ANÁLISE ESTRUTURAL PARA AUMENTAR A EFICIÊNCIA DE PESQUISAS "ONLINE"

George Eduardo Freund
Gerente de Engenharia de Sistemas do Instituto de Pesquisas Tecnológicas do Estado de São Paulo, e professor na Universidade de São Paulo.

RESUMO

A pesquisa online *em bancos de dados continua a ser um desafio para a maioria dos usuários, principalmente no tocante à identificação dos termos mais apropriados para cada base de dados. A análise estrutural,* uma *técnica baseada na análise sintática de termos com o objetivo de identificar palavras semanticamente relacionadas, é proposta como solução para diversos problemas, inclusive o de truncamento arbitrário, ainda não resolvido satisfatoriamente em sistemas* online. *O presente trabalho apresenta resumidamente alguns aspectos teóricos e práticos da pesquisa realizada pelo autor na Universidade de Sheffield, Inglaterra.*

Descritores: Base de dados; Pesquisa online; *Análise estrutural.*

INTRODUÇÃO

O Eusidic Database Guide 1981[1] relaciona mais de 1400 bases de dados disponíveis em forma digital no mundo todo, dividindo-as em 49 categorias, de acordo com o assunto coberto pelas mesmas. Online Review[2] lista 233 bases de dados acessíveis online através dos 12 principais serviços comerciais de recuperação de informação. As 200 principais bases de dados bibliográficos representam hoje mais de 65 milhões de referências disponíveis para consultas online.

Esses números mostram o excepcional crescimento da indústria de informações em forma digital, que praticamente nasceu na década de 70. O objeto de análise deste estudo são todas as bases de dados cujos itens de informação são acessíveis através de palavras-chave. Essas palavras-chave podem ser termos do próprio autor encontrados no texto do item de informação, podendo também ter sido originadas por outra pessoa, caso específico de descritores gerados no processo de indexação. Evidentemente, o sucesso de uma pesquisa depende fundamentalmente da habilidade do usuário em utilizar na busca os mesmos termos presentes na base de dados. O presente trabalho propõe uma técnica destinada a melhorar a eficiência das pesquisas online, sugerindo ao usuário palavras da base de dados possivelmente relevantes à sua pesquisa. Uma breve inspeção do processo de produção e utilização de tais bases de dados pode ilustrar alguns dos aspectos abordados a seguir.

PRODUÇÃO DE BASES DE DADOS

A Figura l expõe de forma resumida as etapas percorridas desde a informação primária até que uma referência seja colocada à disposição do usuário.

As tarefas da fase de identificação são executadas por especialistas em informação e compreendem basicamente a elaboração da referência à informação. O trabalho de identificação do item de informação (estabelecimento de autor, fonte, etc.) pode ser complementado com uma tradução (do título, do resumo, etc.), caso não esteja no mesmo idioma da base de dados. Por vezes é necessário também proceder-se a uma transliteração total ou parcial de informações geradas em outros alfabetos. Os trabalhos de classificação e indexação dependem fundamentalmente de critérios próprios adotados pelos diversos produtores de bases de dados. O mesmo acontece com o resumo nos casos em que não é adotado aquele feito pelo próprio autor. É, portanto, nessa fase de identificação que são tomadas as decisões acerca das palavras que integrarão a base de dados e que deverão ser utilizadas pelo usuário na pesquisa.

A fase de processamento corresponde aos trabalhos de geração da base de dados em forma digital. A

digitalização das informações pode ser obtida através de digitação para fita ou disco magnético, emprego de leitoras óticas ou outro processo qualquer. Um trabalho de revisão é sempre feito para diminuir o número de erros do produto final. Uma vez acumuladas todas as referências a integrarem a base de dados, é realizado um processo de ordenação segundo critério do produtor da base.

