
Avaliação de Sistemas de Recuperação de Informação Breve Retrospectiva dos Principais Projectos

LUIZA BAPTISTA MELO

Biblioteca da Universidade Lusíada, Porto

UMA biblioteca ou centro de documentação têm como principal objectivo responder às necessidades de informação solicitadas pelos seus utilizadores, com pertinência, rapidez e actualidade. A recuperação de informação com estas características requer uma pesquisa efectiva o que implica a utilização de vários instrumentos, para com eficácia, qualidade e sentido de custo se encontrarem as respostas que os utentes desejam obter.

Surge assim a necessidade de ter bons instrumentos de apoio, tais como: catálogos, índices, sistemas de recuperação em linha, etc. Estes instrumentos podem e devem ser avaliados no sentido de se obter um controlo sobre a qualidade da recuperação de informação.

Este trabalho tem como objectivo apresentar uma retrospectiva dos principais projectos que se debruçaram sobre a avaliação de sistemas de indexação. Estabelecendo um encadeamento lógico da evolução do estudo da avaliação de sistemas de recuperação de informação, analisam-se as conclusões de maior interesse e as que influenciaram de algum modo os projectos seguintes.

Apesar de ter sido relativamente fácil adquirir referências bibliográficas sobre o assunto, as dificuldades na obtenção de literatura da especialidade foram bastantes. Tratando-se de um tema pouco desenvolvido a nível nacional, foi com dificuldade que se conseguiu reunir algum material de interesse para a elaboração do presente artigo.

I Projecto Cranfield

De 1957 a 1962, a ASLIB desenvolveu o I Projecto Cranfield¹, que tinha como finalidade saber qual o sistema de recuperação de informação mais eficiente, através de um método objectivo de avaliação.

Este trabalho foi orientado por C. W. Cleverdon, bibliotecário do Royal College of Aeronautics em Cranfield que, sob o título «Uma pesquisa sobre a eficiência comparada de sistemas de indexação», apresentou uma metodologia para a comparação de quatro sistemas: uma classificação facetada, a CDU, cabeçalhos alfabéticos de assuntos e um sistema *Uniterm*.

Na metodologia utilizada foram considerados três diferentes tipos de variáveis:

- três indexadores de grau de experiência variável, um com experiência em indexação e bons conhecimentos de Aeronáutica, outro com experiência apenas em indexação e ainda um terceiro sem experiência em indexação e sem conhecimento da especialidade;
- diferentes períodos de tempo foram utilizados pelos indexadores;
- esta experiência foi totalmente repetida por três vezes, para verificar se a indexação era melhorada à medida que os indexadores adquiriam melhores conhecimentos sobre o assunto.

O material utilizado para a aplicação e estudo destes testes era constituído por 100 relatórios técnicos da área da Aeronáutica, de onde se obtiveram 18 000 termos de indexação e 1200 pesquisas.

Definições de conceitos e critérios básicos para a avaliação da indexação

A qualidade da indexação avalia-se sempre pela exaustividade e pela especificidade. Estes dois critérios a par de outros, apresentados por Cleverdon no I Projecto Cranfield, tornaram-se base da maior parte dos estudos de avaliação da indexação levados a cabo até hoje.

Segundo Cleverdon², a exaustividade na indexação é um termo comparativo; a um nível «elevado» significa que existe uma entrada para todo o conceito possível no conteúdo de um documento, a um nível «baixo» significa que se tem feito uma selecção e é utilizado um pequeno número de conceitos na indexação dos documentos.

A especificidade é também um termo comparativo. Um conceito pode ser traduzido numa linguagem de indexação de tal modo que o termo de indexação é «co-extensivo» com o conceito, isto é, para um valor alto de especificidade o termo de indexação cobre o conceito, mas não mais que o conceito. Alternativamente, a tradução pode ser feita com um descritor menos específico que inclui além do conceito a ser indexado outros conceitos.

A eficácia do funcionamento da linguagem de indexação dependerá do seu desempenho no sistema de recuperação de informação, que pode ser observado pela revocação (*recall*) e pela relevância (*relevance*).

A taxa de revocação é igual a $100R/C$, onde C é igual ao número total de documentos considerados de relevância para uma dada questão, enquanto R é igual ao número de documentos relevantes recuperados numa única pesquisa. A taxa de relevância é igual a $100R/L$, onde L é igual ao número de documentos recuperados numa única pesquisa.

A eficácia do funcionamento é afectada pelas variações da exaustividade e da especificidade da indexação, assim como do programa de pesquisa. O comportamento do sistema pode-se obter traçando uma curva da taxa de revocação *versus* a taxa de relevância.

