

**Avaliação de Sistemas de
Recuperação de Informação da
Web em Português: Uma
Proposta Inicial à Comunidade**

Comunicação do grupo XLDB do LASIGE ao

Avalon'03

Mário J. Silva

Bruno Martins

Miguel Costa

DI-FCUL

TR-03-11

7 de Julho de 2003

Departamento de Informática
Faculdade de Ciências da Universidade de Lisboa
Campo Grande, 1749-016 Lisboa
Portugal

Technical reports are available at <http://www.di.fc.ul.pt/tech-reports>. The files are stored in PDF, with the report number as filename. Alternatively, reports are available by post from the above address.

Avaliação de Sistemas de Recuperação de Informação da Web em Português: Uma Proposta Inicial à Comunidade

Comunicação do grupo XLDB do LASIGE ao Avalon'03

Mário J. Silva, Bruno Martins e Miguel Costa
Faculdade de Ciências da Universidade de Lisboa,
1749-016 Lisboa, Portugal

mjs@di.fc.ul.pt, bmartins@xldb.fc.ul.pt, mcosta@xldb.fc.ul.pt

7 de Julho de 2003

Resumo

Apresentamos uma colecção de documentos extraídos de um conjunto de domínios da Web e uma proposta inicial de uma tarefa de avaliação de motores de busca sobre essa colecção. A proposta baseia-se na abordagem experimentada na Webtrack da TREC. Assenta na definição de uma tarefa inicial de pesquisa de home pages de instituições e páginas pessoais contidas na colecção. A colecção que nos propomos fornecer, designada por TumbaGOVPT, inclui cerca de 1 milhão de documentos e corresponde aos conteúdos recolhidos de uma lista de domínios referenciados como pertencentes a instituições da administração pública central portuguesa.

1 Introdução

Para definir um processo de avaliação para sistemas de recuperação de informação é necessário um grande esforço organizativo, tanto para reunir as partes interessadas como para construir os recursos utilizados durante a avaliação [1, 2]. É também essencial definir o que se deseja medir e, consequentemente, as métricas a utilizar e a metodologia comparativa baseada nessas métricas (como as medições serão feitas, se a forma de julgamento será manual, automática ou uma mistura de ambas, etc).

A nossa proposta destina-se a avaliar sistemas de recuperação de informação (RI) da Web. Propomos uma colecção de documentos extraídos de um conjunto de domínios da Web, denominada TumbaGOVPT e uma tarefa inicial de avaliação sobre esta colecção.

O nosso interesse na avaliação destes sistemas deriva do trabalho de investigação em sistemas de informação Web, enquanto membros da equipa de desenvolvimento e operação do motor de busca tumba! [8]. Além de podermos contribuir para a avaliação conjunta com colecções de documentos recolhidos da Web, temos também interesse em avaliar o nosso sistema nas várias tarefas em que possa participar. A avaliação conjunta permitir-nos-á recorrer a padrões de avaliação referenciados externamente e, por essa via, melhorar a qualidade das avaliações que realizarmos.

Uma das motivações para realizar o trabalho aqui descrito e fazer esta proposta deve-se a que a TREC (Text Retrieval Conference), a conferência/actividade de avaliação mais importante neste domínio, se centra na avaliação de sistemas na língua inglesa. Existem quanto a nós diferenças substanciais no processamento computacional da língua portuguesa que motivam a sua avaliação tendo em consideração estas diferenças, nomeadamente no que respeita aos acentos, cedilhas e formas reflexivas dos verbos.

Há no entanto todo um conjunto de ideias sobre avaliação em RI da TREC e uma larga experiência na sua condução que podem e devem ser levadas em conta. A metodologia seguida é, na nossa opinião, completamente reutilizável até prova em contrário e deve ser na medida do possível reaproveitada.

2 Avaliação de sistemas de recuperação de informação

Na RI textual, as duas medidas mais utilizadas para avaliação são a precisão e a lembrança (*precision* e *recall* na língua inglesa) [7]. Estas medidas são baseadas na noção de documentos relevantes, de acordo com uma determinada necessidade de informação. A lembrança mede a proporção de documentos relevantes de uma colecção que foram recuperados e a precisão a proporção dos documentos recuperados nas pesquisas que são relevantes.

