Igrep - Um Sistema para Busca Aproximada em Textos Indexados*

Márcio Drumond Araújo Nivio Ziviani

Universidade Federal de Minas Gerais
Departamento de Ciência da Computação
{drumond,nivio}@dcc.ufmg.br

Resumo

O presente trabalho apresenta o sistema IGREP para busca aproximada em textos de grande porte desenvolvido através da dissertação [Ara97]. O IGREP faz uso de uma lista invertida, composta por uma tabela contendo o vocabulário de palavras do texto e uma lista de endereços no texto correspondendo às ocorrências no mesmo de cada palavra do vocabulário. O tamanho do vocabulário é menos do que 1% do tamanho do texto como um todo, fazendo com que seja possível a sua manutenção em memória principal durante todo o processo de busca. Para consultas contendo uma palavra, a busca fica restrita apenas ao vocabulário. Para consultas contendo mais de uma palavra, a busca fica restrita ao vocabulário e respectivas listas de endereços, não havendo pois nenhum acesso ao texto na fase de pesquisa. O sistema permite desde a busca com erros ou não de palavras e frases até buscas de seqüências complexas contendo conjuntos de caracteres, caracteres coringa e expressões regulares arbitrárias. O tempo de busca é $O(\sqrt{n})$ para casos típicos. Os resultados experimentais mostram que o sistema funciona bem na prática: para um texto de 1 *gigabyte*, casamentos compostos por 3 palavras com até 1 erro são recuperados em aproximadamente 6 segundos. No caso de busca sem erro, as ocorrências são obtidas em menos de meio segundo. O sistema contém duas ferramentas: o IGREPINDEX destinado a construção da lista invertida e o IGREP responsável pela parte de busca de padrões.

1 Motivação para a Criação do Sistema

O problema da busca em texto dentro da área de recuperação de informação tem ganhado muita popularidade ultimamente. Nesses problemas, o usuário expressa sua necessidade de informação através de seqüências ou padrões fornecidos para serem pesquisadas e o sistema de informação recupera as posições onde tais seqüências ocorrem dentro do texto. Quando o texto é muito grande, normalmente utiliza-se alguma técnica de indexação para uma maior eficiência. Uma técnica simples e popular de indexação é a lista invertida. Ela é adequada quando o padrão a ser pesquisado é formado por palavras comuns. Como esse

*Este trabalho foi realizado com apoio do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e Projeto RITOS/CYTED.

é um tipo de busca extremamente comum, por exemplo, em sistemas de busca utilizados na Internet, a lista invertida tem se tornado bastante popular em tal contexto.

Entretanto, uma fraqueza que os sistemas para busca em textos de grande porte comercialmente disponíveis é a necessidade de se fazer a busca de padrões considerando somente a grafia correta dos mesmos devido à utilização de estruturas de indexação. No presente trabalho é apresentado justamente um sistema eficiente, chamado IGREP, para a recuperação de padrões de forma exata e aproximada em textos de grande porte através do uso de uma lista invertida. Na busca aproximada procura-se todas as seqüências no texto que estão numa dada distância menor ou igual a k do padrão procurado. O sistema IGREP é eficiente tanto na parte de construção do índice quanto na parte de busca. O índice pode ser construído em tempo linear na prática para textos da ordem de 1 *gigabyte*, com complexidade de espaço também $O(n)$. O tempo de busca é proporcional a $O(\sqrt{n})$ e $O(\sqrt{n}\log_2 n)$ nos casos mais típicos.

O sistema tem sido testado de forma bem sucedida em textos da ordem de 1 *gigabyte*. Ele ainda suporta diversas variações para o problema da busca aproximada. Além da busca de palavras e frases comuns, o sistema permite a busca de padrões contendo conjuntos de caracteres (intervalos, conjuntos arbitrários, complementos e caracteres coringa), buscas aproximada e exata sendo feitas num mesmo padrão e expressões regulares arbitrárias.

2 O Processo de Busca e a Lista Invertida

O índice utilizado no presente trabalho é baseado em uma lista invertida que possui dois componentes: o primeiro contém todas as palavras distintas do texto (ou seja, o vocabulário) e o segundo contém os endereços das ocorrências de todas as palavras do vocabulário do texto, na ordem em que cada ocorrência aparece.

A figura 1 ilustra a estrutura da lista invertida para um texto exemplo contendo seis palavras. Cada posição do vocabulário contém uma palavra distinta do texto e uma referência para a última posição de sua respectiva lista de ocorrências.