Dos processos de distribuição de informação secundária, interessam-nos basicamente os serviços de busca online. Para que a base de dados possa ser consultada através de terminais, há um pré-processamento objetivando basicamente a criação de dois arquivos[3], o arquivo principal, contendo toda informação da base de dados, ordenada segundo critério de cada serviço, e o arquivo invertido, contendo, em ordem alfabética, todos os termos (nomes, descritores, etc.) passíveis de serem utilizados na busca. Para cada termo constante desse arquivo invertido, são apontados os itens de informação dos quais o mesmo faz parte.

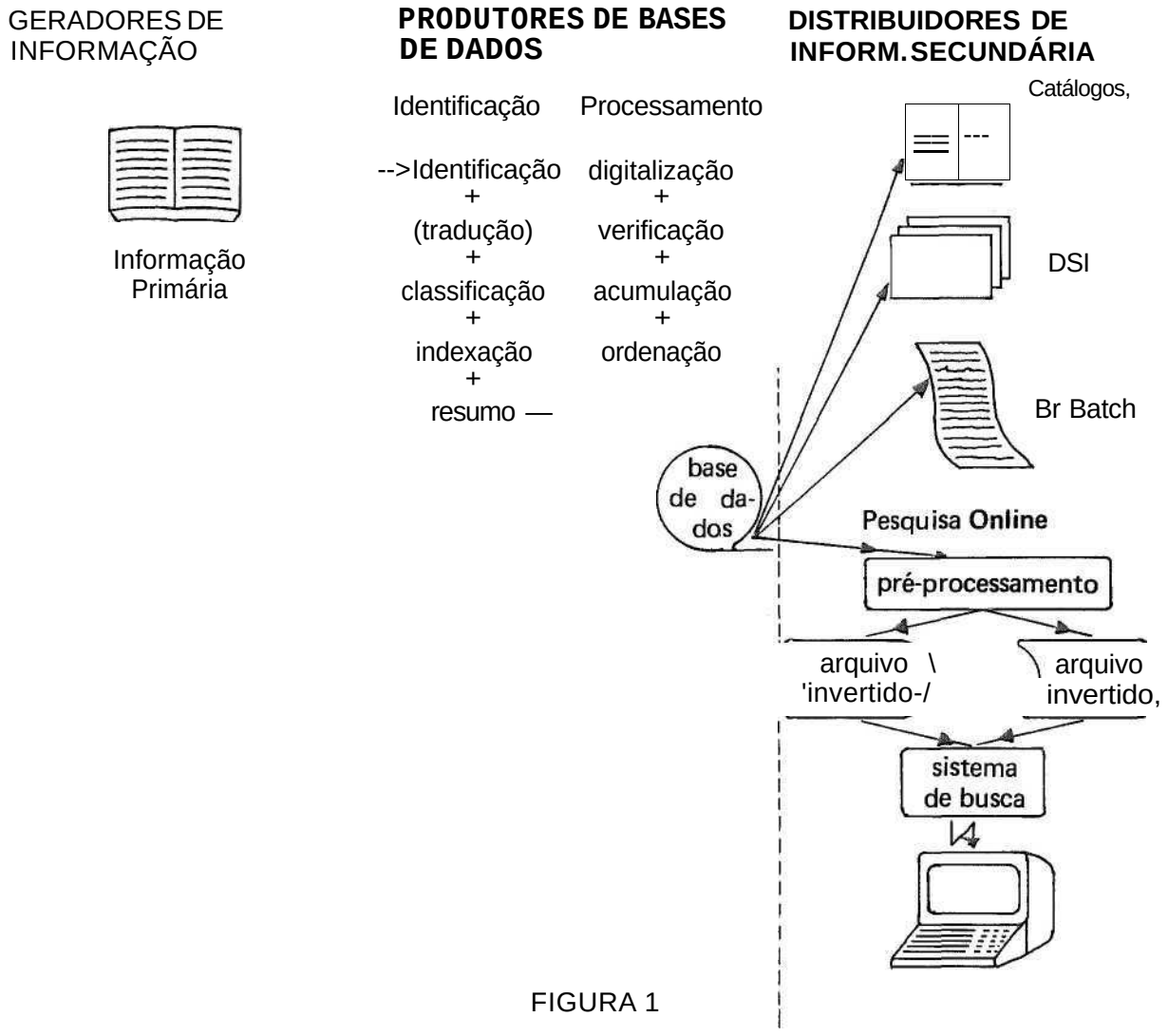

FIGURA 1

O caminho percorrido pela informação desde o autor até o terminal do intermediário é bastante longo, possibilitando o aparecimento de uma variedade de erros e distorções. O intermediário deve procurar conhecer a linguagem utilizada por autores, tradutores, classificadores e indexadores e ainda prever os tipos de erros que serão introduzidos pelos processos manuais, mecânicos e eletrônicos de conversão, armazenamento e transmissão de informações.

FORMULAÇÃO DA PESQUISA

Ao se preparar para realizar uma pesquisa em banco de dados, o intermediário se preocupa basicamente

com dois aspectos: a estratégia de busca e a seleção dos termos a serem utilizados. A participação de um técnico em informação, ou mesmo a realização da pesquisa por um intermediário, é, sem dúvida, uma contribuição positiva na definição da estratégia de busca.

Na escolha dos termos a serem utilizados, o conhecimento da base de dados é de grande valia. O desconhecimento da linguagem técnica do assunto pesquisado, especialmente em áreas técnicas e científicas, é, no entanto, um problema a ser superado. Esse problema, no caso específico do Brasil e em outros casos em que o idioma da base de dados não é o dos interessados, pode afetar até mesmo o usuário final, que muitas vezes desconhece certos termos em