PROGRAMA DE PÓS-GRADUAÇÃO EM SISTEMAS E PROCESSOS INDUSTRIAIS – MESTRADO

Antonio Carlos Alves

GIRS – GENETIC INFORMATION RETRIEVAL SYSTEM

Santa Cruz do Sul, outubro de 2009.

Antonio Carlos Alves

GIRS – GENETIC INFORMATION RETRIEVAL SYSTEM

Dissertação apresentada ao Programa de Pós-Graduação em Sistemas e Processos Industriais – Mestrado, área de concentração em Controle e Otimização de Processos Industriais, Universidade de Santa Cruz do Sul, como requisito parcial para a obtenção do título de Mestre em Sistemas e Processos Industriais.

Orientador: Prof. Dr. Jacques N. C. Schreiber
Co-Orientador: Prof. Dr. João Carlos Furtado

Santa Cruz do Sul, outubro de 2009

Antonio Carlos Alves

GIRS – GENETIC INFORMATION RETRIEVAL SYSTEM

Esta dissertação foi apresentada ao Programa de Pós-Graduação em Sistemas e Processos Industriais – Mestrado, área de concentração em Controle e Otimização de Processos Industriais, Universidade de Santa Cruz do Sul, como requisito parcial para a obtenção do título de Mestre em Sistemas e Processos Industriais.

Dr. Jacques Nelson Corleta Schreiber
Professor Orientador

Dr. João Carlos Furtado
Professor Co-Orientador

Dr. Raul Sidnei Wazlawick

Dra. Rejane Frozza

AGRADECIMENTOS

Aos meus pais Fernando e Teresinha, que tornaram esta trajetória possível, graças as suas dedicação, apoio e educação e sobre tudo com seu carinho durante toda a minha vida.

A minha esposa Andrisa, que soube me apoiar nas horas difíceis, soube compreender a minha ausência, abdicou de minha companhia e sobre tudo me incentivou a seguir adiante nos momentos de desânimo com muito amor e dedicação.

A meus filhos Débora, Fernando e Vitor, que embora precisando de minha atenção, aceitaram a minha ausência, o meu cansaço e muitas vezes deixaram seus brinquedos para que eu pudesse estudar ou fazer meu trabalho.

Aos meus colegas da Vivendas Imobiliária e Seguros e em especial aos meus superiores, que compreenderam o meu esforço e compensaram de uma forma ou de outra a minha ausência em alguns momentos.

Ao professor Jacques Schereiber, que me orientou e estimulou durante a execução desta dissertação, não medindo esforços e acima de tudo me emprestando sua experiência e conhecimento, sem as quais não seria possível a concretização deste trabalho.

Ao professor João Carlos Furtado, que de igual forma foi de grande presteza e dedicação, me ajudando no desenvolvimento deste trabalho.

Aos meus colegas, professores, funcionários e amigos, com quem convivi durante estes últimos anos, em especial a Janaína, que com sua paciência e sempre solícita, tanto me auxiliou no que foi necessário, pois, sem os quais não seria possível atingir este objetivo.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pela bolsa de estudos que tornou possível a realização desta jornada.

Enfim a todos, que de alguma maneira me ajudaram direta ou indiretamente e, principalmente a DEUS por ter me iluminado e colocado todas estas pessoas maravilhosas em meu caminho.

RESUMO

Atualmente as empresas e as corporações têm uma grande quantidade de documentos internos (normas, contratos, resoluções, atas, comunicações, entre outros) gerados no seu cotidiano. Estes documentos são armazenados nos mais diversos locais e formatos, e desta forma, surge à necessidade de recuperar informação constante nestes documentos. O problema é que a quantidade e a diversidade destes documentos armazenados dificultam muito a localização das informações consideradas úteis pelos usuários destas coleções.

Além disso, ainda há uma grande dificuldade de encontrar as informações verdadeiramente relevantes, que muitas vezes ficam perdidas em meio a uma quantidade excessiva de documentos não tão relevantes.

Esta dissertação apresenta uma proposta de um sistema de recuperação de informações onde a relevância dos documentos recuperados evolua a cada interação através do uso de algoritmos genéticos e com a ajuda direta dos usuários deste sistema através de um *feedback* implícito, visando uma melhora na relevância dos documentos recuperados, a cada nova consulta.

Palavras-Chave: recuperação de informação, algoritmos genéticos, motor de busca, retro-alimentação de relevância

ABSTRACT

Currently, companies and corporations deal with a big amount of internal documents (rules, contracts, resolutions, records, communications, others) generated daily. Such documents are stored in several places and formats and, consequently, there is a need to restore information found in these documents. The problem is that the quantity and diversity of documents stored make it difficult for users to find useful information in the files.

Moreover, there is a great difficulty in finding truly relevant information that is often lost among a great number of less important documents.

This thesis brings a proposal of an information retrieval system where the value of restored documents increases in each interaction through genetic algorithms with direct help of the system users and trough an implicit feedback, aiming at the improvement of restored documents' relevance every new search.

Keywords: information retrieval, genetic algorithms, search engine, relevance feedback

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS

RI	Recuperação da Informação
SRI	Sistema de Recuperação da Informação
AG	Algoritmos Genéticos
NP	*Non-Deterministic Polynomial time*
GIRS	*Genetic Information Retrieval System*
TF	*Term frequency*
TF-IDF	*Term Frequency–Inverse Document Frequency*
TREC	*Text REtrieval Conference*
NIST	*National Institute of Standards and Technology*

SUMÁRIO

INTRODUÇÃO

O crescimento do conjunto de documentos no formato digital tem tornado cada vez mais necessária a criação de ferramentas que auxiliem os usuários a encontrar dados e informações dentro das bases que armazenam esses documentos. Esse é o objetivo básico dos mais diversos sistemas de busca, como o Google, Copernic e Yahoo, que respondem às necessidades dos usuários para uma determinada consulta, geralmente feita utilizando-se palavras-chave. Cada sistema utiliza um algoritmo específico para ordenar os documentos que serão sugeridos ao usuário em resposta à consulta efetuada. Nestes buscadores, serão recuperados todos os documentos que contenham as palavras-chave ou expressões, resultando em uma quantidade excessiva de documentos e na dificuldade de selecionar os documentos relevantes ao interesse do usuário.

O problema está na dificuldade de um usuário encontrar informações realmente relevantes e confiáveis para a realização de seu trabalho. Portanto, a recuperação de documentos com a informação relevante ao contexto constitui-se um problema e a busca de informações sobre este acervo uma tarefa árdua. As conseqüências são atrasos e imprecisão.

Diante disto, formulou-se a seguinte hipótese: “É possível criar um modelo de RI onde a qualidade dos resultados (relevância) evolua como consequência do *feedback* dos usuários?”.

Com o objetivo de provar a hipótese, buscou-se estudar e propor um sistema de recuperação de informações onde a relevância dos documentos recuperados evolua a cada interação através do uso de algoritmos genéticos e com a ajuda direta dos usuários deste

sistema através de um *feedback* implícito, visando uma melhora no tocante à relevância dos documentos recuperados, a cada nova consulta.