Para comparar diferentes configurações de sistemas de RI, a precisão e a lembrança são calculadas usando uma colecção de consultas, documentos e julgamentos de relevância conhecidos. Outras medidas utilizadas são a medida F, a medida E e o fallout [7]. A decisão de quais as medidas utilizar numa avaliação depende da aplicação, havendo sempre discussões sobre a confiabilidade de tais medidas [10]. Um exemplo é o artigo de Gwizdka &

Chignell [4], onde se discute como avaliar motores de busca. Não é claro, por exemplo, até que ponto pequenas diferenças na precisão e lembrança têm algum efeito perceptível na usabilidade dos sistemas.

A TREC (Text Retrieval Conference) é uma conferência de avaliação de RI na forma textual. Consiste numa série de workshops cujo âmbito é a avaliação em larga escala de tecnologias de RI, permitindo comparar as diversas técnicas utilizadas pelos grupos participantes. Para cada tarefa (track) há uma base de documentos com cerca de 2 gigabytes de texto (entre um milhão e um milhão e meio de documentos) e algumas consultas que estabelecem o que é a informação procurada e o que constitui um documento relevante.

Existem várias críticas a este tipo de avaliação, relacionadas com i) a sua utilização em ambientes de “laboratório” e não em ambientes reais, ii) a credibilidade a dar aos julgamentos de relevância (já que este é um conceito subjectivo [12]), e iii) a representatividade do conjunto de consultas e de documentos (costumam ser voltados para tópicos de ciência e tecnologia). No entanto, e apesar destas questões, este tipo de avaliação tem sido proveitoso para RI, já que tem indicado como a melhorar através de novas técnicas [5].

2.1 Webtrack

A nossa proposta inicial de avaliação para máquinas de pesquisa operando sobre a língua portuguesa baseia-se na abordagem usada na Webtrack da TREC ¹.

Uma das tarefas centrais da Webtrack, denominada “the home/named page finding task”, consiste em encontrar homepages de instituições e outras páginas web, dado um nome descritivo. A ideia de fazer a avaliação desta forma relaciona-se com o facto de muitas vezes os utilizadores destes sistemas pesquisarem uma página pelo seu nome. Nestes casos, um sistema de recuperação eficaz iria retornar a página que o utilizador procura nos primeiros resultados. Por exemplo, quando um utilizador procura na Web a entidade “Faculdade de Ciências da Universidade de Lisboa”, ao submeter este conjunto de termos como interrogação a um sistema de RI, espera obter a página de entrada da instituição em primeiro lugar numa lista de resultados.

A avaliação consiste em definir à partida uma amostra representativa de nomes de instituições e home pages pessoais e associar a cada uma dessas instituições a sua página “oficial”. A qualidade relativa dos sistemas pode ser medida através da ordem com que os resultados esperados são apresentados. Embora bastante simples, esta tarefa tem a vantagem de ser relativamente

¹http://es.cmis.csiro.au/TRECWeb/guidelines_2003.html

Número de Documentos	1068552
Tamanho da colecção	25720 Mb
Tamanho da colecção (Texto Filtrado)	3750 Mb
Número de documentos HTML	994784
Número de documentos PDF	40237
Número de documentos WORD	8974
Número Total Hyperlinks entre Sites	70404
Tamanho Médio de um Documento HTML	15 Kb
Tamanho Médio de um Documento (Texto Filtrado)	3.7 Kb

Tabela 1: Estatísticas da Recolha TumbaGOVPT

fácil de planear e ser realizável com poucos recursos. Contudo, apesar da simplicidade, pode fornecer pistas bastante úteis para que os investigadores possam melhorar os seus sistemas.

3 A colecção de dados

O corpus proposto como base para a tarefa de avaliação (TumbaGOVPT) corresponde a uma recolha dos conteúdos dos domínios da administração pública central portuguesa (inclui universidades públicas e institutos de investigação do estado). Esta recolha é da ordem de 1 milhão de URLs, sensivelmente 25 Gb. A Tabela 1 sumariza algumas das características mais importantes desta colecção. A recolha é recente (foi efectuada na Primavera de 2003) e representa o tipo de informação que um serviço real de pesquisa nesta colecção de domínios iria utilizar.