outras línguas. Preece[4] afirma que *a* falta de experiência do usuário final em pesquisa a banco de dados é mais claramente sentida na inabilidade em selecionar o vocabulário adequado para explicitar sua necessidade de informação na linguagem de cada base de dados. Alguns poucos produtores de bases de dados têm-se preocupado com esse fato, editando vocabulários técnicos ou thesauri utilizados nos trabalhos de indexação, classificação, etc.

SISTEMAS ONLINE

Além da pouca consistência dos diversos produtores de bases de dados quanto ao vocabulário técnico empregado, a falta de critérios de grafia e o fraco controle de qualidade geram uma grande variedade de termos a serem pesquisados. As principais causas dessas variações são:

a) Erros de grafia: do próprio autor, ou introduzidos nas várias etapas de identificação e processamento das informações. Variações causadas por erros são significativas, chegando, segundo Bourne[5], a 23% do dicionário, para o caso extremo, e representando em média 11% dos termos utilizados pelas bases de dados.

b) Sinônimos ou termos equivalentes, utilizadas por autores ou gerados por indexadores.

c) Variantes como singular x plural, conjugações de verbos ou declinações para certas línguas.

d) Grafias alternativas: caso das variações americano x inglês (como, por exemplo, certas palavras terminadas em OUR e RE que o americano escreve com OR e ER), conceitos multitermo, que podem ser escritos de várias formas (como, por exemplo, ON LINE, ONLINE e ON-LINE), regras diferentes para transi iterações (como, por exemplo, de TCHEBYSHEFF e CHEBYSHEFF), etc.

e) Inconsistências em abreviações como APROX, APROXM, etc.

A preocupação dos vendedores de informação online é grande, pois, além de dificultar o acesso à informação, prejudicando até mesmo sua imagem junto aos usuários, o excessivo número de variantes de palavras aumenta consideravelmente os custos de processamento e armazenamento das bases de dados.

