Infraestrutura modular de teste para 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

É abordado o problema da pesquisa de informação, nos diferentes modelos existentes, dando ênfase à construção de uma plataforma modular flexível, capaz de testar de uma forma controlada diferentes modelos de pesquisa e assim contribuir para o desenvolvimento do tema, evitando assim a construção de diferentes sistemas de pesquisa que seria necessário desemvolver para testar os diferentes modelos.

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 à situação de excesso de informação, incapacitando o utilizador de obter a informação desejada e em tempo util, a qual se torna cada vez mais indispensável e crítica. Por esta e outras razões, torna-se relevante o estudo e desenvolvimento de mecânismos que dada uma necessidade conduzam à obtenção da informação desejada no mais curto espaço de tempo.

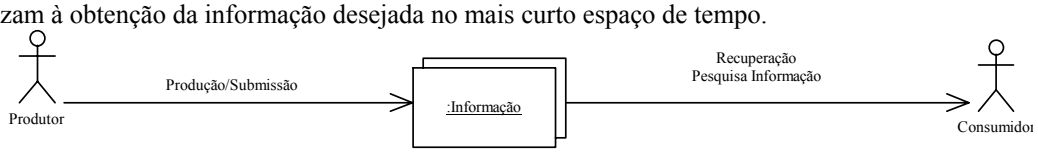


Figura 1.1: Ciclo de vida da informação.

Pretende examinar-se as técnicas habituais de pesquisa, verificar quando é que podem ser aplicadas, explorar novas técnicas e aproximações seleccionando o melhor de cada aproximação e demonstrar que existem ganhos de relevância da informação desejada ao seguirem-se aproximações combinadas de diferentes métodos. Diferentes métodos tipicamente, implicam a construção de diferentes sistemas de pesquisa. Dado este senário surge a necessidade de explorar o tema da construção de uma infraestrutura modular capaz de satisfazer os requesitos dos diferentes métodos. Para o efeito vai-se criar uma infraestrutura modular de teste elaborada pela integração de diferentes produtos, que permite validar de forma controlada conceitos \ técnicas na área da pesquisa de Informação na Internet. O presente artigo pretende explorar o tema da construção de sistemas de pesquisa modulares que possam servir de base à investigação das técnicas de pesquisa de informação evitando a actual proliferação de sistemas de pesqusia. É um tema importante, havendo poucos trabalhos de sistemas modulares de teste desenvolvidos (1). O presente trabalho, encontrando-se divido em 5 capítulos: Primeiro, fazemos a introdução ao problema discutido no artigo. Segundo sistemas de pesquisa, individuais, modelo vectorial, seguimento de ligações e classificação automática de documentos. Terceiro, descrição da estrutura modular, que vai permitir um conjunto de resultados na área da pesquisa de informação. Quarto, avaliação dos resultados. Quinto e último, as conclusões.

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

2 Pesquisa de Informação

Os sistemas de pesquisa tradicionais, de uma forma geral, baseiam-se na comparação por processos pré-estabelecidos, dos termos representativos das necessidades de informação com os termos representativos de cada documento. Desta comparação resulta um conjunto de documentos (habitualmente ordenados), que o sistema considera relevantes para a satisfação dos interesses de informação do utilizador.

Os sistemas de pesquisa são caracterizados pelo esquema de blocos representado na figura 2. Existe um repositório de informação onde são guardados os documentos nos mais variados formatos constituindo um espaço heterogéneo de pesquisa. O conteúdo deste espaço é indexado de forma a criar um espaço de menor dimensão representativo do inicial (I-Índices) onde se farão as pesquisas de acordo com os métodos em questão. As necessidades de informação são habitualmente expressas por um conjunto de termos que o sistema manipula convenientemente para chegar a um conjunto de termos representativos das necessidades de informação (P-Pergunta). Da comparação entre estes dois representativos resulta um conjunto de documentos, que o sistema identifica como relevantes. Dos documentos que o sistema mostra como relevantes o utilizador escolhe quais pretende consultar na base de dados dos documentos disponíveis.