Principais conclusões obtidas no I Projecto Cranfield

A partir da referida metodologia, chegaram-se às seguintes conclusões³:

- A eficiência apresentada pelos quatro sistemas de indexação utilizados foi aproximadamente a mesma. Daí se conclui que, para uma dada colecção de documentos e um dado conjunto de perguntas, haveria um desempenho máximo de recuperação; fosse qual fosse a linguagem de indexação usada, a forma da curva (taxa de revocação *versus* taxa de relevância) não era substancialmente alterada, e a única razão para alguma variação significativa da curva dever-se-ia a inadequações de comportamento intelectual nas decisões sobre os conceitos do assunto, ou na indexação ou ainda na pesquisa. Para cada conjunto de documentos nos diferentes campos da Ciência o resultado do comportamento das curvas podia provavelmente diferir nos valores de taxas de revocação e relevância sem, no entanto, alterar significativamente a forma da curva. De igual modo, quando o mesmo conjunto de documentos foi utilizado para duas situações separadas, desde que houvesse dois conjuntos diferentes de questões, dava origem a diferentes comportamentos. Os testes de

Cranfield realizaram-se com vários ramos da Ciência, Engenharia e Metalurgia. Com estes tipos de documentos e conjuntos de perguntas pareceu que o comportamento de recuperação máxima poderia ser o que se apresenta na Figura 1. Quando o assunto da colecção era Química Orgânica, o resultado foi o que se apresenta na Figura 2, para a Sociologia foi o da Figura 3.

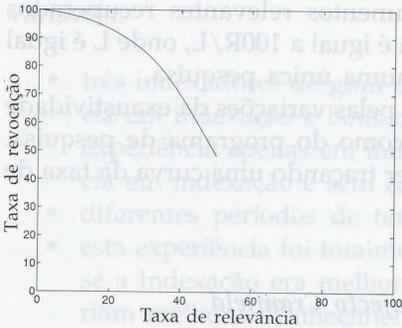


Figura 1

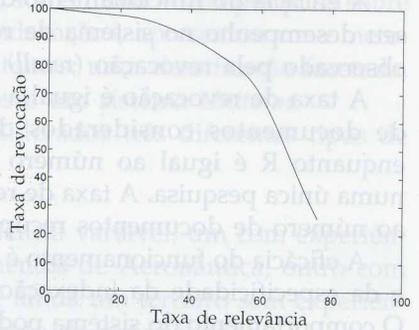


Figura 2

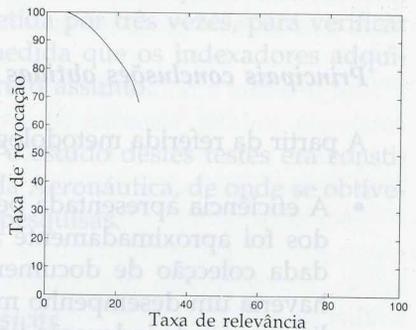


Figura 3

- Não houve diferença significativa entre os diferentes indexadores.
- Não se verificaram grandes alterações entre as três seqüências efectuadas.
- A indexação melhorou à medida que os indexadores adquiriam conhecimentos e experiência.
- O intervalo de tempo de quatro minutos para indexar determinados documentos provou ser o mais adequado. Tempo adicional não melhorou os resultados.
- O erro humano, quer na indexação quer na pesquisa, foi o factor isolado que mais contribuiu para a presença de falhas nos sistemas.

- Finalmente, é importante concluir que a principal contribuição deste projecto na avaliação de sistemas de indexação foi sem dúvida a nível teórico. Surgiram os conceitos de revocação e relevância e a formulação dos critérios de especificidade e exaustividade. O isolamento destes factores básicos estabeleceu uma base satisfatória para a realização de trabalhos experimentais em bibliotecas, com vista à comparação de sistemas de indexação.

A principal crítica ao I Projecto Cranfield refere-se ao grande número de variáveis que apresentava para simular as situações reais.

II Projecto Cranfield

Em consequência dos resultados obtidos e atendendo à principal crítica apontada ao I Projecto Cranfield, surgiu em 1963 um segundo projecto. Este novo trabalho é caracterizado por ser puramente experimental, as condições são criadas no laboratório e a mudança de cada variável é observada na ocasião.

Nestes testes a colecção dos documentos foi analisada previamente em relação a cada um dos assuntos estabelecendo-se assim o conjunto dos documentos relevantes em cada caso.

Efectuando-se uma pesquisa obtêm-se dois conjuntos de documentos: recuperados e não recuperados. Cada um desses conjuntos contém dois subconjuntos de documentos: relevantes (a, c) e não relevantes (b, d).

Recorrendo a uma representação matricial tem-se:

Tabela 1

	RECUPERADOS	NÃO RECUPERADOS	TOTAL
RELEVANTES	a	c	a + c
NÃO RELEVANTES	b	d	b + d
	a + b	c + d	a + b + c + d = N

Para se observar o comportamento do sistema, utilizou-se um método semelhante ao empregue no I Projecto, traçou-se um gráfico taxa de revocação *versus* taxa de precisão onde:

$$\text{Taxa de revocação} = \frac{a}{a + c} \times 100\%$$

$$\text{Taxa de precisão} = \frac{a}{a + b} \times 100\%$$

Por outro lado, Cranfield decidiu substituir a designação «relevância» para «precisão» por achar a primeira demasiado subjectiva.