Propomo-nos a dar acesso a um “dump” desta colecção, num esquema legal semelhante ao mecanismo de cedência dos conteúdos do domínio .GOV americano usado pelos promotores da Webtrack da TREC. Desta forma, quem possua um sistema de RI pode indexar estes documentos e executar a avaliação, tendo a certeza que os documentos mencionados na tarefa foram indexados pelo seu sistema. Esta parece-nos a melhor forma de garantir um teste padrão que possa efectivamente ser utilizado para comparar sistemas distintos, uma vez que será usada sempre a mesma fonte de informação nas várias experiências.

4 A Tarefa de Avaliação

As páginas de referência a usar na tarefa inicial de avaliação que propomos

são as da recolha TumbaGOVPT .

Tal como exposto atrás, propomos uma tarefa de avaliação para máquinas de pesquisa seguindo os moldes da Webtrack da TREC, nomeadamente a tarefa “home/named page finding”. A tarefa consiste em encontrar home-pages de instituições públicas e páginas pessoais presentes na colecção de documentos web TumbaGOVPT .

Para medir o desempenho dos sistemas em tarefas deste tipo elaborámos um conjunto misto de 30 URLs representando tanto páginas pessoais como de instituições públicas. Verificámos manualmente que todas estas páginas de entrada estão catalogadas na colecçãoTumbaGOVPT. As 30 instituições e páginas pessoais seleccionadas encontram-se listadas nas Tabelas 2 e 3. Em algumas situações, observámos que o conteúdo da mesma página se encontra disponível em vários URLs diferentes. Nesse caso, todos eles devem ser considerados correctos e, por essa razão, são também listados.

Embora pouco extenso (e achamos que assim se deva manter, de modo a que o processo de avaliação possa ser conduzido por uma apenas pessoa num curto período de tempo), tentámos colocar no conjunto páginas que reflectissem diversos problemas a ser tratados pelo sistema, nomeadamente páginas com inexistência de meta-informação, vários graus de qualidade do código HTML, como inexistência de títulos descritivos nas páginas, etc. Desta forma, e uma vez que a cada URL se encontra associado uma dificuldade específica, torna-se possível à entidade que desenvolveu um sistema, ao detectar um mau resultado na avaliação para uma dada interrogação, procurar a sua possível causa.

4.1 Pontuação

Para análise comparativa dos sistemas avaliados, será atribuída uma pontuação com base na ordem da primeira resposta correcta a cada interrogação na lista de resultados. Para tal, quem efectua a tarefa deverá submeter ao sistema o conjunto de termos correspondente ao nome de cada entidade no conjunto de URLs. Posteriormente, deverá ser anotado em que posição nos primeiros vinte resultados surge o primeiro URL correspondente – um valor numérico de 1 a 20, ou 21 caso o sistema falhe em encontrar a página pretendida nos primeiros 20 resultados.

Restringimos a avaliação aos primeiros 20 resultados devido ao facto de estudos anteriores [9] comprovarem que os utilizadores raramente vêm mais que as duas primeiras páginas de resultados retornadas. Desta forma, são penalizados os sistemas que não satisfazem este critério.

Uma vez pontuada cada interrogação, é atribuída uma pontuação ao sistema, a qual corresponde à soma dos resultados para as várias pesquisas.