Diversas medidas têm sido tomadas para minimizar esses problemas. Thesauri, quando existentes, são colocados online com as respectivas bases de dados. Comandos como EXPAND e NEIGHBOR, oferecidos por LOCKHEED/DIALOG e SDC/ORBIT, permitem percorrer em ordem alfabética os arquivos invertidos. Nenhuma dessas soluções, no entanto, resolve o problema dos erros de grafia, responsáveis por parcela significativa dos insucessos nas pesquisas.

Apresentamos a seguir o resultado de estudos realizados no sentido da aplicação de técnicas de conflação para aumentar a eficiência de pesquisas online.

TÉCNICAS DE CONFLAÇÃO

O termo 'conflação' é empregado em lugar do inglês conflation, por julgarmos a tradução proposta por dicionários — a saber, 'confusão' —'não adequada para o assunto em questão. Entendemos por conflação de dois ou mais termos seu uso indistinto para uma determinada finalidade.

As técnicas de conflação mais conhecidas podem ser agrupadas em três categorias:

a) Procedimentos para reunir palavras foneticamente semelhantes — esses procedimentos, a exemplo do sistema SOUNDEX[6], começam a ser empregados em maior escala na consulta a catálogos telefônicos. Alguns baseiam-se em reconhecedores acústicos e aceitam entradas faladas, enquanto outros recebem entrada através de terminal cujas teclas são agrupadas de acordo com as características fonéticas do alfabeto.

b) Técnicas de redução de palavras a um radical comum (stemming algorithms) — palavras como AQUECIMENTO, AQUECEDOR, DESAQUECER, etc. são todas reduzidas à sua raiz comum, AQUEC—, para fins de armazenamento e recuperação. As vantagens dessas técnicas para Information Retrieval se resumem em economia de espaço de armazenamento e em maior revocação na recuperação. A desvantagem maior é sua dependência de extensas listas de prefixos, sufixos e regras de remoção dos

mesmos para cada língua, o que torna os algoritmos complexos, ineficientes e imprecisos.

c) Técnicas de Análise Estrutural —essas técnicas, a serem analisadas em maior detalhe a seguir, baseiam-se em um trabalho de Adamson e Boreham[7], que afirmam ser a estrutura das palavras uma boa base para classificação automática, em decorrência de sua forte relação com o conteúdo semântico das mesmas.

ANÁLISE ESTRUTURAL

As técnicas de análise estrutural baseiam-se na decomposição de palavras em unidades menores, a saber: conjuntos de n letras consecutivas, que passaremos a chamar de n-gramas. O número de n-gramas em comum é uma medida de semelhança entre duas palavras. Dessa forma, as palavras CONSTRUIR e DESTRUIÇÃO, por exemplo, que são semanticamente relacionadas, podem ser comparadas:

a) em função de seus digramas

CO ON NS ST TR RU UI IR
DE ES ST TR RU UI IÇ CA AO

tendo 4 digramas em comum, ou

b) em função de seus trigramas

CON ONS NST STR TRU RUI UIR
DES EST STR TRU RUI UIC ICA CÃO

tendo 3 trigramas em comum.

A similaridade entre palavras pode ser calculada por uma série de coeficientes que levam em conta não só o número c de n-gramas em comum, mas também os números a e b de n-gramas de cada uma das palavras. Esses coeficientes, todos normalizados em relação aos comprimentos das palavras, dão valores entre O (sem semelhança) e 1 (máxima similaridade) e são:

Dice : $Sd = \frac{2c}{a+b}$

Overlap : $So = \frac{c}{\min(a,b)}$

Coseno : $Sc = \frac{c}{\sqrt{a.b}}$

Uma série de experimentos foi realizada para verificar a aplicabilidade da análise estrutural na realização de pesquisas online em banco de dados. Para essas experiências foi desenvolvido um sistema computacional interativo, que, para cada termo de pesquisa introduzido pelo usuário, recupera e apresenta no terminal todos os termos do banco de dados que apresentem um coeficiente de similaridade acima de um mínimo especificado. Dessa forma o pesquisador pode expandir sua consulta com termos significativos e realmente utilizados na indexação da base de dados.