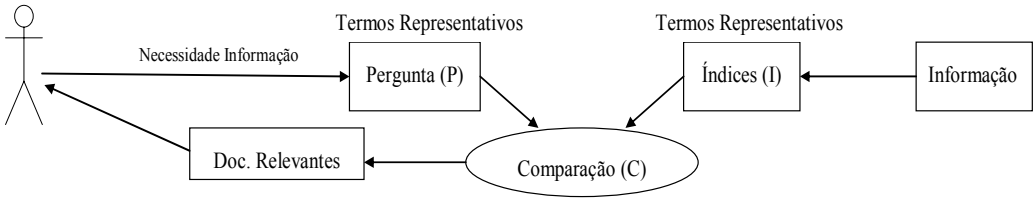


Figura 2: Serviço de recuperação de informação na sua forma mais simples.

Dada a complexidade do problema, são acrescentados mecânismos adicionais como forma de melhorar os resultados, nomeadamente mecânismos de:

Expansão e normalização dos termos introduzidos pelos utilizadores; Normalização geral de termos, usando sistemas de classificação;Retroacção automática e do utilizador face aos resultados; Seguimento das ligações dos documentos; Uso de sistemas de classificação; Combinação de diferentes métodos.

Os principais sistemas de pesquisa são os modelos: **vectorial** [2], de **ligações** [3] e de **classificação** (ver §3).

Modelo Vectorial

O modelo vectorial representa as necessidades de informação e representativos de documentos num espaço n dimensional usando como medida de comparação o produto interno dos vectores. Os principais parâmetros a testar podem ser:

- Fórmulas para calcular os pesos, os representativos dos documentos e a pergunta bem como a fórmula usada para os comparar. (Neste trabalho usamos a medida lnu [4]). Outras medidas podem ser facilmente implementadas (inclusive o modelo booleano).
- Uso de frases e retroação.
- Expansão dos termos das perguntas com base na informação existente nos documentos considerados relevantes pelo sistema.
- Índices construídos a partir dos títulos, documento completo, corpo dos documentos.
- Comprimento das perguntas (pequenas, médias e longas).
- Normalização de termos através de sistemas de classificação.

Dos parâmetros acima referidos podemos ter 1(formula medida)x2(retroação)x2(expansão perguntas) x3(índices)x3(perguntas)x2(normalização)=72 sistemas possíveis, sendo este número multiplicado pelas diferentes forma de medidas que se queira introduzir.

Utilizador

Modelo de Ligações

O modelo de ligações começa por identificar conjuntos semente através do modelo vectorial. Destes conjuntos são calculadas as medidas de hub e autoridades através do algoritmo de HITs modificado, expressas nas fórmulas (1) e (2)

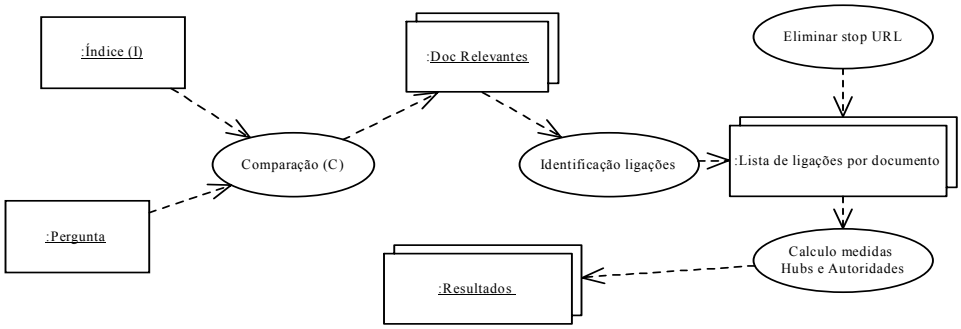
$$a(p) = \sum_{q \to p} h(q) \times auth\_wt(q, p) \quad (1)$$

$$h(p) = \sum_{p \to q} a(q) \times hub\_wt(p, q) \quad (2)$$

onde *auth_wt(q p)* é 1/*m* para a página *q* cujo endereço tem *m* documentos apontando para *p* e *hub_wt(p q)* é 1/*n* para a página *q* a qual é apontada por *n* documentos do endereço de *p*.