Como motivadores dessa dissertação de mestrado, pode-se citar:

- a crescente quantidade de informação a que os usuários têm sido submetidos;
- o desafio de encontrar a informação que interessa a esses usuários;
- a redução do tempo na busca por informações relevante nos documentos da base de dados;
- a melhora na qualidade da informação obtida.

Cabe destacar que esta dissertação concentra seu foco em propor um sistema de recuperação de informação para utilização em *intranets* onde a quantidade de documentos indexados é limitada o que torna possível a utilização de algoritmos genéticos (AG) nesta recuperação, e não na *internet*, onde tal técnica se tornaria inviável devido à quantidade muito grande de documentos e o consequente tempo excessivo para o processamento das buscas.

1.1 Objetivo Geral

O objetivo geral do estudo é analisar e propor um sistema de recuperação de informações onde a relevância dos documentos recuperados evolua a cada interação através do uso de algoritmos genéticos e com a ajuda direta dos usuários deste sistema através de um *feedback* implícito de relevância.

1.2 Objetivos Específicos

Pode-se citar como objetivos específicos deste estudo:

- estudar detalhadamente os Sistemas de Recuperação de Informação já implementados;

- analisar os métodos e técnicas de implementação dos Algoritmos Genéticos e adaptá-lo à um Sistema de Recuperação de Informações;
- estudar e implementar as métricas de validação dos sistemas de recuperação da informação;
- estudar e propor uma interface amigável para a pesquisa e, principalmente, para o *feedback* dos usuários;
- implementar um motor de busca que alie performance e credibilidade na recuperação de informações.

1.3 Organização do Texto

O texto desta dissertação foi dividido em dez capítulos, incluindo os capítulos da Introdução, Conclusão e Bibliografia.

No Capítulo 2 são apresentados os conceitos de Recuperação de Informação, os modelos de recuperação, indexação de documentos, relevância e as métricas precisão e revocação.

O Capítulo 3 apresenta um histórico sobre o desenvolvimento dos Algoritmos Genéticos, sendo descritos os conceitos de AG, seus principais aspectos e explanados os operadores genéticos, cujo entendimento se torna necessário para a compreensão do desenvolvimento desta dissertação.

O Capítulo 4 apresenta alguns aspectos referentes a trabalhos relacionados com os conteúdos desta dissertação.

No Capítulo 5 é apresentado a metodologia de pesquisa e onde é detalhado o protótipo desenvolvido, dividido em duas seções principais, sendo a primeira seção de Configuração do Sistema e a segunda seção de Consultas dos Usuários, onde são explicados os dois tipos de

buscas possíveis pelo sistema, ou seja, Busca Direta ou Busca pelo GIRS, onde são explanados todos os passos realizados pelo protótipo desde a consulta dos usuários até a apresentação final dos resultados. A descrição do protótipo foi publicada parcialmente em (ALVES, 2008).

O Capítulo 6 apresenta as metodologias de avaliação dos sistemas de recuperação da informação, bem como a metodologia utilizada neste trabalho.

O Capítulo 7 apresenta os resultados obtidos com as avaliações do Sistema GIRS.

A seguir são apresentadas as Conclusões, as sugestões de trabalhos futuros e as referências utilizadas nesta dissertação.

2. RECUPERAÇÃO DA INFORMAÇÃO

Em 1951, Calvin Mooers criou o termo "*Information Retrieval"* (Recuperação de Informação - RI), e definiu que "A Recuperação de Informações trata dos aspectos intelectuais de descrição da informação e sua especificação para busca, e também de qualquer sistema, técnicas ou máquinas que são empregadas para realizar esta operação". (Ferneda, 2003)

Segundo Baeza (1999), a recuperação de dados, no contexto da recuperação de informação, consiste em determinar quais os documentos de uma coleção contêm as palavras chaves da consulta que, com frequência, não são o suficiente para satisfazer à necessidade de informação deste usuário. Para ele, a recuperação de dados não resolve o problema de recuperar a informação sobre um assunto, para ser eficaz em sua tentativa de satisfazer à necessidade de informação do usuário, a recuperação de informação deve interpretar de acordo com um grau de relevância a solicitação do usuário. A dificuldade não consiste somente em extrair esta informação, mas também em usá-la para definir a relevância. Assim, a noção da relevância está no centro da recuperação de informação. "O objetivo preliminar de um SRI - Sistema de Recuperação de Informação é recuperar todos os documentos que são relevantes a uma solicitação do usuário com uma quantidade mínima de documentos não-relevantes." (Baeza, 1999)

A Figura 1 apresenta um esquema básico de um Sistema de Recuperação de Informação.

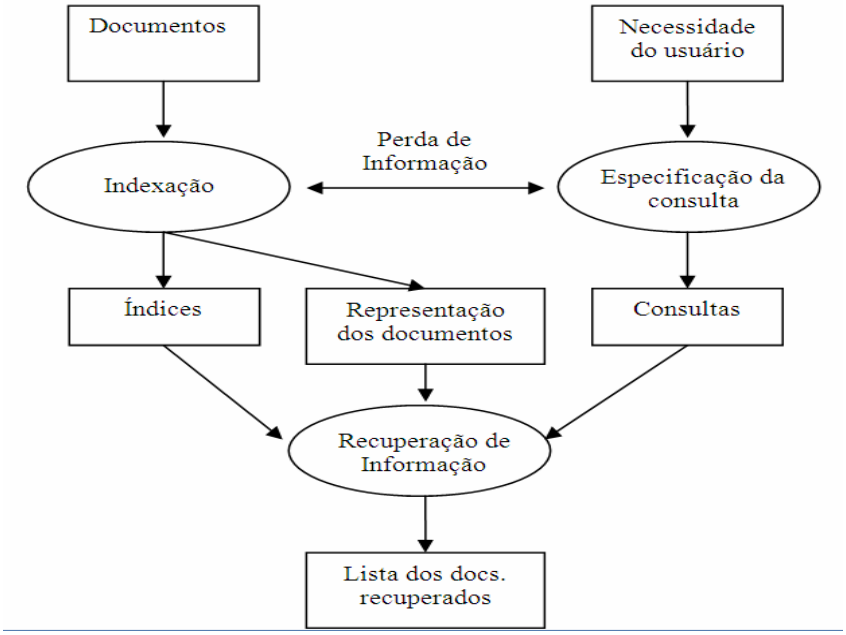

Figura 1 – Esquema básico de um Sistema de Recuperação de Informação
Fonte: Adaptado de Gey (1992)

Esse capítulo está organizado da seguinte forma: a primeira seção apresenta os principais modelos de recuperação de informação. A seção 2.2 trata sobre a indexação dos documentos, os principais tipos e suas peculiaridades. A seção 2.3 introduz as métricas Precisão e Revocação, essenciais para a avaliação dos sistemas de Recuperação de Informações. Na última seção são explanados os conceitos de *feedback* de relevância.