Páginas Institucionais	URL
Segurança Social	www.seg-social.pt/ www.seg-social.pt
Procuradoria Geral Distrital de Lisboa Universidade da Madeira	www.pgdlisboa.pt/ www.uma.pt www.uma.pt/ www.uma.pt/index.html
Comissão Nacional de Protecção de Dados	www.cnpd.pt www.cnpd.pt/ www.cnpd.pt/index.htm
Faculdade de Letras da Universidade de Coimbra	www.fl.uc.pt www.fl.uc.pt/ www.uc.pt/FLUC/ www.uc.pt/fluc/
Conselho de Reitores das Universidades Portuguesas	www.crup.pt www.crup.pt/ www.crup.pt/index.htm
Parlamento	www.parlamento.pt www.parlamento.pt/ www.assembleiadarepublica.pt www.assembleiadarepublica.pt/ www.assembleiadarepublica.pt/index.html
Plano para a Eliminação da Exploração do Trabalho Infantil	www.peeti.idict.gov.pt www.peeti.idict.gov.pt/ www.peeti.idict.gov.pt/default.htm
Centro Cultural de Belém	www.ccb.pt
Universidade de Coimbra	www.uc.pt www.uc.pt/ www.uc.pt/index.html
Ministério da Economia	www.min-economia.pt www.min-economia.pt/ www.min-economia.pt/index.html
Biblioteca Nacional	www.bn.pt www.biblioteca-nacional.pt www.bn.pt/ www.bn.pt/default.asp
Escola Superior de Educação de Setúbal	www.es.e.ips.pt www.es.e.ips.pt/ www.es.e.ips.pt/index.html
Departamento de Informática da Faculdade de Ciências da Universidade de Lisboa	www.di.fc.ul.pt www.di.fc.ul.pt/ www.di.fc.ul.pt/entrada.html www.di.fc.ul.pt/index.html printers.di.fc.ul.pt printers.di.fc.ul.pt/ printers.di.fc.ul.pt/entrada.html webwww.di.fc.ul.pt webwww.di.fc.ul.pt/ webwww.di.fc.ul.pt/entrada.html
Instituto do Consumidor	www.ic.pt www.ic.pt/ www.ic.pt/index.html www.consumidores.pt
Ministério da Cultura	www.min-cultura.pt/ www.min-cultura.pt/ www.min-cultura.pt/index.html
Reitoria da Universidade de Lisboa	www.ul.pt/reitoria.html
Laboratório Nacional de Engenharia Civil	www-ext.lnec.pt www-ext.lnec.pt/ www-ext.lnec.pt/index.phtml
Instituto de Desenvolvimento e Inspeção das Condições de Trabalho	www.idict.gov.pt www.idict.gov.pt/ www.idict.gov.pt/default.asp
Escola Superior de Tecnologia de Setúbal	www.est.ips.pt www.est.ips.pt/

Tabela 2: Lista de páginas de entrada de instituições públicas propostas para a tarefa de avaliação.

Páginas Pessoais	URL
Jorge Pacheco	aloof.cii.fc.ul.pt/Aloof/ aloof.cii.fc.ul.pt/Aloof/aloof.html aloof.cii.fc.ul.pt/Aloof/main.html
Paulo Veríssimo	www.di.fc.ul.pt/~pjuv/
Pedro Veiga	www.di.fc.ul.pt/~pmv/
Mário Gaspar da Silva	xldb.fc.ul.pt/mjs/ xldb.di.fc.ul.pt/mjs/
Daniel Coelho Gomes	xldb.di.fc.ul.pt/daniel/ xldb.di.fc.ul.pt/daniel/index.html xldb.fc.ul.pt/daniel/index.html xldb.fc.ul.pt/daniel/
Helder Manuel Coelho	www.di.fc.ul.pt/~hcoelho/
Paulo Quaresma	www.di.uevora.pt/~pq/ www.di.uevora.pt/~pq/index.html host.di.uevora.pt/~pq/ host.di.uevora.pt/~pq/index.html
José Félix Costa	fgc.math.ist.utl.pt/jfc.htm
Paulo Alexandre Marques	www.deetc.isel.ipl.pt/analisedesinai/Pessoais/PauloMarques
José Paulo Pimentel de Castro Coelho	agricultura.isa.utl.pt/agricultura/docentes/pimentel.asp

Tabela 3: Lista de páginas pessoais propostas para a tarefa de avaliação.

Esta métrica possui a vantagem de dar uma penalização linear com a distância dos resultados obtidos ao resultado óptimo, podendo ainda ser utilizada posteriormente para comparar o mesmo sistema ao longo de experiências de avaliação sucessivas.