Compararam-se 33 diferentes linguagens de indexação, desde termos simples tirados dos próprios documentos indexados até vocabulários extremamente controlados e estruturados.

Com o objectivo de comparar quantitativamente os resultados entre os vários sistemas de indexação utilizados, G. Salton estabeleceu o método de «graduação» (*ranking*) dos documentos relevantes⁴ que fornecia uma medida do sistema em função do nível de coordenação. O nível de coordenação dependia do número de termos associados para cada pesquisa. Com este método foi possível calcular o coeficiente normalizado de revocação que actua como uma medida de precisão. A principal vantagem deste teste de avaliação era a obtenção de um número simples e não de um gráfico revocação-precisão, sendo assim mais fácil de manipular.

Principais conclusões obtidas no II Projecto Cranfield

Através da análise dos resultados obtidos nestes testes chegaram-se às seguintes conclusões⁵:

- A utilização de termos simples da linguagem natural, retirados dos documentos, proporcionaram bons resultados, que foram melhorados quando se eliminaram sinónimos e paradigmas.
- Os piores resultados foram obtidos quando se utilizou a linguagem natural para expressar conceitos com termos compostos. O fracasso dos resultados obtidos nesta etapa pode ser devido à pré-coordenação dos assuntos compostos.
- Os vocabulários controlados apresentaram piores resultados que os termos simples, mas, em geral, muito melhores que os conceitos traduzidos por termos compostos.
- As variações dos factores especificidade e exaustividade nas diferentes linguagens de indexação foram também abordadas neste trabalho.

Apresentam-se alguns dos resultados obtidos na Tabela 2.

A linguagem de indexação II.1 (conceitos simples expressos em linguagem natural) foi a que apresentou piores resultados, possivelmente pelo problema referido anteriormente; utilização de associação de termos para expressar conceitos de modo pré-coordenado.

Tabela 2

Linguagem de indexação	Características da linguagem	N.º de termos	Coefficiente normalizado de acerto
I.3	Termos simples e fundidos os paradigmas	2541	65,82%
I.9	3.º nível de agrupamento hierárquico	306	61,17%
II.1	Conceitos simples expressos em linguagem natural	10 000	44,64%

Por outro lado, a redução a um vocabulário mínimo agrupando hierarquicamente os conceitos simples, não piorou significativamente o desempenho do sistema, (exemplo linguagem de indexação I.9).

Os resultados obtidos nestes testes foram de algum modo inesperados, pois a experiência de grande número de bibliotecários tinha verificado que os vocabulários controlados eram muito úteis na prática.

A principal crítica apontada a este trabalho referia-se às condições criadas em laboratório, que apresentavam um esquema experimental e artificial, para simular o tipo de situações encontradas na vida real. É pois com reservas e atendendo às circunstâncias em que os testes foram elaborados que devemos observar os resultados do II Projecto Cranfield.

Sistemas MEDLARS

A National Library of Medicine (NLM), dos Estados Unidos da América do Norte, situa-se na cidade de Bethesda, no estado de Maryland, e é uma das maiores bibliotecas especializadas do mundo. Nela é publicada, periodicamente, desde 1879, uma lista de referências dos principais artigos que surgem em publicações periódicas das Ciências Biomédicas de todo o mundo. Essa lista é o *Index Medicus*.

Pressionada pelo constante aumento do volume de literatura biomédica, a NLM criou desde 1964 um sistema informatizado de recuperação de

referências a que deu o nome MEDLARS (Medical Literature Analysis and Retrieval System).

A MEDLARS passou a servir para a elaboração do *Index Medicus* e de publicações congêneres. Por outro lado, com o imenso catálogo da MEDLARS, a NLM começou a fornecer bibliografia, sobre os mais diversos temas biomédicos, organizada segundo consultas de profissionais dessa especialidade⁶.

Em 1965, iniciou-se um programa com o objectivo de avaliar e analisar o comportamento do sistema, em relação à recuperação da informação, aos factores que a afectam e às necessidades de maior eficiência e menor custo.

Essa avaliação foi feita a partir do estudo dos requisitos pedidos nas pesquisas dos utilizadores e os testes envolveram cerca de 800 000 documentos que foram indexados com um vocabulário controlado, *Medical Subject Headings (MeSH)*, constituído por cerca de 7000 termos.

Principais conclusões obtidas no sistema MEDLARS

Este grande projecto, orientado por F. W. Lancaster, analisou 299 pesquisas que revelaram que o sistema, como um todo, funcionava com uma revocação média de 57,7% e uma precisão de 50,4%.

Observando-se os resultados obtidos fez-se uma análise pormenorizada das deficiências apresentadas no sistema⁷:

1. Na linguagem de indexação

Os fracassos devidos à linguagem de indexação deveram-se à falta de especificidade, a coordenações falsas e a relações incorrectas entre os termos.