2.1 Modelos de recuperação de informação

A recuperação da informação tem como base o uso de termos de indexação, que representam a unidade básica de acesso à informação. A partir dessa unidade foram desenvolvidos vários modelos para facilitar o acesso à informação e melhorar a precisão do resultado da busca ou consulta.

O modelo quantitativo leva em consideração a frequência com que cada termo de indexação é encontrado no documento, sendo então gerado um arquivo invertido onde para cada termo são associados os documentos com sua respectiva frequência. Na recuperação leva-se em consideração esta frequência e os documentos são apresentados por um *ranking*

decrescente.

O modelo booleano baseia-se na teoria dos conjuntos e na álgebra de Boole. As consultas são elaboradas através de expressões de busca combinando termos de indexação e operadores booleanos (*and, or* e/ou *not*). Segundo Baeza (1999), "a grande vantagem desse modelo é a clareza do seu formalismo e a sua simplicidade".

Por ser baseado simplesmente na teoria de conjuntos, a resposta apresentada por tal modelo é considerada pobre para a maioria das aplicações, pois sua resposta resume-se a um conjunto de documentos que satisfazem à consulta, não indicando qual a qualidade de um documento em relação ao outro. As limitações do modelo levam a sua pouca utilização em grandes coleções de documentos. (PATHAK, 2000)

Outro modelo, também muito utilizado, é o modelo de espaço Vetorial. Segundo Baeza (1999), este modelo baseia-se na comparação parcial (vetorial) entre a representação dos documentos e a da consulta do usuário.

Aos termos das consultas e documentos são atribuídos pesos que especificam o tamanho e a direção de seu vetor de representação. Ao ângulo formado por estes vetores dá-se o nome de Theta (Θ). O cos Θ determina a proximidade da ocorrência. O cálculo da similaridade é baseado neste ângulo entre os vetores que representam o documento e a consulta.

Apesar de ser mais preciso (por indicar a ordem de similaridade dos documentos em relação à consulta) e mais completo (por tratar da busca parcial) que o modelo de busca booleano, este modelo ainda apresenta suposições que fazem com que o resultado deste cálculo de similaridade não seja tão interessante como o esperado. Estamos nos referindo ao fato do cálculo de similaridades ser simplesmente baseado no número de vezes que um termo acontece em um documento. Devemos notar que nem sempre os termos têm o mesmo poder de representatividade em um documento quando estamos observando a coleção como um todo. Por exemplo, um termo que aparece em 900 dos 1000 documentos da coleção é menos discriminante que um termo que aparece em apenas 100 deles, sendo natural que palavras com pouco poder de discriminação tenham seu peso reduzido no cálculo de similaridade, e vice-versa. (SANTOS, 2001)

Para resolver este problema, a similaridade deve ser calculada através da seguinte fórmula, Salton (1988):

$$sim(d,q) = \frac{\sum_{i=1}^{t} w_{id} \times w_{iq}}{\sqrt{\sum_{i=1}^{t} w_{id}^2} \times \sqrt{\sum_{i=1}^{t} w_{iq}^2}} \quad (1)$$

Os pesos servem para determinar a relevância de cada termo para as consultas (*Wiq*) e para os documentos (W_{id}) no espaço vetorial. Segundo Cardoso (2000), para o cálculo dos pesos W_{iq} e W_{id}, utiliza-se uma técnica que faz o balanceamento entre as características do documento, utilizando o conceito de freqüência de um termo num documento. Se uma coleção possui *N* documentos e n_{ti} é a quantidade de documentos que possuem o termo t_i, então o inverso da freqüência do termo na coleção, ou *idf* (*inverse document frequency*) é dado pela equação 2, conforme Salton(1988):

$$idf_i = \log \frac{N}{n_i} \quad (2)$$

Este valor é usado para calcular o peso, utilizando a equação 3, adaptada de Salton(1988):

$$W_{id} = freq(_{ti},d) * idf_i \quad (3)$$

ou seja, é o produto da freqüência do termo no documento pelo inverso da freqüência do termo na coleção.

As principais vantagens do modelo vetorial são a sua simplicidade, a facilidade que ele provê de se computar similaridades com eficiência e o fato de que o modelo se comporta bem com coleções genéricas.

Além desses, outros modelos foram desenvolvidos buscando melhorar a precisão e performance dos sistemas de recuperação de informação. Esses modelos baseiam-se em teorias de disciplinas como probabilidade, estatística e inteligência artificial. O probabilístico utiliza a técnica de determinar um *ranking* do nível de relevância dos documentos em ordem decrescente. Trata-se, portanto, de um método de classificação dos documentos que serão apresentados segundo a ordem de relevância pré-estabelecida.

Existem outros modelos alternativos como o *fuzzy*, com redes neurais, redes bayesianas, modelo booleano estendido, modelos especialistas, mas segundo Baeza (1999), estes não são comumente utilizados, ou apenas em pequenos sistemas de recuperação de informação.

2.2 Indexação

Indexar significa identificar as características de um documento e colocá-las em uma estrutura denominada índice. O índice nada mais é que uma espécie de filtro que é capaz de selecionar os documentos relevantes e manter de fora os documentos irrelevantes.

Os sistemas de busca de documentos mais amplamente empregados (Google, Altavista, Excite) utilizam a indexação automática, que consiste em extrair dos documentos palavras com pouco significado (*stopwords*), normalizar os termos reduzindo-os aos seus radicais (*stemming*) e após fazendo a indexação pelos termos ou frases não considerando o significado que possuem dentro de um determinado contexto (Figura 2). Segundo Kuramoto (2002), neste tipo de indexação é visível a despreocupação com a relevância da recuperação de informação que o sistema de busca irá fornecer ao usuário.

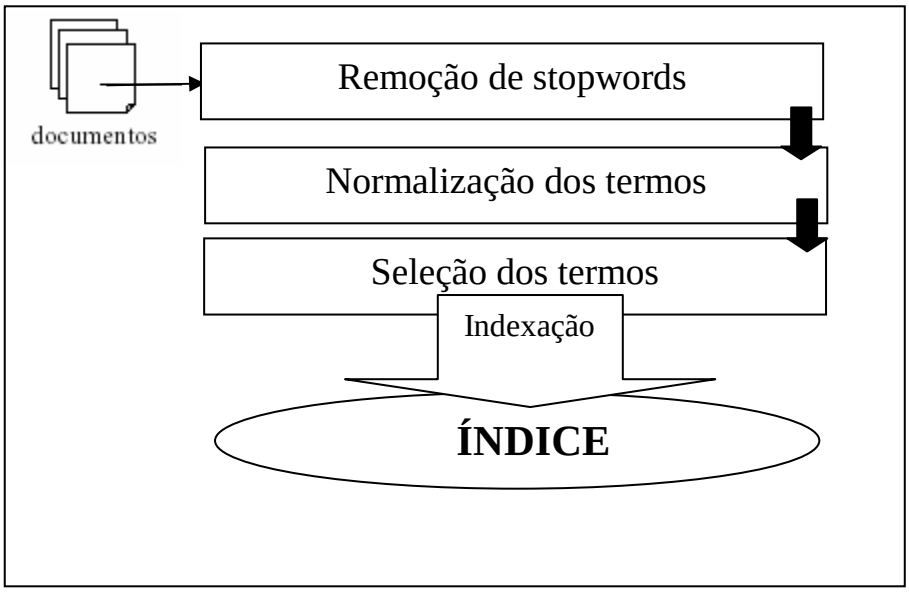

Figura 2 – Etapas do processo de indexação automática
Fonte: Adaptado de Kuramoto (2002)

Conforme Wives (2002), a normalização dos termos (*stemming*) muitas vezes não é utilizada, pois este processo elimina a discriminação entre os documentos. Nestes casos, a indexação ocorre com as variações morfológicas como elas se apresentam nos documentos, cabendo ao usuário solicitar as variações e formas que desejar, durante as consultas.

