combinação é uma extensão do método WS numa tentativa de resolver alguns dos problemas apresentados pelas fórmulas SM e WS.

2.2 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. Os sistemas VSM medem a semelhança entre perguntas e documentos, nos sistemas HITS representam as autoridades das ligações de um documento em relação ao assunto da pergunta e os sistemas medem a probabilidade do documento pertencer à mesma categoria da pergunta. Neste cenário é útil combinar as ordens dos documentos em vez das medidas.

Para compensar as diferenças entre a combinação das componentes dos sistemas surge a fórmula Soma das Ordens Pesadas (*Weighted Rank Sum* (WRS)), a qual usa medidas baseadas em ordens (i.e. 1/ordem) no lugar das medidas dos documentos na fórmula WS:

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

onde: FS = medida de combinação do documento; 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, (sumario 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 despresa é a possibilidade da 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.

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

onde: FS = medida de combinação do documento; 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 pelo qual um documento é pesquisado. A fórmula F5 descreve a fórmula ROWRS:

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

onde: FS = medida de combinação do documento; 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 de somas pesadas os tópicos 451 a 500 são usados como dados de treino para determinar os pesos. A 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 relevantes foi usada para determinar os pesos na fórmula WRS (fórmula F3). A fórmula OWRS (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. De outra forma, o conjunto de resultados de um sistema é particionado 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.

Para a fórmula ROWRS (Fórmula F5), a qual necessita estimar o desempenho numa dada ordem, a precisão média global não é adequada. Assim três medidas de ordem *Precisão (P)* Eficiência *(F)* e *Successo/Falhas (sf)* em cada ordem são usadas para calcular os pesos das três versões da fórmula ROWRS. O valor F é o coeficiente de semelhança de Dice para um conjunto de documentos relevantes para um pergunta, o valor de precisão aumenta tendo em conta a cobertura numa dada ordem (Shaw 1986):

$$F = \frac{2r}{n + N_r} = \frac{2}{1/R + 1/P} \qquad \text{(F6)}$$

onde: r = número de documentos relevantes pesquisados; n = número de documentos pesquisados; Nr = número total de documentos relevantes; R = cobertura (r/Nr); P = precisão (r/n)

Como os pesos das medidas de ordem são sensíveis à ordem exacta do documento eles são aplicados em 'blocos de ordem' (i.e. ordens de 1 a 10, 11 a 20 etc.). Por outras palavras, as medidas de componentes combinadas (RS_i na fórmula F5) num dado bloco de ordem têm todo o mesmo peso e são determinados pela média das medidas sobre todos os blocos ordenados.

Como as medidas P e F são baseadas no desempenho para uma determinada ordem k (i.e. o número de documentos relevantes nos k primeiros resultados de topo) *sf* é a medida baseada no sucesso/falhas da pesquisa em cada ordem k (i.e. 1/k se o documento na ordem k é relevante ou 0 caso contrário). A medida *sf* estima o desempenho do sistema numa dada ordem do intervalo sem ter em conta o seu desempenho nos piores intervalos de ordem, numa tentativa de aumentar a probabilidade do sistema pesquisar documentos relevantes em ordens baixas. Por exemplo um documento não-relevante na ordem 101 com 100 documentos relevantes na ordem 1 a 100 (doc-A) terá maior P e F que um documento relevante na ordem 101 com 0 documentos relevantes na ordem 1 a 100 (doc-B) mas o *sf* do doc-B será maior que o *sf* do doc-A. Quando as componentes combinadas incluem sistemas que pesquisam documentos relevantes a baixas ordens esta aproximação é benéfica.

Ambas as fórmulas WRS e OWRS têm três variações que amplificam a contribuição do sistema com melhor desempenho analisado. As variações por ordem crescente que dão ênfase aos sistemas de topo são: (1) sistema topo 1 (st1); (2) sistema topo 2 (st2); (3) aumento da sobreposição (olpboost).