Figura 3: Identificação dos módulos principais do sistema de seguimento das ligações.

A definição de endereço foi criada a partir do URL do documento cortando-o na primeira ocorrência da marca da barra de divisão (i.e. '/') (endereço pequeno) e a forma longa até a última ocorrência da barra de



divisão (endereço longo).

Os parâmetros principais do sistema consideram-se a escolha do método para originar o conjunto semente, a definição de endereços e a forma de calcular as medidas de hubs e autoridades. O número de sistemas possíveis é 2x3=6.

Classificação

Na classificação de documentos, faz-se a comparação entre as necessidades de informação e os documentos através de um sistema de classificação (hiearquia de conhecimento previamente definida). Os sistemas de classicação têm um conjunto de termos em cada categoria, os quais podem ser usados, para expandir e normalizar os termos das perguntas ou então catalogar documentos de forma automática (i.e., identificação dos documemtos com maiores semelhanças aos termos de cada categoria). Na catalogação automática o utilizador pode navegar no espaço classificado (com interface apropriada) identificando os temas que acha que satisfazem as suas necessidades de informação, tal como se percebe pela representação esquemática apresentada na figura 4.

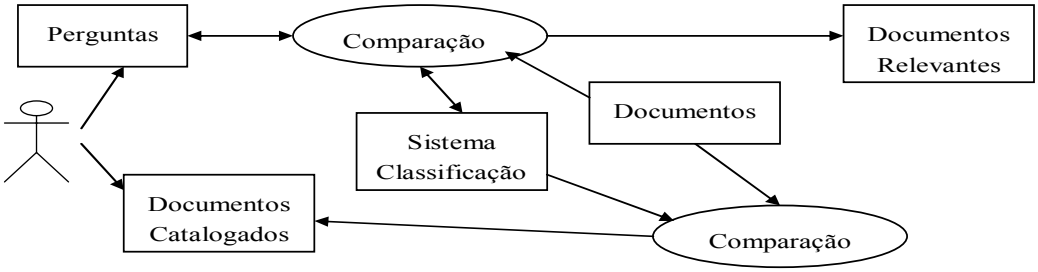


Figura 4:Pesquisa de informação através Sistemas de Classificação.

Combinação

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).

A **fórmula de combinação de semelhanças** [5] 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** [6,7,8 e 9] 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 das componentes de cada sistema as quais tendem a medir as semelhanças das perguntas e dos documentos, numa escala ordenada. A fórmula das somas pesadas tem as seguintes variações:

1-tem em conta a sobreposição das medidas, 2-ordem pelo qual um documento é pesquisado, dar amaior importância ao sistema(s) de topo, originando maiores número de combinações possíveis.

3 Requisitos para a infraestrutura de teste

O objectivo é a contrução de um sistema modular que permita testar de forma controlada diferentes metodologias de pesquisa de informação. As componentes básicas de um sistema de pesquisa organizam-se de acordo com o diagrama de blocos apresentado na figura 5:

1. Informação (colecção de documentos) obtida por motores de pesquisa ou colecções elaboradas para o efeito (i.e., WT10g, GOV [10].
2. Pergunta (necessidades de informação). Interface para o utilizador formular a sua necessidade de informação e receber a retroação. Pode ser adicionado um *speller* (Jspell) de modo a corrigir eventuais erros ortográficos.
3. Tsearch2 [11], modulo a integrar no openfts de modo a permitir a indexação total dos documentos e interação com a base de dados Postgresql.
4. Toda a informação é guardada numa base de dados Postgresql [12]. Para permitir diferentes formas de pesquisa, são guardados diferentes tipos de dados. Índices de Títulos dos documentos, frases, documento sem título, documento geral, bem como endereços.
5. Algorimtos de *Information Retrieval* (IR). Toda a variedade de algoritmos pode ser integrada no sistema de forma modular tais como: diferentes formas de calcular o produto interno; determinação de frases com ajuda ou não de dicionários; redução das palavras à sua forma base, disponível em varias línguas com modulo Snowball [13]. No caso da língua Inglesa é aplicado o algoritmo de Porter [14], havendo também disponiveis algoritmos para a língua Portuguesa; mecânismos de retroação automática [4].
6. Fórmulas de combinação, são implementadas diferentes medidas de combinação tendo como base os resultados obtidos.
7. Classificação de documentos (DC e TM). Estes métodos encontram-se definidos na secção 3.1 e esquematizados na figura 6.
8. Cálculo das medidas de hubs e autoridades, tendo sido exploradas as fórmulas (1) e (2).
9. *OpenFts*, os documentos obtidos são formatados de modo ao modulo openfts [15] os puder indexar. Neste processo são removidas *stop words* de acordo com uma lista previamente difinida
10. Resultados, o sistema devolve uma lista ordenada de documentos de acordo com as métricas definidas no modelo de comparação.
11. Sistema de classificação. Podem ser implementados aproveitando os existentes em áreas específicas (exemplo, ACM na área da ciência dos computadores) ou criado a partir de outro existente.
12. Avaliação dos resultados. É feita com base nas medidas de precisão e cobertura. Para a determinação desta e de outras medidas é necessário conhecer o conjunto de documentos relevantes para uma dada pergunta.

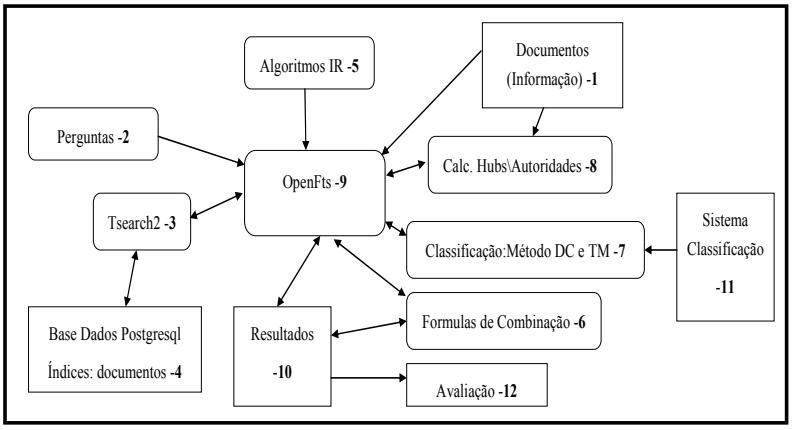


Figura 5: Estrutura modular do sistema de pesquisa.

Classificação de documentos

Criamos um sistema de classificação com base na informação do Yahoo <http://dir.yahoo.com>, no Inverno de 2002. Para além de reproduzir uma versão simplificada local do Yahoo (o site completo do Yahoo (i.e. Estrutura das páginas Web e directorias) foi reproduzido numa máquina local. Apenas a classificação da informação foi mantida em cada página do Yahoo), o motor de pesquisa criou ficheiros de mapa de endereços, em cada uma das categorias principais do Yahoo para encapsular a informação classificada. Os ficheiros de mapa de sites contêm essencialmente uma classificação hierárquica de categorias de termos, títulos de endereços e descrições da pergunta em causa.

Foram elaboradas duas aproximações baseadas em: estimativa da probabilidade da associação termo-categoria; baseada na semelhança entre termos da pergunta e da categoria. A principal distinção entre os dois métodos reside na forma de classificar informação do Yahoo que é influenciada pela forma de encontrar as melhores semelhanças entre uma categoria e uma pergunta.