A indexação manual é realizada por uma ou mais pessoas encarregadas, as quais devem analisar o conteúdo de cada documento e identificar palavras-chave que o caracterizem. Geralmente, há um conjunto de termos pré-definidos e específicos para cada assunto da área em questão. As pessoas encarregadas pela indexação dos documentos devem identificar a que assunto cada um deles pertence e utilizar os termos adequados.

Segundo Moura (2001), os sistemas de busca que trabalham com diretórios, como o "*Yahoo"*, utilizam a indexação manual, onde algumas hierarquias de assuntos são definidas e a indexação se dá com base nestes assuntos. Com isso, se novos documentos forem acrescentados à base de dados eles serão indexados pelos assuntos existentes, caso não se adaptem a estes assuntos, novos assuntos deverão ser acrescentados para haver a indexação.

Alguns sistemas optam por manter todo o texto dos documentos, o que se chama de sistemas de texto completo ou de texto integral. Com a indexação de texto completo, documentos podem ser encontrados utilizando-se qualquer palavra ou frase que possa estar presente no documento. Este tipo de sistema torna-se inviável quando a quantidade de documentos é muito grande, devido ao grande espaço para armazenagem e o grande custo computacional para a sua recuperação.

Segundo Baeza (1999), uma das principais técnicas de indexação é o Arquivo Invertido que é uma estrutura orientada por palavra baseado em listas de palavras-chave ordenadas, sendo que cada palavra-chave possui *links* para os documentos contendo aquela palavra-chave. Cada documento é associado a uma lista de palavras-chave ou de atributos. A lista é invertida e passa a não ser mais ordenada pela ordem de localização, mas sim por ordem alfabética. (FAGUNDES, 2007)

A principal vantagem deste tipo de estrutura é sua facilidade de implementação e a principal desvantagem o alto custo para atualização do índice, devido à necessidade de uma nova indexação completa do sistema para atualizá-lo. (BAEZA, 1999)

Cada termo é associado a um peso, que pode ser definido como um indicador de importância da palavra em relação ao documento. Este indicador serve para a avaliação da relevância no momento da recuperação, ou seja, na consulta do usuário. Um exemplo comum de peso é o cálculo da razão entre o número de ocorrências de uma palavra em um documento e o número total de palavras do documento.

O cálculo do valor da freqüência (F) é obtido pela razão entre a quantidade de vezes que o termo aparece no documento (QR) e o valor resultante da subtração entre a quantidade total de palavras extraídas (TP) e da quantidade de *stopword* existente no documento (QS), conforme equação 4. (FAGUNDES, 2007)

$$F = QR/(TP - QS) \tag{4}$$

2.3 Relevância e as estimativas Precisão e Revocação

A noção da relevância está no centro da recuperação de informação. “O objetivo preliminar de um sistema de recuperação de informação é recuperar todos os documentos que são relevantes a uma solicitação do usuário com uma quantidade mínima de documentos não-relevantes.” (BAEZA, 1999)

Documentos relevantes são aqueles que estão inseridos no contexto da pesquisa realizada pelo usuário e que têm alguma relação com a informação procurada. (AIRES, 2006)

As medidas usuais para avaliar a eficiência dos sistemas de recuperação de informações são precisão *(precision)* e revocação *(recall)* que são medidas baseadas na noção de documentos relevantes de acordo com uma determinada necessidade de informação.

Conforme a Figura 3, R é o conjunto de todos os documentos relevantes contidos na base de dados, normalmente desconhecido pelo usuário, A é o conjunto de todos os documentos retornados pela consulta e Ra é o conjunto de todos os documentos relevantes retornados pela consulta.

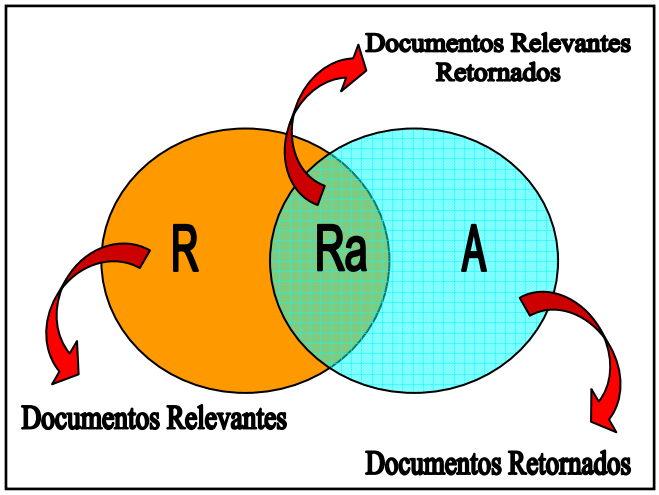

Figura 3 – Conjuntos de documentos usados na estimativa de precisão e revocação
Fonte: adaptado de Fagundes (2007)

Conforme Fagundes(2007), considerando um exemplo de uma consulta C sobre um conjunto de dados D, sendo R o conjunto de documentos relevantes que devem ser retornados pela consulta. Portanto, |R| é o número de documentos que devem ser retornados pela consulta C. Suponha-se que essa consulta C seja processada produzindo um conjunto A de documentos recuperados, sendo que |A| é o número de documentos nesse conjunto e |Ra| é o número de documentos relevantes retornados. Assim, as medidas de revocação (*recall*) e precisão (*precision*) podem ser definidas, segundo Baeza(1999), como:

Revocação: é a razão entre o número de documentos relevantes retornados (intersecção entre os conjuntos R e A) em relação a todos os documentos relevantes da base de dados (conjunto R), conforme equação 5, conforme Fagundes (2007):

$$\text{Revocação:} \quad \frac{Ra}{R} \qquad (5)$$

Precisão: é a razão entre o número de documentos relevantes retornados (intersecção entre os conjuntos R e A) em relação ao número de documentos recuperados (conjunto A), conforme equaçao 6, conforme Fagundes (2007):

$$\text{Precisão: } \frac{Ra}{A} \qquad (6)$$

Em geral, precisão e revocação são calculados usando uma coleção de consultas, documentos e julgamentos de relevâncias conhecidos e, supondo-se que todos os documentos do conjunto A foram examinados. Essas medidas são inversamente proporcionais, ou seja, quando uma medida aumenta a outra tende a diminuir. (AIRES, 2006)

Existem outras formas de medir a eficiência de um Sistema de Recuperação de Informação, mas geralmente estas medidas são mais difíceis de serem interpretadas, além de exigirem informações que não são obtidas sem uma análise mais detalhada dos documentos, o que demanda um esforço computacional maior. Devido a estes fatores, sugere-se a utilização da revocação e a precisão por serem mais facilmente interpretadas (WIVES, 1997).