A ideia básica é ultrapassar o resultado do sistema com melhor desempenho usando uma função de combinação de pesos que amplifique a medida de ordem do documento pesquisado pelos sistemas de topo e ao mesmo tempo baixe as contribuições dos sistemas com desempenho inferior. Uma fórmula generalizada de *st1* st2 e *olpboost* pode ser expressa:

$$FS = \sum (wf_{kj}(L_i) * RS_i), \qquad \text{(F7)}$$

onde: FS = medida de combinação do documento; $w_{kj}(L_i)$ = peso da função do sistema do grupo L_i na sobreposição da partição k na ordem j ; L_i = grupo do sistema i baseado no desempenho; RS_i = medida baseada na ordem do documento pelo sistema i.

As fórmulas F8, F9 e F10 descrevem o peso das funções de *st1, st2* e *olpboost*:

$$wf_{kj}(L_i) = \begin{bmatrix} avgp(i) * w_{ikj} & se & L_i = st1 \\ fsc * w_{ikj} & se & L_i \neq st1 \;\; \& \;\; olp1 \\ 0 & caso - contra'rio & \end{bmatrix} \qquad \text{(F8)}$$

$$wf_{kj}(L_i) = \begin{bmatrix} avgp(i) * w_{ikj} & se & L_i = st1 & \\ fsc * w_{ikj} & se & L_i = st2 & e \;\; olp1 \\ fsc * avgp(i) * w_{ikj} & se & L_i \neq (st1 \,||\, st2) & e \;\; olp1 \\ 0 & caso - contra'rio & & \end{bmatrix} \qquad \text{(F9)}$$

$$wf_{kj}(L_i) = \begin{bmatrix} avgp(i) * w_{ikj} * ocnt & se & L_i = st1 & \\ fsc * w_{ikj} & se & L_i = st2 & e \;\; olp1 \\ fsc * avgp(i) * w_{ikj} & se & L_i \neq (st1 \,||\, st2) & e \;\; olp1 \\ 0 & caso - contra'rio & & \end{bmatrix} \qquad \text{(F10)}$$

onde: *avgp*(*i*) = média da precisão geral do sistema *i* no conjunto de treino; *ocnt* = número de sistema que pesquisaram o documento; *st1* = melhor sistema; *st2* =egundo melhor sistema; *olp1* = verdadeiro se o documento foi pesquisado por st1; *fsc* = medida provisória de combinação de um documento. Quando calculada a medida de combinação, csão somadas as medidas dos componentes no sistema pela ordem de desempenho (i.e. medidas de *st1* são adicionadas antes de *st2*) para assegurar um calculo consistente de *fsc*.

As equações acima expostas reordenam os resultados dos sistemas de topo apenas pela introdução de medidas nos documentos pesquisados por sistemas não de topo. Ao usar a medida de *fsc*, que se torna progressivamente maior com a sobreposição, estas fórmulas adicionam maior enfase ao factor de sobreposição o qual influência implicitamente antes do processo de sumarização. A fórmula F9 (*st2*) adiciona mais granulidade à função de pesos ao permitir variações nos níveis de contribuições dos sistemas sobrepostos enquanto que a fórmula F10 (*olpboost*) adiciona ainda outro aumento nos sistemas de topo pela multiplicação da sua medida com a sobreposição contada.

Teremos 11 fórmulas distintas para combinar resultados, dos quais serão analisados os resultados resultante da combinação dos melhores sistemas.

3 Sistemas de Pesquisa

Iremos ter três sistemas principais de pesquisa:

- VSM (modelo vectorial) do qual usamos como parâmetros: (1) comprimento da pergunta (pequena (p), media (m) e longa (l)); (2) índices (documento completo (d), título (t) e corpo do documento (c); (3) uso de frases (1-sim, 0-não); (4) uso de retroacção (1-sim, 0-não). A nomenclatura é v$comprimento_pergunta $ indice$ frase $ retroacção (v$l/m/p$d/t/c$1/0$1/0)
- Hits (modelo de seguimento de ligações, com base no algoritmo HITS (Kleinberg 1997), modificado com a implementação do algoritmo ARC (Chakrabarti et al. 1998b). São parametros deste sistema: (1) comprimento do endereços (longo (l), pequenos (p)); (2) conjunto semente baseado no melhor sistema simples VSM identificado (v*c10), do qual variou-se o comprimento das perguntas. A nomenclatura é h$comprimento_endereço$ comprimento_pergunta_conjunto_semente (hl/pl/m/p).
- Classificação: Iremos usar o método *Term Match* (TM) (Ferreira, 2004). Teremos como parametros: (1) categorias de topo (# 1,2,3) usadas para ; (2) indices (corpo documento sem frases -1; corpo documento com frases -2; documento completo sem frases -3; documento completo sem frases -4; (3) Retroacção (0-não, 1-sim). A nomenclatura é t$#cat. topo$ indice$ retroacção (t$1/2/3$1/2/3/4$1/0).

3.1.1 Sistemas simples

O melhor sistema VSM medido pela precisão média foi *vlc00*. O melhor sistema HITS *hpl*. O melhor sistema TM *t121*. De facto a precisão média dos sistemas de topo diminui sensívelmente para metade quando se passa de uns métodos para outros por esta ordem VSM, TM e HITS indicando a vantagem do método textual VSM sobre os outros. A análise dos resultados (sistemas simples), sugere que perguntas longas bem como a definição dos endereços são os parâmetros mais influêntes no desempenho do sistema de pesquisa.

VSM	v$pergunta (l,mp)$frase(0,1)
F2 (vlc00 e vlc10 combinados)	
F (v*c*0 combinados-1º * é a pergunta (l,m,p) e o 2º é o uso de frases (0,1))	
Hits	h$comprimento endereço (l,p)$pergunta(l,m,p)$frase(0,1)
F todos os sistemas combinados	
TM	t$#categoria(1,2,3)$frase(0,1)
F2 t100 e t110 combinados	ultimo parametro é a retroação 0- s/ uso
F3 t100, t110, t200 e t210 combinados	
F4 t110, t210 e t210 combinados	

		s/ Sist. topo	st1	st2	olpboost
Fórmula		0	1	2	3
WRS	b	b0			
OWRS	c	c0	c1	c2	c3
ROWRS-sf	d	d0	d1	d2	d3
ROWRS-P	e	e0			
ROWRS-F	f	f0			

Tabela 1: Nomenclatura geral dos sistemas usados (esquerda) e dos Sufixos (direita) a acrescentar à nomenclatura geral dos sistemas usados, os quais identificam a fórmula de combinação usada.

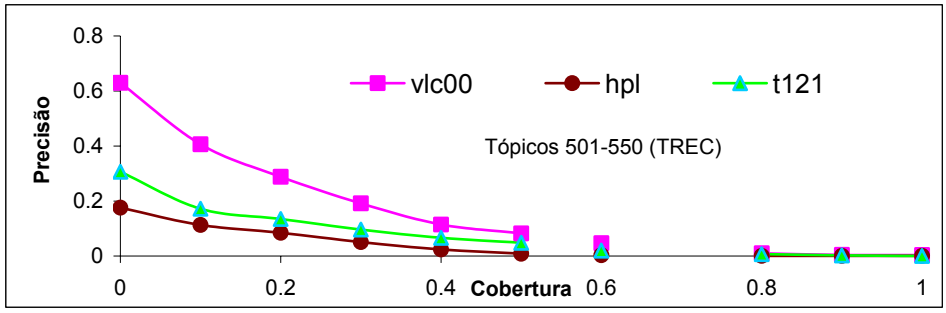

Figura 1: Curvas precisão-cobertura para sistemas simples de pesquisa para as perguntas tópicos 501-550.