As falhas relacionadas com a exaustividade foram de dois tipos: aquelas em que a indexação foi demasiado exaustiva, e aquelas em que a exaustividade não foi suficiente. Nestes testes consideraram-se de maior importância as publicações periódicas. Assim, a estas foram atribuídos uma média de dez termos, enquanto a outros documentos só se atribuiu uma média de quatro termos. Estas circunstâncias reflectiram em parte os resultados obtidos. Optimizações na revocação da ordem dos 30% a 40% seriam alcançadas se ao indexar se atribuísse uma média de dez termos a cada um dos documentos da colecção analisada.

2. No processo de indexação

Os erros humanos deram origem a uma proporção razoável de falhas; os erros devido a atribuição de termos inapropriados na inde-

xação dos documentos contribuíram com 1%, enquanto as omissões contribuíram para 10% das falhas da revocação.

3. No processo de pesquisa

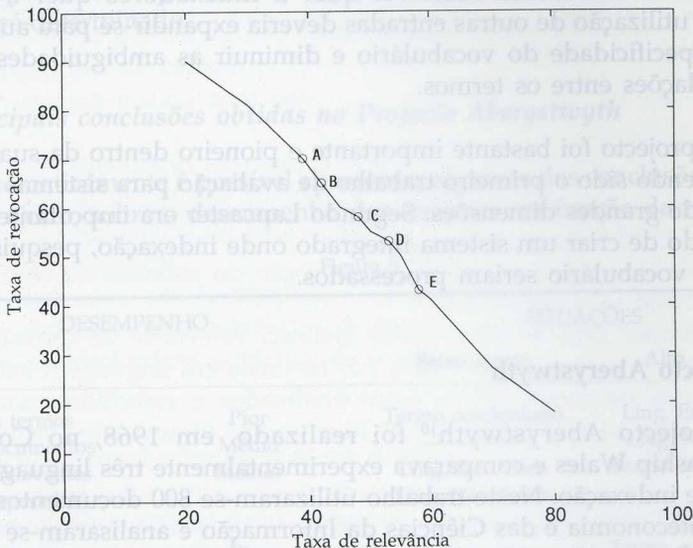
O processo de pesquisa contribuiu com uma proporção muito alta de falhas. Neste passo da recuperação de informação é de salientar que ou a pesquisa era realizada no nível errado de generalidade ou o pesquisador deixava de incluir todas as abordagens razoáveis.

4. Na interação entre o pesquisador e o sistema

Verificou-se finalmente que as pesquisas heurísticas (o utilizador modificava a sua estratégia de acordo com o resultado que ia obtendo) formuladas em linguagem natural obtiveram melhores resultados do que pesquisas interativas. Neste tipo de pesquisa, o diálogo com o bibliotecário ou o analista de pesquisa originou distorções ao tentar-se um enquadramento num molde imposto por quem respondia ao utilizador ou pelo MeSH.

Figura 4

Desempenho global da MEDLARS representado pela extrapolada curva de desempenho dos 5 centros de pesquisa.



A Figura 4^a apresenta um gráfico de taxa de revocação versus a taxa de precisão obtido com resultados dos testes de pesquisas efectuadas em cinco

centros de processamento de informação da MEDLARS. Extrapolando estes cinco pontos foi possível obter o comportamento do sistema MEDLARS. Pode-se verificar que a média a que o sistema estava a operar apresenta uma taxa de revocação de 58% e uma taxa de precisão de 50%. Isto significa que se recuperam, em média, 175 citações por pesquisa.

Com o objectivo de otimizar a eficácia do sistema, e com base num diagnóstico feito a partir da análise das pesquisas, Lancaster propôs os seguintes melhoramentos⁹:

- Aumentar a atenção no pedido do utilizador através da reformulação do pedido de pesquisa. A finalidade era proporcionar uma maior participação do utente de forma que este pudesse transmitir ao sistema a sua questão com todos os requisitos necessários e com grande precisão.
- O pedido de pesquisa deveria incluir o grau de precisão e revocação, assim como a tolerância destes valores desejados por cada utente. Estes dados seriam levados em conta na estratégia utilizada no decorrer da pesquisa.
- O vocabulário utilizado pela MEDLARS deveria ser aumentado e tornar-se facilmente acessível quer a indexadores quer a utentes. A utilização de outras entradas deveria expandir-se para aumentar a especificidade do vocabulário e diminuir as ambiguidades e falsas relações entre os termos.

Este projecto foi bastante importante e pioneiro dentro da sua especialidade, tendo sido o primeiro trabalho de avaliação para sistemas informatizados de grandes dimensões. Segundo Lancaster era importante avançar no sentido de criar um sistema integrado onde indexação, pesquisa e controlo de vocabulário seriam processados.