2.4 *Relevance Feedback* ou Retro-alimentação de Relevância

A definição da relevância ocorre, normalmente, através de um processo denominado retro-alimentação. Neste processo, o usuário informa, implícita ou explicitamente, quais documentos são de seu interesse (AIRES, 2006). A forma implícita pode ser exemplificada pelo simples acesso do usuário a um determinado documento. Um exemplo para a forma explícita é aquele na qual o usuário seleciona os documentos que julga importantes a sua busca e submete essa informação ao sistema. (FAGUNDES, 2007)

Uma forma de aumentar a eficiência de um sistema de recuperação de informação é reduzir as diferenças linguísticas, sociais ou culturais existentes entre usuários e indexadores, incorporando as decisões de ambos na forma de representar os documentos. Este processo, segundo Ferneda (2003), é conhecido como *Relevance Feedback* ou retro-alimentação de relevância e consiste em alterar sucessivamente o peso da expressão de busca em função dos

termos de indexação utilizados na representação dos documentos considerados relevantes pelo usuário após a execução de uma busca.

Através desta retro-alimentação, caso o documento receba uma avaliação como relevante para os termos da consulta, os pesos armazenados para estes termos serão reajustados positivamente.

Segundo Crestani (1997), *feedback* de relevância (*relevance feedback*) é uma técnica que permite ao usuário expressar de modo melhor sua necessidade de informação, adaptando sua consulta original.

Essa retro-alimentação (*feedback*) pode ser feita pelo usuário através de informações de sua satisfação sobre o resultado apresentado, ou automaticamente pelo sistema. (FAGUNDES, 2007)

2.5 Considerações

Embora a literatura apresente diversos modelos de recuperação da informação e diversos sistemas buscam resolver o principal problema em RI que é a relevância dos documentos recuperados, ainda existem muitas dúvidas a respeito da eficácia destes modelos e sistemas. No intuito de ajudar a resolver este problema optou-se em utilizar neste estudo um modelo de recuperação da informação baseado nos algoritmos genéticos e no *feedback* dos usuários do sistema.

O próximo capítulo apresenta um histórico sobre o desenvolvimento dos Algoritmos Genéticos, sendo descritos os conceitos de AG, seus principais aspectos e explanados os operadores genéticos, cujo entendimento se torna necessário para a compreensão do desenvolvimento desta dissertação.

3. ALGORITMOS GENÉTICOS

Desde a antiguidade, a humanidade vem procurando a imitação de mecanismos existentes na natureza ou para associá-los a tecnologias desenvolvidas pelo próprio homem ou na busca do maior entendimento de seu funcionamento. Em especial no caso de fenômenos biológicos, a diversidade dos empreendimentos humanos visando compreendê-los é tamanha que é parte de nossa própria história de civilização.

Nos séculos XIX e XX, fruto das diversas fases das revoluções tecnológicas, esses esforços, em certo sentido, pela primeira vez ganharam condição científica independente, separando-se das análises religiosas e filosóficas sob a condição humana, para se constituírem em áreas de conhecimento próprias. Novas técnicas têm sido inspiradas na natureza ou na biologia de um modo geral, como os algoritmos genéticos (AG) ou as redes neurais. (SOBRINHO, 2003)

Lobo (2005) afirma que na segunda metade do século XIX é proposta, por Darwin, a Teoria da Seleção Natural, um dos mais importantes princípios no ramo da evolução, pois defendia a idéia de que na natureza, aqueles seres vivos que melhor se adaptassem tenderiam a sobreviver. A partir desse marco, os fundamentos da teoria evolucionista foram lançados e se, constituem, nos dias de hoje, nos princípios que norteiam as pesquisas sobre o desenvolvimento das espécies ao longo da vida na Terra. Os algoritmos genéticos se inspiram nas análises de comportamento de populações de indivíduos e sua capacidade de adaptação ao meio no qual estão inseridos. A possibilidade de imitação desse comportamento através de algoritmos computacionais deu margem ao surgimento destas novas técnicas de otimização de problemas, passíveis de serem aplicadas a problemas diversos, sem uma formulação matemática clara.

A origem desses algoritmos pode ser buscada na formulação de Darwin a respeito da evolução das espécies. As teses evolucionistas e a compreensão dos fenômenos da hereditariedade a partir dos estudos de Mendel são os elementos-chave para o desenvolvimento dos algoritmos genéticos. Eles transformam uma população de indivíduos, cada um com um valor associado de adaptabilidade, chamado de aptidão, numa nova geração de indivíduos, usando os princípios de reprodução e sobrevivência dos mais aptos, pela aplicação de operações genéticas como cruzamento e mutação.

Esse capítulo está organizado da seguinte forma: a primeira seção apresenta um histórico do desenvolvimento dos algoritmos genéticos. A seção 3.2 introduz os principais conceitos necessários ao entendimento da metodologia de algoritmos genéticos. A seção 3.3 discute os aspectos principais da construção e implementação dessa classe de algoritmos.

3.1. Histórico do Desenvolvimento de Algoritmos Genéticos

A Computação Evolucionária foi introduzida em 1960 por Ingo Rechenberg com seu trabalho "*Estratégias de Evolução*" (*Evolutionsstrategie* no original). (SOUZA, 2000)

A partir deste trabalho, diversos cientistas começaram a estudar sistemas computacionais que buscavam imitar a idéia da evolução das espécies. Sistemas computacionais evolucionários, onde a evolução poderia ser usada como uma ferramenta de otimização para a solução de diversos tipos de problemas começam a ser desenvolvidos.

Segundo Lobo (2005), o cientista americano John Holland, em conjunto com seus alunos e colegas da Universidade de Michigan, propôs o desenvolvimento de programas computacionais que incorporassem os princípios da evolução, de modo a possibilitar a solução, via simulação, de problemas complexos, justamente como a natureza o fazia, ou seja, produzindo organismos complexos para resolver o problema de sua sobrevivência. Seu trabalho iniciou manipulando vetores binários que representavam cromossomos, e cada indivíduo gerado seria uma tentativa de solução do problema. (SOBRINHO, GIRARDI 2003)

O algoritmo de Holland conseguia resolver problemas complexos de uma maneira muito simples. A exemplo do que acontece na natureza, o algoritmo não sabia o tipo do problema que estava sendo resolvido. Uma simples função de adequação fazia o papel da medida da adaptação dos organismos (cromossomos) ao meio ambiente. Assim, os cromossomos com uma melhor adaptação, medida por essa função, tinham melhor oportunidade de reprodução do que aqueles com má adequação, imitando o processo evolucionário da natureza. (LOBO, 2005)

Os primeiros trabalhos de John Holland são melhor explicados em seu livro "*Adaptation in Natural and Artificial Systems*", publicado em 1975. A partir daí, os algoritmos genéticos têm sido aplicados nos mais variados tipos de problemas e em diversas áreas do conhecimento, como na otimização de sistemas, metodologias de tomada de decisão, simulação, sistemas de aprendizado e sistemas econômicos.