5 Considerações sobre a Tarefa de Avaliação

As páginas da colecção TumbaGOVPT foram por nós indexadas e podem ser procuradas usando o software do tumba!, estando acessíveis presentemente no URL <http://194.117.20.250>. Verificámos também que todas as páginas incluídas na lista proposta para a experiência se encontram também indexadas presentemente pelo motor de busca Google, sendo assim possível usar a metodologia e conjunto de pesquisas proposto para comparar sistemas de RI sobre o Português como o tumba!, o Google ou outros motores de busca disponíveis para utilização pública.

Tal como sublinhado anteriormente, esta é apenas uma ideia inicial para a realização de experiências de avaliação em RI sobre o Português, podendo ser estendida consoante o que for entendido como mais apropriado. Os participantes poderão explorar outras questões relacionadas com a recuperação de informação particulares dos seus sistemas, nomeadamente o tratamento de pesquisas com erros ortográficos. Contudo, para que os resultados possam efectivamente ser usados em comparações entre sistemas, cremos que na tarefa de avaliação não deverão ser incluídos mecanismos de expansão e modificação manual ou interactiva de interrogações. Parece-nos porém não fazer sentido haver qualquer restrição relacionada com a indexação dos documentos. A indexação total ou parcial dos elementos dos documentos (meta-tags,

títulos, etc) deverá depender apenas do critério dos participantes.

Embora desta forma a comparação entre sistemas seja discutível visto a base de conhecimentos usada ser diferente, torna-se possível a participantes que não tenham os recursos necessários para operar sobre a colecção TumbaGOVPT ter uma ideia do desempenho dos seus sistemas, havendo ainda a vantagem de se levar em conta na avaliação não só a indexação como também o componente de crawling [6, 3] de um sistema desta natureza.

6 Justificação da abordagem proposta

Dada a dificuldade associada à realização de avaliações de sistemas de RI, procurámos desde logo iniciar o trabalho com uma proposta simples, pouco automatizada e pouco formal que, tanto quanto possível, use o que já se aprendeu com a avaliação noutras línguas. Posteriormente refinaremos o processo à medida que a experiência o permita (à semelhança do que aconteceu para a avaliação de sistemas em inglês).

Qualquer que seja a metodologia utilizada, a tarefa de avaliação não deve também ser nem muito fácil nem muito difícil para a tecnologia actual [11]. Se for muito fácil, todos os sistemas terão um bom comportamento e muito pouco é aprendido. Por outro lado, se for muito difícil, todos os sistemas terão baixo desempenho e novamente muito pouco é aprendido.

Os grandes objectivos e as razões da nossa escolha desta tarefa de avaliação foram, em síntese:

Contribuir para o lançamento de uma actividade continuada de avaliação em RI de informação da Web em Português, de forma a permitir aos investigadores avaliar os seus sistemas face a outros e melhorar o nível técnico global dos seus trabalhos.

Aumentar a base de conhecimento disponível sobre a Web portuguesa a partir de um subconjunto compreensivo da mesma.

Ajudar a estabelecer um processo que conduza a uma especificação clara de critérios de avaliação e de progresso para experiências continuadas desta natureza, baseados em recursos e resultados obtidos cooperativamente.

7 Conclusões

Embora simples, cremos que a nossa proposta constitui um ponto de partida a considerar e oferece um esquema de avaliação realizável com os recursos disponíveis. Permitir-nos-á começar de maneira pouco automatizada

e informal, e ir refinando o processo à medida que for ganha experiência sobre a avaliação destes sistemas.

Uma avaliação em grande escala, realizada por uma comunidade abrangente, não é estática. Deve continuar a evoluir, adaptar e desafiar a comunidade de pesquisa, com os pés firmemente plantados em aplicações realistas. Desta forma, pensamos também em no futuro alargar a metodologia de avaliação para incluir outras tarefas, nomeadamente uma de destilação de tópicos (Topic Distillation Task) à semelhança do TREC. Para avaliar o comportamento face às necessidades de informação e satisfação dos utilizadores, objectivos últimos de um sistema de RI, há outros métodos mais eficazes como entrevistas, observações e experiências “think-aloud”. Este tipo de avaliação é contudo bem mais caro e demorado, sendo de tentar apenas quando a avaliação conjunta estiver mais estabelecida na nossa comunidade.