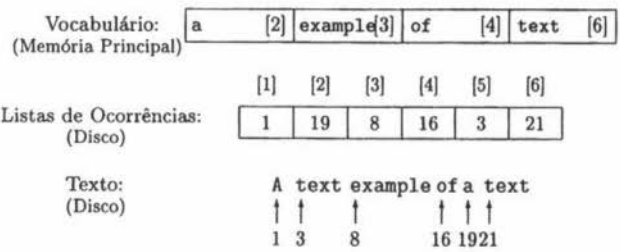

Figura 1: Estrutura da lista invertida

Uma lista invertida como a utilizada pelo sistema geralmente representa um espaço adicional em disco da ordem de 60% do tamanho do texto, sendo que o algoritmo utilizado é o desenvolvido por Moffat e Bell [MB95], sem uso de compressão. Outro resultado prático obtido é que apesar de tal algoritmo ser $O(n\log_2 n)$, para textos de até um 1 *gigabyte* ele tem apresentado um comportamento linear na prática.

Para responder a uma busca do usuário, apenas o vocabulário e as listas de ocorrências são necessárias, não havendo necessidade de acessos a texto em nenhum momento. Para padrões compostos por uma única palavra, a busca se restringe ao vocablário. Se algum casamento for detectado no vocabulário, automaticamente o sistema já obtém todas as ocorrências daquela palavra no texto através das listas de ocorrências. Quando o padrão contém mais de uma palavra, o sistema procura no vocabulário por cada palavra separadamente recuperando as respectivas listas de ocorrências. Com as listas em mãos, o IGREP realiza o que é chamado interseção das listas, procurando por referências que têm o mesmo posicionamento relativo encontrado dentro do padrão. A cada ocorrência em que a distância relativa entre as diferentes palavras casadas com as palavras do padrão satisfazem o mesmo distanciamento encontrado entre as palavras no padrão, tem-se um casamento completo.

Para ilustrar o processo de busca no índice são apresentados dois exemplos. A busca exata do padrão "`text`" na figura 1 envolve uma busca binária no vocabulário, obtendo a lista de ocorrências equivalente ao intervalo $[5,6]$ no arranjo de listas. O outro exemplo considera a busca com 2 erros do padrão "`text sample`". Primeiro o sistema busca a palavra "`text`" obtendo o intervalo $[5,6]$ no arranjo de listas. Depois, ele busca a palavra "`sample`" retornando como casamento a palavra "`example`" com dois erros correspondente ao intervalo $[3,3]$. As duas listas contém os endereços $\{3,21\}$ e $\{8\}$ respectivamente, correspondendo aos endereços contidos nos intervalos $[5,6]$ e $[3,3]$ do arranjo de listas. Através da interseção entre as duas listas de ocorrências, o sistema indica que na posição 3 há um casamento do padrão como um todo, levando em conta que a seqüência "`text example`" apresenta a mesma distância relativa entre suas palavras em relação ao padrão "`text sample`".

3 A Sintaxe do Padrão

A seguir são mostrados alguns exemplos que ilustram a sintaxe permitida para o padrão.

- "`t[a-z]xt`": palavras iniciadas por "`t`", seguido de qualquer caractere entre "`a`" e "`z`", seguido de "`xt`".
- "`t[aeiou]xt`": palavras iniciadas por "`t`", seguido de qualquer vogal, seguido de "`xt`".
- "`t[^aeiou]xt`": palavras iniciadas por "`t`", seguido de qualquer caractere diferente de vogal, seguido de "`xt`".
- "`t.xt`": palavras iniciadas por "`t`", seguido de qualquer caractere, seguido de "`xt`".
- "`t#xt`": palavras iniciadas por "`t`", seguido de qualquer caractere do alfabeto zero ou mais vezes em qualquer ordem, seguido de "`xt`".
- "`t(e|ai)xt`": palavras iniciadas por "`t`", seguido de "`e`" ou de "`ai`", seguido de "`xt`". Nesse caso, o padrão é visto como uma expressão regular.
- "`t(e|ai)*xt`": palavras iniciadas por "`t`", seguido de "`e`" ou de "`ai`" zero ou mais vezes em qualquer ordem, seguido de "`xt`". Esse padrão também é uma expressão regular.

- "<te>xt": o sistema busca as ocorrências da palavra "text" considerando que o prefixo "te" tem de ocorrer de forma exata e o sufixo "xt" pode ocorrer de forma aproximada.