3.2 Conceitos de Algoritmos Genéticos

Algoritmos genéticos são um conjunto de modelos computacionais inspirados na genética. Estes algoritmos modelam uma solução para um problema específico em uma estrutura de dados como a de um cromossomo e aplicam operadores que recombinam estas estruturas, preservando informações críticas.

Algoritmos genéticos é um ramo dos algoritmos evolucionários e podem ser definidos como uma técnica de busca baseada numa metáfora do processo biológico de evolução natural. (LINDEN, 2006)

Um algoritmo genético é um procedimento que mantém uma população de estruturas (chamadas indivíduos), representando possíveis soluções de um determinado problema. Estas estruturas são, então, avaliadas, para gerar oportunidades reprodutivas, de forma que cromossomos que representam uma solução "melhor" tenham maiores chances de se reproduzirem do que os que representam uma solução "pior". A definição do que seja uma solução "melhor" ou uma solução "pior" é tipicamente relacionada à população atual. (LOBO, 2005)

Conforme Lobo (2005), um algoritmo genético é, assim, qualquer modelo computacional baseado em população que utiliza operadores de cruzamento e mutação para gerar novos pontos amostrais em um espaço de busca. O maior interesse no algoritmo genético está em usá-lo como ferramenta de otimização, pois se trata de uma poderosa técnica para busca de soluções de problemas de alta complexidade.

Os algoritmos genéticos baseiam-se na analogia com o processo de evolução biológica das espécies, os algoritmos genéticos mantêm uma determinada informação sobre o ambiente e a acumulam durante o período de adaptação. Posteriormente, utilizam tal informação acumulada para minimizar o espaço de busca e gerar novas e melhores soluções dentro de um domínio.

Deve ser observado que cada cromossomo, chamado de indivíduo no algoritmo genético, corresponde a um ponto no espaço de soluções do problema de otimização. O processo de solução adotado nos algoritmos genéticos consiste em gerar, através de regras específicas, um grande número de indivíduos (população).

A seguir, são apresentados alguns conceitos relacionados aos algoritmos genéticos e necessários ao seu entendimento, Romanhuki (2008) e Lobo (2005):

• cromossomo ou genótipo: cadeia de caracteres, representando alguma informação relativa às variáveis do problema. Desta forma, cada cromossomo representa uma solução do problema.

• gen ou gene: é a unidade básica do cromossomo. Cada cromossomo tem certo número de genes, cada um descrevendo certa variável do problema. Podem ser do tipo binário, inteiro ou real.

• população: conjunto de cromossomos ou soluções.

• fenótipo: cromossomo decodificado.

• geração: o número da iteração que o algoritmo genético executa para gerar uma nova população.

• operações genéticas: operações que o algoritmo genético realiza sobre os seus cromossomos.

• espaço de busca ou região viável: o conjunto, espaço ou região que compreende as soluções possíveis ou viáveis do problema a ser otimizado. Deve ser caracterizado pelas funções de restrição, que definem as soluções viáveis do problema a ser resolvido.

• função objetivo ou de aptidão: construída a partir dos parâmetros envolvidos no problema. Fornece uma medida da proximidade da solução em relação a um conjunto de parâmetros. A função de aptidão permite o cálculo da aptidão de cada indivíduo e fornecerá o valor a ser usado para o cálculo de sua probabilidade de ser selecionado para reprodução.

• aptidão bruta: saída gerada pela função de aptidão para um indivíduo da população.

• aptidão máxima: melhor indivíduo da população.

Esses são os termos básicos da nomenclatura adotada para o estudo de algoritmos genéticos.

Segundo Carvalho (2008), é importante também, analisar de que maneira alguns parâmetros influem no comportamento dos Algoritmos Genéticos, para que se possa estabelecê-los conforme as necessidades do problema e dos recursos disponíveis.

• Tamanho da População: O tamanho da população afeta o desempenho global e a eficiência dos Algoritmos Genéticos (AGs). Com uma população pequena o desempenho pode cair, pois deste modo a população fornece uma pequena cobertura do espaço de busca do problema. Uma grande população geralmente fornece uma cobertura representativa do domínio do problema, além de prevenir convergências prematuras para soluções locais ao invés de globais. No entanto, para se trabalhar com grandes populações, são necessários maiores recursos computacionais, ou que o algoritmo trabalhe por um período de tempo muito maior.

• Taxa de Cruzamento: Quanto maior for esta taxa, mais rapidamente novas estruturas serão introduzidas na população, mas com valores muito altos pode ocorrer perda de estruturas de alta aptidão. Com um valor baixo, o algoritmo pode tornar-se muito lento.

• Taxa de Mutação: Uma baixa taxa de mutação previne que uma dada posição fique estagnada em um valor, além de possibilitar que se chegue a qualquer ponto do espaço de busca. Com uma taxa muito alta a busca se torna essencialmente aleatória.

• Intervalo de Geração: Controla a porcentagem da população que será substituída durante a próxima geração. Com um valor alto, a maior parte da população será substituída, mas com valores muito altos pode ocorrer perda de estruturas de alta aptidão. Com um valor baixo, o algoritmo pode tornar-se muito lento.

3.3 Principais Aspectos dos Algoritmos Genéticos

Conforme Ferreira (2003), o primeiro aspecto a ser considerado é a representação do problema, de maneira que os algoritmos genéticos possam trabalhar adequadamente sobre eles.

A representação das possíveis soluções de um problema podem ser apresentadas no formato de um código genético, que irá definir a estrutura do cromossomo a ser manipulado pelo algoritmo. Essa representação do cromossomo depende do tipo de problema e do que, essencialmente, se deseja manipular geneticamente.

Os principais tipos de representação e os problemas aos quais são tipicamente aplicados são mostrados na Tabela 1, de acordo com Pacheco (2005):

Tabela 1 - Tipos de Representação de Cromossomos.
Fonte: elaborado pelo autor

Representação	**Numéricos**
Binária	Numéricos, Inteiros
Números Reais	Numéricos
Permutação de Símbolos	Baseados em Ordem
Símbolos repetidos	Grupamento

Tradicionalmente, os indivíduos são representados genotipicamente por vetores binários, nos quais cada elemento de um vetor denota a presença de (1) ou ausência (0) de uma determinada característica, ou seja, o seu genótipo. (FERREIRA, 2003)

Como exemplos, na literatura são descritas as seguintes formas de representação dos indivíduos (PACHECO, 2005):

• Vetores de números inteiros ou de números reais (2,345; 4,3454; 5,1; 3,4);

• Cadeias de bits (111011011).