3.1.2 Combinações internas

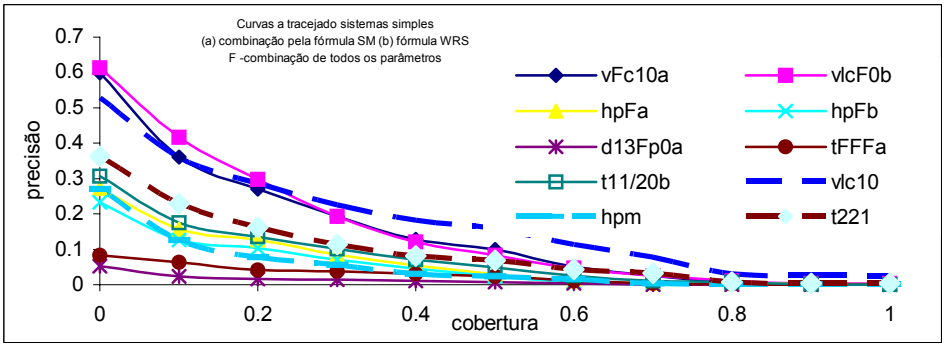

Figura 2: Curvas precisão-cobertura para combinação de parametros internos de cada sistema, para os tópicos 501-550.

As combinações do método-Internos de sistemas VSM e TM têm comportamento semelhante diminuindo o desempenho da linha de base embora a combinação dos sistemas TM degradem mais as combinações dos sistemas VSM usando a fórmula SM.

Nas combinações de sistemas HITS, a formula SM produz melhores resultados que a formula WRS baseada na ordem, a qual é oposta aos resultados da combinação dos sistemas VSM e TM. Para se determinar a razão do aumento do desempenho das combinações quando se integram sistemas HITS e se degradam quando integramos sistemas VSM e TM, foi examinada a sobreposição de documentos relevantes nos sistemas HITS com sistemas VSM e TM e verificou-se que os sistemas HITS pesquisam uma maior diversidade de documentos relevantes que os sistemas VSM ou TM e por isso tendem a ganhar mais com a combinação.

3.1.3 Combinações externas de sistemas de topo

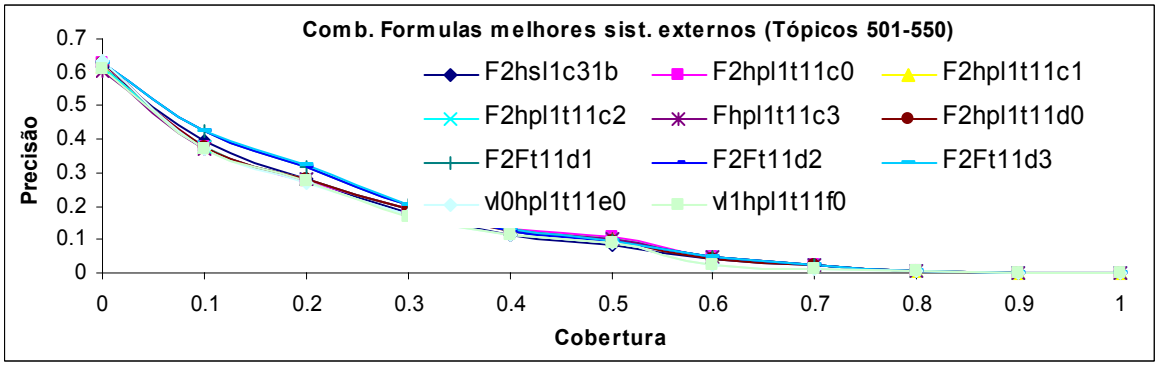

Figura 3: Curvas precisão-cobertura para combinação extenas de sistemas, para os tópicos 501-550.