Projecto Aberystwyth

O Projecto Aberystwyth¹⁰ foi realizado, em 1968, no College of Librarianship Wales e comparava experimentalmente três linguagens diferentes de indexação. Neste trabalho utilizaram-se 800 documentos da área de Biblioteconomia e das Ciências da Informação e analisaram-se cerca de 63 consultas verídicas de informação que tinham sido feitas à ASLIB ou à Library Association. Foram estudadas três linguagens pós-coordenadas de indexação: termo condensado, linguagem natural não controlada e linguagem estruturada hierarquicamente. A linguagem que tomou a designação

de termo condensado, correspondia a um vocabulário mínimo baseado numa lista compilada pelo Departamento de Pesquisa da ASLIB.

A linguagem natural não controlada era constituída por cerca de 1200 palavras simples derivadas dos documentos pelos indexadores.

A linguagem estruturada hierarquicamente era baseada numa classificação de Biblioteconomia elaborada pelo CRG (Classification Research Group). Esta classificação apresentava 800 facetas, sem ordem de citação prefixada e tinha a mesma notação e a mesma ligação hierárquica que caracterizavam a classificação que lhe serviu de base.

Neste processo simulou-se tanto quanto possível situações reais. Para isso, as pesquisas para cada consulta em cada sistema foram feitas de modo a dar o equivalente a 4 pesquisas, englobando várias possibilidades de revocação. Assim, a revocação mínima correspondia a um documento «altamente» relevante por consulta satisfeita (utilizador de alta precisão), enquanto a revocação máxima seria dada por todos os documentos «altamente» ou parcialmente relevantes numa biblioteca (representava as necessidades de um utilizador de «alta» revocação).

Foi ainda levada em conta a variável tempo, isto é, definiu-se para a duração das pesquisas um intervalo de tempo que terminava quando ele era ultrapassado ou quando os pesquisadores atingiam o objectivo de revocação pré-determinado.

Principais conclusões obtidas no Projecto Aberystwyth

Esquemáticamente é possível apresentar algumas das conclusões obtidas neste projecto, sobre o desempenho dos sistemas em função da revocação.

Figura 5

DESEMPENHO		SITUAÇÕES	
		Baixo acerto	Alto acerto
Em termos de documentos não relevantes recuperados	Pior	Termo condensado	Ling. Estr. Hier.
	Médio	_____	Ling. Não Estr.
	Melhor	Ling. Estr. Hier.	Termo condensado
Em termos de não recuperação de documentos não relevantes	Pior	_____	Termo condensado
	Melhor	_____	Ling. Estr. Hier.
		Ling. Não Contr.	

- Em função do tempo de duração da pesquisa o melhor sistema foi o termo condensado enquanto a linguagem estruturada hierarquicamente se apresentou ligeiramente pior.
- O termo condensado não era bastante específico.
- Não revelou desvantagem significativa o facto da linguagem não controlada não ter estrutura.
- Fizeram-se também comparações entre uma linguagem pós-coordenada e uma linguagem pré-coordenada (a Classificação Facetada Estruturada Hierarquicamente, com ordem de citação prefixada, índices de classificação sintetizados e um índice em cadeia). Concluiu-se que o desempenho numa linguagem pré-coordenada era pior, numa situação de máxima revocação, em parte, devido ao intervalo de tempo da pesquisa ser mais longo.
- Outro teste realizado utilizou o sistema de indexação relacional de Farradane¹¹ (sistema que adoptou operadores relacionais). Indexaram-se 250 documentos e efectuou-se um teste comparativo utilizando 60 questões para medir o desempenho do sistema com e sem operadores. Os operadores funcionaram de modo a melhorar a precisão piorando a revocação. Cerca de 13% das pesquisas foram afectadas, 10% tiveram a precisão melhorada, 3% pioraram a revocação, mas a optimização em termos da supressão de documentos não relevantes foi pequena.

Sistema SMART

O sistema SMART (Salton's Magical Automatic Retrieval of Text)¹², orientado por Gerard Salton em 1968, nos Estados Unidos, englobou um projecto que se preocupou com o estudo da avaliação de sistemas de recuperação de informação informatizados.

A principal característica deste sistema referia-se às entradas dos documentos que eram constituídas por resumos em linguagem natural ou títulos. Estes pequenos textos eram analisados e estudados por vários métodos. Para isso o sistema SMART incorporava as seguintes facilidades¹³:

1. Um sistema para separar os prefixos e os sufixos das palavras inglesas, (método dos sufixos e do *thesaurus* de palavras raiz), que construía «identificadores» dos documentos a partir das palavras contidas nos próprios documentos.
2. Um *thesaurus* que podia ser utilizado para reconhecer sinónimos das palavras raiz e substituí-las por um ou mais «números de conceito»

(*concept numbers*). Estes «números de conceito» serviam como «identificadores» de conteúdo em lugar das palavras raiz originais.