O primeiro método, e feito com base num **dicionário de classificação** (DC) o qual mede a semelhança de termos das perguntas e as categorias com uma probabilidade de associação. Este método, primeiro ordena categorias em relação à pergunta usando os pesos de associação dos termos-categorias do DC e depois ordena os documentos em relação ao centro da classe que representa a melhor categoria. Os parâmetros que se podem testar nos sistemas DC são:

- Categorias dos termos universais. Existem cinco variações do universo de termos (i.e. termos que descrevem ou pertencem a uma dada categoria) que são usados para construir DC e centros de classes.
- Número de categorias de topo (1 ou 3)
- Comprimento das perguntas (pequenas, médias, longas)
- Termos indexados WT10g, são combinações de termos/frases com corpo/título/documento, resultam 6 combinações de termos
- Retroacção.

Assim vem 5x2x3x6x2=360 sistemas possíveis.

O segundo método de classificação, *Term Match Method* (TM) produz uma lista ordenada de categorias para verificar as semelhanças dos termos das perguntas com os termos dos ficheiros de mapas do Yahoo (i.e. etiquetas de categorias, de títulos e descrições de endereços do Yahoo; estes ficheiros de endereços de sites são concatenados e removidos da mesma forma que as perguntas; URLs no ficheiro de mapa de endereços são deixados intactos). Desta operação resulta um conjunto de nós semelhantes numa classificação hierárquica e uma lista ordenada de categorias. A segunda fase do método de TM é comparável ao método DC, no qual o vector da pergunta é expandido (centro da classe no método de TM) construindo as melhores categorias e produzindo uma lista ordenada de documentos da colecção WT10g. Uma das diferenças entre os dois processos de semelhança categoria-documento baseia-se na criação de termos expandidos da pergunta:

O centro da classe no método DC é um vector com todos os termos da categoria expândidos com peso *ltc*. No método TM o centro da classe é um termo do vector da categoria seleccionada, normalizado.

Os parâmetros testados são:

- Número de categorias de topo usadas na colecção {1,2,3}.
- Índices de termos WT10g (frases e termos de títulos de documentos, corpo de documentos e documentos completos).
- Uso de pseudo-retroação.

Temos assim 3x6x2=36 sistemas.

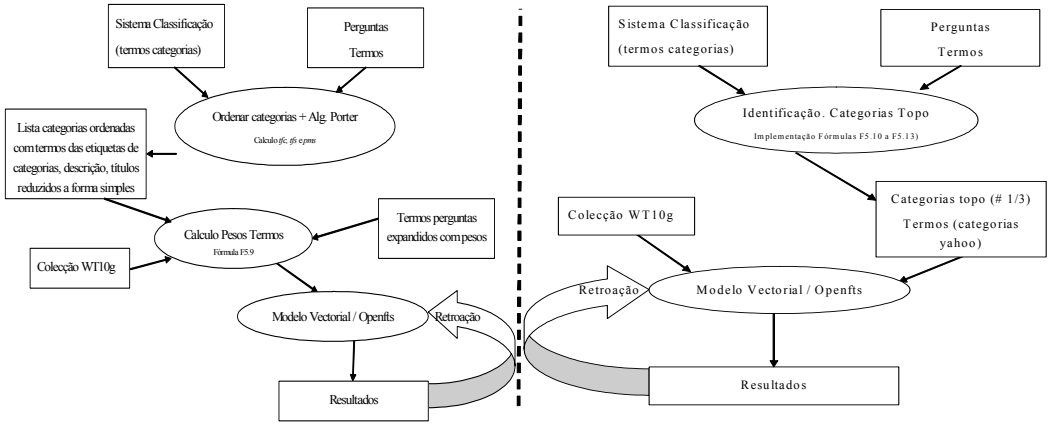


Figura 6: Método de classificação DC (esquerda) e TM (direita).

4 Análise dos Testes

Assim iremos focar a melhor forma de avaliar os resultados. Dos diversos parâmetros de cada sistema resultam 72 sistemas VSM, 6 HITS, 360 DC e 36 TM. Ou seja 474 sistemas, os quais podem ser combinados de 6 maneiras (6 variantes das formulas de combinações). Assim temos 474x6=2844 sistemas possíveis. Esta infraestrutura deu origem a um conjunto de resultados que se encontram descritos em [4].