O sistema desenvolvido baseia-se em um arquivo invertido de n-gramas; cada registro do mesmo contém um n-grama e um apontador para uma lista ligada que contém o número de cada termo que contém o dito n-grama e sua frequência de ocorrência.

Cada termo submetido ao sistema é decomposto em seus n-gramas, e, através do arquivo invertido, são identificadas as palavras de possível interesse. Para cada uma delas é então calculada a similaridade através do coeficiente de Dice. Os experimentos foram realizados para digramas e para trigramas, e foram utilizadas 2 coleções de documentos, em inglês, com dicionários de aproximadamente 5000 e 12000 termos.

OS EXPERIMENTOS

Na Figura II é apresentado um exemplo de utilização do sistema. Nos experimentos realizados, foi determinada, para cada termo apresentado pelo sistema, sua relevância ou não em relação ao termo de pesquisa. Diversos foram os critérios adotados para estabelecer relevância, destacando-se os seguintes:

a) Palavras com a mesma estrutura básica do termo de pesquisa e semanticamente relacionadas.
Ex.: ELÁSTICO e INELÁSTICO, ELASTICIDADE, etc.

b) Casos certos de erros de grafia.

Não foram consideradas relevantes à pesquisa:

a) Palavras com a mesma estrutura básica do termo de pesquisa, mas semanticamente não relacionadas.
Ex.: INFLAÇÃO e INFRAÇÃO

b) Palavras semanticamente relacionadas, mas recuperadas em decorrência da similaridade de afixos.
Ex.: LIQUIDO e FLUIDO

c) Casos duvidosos de erros de grafia. Ex.: o termo do dicionário difere em uma letra de duas ou mais palavras distintas.

Nos experimentos foram utilizados cerca de 270 termos de pesquisa, sendo estabelecidos, para cada um, níveis mínimos de similaridade variando entre 0.80 e 0.40 com intervalo de — 0.05. Obviamente, à medida que se reduz o nível mínimo de similaridade, aumenta o número de palavras recuperadas.

```
 INPUT QUERY TERM
>DIMENSION
 INPUT TERM =  DIMENSION#
MAXIMUM MATCH LEVEL= 9TRIGRAMS IN
COMMON
 ENTER LOWEST DESIRED MATCH COEF. ***DO
not ENTER ZERO***
>0.55


BONEDIMENSIONAL<#=                0.58
DIMENSION#                        1.00
DIMENSIONAL=#=                    0.80
DIMENSIONLESS#=                   0.73
DIMENSIONS#                       0.84
EXTENSION#                        0.56
NONDIMENSIONAL#-                  0.61
ONEDIMENSION#                     0.76
ONEDIMENSIONAL#                   0.61
TENSION#                          0.63
THREEDIMENSIONAL#                 0.56
TWODIMENSION#                     0.76
TWODIMENSIONAL#=                  0.61
TWODIMENSIONS#                    0.64

ENTER: 0 TO STOP PROGRAM
          1 FOR ANOTHER QUERY


 INPUT QUERY TERM
> *PLANE
 INPUT TERM = PLANE#
MAXIMUM MATCH LEVEL = 4 TRIGRAMS IN
COMMON

AEROPLANE#                        1.00
AIRPLANE#                         1.00
BIPLANE#                          1.00
FOREPLANE#                        1.00
INPLANE#                          1.00
MIDPLANE#                         1.00
MONOPLANE#                        1.00
PLANE#I                           1.00
TAILPLANE#                        1.00

ENTER:  0  TO STOP PROGRAM
          1 FOR ANOTHER QUERY
```

```
INPUTQUERYTERM
>  STABILIZE
INPUT TERM =
MAXIMUM MATCH LEVEL= 8 TRIGRAMS IN
COMMON
ENTER LOWEST DESIRED MATCH COEF. ***DO
not ENTER ZERO***
>0.50


DESTABILIZING#                    0.55
INSTABILITY#                      0.50
STABILISE#                        0.67
STABILISED#                       0.63
STABILISER#                       0.63
                                  0.60
                                  0.67
STABILIZATION#                    0.64
STABILIZED#                       0.84
STABILIZER#                       0.84
STABILIZING#                      0.70


ENTER:  0  TO STOP PROGRAM
          1 FOR ANOTHER QUERY


INPUT QUERY TERM
>  *FORM
INPUT TERM = FORM#
MAXIMUM MATCH LEVEL = 3 TRIGRAMS IN
COMMON

CLOSEDFORM#                       1.00
CRUCIFORM#                        1.00
FORM#                             1.00
FUSIFORM#                         1.00
NONUNIFORM#                       1.00
PLANFORM#                         1.00
STRATIFORM#                       1.00
UNIFORM#                          1.00
WAVEFORM#                         1.00


ENTER: - 0 TO STOP PROGRAM
          1 FOR ANOTHER QUERY

>0
     END OF PROGRAM
```