3. Um arranjo hierárquico de conceitos incluídos no *thesaurus* que tornava possível, atribuindo qualquer «número de conceito», descobrir a posição do conceito na hierarquia ou num conjunto possível de relações cruzadas.
4. Procedimentos estatísticos que processavam coeficientes similares baseados na ocorrência dos conceitos que estavam incluídos nas frases de uma dada colecção. A relação dos conceitos, determinada por associação estatística, podia então ser adicionada aos conceitos originais disponíveis para identificar os vários documentos.
5. Métodos de análise sintáctica, que tornavam possível fazer uma comparação da análise sintáctica da frase, dos documentos e das pesquisas dos utilizadores, com uma pré-codificação das estruturas sintácticas feitas a partir de um dicionário (critério em árvore). Essa comparação era efectuada de tal modo que o mesmo «número de conceito» era atribuído para um grande número de equivalências semânticas, mas com construções sintácticas diferentes.
6. Métodos de tratamento estatístico de frases que operavam como o método anterior (análise sintáctica da frase) mas utilizavam um dicionário pré-construído para analisar frases utilizadas como identificadores de conteúdo.
7. O sistema incluía para isso:
 - um dicionário de palavras raiz
 - um dicionário de prefixos e sufixos
 - um dicionário de palavras
 - um *thesaurus*
 - um «conceito hierárquico»
 - um dicionário estatístico de frases
 - um dicionário sintáctico de frases

Principais conclusões obtidas no sistema SMART

As conclusões obtidas pelas principais experiências, feitas no âmbito da avaliação do sistema SMART, foram as seguintes:

- A análise simples e automática da linguagem, consistindo na atribuição de «pesos» para as palavras raiz contidas nos documentos, pro-

duzia uma eficácia de recuperação quase equivalente à que se obtinha por indexação intelectual efectuada manualmente sobre condições controladas.

- Utilizando o procedimento com o *thesaurus*, para reconhecer os sinónimos e outros termos relacionados e substituir pelas palavras raiz com as correspondentes categorias do *thesaurus*, melhorava-se a eficácia na recuperação cerca de 10% na revocação e 10% na precisão.
- Com métodos de análise de linguagem mais sofisticados (atribuição de frases em vez de termos adicionais, conceitos hierárquicos, determinação de relações sintácticas entre os termos, etc.) em média não se verificaram melhoramentos em relação aos processos normais de utilização de *thesaurus*.
- O projecto de avaliação deste sistema englobou também a verificação do comportamento do sistema com bibliografia de origem alemã. Para esse efeito, foi feita uma experiência utilizando um *thesaurus* multilingue em conjunção com duas colecções diferentes de documentos, uma de expressão alemã e outra de língua inglesa. Verificou-se que o processo de linguagem cruzada (pedidos de documentos em Alemão *versus* pedidos em Inglês) era quase tão efectivo como o processamento de uma só língua. Em particular, diferenças de morfologia (por exemplo sufixos tirados sem regra) e ambiguidades na língua não pareceram causar substancial degradação quando se mudava de uma língua para a outra. Por estas razões, os métodos automáticos de recuperação usados no sistema SMART para o Inglês pareceram ser aplicáveis a material expresso noutras línguas.

Finalmente e numa tentativa de resumir as conclusões obtidas nestes testes, apresenta-se uma escala de ordem de valores para a eficácia da recuperação da informação no sistema SMART¹⁴:

1. Mais eficaz Método do processamento do resumo com reconhecimento da frase e dos sinónimos
2. Eficaz Utilização da junção, palavra chave pesada e estatística de associação de palavras, na análise dos resumos
3. Menos eficaz Método do emparalhamento lógico de palavras chave desprezando termos pesados
4. Muito menos Processamento do título apenas com o título do documento para eficaz análise.

Conclusões

Só nos finais da década dos anos 50 é que o problema da avaliação da indexação suscitou maior interesse aos estudiosos desta área.

Uma das primeiras experiências efectuadas com objectivo de avaliar as eficiências de vários sistemas de indexação foi levada a cabo em 1953 pela ASTIA (Armed Service Technical Information Agency)¹⁵. A comparação de um conjunto de cabeçalhos de assuntos que a própria ASTIA tinha criado, e um sistema *Uniterm* elaborado por Mortimer Taube não levou, no entanto, a qualquer conclusão importante.

Em 1957 é C. W. Cleverdon, com o I Projecto Cranfield, o pioneiro em termos de uma contribuição efectiva para este problema. A sua metodologia, que engloba a definição dos conceitos de revocação e relevância assim como a formulação dos critérios de especificidade e exaustividade, estabelece uma base para a realização de trabalhos experimentais, em bibliotecas e centros de documentação, com o objectivo de comparar sistemas de indexação. Objecto de alguma crítica, tal como a utilização de um grande número de variáveis para a simulação de situações reais, é na verdade este trabalho a génese de um conjunto de parâmetros que até hoje são considerados nos processos de avaliação de sistemas de recuperação de informação.