O problema da avaliação de resultados na Web, tem como principal dificuldade conhecer o número de documentos relevantes num determinado tópico, daí a necessidade de avaliar resultados numa colecção controlada (como a WT10g).

A avaliação é feita com base nas técnicas habituais de medida de desempenho como a cobertura, precisão, R-precisão e precisão média. Precisão é importante em altas ordens porque os utilizadores típicamente avaliam os primeiros 10 a 20 resultados [16, 17, 18 e 19]. R-Precisão compensa a falta de sensibilidade da precisão para o tamanho da colecção de documentos relevantes a ordem R para uma determinada pergunta.

A precisão média é a soma de todas as precisões nas ordens com documentos relevantes pesquisados, divididos pelo número total de documentos relevantes, sendo uma medida simples que reflete o desempenho do sistema para todos os documentos relevantes. O óptimo *F* é o máximo do valor *F* (3) sobre todas as ordens e é outro valor simples que mede o desempenho do sistema tendo em conta a precisão e a cobertura. *F* é o coeficiente de Dice de semelhanças para um conjunto de documentos relevantes pesquisados em relação a uma determinada pergunta que aumenta o valor da precisão tendo em consideração o valor de cobertura numa dada ordem [20] dado pela equação (3):

$$F = \frac{2r}{n + N_r} = \frac{2}{1/R + 1/P} \qquad (3)$$

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*).

A precisão interpolada em 11 pontos padrão ({0, 0.1, 0.2, …, 1.0}) é calculada num gráfico de cobertura-precisão onde indicadores de desempenho do sistema podem ser visualizados e comparados.

Para investigar o significado da sobreposição na combinação (i.e. o número de métodos que pesquisam um determinado documento) foi feita uma análise de sobreposição de resultados combinados onde se analisaram os resultados das partições sobrepostas (i.e. documentos pesquisados pelo método *i* apenas, documentos pesquisados por nenhum dos métodos *i* e *j*, documentos pesquisados por todos os métodos) e a granulidade do grau de documentos relevantes foi feita. Como os documentos relevantes pesquisados por diferentes fontes de evidência têm maior probalidade de serem relevantes [21, 22, 23, 24 e 25] é de esperar mais documentos relevantes nas partições sobrepostas.

Resultados obtidos podem ser consultados em [4], no entanto podemos mostrar os melhores resultados dos sistemas simples, figura 7. O melhor sistema VSM medido pela precisão média foi *vlc00* (perguntas grandes – corpo com frases textuais – sem retroação. O melhor sistema HITS *hpl* (pequeno endereço conjunto semente do sistema *vlc00*). O melhor sistema DC foi dc13dm0 (uma categoria de topo, descrição endereço, frases do documento, perguntas médias sem retroação). O melhor sistema TM *t121* (uma categoria de topo, frases do corpo do documento comretroação). 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, HITS e DC indicando a vantagem do método textual VSM sobre os outros.

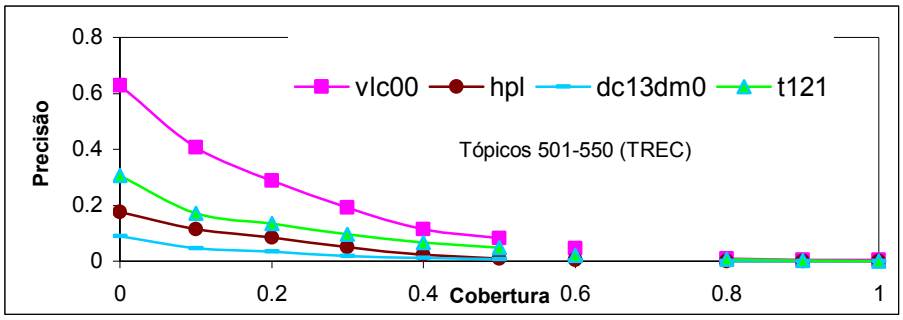


Figura 7: Curvas precisão-cobertura para sistemas simples de pesquisa para as perguntas tópicos 501-550.