FIGURA II

O cálculo do coeficiente de precisão é imediato, a partir da decisão quanto à relevância ou não dos termos recuperados. Para o cálculo da revocação, foi necessário fazer uma aproximação em virtude do desconhecimento do número total de termos relevantes no dicionário. Em vez deste, foi adotado para os cálculos o número de termos relevantes recuperados para um nível mínimo de similaridade igual a 0,40. Alguns testes manuais mostraram que essa aproximação não introduz um erro significativo nos cálculos.

O objetivo do cálculo da precisão e revocação foi determinar um ponto de equilíbrio para o nível mínimo de similaridade, de modo a ter um número significativo de termos relevantes, sem com isso aumentar excessivamente o número total de palavras recuperadas.

Em cada nível de similaridade (ns) foram determinados o número de termos relevantes recuperados (RR_{ns}) e o número de termos não-relevantes recuperados (NR_{ns}), calculando-se precisão e revocação, respectivamente, por:

$$P = \frac{RR_{ns}}{RR_{ns} + NR_{ns}}$$

$$R = \frac{RR_{ns}}{RR_{0.40}}$$

Na Tabela I e Figura III são apresentados os resultados obtidos para os experimentos com um dicionário de 12000 termos.

TRUNCAMENTO ARBITRÁRIO

O mesmo sistema desenvolvido para os experimentos citados anteriormente foi utilizado para avaliação da aplicabilidade da técnica de análise estrutural para casos de truncamento arbitrário em termos de pesquisa. Trata-se do problema de expandir, no dicionário, termos de pesquisa truncados à direita, à esquerda ou ambos:

AMPLIF* deve recuperar AMPLIFICADOR
AMPLIFICADORES
AMPLIFICAÇÃO, etc.

*STAVE L deve recuperar ESTÁVEL
INSTÁVEL
ASTÁVEL, etc.

*SIMETR*deve recuperar ASSIMÉTRICO
ASSIMETRIA
SIMÉTRICO, etc.

O truncamento à direita (pesquisa por prefixos) é um caso trivial em sistemas **online.** Todos os sistemas apresentam essa facilidade, pois a pesquisa por prefixos é extremamente simples em dicionários

TABELA I

	no. of retrieved words		
treshold value	related	No-relat	total
0.80	1.74	0.03	1.77
0.75	2.90	0.13	3.03
0.70	4.30	0.36	4.66
0.65	5.18	0.92	6.10
0.60	6.28	2.25	8.53
0.55	7.48	4.95	12.43
0.50	8.39	11.31	19.70
0.45	8.93	18.88	27.81
0.40	9.85	46.87	56.72