Atendendo à principal crítica a que foi sujeito surge, em 1963, o II Projecto Cranfield. Puramente experimental, este trabalho chegou a conclusões um tanto ou quanto surpreendentes: a utilização de termos simples da linguagem natural retirados dos documentos para a indexação revela bons resultados, ao contrário do que se verificava em grande número de bibliotecas, que utilizando vocabulários controlados obtiveram grande êxito. Claro que o esquema experimental artificial de Cranfield teve, em contraste com o tipo de situações encontradas na vida real, grande contribuição para as conclusões alcançadas. No entanto, estes resultados apresentaram os seus reflexos, levando mesmo Klingbiel¹⁶ a afirmar que «a linguagem natural da literatura científica é perfeitamente adequada na indexação e na recuperação».

Esta época é especialmente caracterizada pelo grande desenvolvimento das técnicas de construção de *thesauri*. De acordo com Gilchrist¹⁷ «foi na verdade o apogeu».

O Projecto Aberystwyth, em 1968, é mais uma contribuição para o estudo da avaliação. Abarcando a análise de três sistemas de indexação particulares, é um trabalho importante e revela que os ganhos decorrentes do emprego de linguagens complexas de indexação são insignificantes.

É pois de todo o interesse, em etapas futuras, tentar avaliar a medida em que os resultados destes testes podem ser extrapolados para situações mais amplas. Caso se confirmem estes resultados, isto é, no sentido dos ganhos decorrentes do emprego de linguagens complexas de indexação serem pouco significativos, será difícil justificar o tempo adicional exigido para o seu uso.

Os meados da década de 60 são caracterizados pelo enraizamento da informática, os problemas da avaliação da indexação debruçam-se naturalmente sobre os grandes sistemas de recuperação de informação informatizados. Agora, os sistemas apresentam características muito diferentes quer quanto às dimensões quer quanto aos acessos. A ligação em linha torna importante outros critérios, a interacção do utilizador com o sistema é um dado a ser analisado com a máxima atenção. Pesquisa heurística e pesquisa interativa são comparadas em diversos testes.

A National Library of Medicine, nos Estados Unidos, elaborou em 1965 um importante trabalho de avaliação do seu sistema MEDLARS. Basicamente, as conclusões obtidas têm em vista a melhoria da eficácia da recuperação da informação. O aumento da atenção sobre o pedido da pesquisa, a indicação pelo utente do grau desejado de precisão e revocação e suas tolerâncias e a expansão do vocabulário *MeSH*, para otimizar a especificidade da linguagem de indexação, são os principais melhoramentos aconselhados por Lancaster após ter realizado e analisado os testes de avaliação do sistema.

Em 1968, outro grande sistema informatizado, SMART, envolveu no seu trabalho testes de avaliação. Este poderoso sistema tem a particularidade das entradas dos documentos serem constituídas pelo título ou por resumos. As potencialidades são tais que a análise dos pequenos textos podem ser feitas em Inglês ou em Alemão.

A partir dos anos 70 é difícil generalizar conclusões. A par dos avanços tecnológicos a nível informático aparecem numerosos sistemas de recuperação de informação. Surge periodicamente bibliografia estrangeira da área divulgando novos testes e novas pesquisas. No entanto, é importante salientar que todos esses trabalhos assentam nos conceitos básicos, estabelecidos nos projectos pioneiros de Cranfield, para realizar comparações entre a eficácia dos diversos sistemas de indexação.

O processo de indexação, envolvendo basicamente as etapas de análise do texto por assuntos, extracção dos conceitos e tradução dos conceitos em termos de indexação, inspira investigadores da área que elaboram múltiplas metodologias para avaliação de sistemas de recuperação de informação¹⁸.

Muito recentemente, em 1993, a Graduate School of Library and Information Science, University of California (USA), em colaboração com o Departamento de Information and Library Studies, Loughborough University (Reino Unido) levaram a cabo mais um estudo de avaliação¹⁹.

A metodologia seguida prestou particular atenção às primeiras duas etapas que envolvem o processo de indexação e apresentou resultados interessantes em relação à análise por assuntos, a natureza factual ou subjectiva, a complexidade (simples ou composta), a presença de título, subtítulo, resumo, ordenação de parágrafos num texto apresentam particular importância no desempenho da análise por assuntos. Foi evidente que documentos representativos das áreas das Ciências Sociais e das Ciências, geralmente associados a textos factuais, apresentam grande facilidade na identificação do conceito geral e do tópico principal quando este é claramente central. Os documentos da área de Humanísticas, associados a textos subjectivos, apresentam pobres e difíceis desempenhos na tarefa de análise de assuntos.

Em Portugal os estudos sobre testes de avaliação de sistemas de recuperação de informação não têm sido particularmente divulgados pelos técnicos da área.

Actualmente, existe já a nível nacional *software* especialmente concebido para facilitar a recuperação de informação. Seria interessante testar algumas bases de dados bibliográficos nesses suportes, uma vez que a avaliação é fundamental na optimização do desempenho dos sistemas.