5 Conclusões

Pretendeu-se descrever uma infraestrutura modular para testar os sistemas e métodos de pesquisa na Web, bem como a discussão conceptual dos diferentes sistemas de pesquisa e a forma de os implementar. Descreveu-se a forma de implementar na mesma plataforma sistemas de pesquisa baseados em modelos vectorial, de ligação e de classificação, bem como a forma de combinar resultados de modo a descubrir os parâmetros mais importantes da pesquisa. O sistema devido a sua estrutura modular permite adaptar diferentes métodos de pesquisa, permitindo construir sistemas de recuperação de informação nas áreas da filtragem, pesquisa e classificação. A partir deste sistema foram criados outros, nas áreas da [26]:

- Filtragem de informação: (1) MyNewsPaper –Jornal personalizado; (2) MyTv – Serviço de Filtragem sobre programas da TvCabo Portugal; (3) MyEnterpriseNews – Serviço de identificação de informação na Web sobre determinado produto ou empresa.
- Classificação de infornação; (1) MyDocument – Serviço de gestão empresarial de documentos; (2) MyClassificator – Serviço de classificação de informação.
- Pesquisa de informação, foram construídos cerca de 1900 sistemas de pesquisa, variando parâmetros e combinanado sistemas.

Facto importante a realçar do conjunto de experiências realizadas é a existência de um ambiente controlado de forma a puder obter métricas do desempenho dos sistemas criados. O presente trabalho encontra-se a ser completado com um conjunto de interfaces que permitam disponibilizar o serviço on-line.

6 Referências

[1] IRIS -<http://www.ils.unc.edu/iris/>

[2] Salton G. (1968). The evaluation of computer-based retrieval systems. Automatic Information Organization and Retrieval (pp. 361-349). New York: McGraw-Hill.

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

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

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

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

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

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

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

[10] Colecção Wt10g - http://es.cmis.csiro.au/TRECWeb/ access_to_ data.html.

[11] Tsearch - www.sai.msu.su/~megera/ postgres/gist/tsearch/V2/

[12] Base de dados, Postgresql - www.postgresql.org

[13] Snowball - snowball.tartarus.org/

[14] Porter M. (1980). An algorithm for suffix stripping. Program 14 130-137.

[15] OpenFts - http://openfts.sourceforge.net/

[16] Jansen M. B. Spink A. e Saracevic T. (1998). Failure analysis in query construction: data and analysis from a large sample of Web queries; Proceedings of the third ACM Conference on Digital libraries 289-290.

[17] Pollock, A., & Hockley, A. (1997). What's wrong with Internet searching? D-Lib Magazine [On-line]. http://www.dlib.org/dlib/march97/bt/03pollock.html

[18] Jansen, M. B., Spink, A., Bateman, J., & Saracevic, T. (1998). Real life information retrieval: a study of user queries on the Web. SIGIR Forum, 32(1).

[19] Silverstein C. Henzinger M. Marais H. e Moricz M. (1998). Analysis of a very large AltaVista query log. Technical Report 1998-014 COMPAQ System Research Center

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

[21] Belkin, N. J., Cool. C., Croft, W. B., & Callan, J. P. (1993). The effect of multiple query representations on information retrieval system performance. Proceedings of ACM SIGIR, 339-346.

[22] Katzer, J., McGill, M. J., Tessier, J. A., Frakes, W., & DasGupta, P. (1982). A study of the overlap among document representations. Information Technology: Research and Development, 1, 261-274.

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

[24] Saracevic, T., & Kantor, P. (1988). A study of information seeking and retrieving. III. Searchers, searches, overlap. Journal of American Society for Information Science, 39, 197-216.

[25] Turtle H., & Croft, W. B. (1991). Evaluation of an inference network-based retrieval model. ACM Transactions on Information Systems, 9, 187-222.

[26] Ferreira, J., Silva, A., Delgado, J. (2004), MyWebsearch: How to build Classification, Filtering and Retrieval Systems in the same experimental platform. Submited to ECIR05 (27th European Conference on Information Retrieval, 21-23 Março, Santiago Compostela, Espanha).