Determinantes da Eficácia em Pesquisas em CD-ROM

ANA GONÇALVES

Universidade do Porto, Faculdade de Psicologia e Ciências da Educação

GOSTAVA de começar por agradecer a oportunidade que me foi dada para apresentar esta comunicação, que me obrigou e reflectir sobre procedimentos que facilmente transformamos em rotinas e sobre frágeis teorias em que por vezes consolidamos o nosso saber e as nossas práticas.

A experiência analisada nesta intervenção colhi-a num ano de direcção do Serviço de Documentação e Informação (SDI) da Faculdade de Psicologia e Ciências da Educação da Universidade do Porto.

É um serviço que integra cerca de 5000 monografias, a assinatura de 200 periódicos, uma testoteca e uma mediateca. O fundo documental está em livre acesso. Informatizámos o sistema de aquisições, a catalogação e a pesquisa, utilizando a versão POR-BASE 3.0.

Temos um plano de conversão retrospectiva de todo o fundo documental para um ano. Possuímos, desde Setembro, três bases de dados em CD-ROM. A *Eric* para a área de Ciências da Educação, a *Psyclit* para a área de Psicologia e a *Sociofile* para a área de Ciências Sociais.

A assinatura destas bases trouxe alterações importantes à qualidade dos serviços prestados pelo SDI. De gestor de prazos de leitura domiciliária, passou a fornecedor de informação pertinente e actualizada, facilitando a investigação a docentes e alunos. Assistiu-se a uma revalorização do Serviço de Documentação, tendo-lhe sido cometidas funções e dinâmicas que decididamente o afastam da biblioteca estática e o aproximam de um centro de pesquisa e difusão de Informação Científico-Técnica.

CADERNOS BAD, Lisboa (1) 1991, p. 79-133

A resposta às expectativas criadas à volta do Serviço de Documentação e Informação, passa necessariamente pela optimização dos Sistemas de Recuperação de Informação (SRI) existentes. No caso concreto das bases em CD-ROM, passa pela normalização de um conjunto de procedimentos que garantam a eficácia na recuperação da informação.

É objecto desta comunicação a reflexão sobre os factores que condicionam a eficácia dos sistemas automáticos de recuperação de informação, incidindo especialmente sobre a avaliação das diferentes estratégias de pesquisa.

A questão da eficácia dos SRI não é de hoje, não é problema exclusivamente nosso, nem é problemática específica das bases em CD-ROM.

Está na base, como é do conhecimento geral, de múltiplos projectos de avaliação já realizados. Importa desde já reter os contributos de alguns destes projectos:

— O Projecto Cranfield, desenvolvido na década de 50 pela ASTIA (Armed Services Information Agency) e continuado depois pela ASLIB, que, sob orientação de Cleverdon, deu um precioso contributo pela clarificação dos conceitos de revocação e de precisão e pela apresentação de matrizes para avaliação dos valores a eles referentes;

— A avaliação da Base MEDLAR'S, realizada por Lancaster em 1965 que pretendia medir objectivamente o desempenho do sistema, determinando quais os factores que o afectavam negativamente e apontando processos de servir os utilizadores de forma mais eficaz e económica. Apurou que o sistema funcionava com uma média de 57,7 por cento de revocação e 50,4 por cento de precisão, imputando ao subsistema interrogação cerca de 35 por cento das falhas detectadas a estes níveis;

— O trabalho de Salton na avaliação do sistema SMART que, partindo de bases com uma configuração diferente daquela a que nos referimos hoje neste *forum*, nomeadamente no aspecto dos procedimentos para a indexação automática, acaba por apresentar conclusões importantes quanto à utilização de um *thesaurus* para o controle de sinonímia;

— O estudo comparativo entre indexação manual e indexação automática realizado recentemente por uma equipa do CNRS, que utilizando o *thesaurus* EDF e o sistema de indexação automática LEXINET, evidencia as insuficiências e vantagens de cada método e acaba por advogar a vantagem da combinação dos dois.