É importante referir, no entanto, um teste de avaliação da eficácia de uma base de dados com material arquivístico²⁰, levado a cabo por Fernanda Ribeiro, em 1992. Este estudo, utilizando os conceitos clássicos de avaliação, envolveu 22 utilizadores que elaboraram pesquisas a duas bases de dados do Arquivo Histórico Municipal do Porto. O suporte informático foi o programa Mini-Micro CDS/ISIS na parametrização ARQBASE.

Neste trabalho a autora conclui que, o controlo de autoridade na indexação por assuntos utilizado numa das bases apresentou melhor desempenho do que a base de dados na qual se usou linguagem de indexação não-controlada. No entanto, nem só a linguagem controlada deverá estabelecer pontos de acesso. É importante existir sempre a possibilidade de pesquisar em linguagem não-controlada aumentando assim a capacidade de revocação do sistema. Verificou-se ainda que testes desta natureza permitem não só medir a eficácia e eficiência de um sistema de recuperação de informação, mas também possibilitam detectar e analisar falhas e consequentemente corrigir deficiências encontradas.

Notas

- ¹ FOSKETT, A. C. T. — *A abordagem temática da informação*. São Paulo: Editora Universidade de Brasília: Editora Polígono, 1973.
- ² CLEVERDON, C. W. — *ASLIB Cranfield Research Project — report on the testing and analysis of an investigation into the comparative efficiency of indexing systems*. Cranfield: College of Aeronautics, 1962.
- ³ CLEVERDON, C. W.; MILLS, J.; KEEN, E. M. — *Factors determining the performance of indexing systems*. Cranfield: Aslib-Cranfield Research Project, 1966.
- ⁴ ATTCHISON, J.; CLEVERDON, C. W. — *A report on a test of the index of metallurgical literature of Western Reserve University*. Cranfield: College of Aeronautics, 1963.
- ⁵ SWANSON, D. — «The evidence underlying the Cranfield results», *Library Quarterly*, 35 (1) 1965, p. 1-20.
- ⁶ CLEVERDON, C. W. — «The Cranfield tests on index language devices», *Aslib Proceedings*, 12 (1) 1966, p. 100-113.
- ⁷ CLEVERDON, C. W.; MILLS, J. — «The testing of index language devices», *Aslib Proceedings*, 19 (6) 1967, p. 173-194.
- ⁸ IDEM, *ibidem*.
- ⁹ SALTON, G. — *Automatic information organization and retrieval*. London: McGraw-Hill, 1968.
- ¹⁰ REES, A. M. — *Review of a report of the Aslib-Cranfield test of the index of metallurgical literature of Western Reserve University*. Cleveland — Ohio: Western Reserve University School of Library Science, 1963.
- ¹¹ GODINHO, José António M. — *MEDLARS e MEDLINE: duas técnicas de recuperação automática de informação biomédica*. Coimbra, 1978.
- ¹² LANCASTER, Frederik W. — «Evaluating the performance of large computerized information system», *Journal of American Medical Association*, (207) 1969, p. 114-120.
- ¹³ IDEM, *ibidem*.
- ¹⁴ IDEM, *ibidem*.
- ¹⁵ KEEN, E. Michael; DIGGER, Jeremy A. — *Report of an information science index language text*. Aberystwyth: College of Librarianship Wales, 1972.
- ¹⁶ FARRADENE, J. E. L. — «Analysis and organization of knowledge for retrieval», *Aslib Proceedings*, 22 (12) 1970, p. 607-616.
- ¹⁷ SALTON, G. — «The evaluating of automatic retrieval procedures — selected test results using the SMART system», *American Documentation*, 16 (3) 1965, p. 209-222.
- ¹⁸ SALTON, G. — «The automatic processing of foreign language documents», *Journal of the American Society for Information Science*, 21 (3) 1970, p. 187-194.
- ¹⁹ VICKERY, B. C. — *Information system*. London: Butterworths, 1973, p. 126.
- ²⁰ MARTIN, W. A. — «Methods for evaluating the number of relevant documents in a collection», *Journal of Information Science*, North-Holland, (6) 1987, p. 173-177.
- ²¹ KINGBIEL, P. H. — *Machined-aided indexing*. Alexandria: Defence Documentation Center, 1969.
- ²² GILCHRIST, A. — *The thesaurus in retrieval*. London: Aslib, 1971.
- ²³ CHAN, Lois Mai; RICHMOND, P. A.; SVENONIUS, E., ed. — *Theory of subject analysis: a sourcebook*. Colorado: Libraries, 1985.
- ²⁴ CHU, Clara M.; O'BRIEN, Ann — «Subject analysis», *Journal of Information Science*, North-Holland, (19) 1993, p. 439-454.
- ²⁵ RIBEIRO, Cândida Fernanda Antunes — *Indexação e controlo de autoridade em arquivos*. Porto: Faculdade de Letras da Universidade do Porto, 1992.