1.1 -DIGRAMS

	no. of retrieved words		
threshold value	related	non-relat	total
0.80	1.26	0.01	1.27
0.75	1.95	0.04	1.99
0.70	3.23	0.12	3.35
0.65	4.10	0.19	4.29
0.60	5.09	0.61	5.70
0.55	6.05	1.17	7.22
0.50	7.15	2.26	9.41
0.45	7.73	3.39	11.12
0.40	8.69	7.86	16.55

I.2-TRIGRAMS

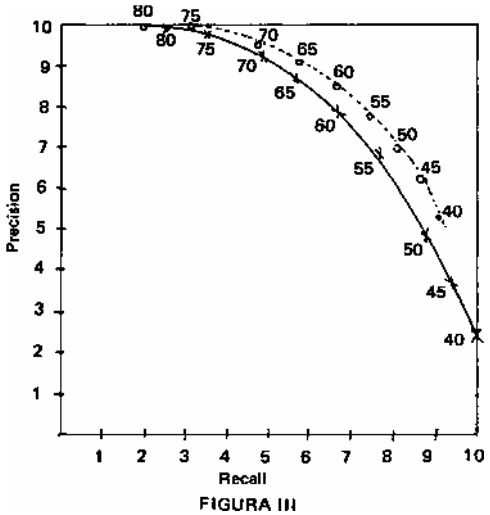

FIGURA III

ordenados alfabeticamente. Um termo truncado à esquerda, porém, poderá estar em qualquer posição do dicionário, dificultando em muito a pesquisa. O sistema DIALOG da Lockheed[8] não permite truncamento à esquerda. O sistema ORBIT da SDC[9] realiza uma pesquisa caráter-a-carâter no texto completo, para resolver possíveis casos de truncamento à esquerda.

Os sistemas que permitem truncamento à esquerda com pesquisa em dicionário não apresentam soluções eficientes. Já foi proposta a inclusão de todas as palavras do dicionário em ordem inversa, invertendo-se também, na pesquisa, o termo truncado. Bratley e Choueka[1 °] propõem uma técnica para resolver truncamentos genéricos baseada numa permutação cíclica de todas as palavras do dicionário, caráter por caráter. Isso implica um overhead médio de 800% do dicionário. A mesma técnica é utilizada por 3RIP[11], um sistema sueco para pesquisa interativa em bancos de dados. Todos esses sistemas propostos são, porém, ineficientes em termos de tempo de pesquisa ou armazenamento do dicionário.

A utilização da análise estrutural mostrou experimentalmente ser uma solução adequada para o problema. No sistema desenvolvido fez-se uma pequena alteração no cálculo do coeficiente de similaridade para expandir termos truncados. No caso de termos truncados, o coeficiente de similaridade é calculado por

$$S_{„} = \frac{Nc}{Np}$$

onde Nc é o número de n-gramas em comum, e
Np é o número de n-gramas do termo de pesquisa (truncado)

Nesse caso também, o nível mínimo de similaridade é fixado em 1,0 (um), o que assegura a recuperação de todos os termos com o radical desejado.

Para os experimentos de truncamento, portanto, o índice de revocação foi sempre igual a 1,0, restando a determinação da precisão para avaliar a técnica. A Tabela II mostra os resultados obtidos trabalhando com digramas e com trigramas.

TABELAM

TRUNCAMENTO ARBITRÁRIO

	Termos	Prec=1.0	Total	Média	OK	Prec.
dig.	51	44	607	11.9	582	0.95
trig.	51	51	582	11.4	582	1.00

CONCLUSÕES

Os resultados experimentais foram bastante satisfatórios e encorajadores. Um sistema de expansão de pesquisas online baseado em análise estrutural poderá se tornar realidade em sistemas comerciais, pois:

1 — Os valores de precisão e revocação indicam ser possível obter-se um fndice muito bom de termos relevantes, mesmo com dicionários maiores, de 50.000 a 100.000 termos, sem necessidade de exame visual de um número excessivo de palavras. Em média, foram apresentados na tela menos de 10 termos, trabalhando com nível de similaridade igual a 0.50, com índice de precisão igual a 0.82.