Se só agora surge entre nós esta preocupação pela eficácia na pesquisa é porque só agora enfrentamos a possibilidade de explorar nos nossos serviços grandes bases de dados e nos confrontamos com a necessidade de dominar os instrumentos de pesquisa.

Que sistemas temos então ao nosso dispôr?

Comecemos por abordar a configuração do sistema em CD-ROM a que temos acesso via *Silver Platter* — na vertente hipóteses de pesquisa.

Estamos perante um *software* de recuperação de informação extremamente elaborado e completo. Podemos dizer que incorpora as recomendações feitas até hoje nos domínios da indexação automática e semi-automática, ou seja:

— recuperação de termos do título, resumo e outros campos,
— recuperação por radicais etimológicos pela técnica de truncatura;
— recuperação de expressões significativas por operadores de adjacência — operador *near;*
— utilização dos operadores booleanos *or, and* e *not* para combinação entre elementos de pesquisa;
— pesquisa por descritores, utilizando a linguagem documental;
— pesquisa por frases-chave;
— combinação de termos dentro do mesmo campo — operador *with;*
— pesquisa por campos específicos — operador *in;*
— limitação da pesquisa por ano de publicação, população, grupos etários;
— selecção de termos para pesquisa, do index ou do texto de visualização do conteúdo de registos;
— O uso de códigos de categorias na limitação das pesquisas.

Afigura-se-nos haver só duas lacunas cujo preenchimento iria, suponho, melhorar o desempenho do sistema: *key-word-in-context* no index e a técnica *zoom.*

Oferecendo aos utilizadores estas múltiplas hipótese — linguagem natural, linguagem documental; alargamento ou redução do âmbito de pesquisa; utilização de vários processos de selecção de termos; uso de diferentes técnicas, as chamadas *exact, range searching, wild card characters* — o sistema *Silver-Platter* fá-los enfrentar por um lado, a insegurança da escolha da estratégia a seguir na construção de expressões de pesquisa e, por outro, o desafio de escolher livre e criativamente o caminho de exploração a seguir.

A teorização feita a partir dos trabalhos de avaliação já referidos, dá-nos alguma ajuda neste domínio. Comecemos por abordar o já clássico e mais que debatido problema da linguagem a utilizar. Linguagem natural ou linguagem documental?

Quanto à primeira afigura-se-nos como vantagens:

— a possibilidade de usar sintaxe na expressão de pesquisa;
— a sua permanente actualização, acompanhando melhor o estádio de desenvolvimento das diferentes ciências;
— uma maior coerência, fruto da indexação automática que elimina a subjectividade na indexação;
— maior especificidade terminológica.

Como desvantagens surgem:

— a ambiguidade, pelo uso das palavras fora do contexto;
— a falta de controle da polissemia;
— a ausência de termos sintéticos que permitam cobrir um dado domínio;
— a maior revocação e menor precisão nas respostas;

Quanto à linguagem documental apresentam-se como elementos positivos:

— o contrôle de sinonímia, a perda de ambiguidade;
— o chamado *«efeito de sintetização»* pelo uso de termos sintéticos que cobrem uma certa categoria de informação;
— uma melhor exaustividade na representação do conteúdo do documento, já que a mesma é feita a partir da totalidade do texto e não só do título ou resumo;
— a menor revocação e maior precisão nas respostas.

Como aspectos negativos salientam-se:

— o desfasamento entre o *thesaurus* e a realidade;
— a diminuição da coerência na indexação;
— o efeito de generalidade dos descriptores;
— as dificuldades na actualização de base quando surgem novos descriptores.

A avaliação da MEDLAR'S feita por Lancaster atribui 35 por cento das falhas de revocação e precisão ao subsistema de interrogação. Que factores influenciam, então, a interrogação?