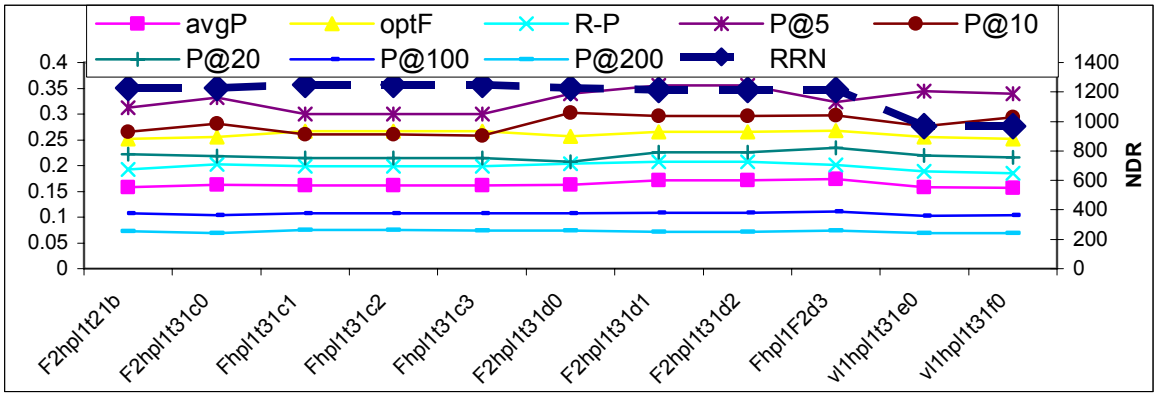

Figura 4: Resultados dos sistemas combinados de topo para os tópicos 501-550. NDR = Número total de documentos relevantes pesquisados; avgP = Precisão média sobre as perguntas; optF = F óptimo; R-P = Precisão R; P@k = Precisão na ordem k; PRj = Precisão de cobertura no nível j.

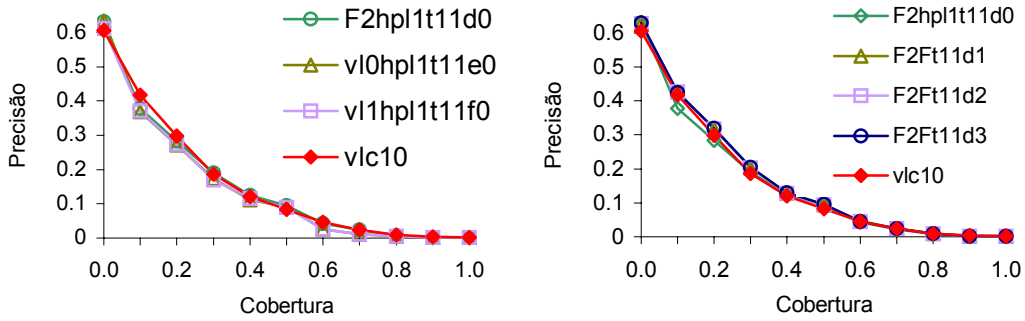

Figura 5: Curvas precisão-cobertura para tópicos 501-550. Gráfico esquerda resultados principais das fórmulas ROWRS-*sf*, ROWRS-*F*, ROWRS-*P e gráfico direita resultados para as variantes das fórmulas* ROWRS-*sf.*

O ganho no desempenho das melhores combinações, é apenas marginal pois os resultados são muito semelhantes aos resultados dos sistemas simples. Todos os três sistemas de topo combinados pesquisam menos documentos relevantes tendo maior precisão, o que sugere que o aumento de desempenho vem do aumento de documentos relevantes a baixas ordens. A perda no número de documentos relevantes pesquisados pode ser atribuído para a tendência da fórmula ROWRS pesquisar exclusivamente documentos do conjunto de resultados. Mesmo sem documentos relevantes pesquisados, ROWRS ultrapassa OWRS em relação às variações dos sistemas de topo.

A comparação das medidas de ordem, mostra que a medida sucesso/falha é superior à precisão ou eficiência baseadas nas fórmulas ROWRS. As variações dos sistemas de topo, as fórmulas ROWRS parecem trabalhar melhor com a contribuição do sistema topo (*olpboost*), em contraste com a fórmula OWRS a qual mostra melhores resultados sem qualquer ênfase em sistema de topo (st1 e st2), figura 5, lado direito.