2 — O sistema pode ser implantado sem grandes transtornos, até mesmo em minicomputador, com um overhead de cerca de 110% no dicionário, o que é bastante aceitável em sistemas online. Durante os experimentos, trabalhando com minicomputador, observou-se um tempo de resposta de cerca de 1 a 2 segundos.

Os resultados experimentais mostram um significativo aumento de eficiência trabalhando com trigramas. No caso de digramas, o nível de similaridade mais indicado situa-se entre 0.55 e 0.60, dependendo do tamanho do dicionário. Já no caso de trigramas, esse nível de similaridade pode ser baixado para 0.45 a 0.55, sem comprometimento da precisão.

A análise estrutural revelou-se uma solução possível para os problemas de truncamento arbitrário em sistemas online. Trabalhando com trigramas, obteve-se um índice de precisão de 100%, e tudo indica que um desempenho semelhante será obtido para dicionários maiores.

No presente artigo procuramos apresentar a técnica de análise estrutural aplicada à pesquisa online em bancos de dados. Ao longo do texto apresentamos exemplos em língua portuguesa para facilitar a compreensão do mesmo. Os experimentos reais, porém, foram realizados com dicionários de termos em língua inglesa. Acreditamos que resultados semelhantes serão obtidos com o português ou qualquer outra língua ocidental, recomendando essa técnica para aplicações em que é necessária a formulação ou expansão de uma pesquisa online em quaisquer bases de dados. A técnica de análise estrutural é particularmente indicada para usuários cuja língua-mãe não é a da base de dados, como é o caso da grande maioria das pesquisas realizadas no Brasil, por ajudar na identificação de palavras e de grafias empregadas.

Foram apresentados alguns dos resultados obtidos nas pesquisas efetuadas com análise estrutural. As conclusões sobre aplicabilidade e eficiência da técnica baseiam-se também em outros resultados não apresentados neste artigo[12].

REFERÊNCIAS BIBLIOGRÁFICAS

[1] TOMBERG, Alex, ed. Eusidic Database Guide 1981. Eusidic, 1980.

2 DATABASES Online-Online Review, 5 (1), 1981.

3 LYNCH, M.F. Computer-based Information services in Science and technology — principies and iechniques. Stevenage, Peter Peregrinus, 1974.

[4] PREECE, S. E. An online associative query modification methodology. Online Review, 4 (4), 1980.

5 BOURIME, C.P. Frequency and impact of spelling errors in bibliographic databases. Information Processing and Management, 13,1977.

6 ODELL, M.K. & RUSSEL, R.C. The soundex system. Patentes americanas n. 1, 261, 167, 1918 e 1,435, 663, 1922.

[7] ADAMSOIM, G.W. e BOREHAM, J. The use of an association measure based on character structure to identify semantically related pairs of words and document titles. Information Storage and Retrieval, 10, 1974.

8 GUIDE to DIALOG searching. Paio Alto, Lockheed Dialog Information Retrieval Service, 1979.

[9] ORBIT User Manual. Santa Monica, SDC Search Service, 1979.

10 BRATLEY, P. e CHOUEKA, Y. Processing truncated terms in document retrieval systems. A ser publicado.

[11] 3RIP — A System Manual. Estocolmo, Paralog AB, 1980.

12 FREUND, G.E. An investigaiion into word confation techniques as an aid to online searching. Tese de Mestrado, University of Sheffield, Inglaterra, 1981.

ABSTRACT

Online searches are still a challenge for most data base users. The main problem is to identify the most suitable keywords in each data base. Structural analysis, a technique based on the syntactic analysis of words with the objective of identifying words semantically related, is presented as a solution to various problems, including the problem of the arbitrary truncating of words, not yet well solved in online systems. Theoretical and practical aspects of a research by the author at the University of Sheffield are also presented. (JMK)