— conhecimento que o utilizador tem da área científica em causa e o domínio sobre a terminologia usada;
— domínio das hipóteses de exploração que o sistema oferece;
— nível de generalidade da pergunta: quanto mais geral a pergunta, maior é a taxa de revocação, menor a precisão;
— uso inapropriado de termos;
— perícia do utilizador na formulação das perguntas;

— perícia do técnico na construção de expressões de pesquisa adequada.

É exactamente a estes dois níveis, linguagem a utilizar e estratégia de interrogação, que os técnicos de informação e documentação detêm maior poder e responsabilidade.

Extraindo o que é relevante, rejeitando o que não interessa, minimizam o esforço, o tempo e os custos de pesquisa e interferem na eficácia do sistema e na eficiência do Serviço.

À luz desta prévia reflexão passarei agora a abordar a nossa experiência de exploração de bases de dados em CD-ROM da *Silver-Platter*. Uma experiência reduzida, já que só há dois meses possuímos as bases.

Estes dois meses foram um período de aprendizagem, sempre necessário, inevitável e rico.

Ensaiamos procedimentos, estratégias e comparamos resultados.

Neste momento, estabilizamos a nossa prática nalgumas intuições:

1 — A importância de explorar as potencialidades da interacção sistema-investigador-técnico, tornando a pesquisa um processo dinâmico, a ser construído com criatividade e perícia. As experiências que fizemos de ausência do técnico ou do investigador deram maus resultados.

2 — A preparação de pesquisa é importante. Abandonámos o método «*sentarmo-nos à frente do computador e diga lá o que quer*». Experimentámos outro que acabámos também por abandonar e que consistia na cedência do *thesaurus* ao investigador para que ele próprio preparasse a pesquisa. Concluímos que este método só aparentemente era eficaz, já que coarctava à partida a pesquisa devido às restrições linguísticas. Neste momento, ao marcar a pesquisa, o utilizador escreve num formulário um pequeno texto onde explica as suas necessidades de informação. Ao escrever, o utilizador é obrigado a pensar e a precisar o que procura sem ser influenciado pelas restrições linguísticas e pelas lógicas do sistema. É-lhe também pedido que dê algumas sugestões de termos ou expressões significativas, em inglês.

A pesquisa é posteriormente preparada pelo técnico que procura os descriptores apropriados e ensaia algumas estratégias.

Quando o investigador vem fazer a pesquisa são-lhe fornecidos já alguns resultados que ele irá avaliar. Passa-se depois à fase exploratória propriamente dita.

3 — Que estratégias temos utilizado?

Em 70 pesquisas avaliadas, o que corresponde mais ou menos a 2/3 da totalidade das pesquisas efectuadas nestes dois meses, temos como resultados relevantes, ou seja, resultados que deram origem a impressões de registos com informação pertinente

para o investigador, os seguintes valores:

— expressões que utilizam ao mesmo tempo descriptores e vocábulos (termos em linguagem natural): 59 %
— expressões que usam só descriptores: 28 %
— expressões que usam só vocábulos: 13 %
— ocorrência de operadores booleanos
and = 80 %
or = 13 %
not = 16 %
— ocorrência de operadores de adjacência
with e *near* = 20 %

Podemos concluir que no conjunto das pesquisas realizadas, a combinação das duas linguagens, utilizando os operadores booleanos, resulta como a estratégia usada com mais frequência. Convém ainda acrescentar que, em todos os casos, utilizámos de início só descriptores.

4 — Se bem que esta primeira avaliação seja bastante limitada para se tirarem conclusões definitivas, ousamos mesmo assim explicitar algumas hipóteses:

4.1 — A introdução do não-descriptor é importante quando lidamos com investigações originais, com linhas de investigação muito recentes, ou quando pretendemos fazer pesquisas retrospectivas.

Para este último caso, darei o exemplo de uma pesquisa na *Psyclit* sobre vinculação que no disco referente ao período 1983-1990 deve ser realizada com o descriptor *ATTACHMENT* e no disco referente a 1974-1982 deve ser realizada cruzando o descriptor *PARENTE-CHILD-RELATIONS* e *ATTACHMENT* em linguagem natural.