Embora a tarefa de avaliação esteja bastante focada, os participantes poderão também explorar outras questões relacionadas com recuperação de informação, nomeadamente pesquisas com erros ortográficos, crawling e indexação eficiente. Estas ideias constituem apenas um ponto de partida para a realização de experiências de avaliação em RI sobre o Português, podendo ser estendidas consoante o que as entidades envolvidas entenderem mais apropriado.

Convém salientar que, no nosso entender, a preocupação central na avaliação destes sistemas não deve ser com questões de desempenho (velocidade, memória, portabilidade), mas sim com questões de adequação linguística e adequação à tarefa de recuperação de informação eficiente do ponto de vista dos utilizadores.

A recolha de documentos **TumbaGOVPT** parece-nos um recurso interessante para a realização de outras experiências de avaliação em recuperação de informação. Julgamos ser desejável virmos a ter uma colecção de documentos na língua Portuguesa sobre a qual existam bastantes julgamentos de relevância. A partir desta colecção poder-se-ão especificar novas tarefas, como a já referida destilação de tópicos, a sumarização automática e muitas outras.

8 Agradecimentos

A participação dos autores neste trabalho é financiada pela Fundação para a Ciência e Tecnologia, projecto XMLBase POSI/SRI/40193/2001. O tumba! é alojado e suportado parcialmente pela FCCN – Fundação para a Computação Científica Nacional. Gostaríamos ainda de agradecer ao nosso colega Daniel Gomes pelos seus comentários preciosos neste trabalho e pelo desen-

volvimento de alguns dos componentes de software em utilização no tumba!.

Referências

- [1] R. Aires, S. Aluísio, P. Quaresma, D. Santos, and M. J. Silva. An initial proposal for cooperative evaluation on information retrieval in portuguese. In *Propor'2003 - VI Encontro para o Processamento Computacional da Língua Portuguesa Escrita e Falada*. Springer LNCS, Junho 2003.
- [2] P. Bailey, N. Craswell, and D. Hawking. Engineering a multi-purpose test collection for web retrieval experiments. *Information Processing and Management*, to appear.
- [3] J. Cho and H. Garcia-Molina. Parallel crawlers. In *Proc. of the 11th International World-Wide Web Conference*, 2002.
- [4] J. Gwizdka and M. Chignell. Towards information retrieval measures for evaluation of web search engines. Artigo publicado na Web pelos autores em http://imedia.mie.utoronto.ca/jacekg/pubs/webIR_eval1_99.pdf, 1999.
- [5] D. Hawking, N. Craswell, P. Thistlewaite, and D. Harman. Results and challenges in Web search evaluation. *Computer Networks (Amsterdam, Netherlands: 1999)*, 31(11-16):1321-1330, 1999.
- [6] M. Najork and A. Heydon. On high-performance web crawling. Technical Report Research Report 173, Compaq Systems Research Center, Setembro 2001.
- [7] C. J. V. Rijsbergen. *Information Retrieval, 2nd edition*. Dept. of Computer Science, University of Glasgow, 1979.
- [8] M. J. Silva. The case for a portuguese web search engine. Technical Report Technical Report DI/FCUL TR-03-3, Faculdade de Ciências da Universidade de Lisboa, Março 2003.
- [9] C. Silverstein, M. Henzinger, H. Marais, and M. Moricz. Analysis of a very large altavista query log. Technical Report 1998-014, Digital SRC, 1998.
- [10] L. T. Su. Value of search results as a whole as the best measure of information retrieval performance. *Information Processing and Management*, 34(5):557-579, 1998.
- [11] E. M. Voorhees and D. M. Tice. Implementing a question answering evaluation. In *Using Evaluation Within HLT Programs: Results and Trends*, 2000.
- [12] M.-M. Wu and D. H. Sonnemwald. Reflections on information retrieval evaluation. In *PNC 1999 Annual Conference Proceedings*, 1999.