Após a definição da representação do problema, a execução do algoritmo pode ser resumida nos seguintes passos, conforme Lobo (2005):

• Gera-se uma população inicial, normalmente formada por N indivíduos (soluções) criados aleatoriamente.

• Avalia-se toda a população de indivíduos segundo algum critério, determinado por uma função, que avalia a qualidade do indivíduo (função de aptidão ou "*fitness*").

• Seleciona-se então os indivíduos que terão suas características (genes) misturadas através dos operadores de cruzamento (*crossover*) e mutação, gerando assim uma nova população.

• São escolhidos os indivíduos de melhores avaliações (obtidas pela função de aptidão) que serão a base para a criação de um novo conjunto de possíveis soluções, chamado de nova geração.

• Estes passos são repetidos até que uma solução aceitável seja encontrada ou até que o número predeterminado de passos seja atingido ou, então, até que o algoritmo não consiga melhorar a solução já encontrada.

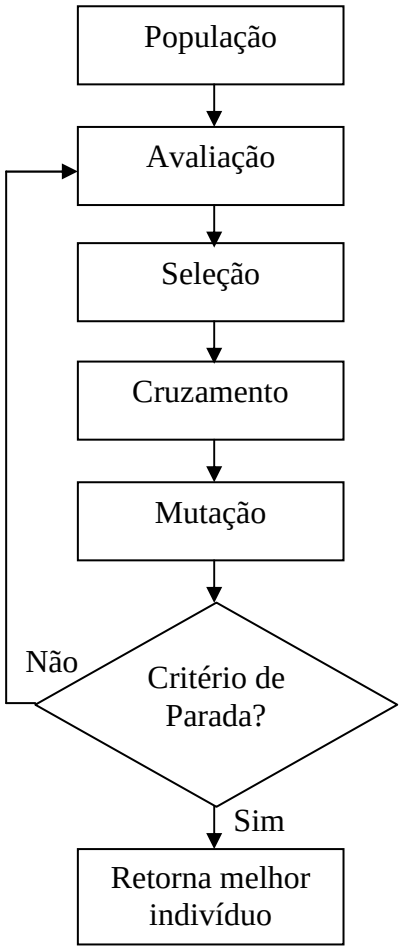

Figura 4 - Estrutura básica de um Algoritmo Genético.
Fonte: elaborado pelo autor

A estrutura funcional do algoritmo está representada na Figura 4, sendo mais detalhada a seguir.

3.3.1 População

A população de um algoritmo genético é o conjunto de indivíduos que estão sendo cogitados como solução e que serão usados para criar o novo conjunto de indivíduos para

análise. O tamanho da população pode afetar o desempenho global e a eficiência dos algoritmos genéticos. Por exemplo, segundo Pacheco (2005), populações que são muito pequenas têm grandes chances de perder a diversidade necessária para convergir a uma boa solução, pois fornecem uma pequena cobertura do espaço de busca do problema. Por outro lado, uma grande varredura do espaço de soluções gera uma grande população, que pode prejudicar o comportamento computacional do problema. Segundo Miranda (2008), "*uma implementação de um algoritmo genético começa com uma população inicial de cromossomos formada de forma aleatória. Essas estruturas são avaliadas e associadas a uma probabilidade de reprodução, de tal forma que as maiores probabilidades são associadas aos cromossomos que representam uma melhor solução para o problema de otimização do que àqueles que representam uma solução pior*".

3.3.2. Avaliação de Aptidão (*Fitness*)

Na avaliação de aptidão é calculado, por intermédio de uma determinada função, o valor de aptidão de cada indivíduo da população. Ainda nessa fase, os indivíduos são ordenados conforme a sua aptidão. Segundo Miranda (2008), este é o componente mais importante de qualquer algoritmo genético, pois é através desta função que se mede quão próximo um indivíduo está da solução desejada ou quão boa é esta solução.

A avaliação é feita através de uma função que melhor representa o problema e tem, por objetivo, fornecer uma medida de aptidão de cada indivíduo na população corrente, que irá dirigir o espaço de busca. Desse modo, a função de avaliação é específica para cada problema. No exemplo a seguir, a função matemática $f(x) = x^2$ mede a aptidão de cada indivíduo. Na Tabela 2 extraída de Pacheco (2005), C1 é um indivíduo mais apto que C2.

Tabela 2 - Exemplo de Função de Aptidão
Fonte: adaptado de Pacheco (2005)

Indivíduo	**Cromossomo**	***X***	***f(x)***	***f rel(x)***
C1	001001	9	81	0,835
C2	000100	4	16	0,165

Em alguns métodos de seleção, é desejável que o valor de aptidão de cada indivíduo seja menor que 1 e que a soma de todos os valores de aptidão seja igual a 1 (*f apt* < 1 e Σ (*f apt*) = 1). Portanto, para cada indivíduo, é calculada a aptidão relativa (*f rel*). A aptidão relativa para um dado indivíduo é obtida dividindo-se o valor de sua aptidão pela soma dos valores de aptidão de todos os indivíduos da população, conforme Lobo (2005).

Para que o algoritmo genético tenha um bom desempenho, é essencial que a função de aptidão seja muito representativa e diferencie, na proporção correta, as “más” soluções das “boas” soluções. Se houver pouca precisão na avaliação, uma ótima solução pode ser posta de lado durante a execução do algoritmo.

3.3.3. Seleção

Dada uma população em que a cada indivíduo foi atribuído um valor de aptidão, o processo de seleção escolhe, então, um subconjunto de indivíduos da população atual, gerando uma população intermediária. Estes indivíduos selecionados serão submetidos aos operadores genéticos.

Segundo Sobrinho (2003), a idéia principal do operador de seleção em um algoritmo genético é oferecer aos melhores indivíduos da população corrente, preferência para o processo de reprodução, permitindo que estes indivíduos passem as suas características às próximas gerações. Isto funciona como na natureza, onde os indivíduos altamente adaptados ao seu ambiente possuem naturalmente mais oportunidades para reproduzir do que aqueles indivíduos considerados mais fracos.

Existem vários métodos de seleção de indivíduos, dentre eles, serão descritos os mais utilizados que são: o método de seleção por Roleta; o método de seleção por Torneio; e o método de seleção Elitista.

3.3.3.1. Método da Seleção por Roleta

O método de seleção mais amplamente utilizado é o método da roleta. Neste método os indivíduos de uma geração (ou população) são escolhidos para fazer parte da próxima geração, através de um sorteio de roleta.

