1
Motivação

A presente pesquisa apresenta as primeiras hipóteses para o desenvolvimento de um novo padrão para mecanismo de buscas de arquivos de texto. Através desse mecanismo, o usuário escreverá a informação que ele quer encontrar e o programa lhe retornará os documentos em que essa informação apareça, sem que seja necessário ao usuário prever quais palavras o autor do texto usou para expressar a informação requerida. A esse novo padrão de busca dei o nome de orientado a idéia, em contraposição ao padrão mais comum de busca que estamos chamando orientado a palavra.

A área de conhecimento em que essa pesquisa se insere é a da Recuperação de Informação (RI), para a qual advogamos a necessidade de conhecimento lingüístico. A Recuperação de Informação normalmente trata de duas problemáticas: i) a identificação de um documento dentre outros num acervo, a partir do pedido de um usuário na forma de texto, e ii) a catalogação dos documentos no acervo conforme assunto, resumo ou outros critérios de indexação.

Para nós só será interessante tratar a primeira problemática.

Normalmente a recuperação de informação se limita a recuperar arquivos e não informações (Ferneda, 2003). Mas muitas vezes é informação aquilo que o usuário pensa estar recuperando, ou que gostaria de estar recuperando. Portanto, recuperar a informação é o requisito primordial a ser atendido pelo mecanismo de busca.

Essa limitação gera a imensa dificuldade que o usuário tem em tentar adivinhar quais palavras deve usar para conseguir o resultado desejado. Imagine-se uma situação corriqueira como usar o terminal de computador de uma biblioteca que rode um mecanismo de busca para os livros de seu acervo. Se o usuário resolver encontrar um livro que contenha uma determinada informação, terá muita dificuldade para descobrir com que palavras essa informação terá sido classificada pelos bibliotecários, se é que a informação foi catalogada.

Essa dificuldade fica bem mais evidente quando os arquivos buscados não se limitam a itens de um banco de dados, como no exemplo da biblioteca, mas quando o documento procurado é um arquivo de texto. As páginas HTML da internet são

exemplos disso. Os arquivos de texto permitem um campo mais profícuo para a Recuperação de Informação, pois a busca pode recair não apenas nos recursos usados para a catalogação (títulos, resumos ou outras formas de indexação), mas, como o próprio conteúdo do arquivo é composto por palavras, tal conteúdo também pode ser alvo do mecanismo de busca.

Nossa pesquisa se limita a procurar soluções, com base em informações lingüísticas, para buscadores cuja pesquisa se faz sobre documentos do tipo arquivo de texto.

1.1
A busca orientada a palavra

Os buscadores de páginas da internet não recuperam informação, recuperam palavras. O mecanismo de busca "enxerga" uma página simplesmente como uma seqüência de palavras (entendidas como seqüências de caracteres) e sua tarefa é encontrar os documentos que contenham as mesmas palavras que o usuário requisitou em sua busca.

O usuário não pode simplesmente requisitar uma informação, que pode estar expressa através de diferentes seqüências de palavras.

Entre o que o usuário deseja e o que ele requisita, portanto, há uma série de estratégias que ele deve utilizar para melhorar os acertos nos resultados apresentados pelo buscador. O usuário tem de prever as palavras que não podem faltar aos documentos que apresentem a idéia buscada, sendo que ainda deve evitar, se possível, as palavras que possam ser comuns a textos que não apresentem a informação desejada.

Os principais problemas desse padrão de busca são: i) o usuário pode perder muito tempo tentando prever as palavras que compõem os documentos que satisfaçam sua busca, e ii) podem ser retornados como resultados muitos documentos irrelevantes por apresentarem as mesmas palavras usadas pelo usuário em sua busca.

O que leva os buscadores ao primeiro problema é não relacionar sinônimos ou paráfrases, o que obriga o usuário a tentar uma série de formas diferentes para expressar a mesma idéia do que quer encontrar. O outro problema é causado pela

incapacidade dos buscadores de perceber os significados que estão sendo expressos tanto pelo usuário em seus requisitos de busca quanto pelos textos contidos nos sites.

A funcionalidade esperada para um sistema de recuperação da informação contida em textos seria recuperar informações, não palavras. Idealmente, o usuário poderia digitar o que espera encontrar, sem se preocupar em prever com exatidão a forma como essa idéia estará escrita nos documentos que poderiam servir como resultados e sem se preocupar com a possibilidade de que as palavras que ele usou em sua pesquisa possam aparecer em outros documentos sem expressar a idéia desejada. Daí a urgência para a criação de um mecanismo de busca orientado a idéia.

1.2 A busca orientada a idéia

Os índices de *precision* e de *recall*[1] de um mecanismo de busca em textos, sugere-se aqui, podem ser aumentados a partir do tratamento lingüístico dos documentos pesquisados e das requisições dos usuários.

A *precision* pode estar ligada aos significados das palavras. Os documentos que são encontrados pelo buscador e que não nos interessam são aqueles que apresentam as palavras usadas pela pesquisa, mas com significados diferentes daquele que queremos. Então, para melhorarmos os índices de *precision* das buscas, o buscador deveria ser capaz de definir qual é o significado de cada palavra nos documentos e nas requisições de busca.

O *recall* pode estar relacionado à paráfrase. Os documentos que têm as informações desejadas, mas que não aparecem listados como resultados pelo buscador, possuem as mesmas idéias que o usuário procura, mas expressas com palavras diferentes daquelas preenchidas como pedido de busca pelo usuário. Para aumentar o *recall*, a busca não deveria se limitar a procurar as palavras exatamente

[1] Dentre os métodos usados para avaliar um programa de computador que recupere informações, como é o caso dos mecanismos de busca, existem duas medidas de avaliação muito utilizadas, a *precision* e o *recall*. A *precision* se refere à quantidade de resultados satisfatórios apresentados pela busca em relação ao número total de resultados apresentados. O *recall* se refere à quantidade de resultados satisfatórios apresentados pela busca em relação ao total de documentos satisfatórios que deveriam ter sido apresentados pelo buscador.

como o usuário as digitou, mas deveria ser capaz de encontrar textos que expressem a mesma idéia que o usuário procura, independentemente da forma expressa.

Nesse novo padrão, os textos não deverão ser entendidos como seqüências de palavras, mas como conjuntos de idéias relacionadas entre si. E o mesmo ocorre com os pedidos de busca do usuário. Como o buscador deverá ser capaz de "compreender" a idéia buscada pelo usuário, seus pedidos não podem ser expressos por palavras soltas, sem nenhuma relação de sentido entre elas. A forma pela qual as palavras estabelecem uma relação de sentido entre si, como veremos mais adiante, é através da relação sintática. Portanto, para que os pedidos de busca também sejam entendidos como idéias relacionadas, eles deverão ser expressos como pequenos textos (uma frase, uma expressão etc.). E a busca deverá se concentrar em encontrar os documentos que contiverem as idéias pedidas, de tal forma que seja uma busca por um pequeno conjunto de idéias (os pedidos), dentro de um conjunto maior de idéias (os arquivos de texto).

A partir de agora, será apresentada uma proposta de modelo de representação de textos para a utilização em buscadores orientados a idéia. Este modelo, ainda embrionário, visa reduzir um texto a um conjunto de idéias tratáveis computacionalmente.