O grau de generalidade ou especificidade da pergunta influencia também decisivamente a estratégia a seguir. Quanto mais específica, mais necessário é normalmente conjugar descriptores com vocábulos. Os descriptores ajudam a restringir ao nível dos macroconceitos, os vocábulos ajudam a deambular pelos termos e expressões muito específicas.

É possível construir expressões de pesquisa de tal modo específicas que podemos chegar a resultados deste tipo: alguém que realiza uma pesquisa sobre um assunto muito específico que corresponde à sua tese de doutoramento, recebe como única resposta os registos de artigos por si escritos.

É evidente que em pesquisas muito gerais do tipo *DRUG-ABUSE and THERAPY* ou do tipo (*STRESS* or *ANXIETY*) *and WORK* devemos usar sempre só descriptores. Muitas vezes, neste caso, o utilizador conclui que necessita de precisar o objecto da sua investigação.

4.2 — Passando agora para o campo dos operadores disponíveis apraz-me dizer que os operadores de adjacência WITH e NEAR são extrema-

mente úteis quando pesquisamos por vocábulos, permitindo o aumento da precisão nas respostas.

Se *FAMILY and ATTACHMENT* recupera a ocorrência destes termos em todos os campos, incluindo título e instituição diminuindo a precisão da resposta, *FAMILY with ATTACHMENT* recupera só registos com a ocorrência dos dois termos no mesmo campo. Exemplificando: (*FAMILY and ATTACHMENT*) *in ab* recupera a ocorrência dos dois termos no campo resumo.

4.3 — Um outro factor a considerar é o tempo. Por um lado, o tempo de duração da fase interactiva de pesquisa, por outro, o tempo que gastamos na construção das expressões de pesquisa.

Convém não ultrapassarmos períodos de uma hora na fase interactiva, já que esta requer muita atenção e «*mente fresca*» para as decisões a tomar. Em pesquisas muito alongadas perdemos muitas vezes a memória do que fomos fazendo e entramos em repetições e em estratégias ilógicas.

Convém ainda referir que desde que não necessitemos de resultados parcelares, havendo vários termos a cruzar e a excluir, é mais rápido construir à partida grandes expressões do que ir por sucessivas etapas elementares.

4.4 — Por fim, a questão da avaliação. Se bem que não haja normas a que nos possamos agarrar e se é certo que a estratégia de pesquisa depende bastante de factores subjectivos como o controle terminológico, a experiência ou a paciência, não devemos no entanto deixar de tentar perceber e explicar os resultados que vamos tendo.

No Serviço de Informação e Documentação do FPCE estamos neste momento a ensaiar uma metodologia de avaliação de todo o Sistema nos diferentes componentes apresentados por Salton e utilizando as matrizes que Cleverdon utilizou no projecto de Cranfield.

Não é um trabalho para se fazer. É um trabalho para se ir fazendo e para estar permanentemente presente como preocupação na definição das nossas linhas de acção.

Bibliografia

CHARTRON, Ghislaine [et al.] — *Indexation manuelle et indexation automatique: dépasser les oppositions*. Documentaliste. 1989, 26 (4-5).

CLEVERDON, C. N. — *Report on the first stage of an investigation into the comparative efficiency of indexing systems*. Cranfield: College of Aeronautics, 1960.

INGWERSEN, Peter; WORMELL, Irene — *Modern indexingand retrieval techniques matching different types of informations needs*. Inf. Forum Inf. and Docum. 1989, 14 (3).

KANTOR, Paul B. — *Evaluation of and feedback in information storage and retrieval systems*. Annual Review of Information Science and Technology. 1982, 17, pp. 99-120.

LANCASTER, F. N. — *Aftermath of an evaluation*. Journal of Documentation. 1971, 27 (1).

SALTON, G. — *The small project*. Londres: Prentice-Hall, 1971.

SALTON, Gerard; MCGILL, Michael J. — *Introduction to modern information retrieval*. Lisbon: McGraw-Hill, 1984.