As diferenças do efeito das fórmulas OWRS e ROWRS no sistema de topo indicam a relação entre ordem e relevância dos documentos de topo. A figura 5 compara a distribuição de documentos relevantes pesquisados pelo sistema de topo st* (i.e. o sistema com melhor desempenho) com sistemas sem sistemas de topo. O declive, representa a densidade de documentos relevantes pesquisados numa dada ordem, para ambos os sistemas indicando uma distribuição desigual de documentos relevantes sobre as ordens. O declive mais acentuado do sistema topo (*vlc10*), nas primeiras ordens, reflecte a maior concentração de documentos relevantes nos sistemas topo que fora destes, o que sugere que as contribuições dos sistemas de topo são benéficas.

Distribuições desiguais de documentos relevantes sobre ordens significa que os pesos baseados na ordem são mais efectivos que os pesos baseados no desempenho, evidênciado pelos melhores resultados da fórmula ROWRS sobre OWRS. Não é claro porque é que os sistemas de topo (st*) aumentem o desempenho quando

usados com pesos baseados na ordem e degradam o desempenho quando aplicados uniformemente sobre as ordens. É possível que nos sistemas de topo (st*) os pesos baseados nas ordens aumentem as contribuições dos sistemas de topo quando são mais benéficos, no entanto o aumento indiscriminado das contribuições de sistema de topo sobre todas as ordens pode ajudar a degradar o desempenho.

4 Conclusões

Foi feita uma analise dos resultados dos principias sistemas de pesquisa de informação (VSM, HITS e Classificação) num ambiente controlado, onde foi possível estabelecer um conjunto de métricas capazes de aferir e avaliar o desempenho dos sistemas de pesquisa. Foi estudado o potencial da combinação de métodos textuais, de ligações e de classificação com o objectivo de melhor o desempenho dos sistemas de pesquisa na *Web* usando a colecção de teste WT10g e informação dos directórios do Yahoo. Os melhores resultados de cada sistema foram combinados das mais variadas formas, explorando a combinação de parâmetros, sistemas e de fórmulas com base em medidas e ordem. Adicionalmente, os melhores sistemas foram combinados para explorar as variações das fórmulas baseadas na ordem. Do conjunto de experiências feitas resultaram cerca de 2000 sistemas de pesquisa diferentes. Como trabalhos futuros possíveis temos: (1) algoritmo dos sistemas HITS poderia ser melhorado com a 'poda' das ligações (Bharat e Henzinger, 1998) ou com o uso de âncoras (Chakrabarti et al., 1998b); (2) agrupamentos dos resultados dos sistemas HITS poderiam ser explorados para diferenciar entre comunidades centrais e secundárias; (3) verificação da suspeição da topologia incompleta das ligações da colecção WT10g; (4) completar a construção do dicionário baseado nas categorias e grupos centrais do Yahoo, deveria ser construído de uma forma mais selectiva orientada para os sistemas DC; (5) nos sistemas DC e TM, a selecção manual das melhores categorias deveria ser explorada; (6) pesquisa com base em diferentes sistemas de classificação de acordo com o assunto (tópico); (7) implementação de interfaces para utilizadores e disponibilização do sistema na Internet.

Referencias

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.

Chakrabarti S. Dom B. Raghavan P. Rajagopalan S. Gibson D. e Kleinberg J. (1998b). Automatic resource list compilation by analyzing hyperlink structure and associated text. Proceedings of the 7th Externosnational World Wide Web Conference.

Ferreira, J., Silva, A., Delgado, J. (2004), How to Improve Retrieval effectiveness on the Web, IADIS E-commerce 2004 Conference

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).

Kleinberg J. (1997). Authoritative sources in a hyperlinked environment. Proceeding of the 9th ACM-SIAM Symposium on Discrete Algorithms.

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.

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.

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

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

Shaw W. M. Jr. (1986). On the foundatin of evaluation. JASIS 37 346-348.

Thompson. P. (1990). A combination of expert opinion approach to probabilistic information retrieval part 1: The conceptual model. Information Processing e Management 26(3) 371-382.