O sistema ainda permite a busca insensível ao caso, sem diferenciar letras maiúsculas de minúsculas (o normal é diferenciá-las). O usuário também tem a possibilidade de atribuir pesos diferentes a cada um dos erros considerados pelo conceito de distância de edição. Normalmente, o sistema considera cada um dos erros com peso unitário.

4 Desempenho do Sistema

A tabela 1 mostra a razão entre os tempos de busca do IGREP e do GLIMPSE versão 3.0 [MW93], um outro sistema voltado para a busca aproximada através de uma estrutura de indexação semelhante à utilizada no presente trabalho. Os testes consideram diferentes taxas de erros ($k = 0, 1, 2$) em frases compostas por 1, 2, 3, 4 e 5 palavras em um texto da ordem de 250 *megabytes*. O texto utilizado engloba as edições 1987, 1988 e 1989 do Wall Street Journal (WSJ), obtido da coleção TREC [Har95]. Mesmo em textos da ordem de 1 *gigabyte* os tempos são bem satisfatórios, sendo que as buscas exatas nessa situação sempre são feitas em menos de meio segundo. A busca aproximada com 1 erro de um padrão contendo 3 palavras leva em torno de 6 segundos.

	Número de palavras da consulta									
	1		2		3		4		5	
k	Igrep	$\frac{\text{Igrep}}{\text{Glimpse}}$	Igrep	$\frac{\text{Igrep}}{\text{Glimpse}}$	Igrep	$\frac{\text{Igrep}}{\text{Glimpse}}$	Igrep	$\frac{\text{Igrep}}{\text{Glimpse}}$	Igrep	$\frac{\text{Igrep}}{\text{Glimpse}}$
0	0.08	0.3%	0.23	0.9%	0.24	1%	0.28	1%	0.34	1%
1	0.58	0.4%	1.99	1.5%	2.15	1.6%	2.59	1.9%	3.16	*
2	0.85	0.5%	8.27	5.1%	4.26	2.6%	4.65	2.9%	5.06	*
3	1.30	0.7%	34.1	17.9%	14.6	7.5%	11.2	*	8.97	*

* No GLIMPSE não é possível fazer a busca aproximada de um padrão com mais de 32 caracteres

Tabela 1: Tempos de busca (em segundos) e a razão dos tempos do IGREP sobre os do GLIMPSE para um texto com cerca de 250 *megabytes*

5 Plataformas Disponíveis

O sistema IGREP já foi testado de forma bem sucedida nas plataformas SunOS 4.1.4 e Solaris 2.5.1 para arquitetura Sun Sparc e no Linux 2.0 para arquitetura Intel i386.

6 Características do Código

O IGREP foi implementado na linguagem *ANSI C*, contando hoje com cerca de 390 *kbytes* de código fonte (cerca de 33 mil linhas de código). O compilador utilizado em todas as plataformas foi o gnu gcc 2.7.2.

7 Disponibilidade do Código

Atualmente, o sistema está na versão 1.0, disponível através de FTP anônimo no endereço `ftp://ftp.dcc.ufmg.br/pub/research/nivio/igrep`.

8 Utilização

O sistema IGREP foi desenvolvido primordialmente para plataformas UNIX. Tanto o IGREPINDEX quanto o IGREP podem ser instanciados via linha de comando a partir do interpretador de comandos (*shell*). Foram também desenvolvidos duas páginas de manual em formato UNIX explicando a utilização das ferramentas do sistema, disponíveis junto com o sistema no endereço FTP citado acima.

Referências

[Ara97] M. D. Araújo. Igrep - um sistema para busca aproximada em textos indexados. Master's thesis, Department of Computer Science, Universidade Federal de Minas Gerais, March 1997. (Supervisor: N. Ziviani).

[Har95] D. K. Harman. Overview of the third text retrieval conference. In *Proc. Third Text REtrieval Conference (TREC-3)*, pages 1–19, Gaithersburg, Maryland, USA, 1995. National Institute of Standards and Technology Special Publication.

[MB95] A. Moffat and T. Bell. In situ generations of compressed inverted files. *Journal of the American Society for Information Science*, 46(7):537–550, 1995.

[MW93] W. Manber and S. Wu. Glimpse: A tool to search through entire file systems. Technical Report 93-34, Dept. of Computer Science, The University of Arizona, Oct. 1993.