Criamos uma roleta (virtual) na qual cada cromossomo recebe uma porção proporcional à sua avaliação (a soma das porções não pode ultrapassar os 100%). Dessa forma, para indivíduos com alta aptidão, é dada uma porção maior da roleta, enquanto aos indivíduos de aptidão mais baixa é dada uma porção relativamente menor. A roleta é girada um determinado número de vezes, dependente do tamanho da população. A cada giro da roleta, um indivíduo é apontado pela seta e selecionado. Aqueles indivíduos sorteados na roleta são escolhidos como indivíduos que participarão da próxima geração e são inseridos na população intermediária. (LINDEN, 2006)

O método da roleta tem a desvantagem de possuir uma alta variância, podendo levar a um grande número de cópias de um bom cromossomo, diminuindo a variabilidade da população. Uma alternativa seria utilizar somente a posição ("*ranking*") de cada indivíduo na população. Mantendo a população ordenada por valores decrescentes da aptidão, a probabilidade de seleção de um indivíduo para a etapa de recombinação cresce com o seu "*ranking*", ou seja, o primeiro do "*ranking*" tem maior probabilidade de seleção. (Lobo, 2005 apud Silva, 2001)

3.3.3.2. Método da Seleção por Torneio

O método de seleção por Torneio consiste em escolher aleatoriamente um número *n* de indivíduos da população para formar uma sub-população temporária.

O indivíduo que apresentar a melhor aptidão dentre estes *n* indivíduos é selecionado para o cruzamento. O processo se repete até que a nova população seja preenchida.

Este método é bastante utilizado, pois oferece a vantagem de não exigir que a comparação seja feita entre todos os indivíduos da população e sim apenas entre os escolhidos para o torneio.

O método possui a grande vantagem da não-geração de super-indivíduos, pois a chance do indivíduo com maior grau de aptidão ser selecionado é a mesma, independentemente de seu grau de aptidão ser alto. Já no método da roleta, ao contrário, o intervalo de seleção iria aumentar muito e por isso a chance do indivíduo ser selecionado também iria ser bem maior.

3.3.3.3 Seleção Elitista

O modelo de seleção elitista normalmente está associado a outros métodos de seleção, na tentativa de se aumentar a velocidade de convergência do algoritmo, bem como em aplicações nas quais possa ser necessário o seu emprego isoladamente. Esta técnica consiste em substituir os piores indivíduos da nova geração pelos melhores indivíduos da população antiga.

O processo simplesmente copia os *n* melhores indivíduos da população corrente para a próxima geração, garantindo que estes cromossomos não sejam destruídos nas etapas de recombinação e mutação.

3.3.4 Operadores Genéticos

O princípio básico dos operadores genéticos é transformar a população através de sucessivas gerações, estendendo a busca até chegar a um resultado satisfatório. Os operadores genéticos são necessários para que a população se diversifique e mantenha características de adaptação adquiridas pelas gerações anteriores.

Segundo Ferreira (2003), o princípio básico dos operadores genéticos é, então, transformar a população por meio de sucessivas gerações, estendendo a busca até chegar a um resultado satisfatório.

Os principais operadores são o de cruzamento e de mutação, os quais têm um papel fundamental em um algoritmo genético.

A seguir, é mostrado um exemplo de algoritmo genético, extraído e adaptado de Carvalho (2008). Durante esse processo, os melhores indivíduos podem ser coletados e armazenados para avaliação. Nesse algoritmo, as seguintes variáveis são utilizadas:

t - tempo atual;
d - tempo determinado para finalizar o algoritmo;
P - população

```
Procedimento AG
    { t = 0;
       inicia_população (P, t)
      avaliação (P, t);
      repita até (t = d)
            { t = t +1;
              seleção_dos_pais (P,t);
              cruzamento (P, t);
              mutação (P, t);
              avaliação (P, t);
              sobrevivem (P, t)
            }
    }
```

Estes algoritmos, apesar de serem computacionalmente muito simples, são bastante poderosos. Além disso, eles não são limitados por suposições sobre o espaço de busca, relativas à continuidade e a existência de derivadas.

3.3.4.1. Operador Cruzamento (*Crossover*)

Uma das principais características dos algoritmos genéticos que os distinguem das demais técnicas de busca é o operador cruzamento. A idéia central do cruzamento é a

propagação das características dos indivíduos mais aptos da população. O operador cruzamento é utilizado após o de seleção. As formas mais comuns de reprodução em algoritmos genéticos são de um ponto de cruzamento, de dois pontos de cruzamento e de cruzamento uniforme. (SANTOS, 2001)

Esta fase é marcada pela troca de segmentos entre "casais" de cromossomos, selecionados para dar origem a novos indivíduos, que formarão a população da próxima geração. Esta mistura é feita tentando imitar a reprodução de genes em células. Trechos das características de um indivíduo são trocados pelo trecho equivalente do outro. O resultado desta operação é um indivíduo que, potencialmente, combine as melhores características dos indivíduos usados como base.

A combinação dos genes responsáveis pelas características do pai e da mãe possibilita o surgimento de infinitas possibilidades de tipos diferentes, fornecendo um vasto campo de ação para a seleção e aumentando a velocidade do processo evolutivo.

O *crossover* consiste em dividir aleatoriamente os cromossomos, produzindo segmentos anteriores e posteriores que realizam um intercâmbio para obter novos cromossomos (descendentes).

As três formas mais comuns de reprodução em algoritmos genéticos são o cruzamento em um ponto, o cruzamento em dois pontos e o cruzamento uniforme, que serão detalhados a seguir.

3.3.4.1.1. Cruzamento em um ponto

Na reprodução baseada em um ponto de cruzamento (*single-point crossover*), o ponto de quebra do cromossomo é escolhido de forma aleatória sobre a longitude da *string* que o representa e a partir desse ponto se realiza a troca de material cromossômico entre os dois indivíduos.

Cada um dos dois descendentes recebe informação genética de cada um dos pais. Um exemplo nesse sentido pode ser observado na Figura 5, utilizando cromossomos de 10 bits. A partir de um número aleatório, divide-se o cromossomo.

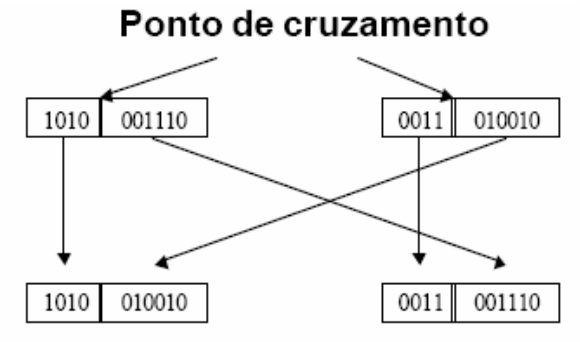

Figura 5 – Esquema com um ponto de cruzamento
Fonte: adaptado de Santos (2001)

Uma observação importante a respeito do cruzamento é que podem ser gerados filhos completamente diferentes dos pais e, mesmo assim, contendo diversas características em comum. Outra questão é que o cruzamento não modifica um bit na posição em que os pais têm o mesmo valor, considerada uma característica cada vez mais importante com o passar das gerações.

3.3.4.1.2. Cruzamento em dois pontos

Na reprodução baseada em dois pontos de cruzamento (*two-point crossover*), procede-se de maneira similar ao cruzamento de um ponto, mas a troca de segmentos é realizada a partir de dois pontos.

Um dos descendentes fica com a parte central de um dos pais e as partes extremas do outro pai e vice versa, como representado na figura 6.

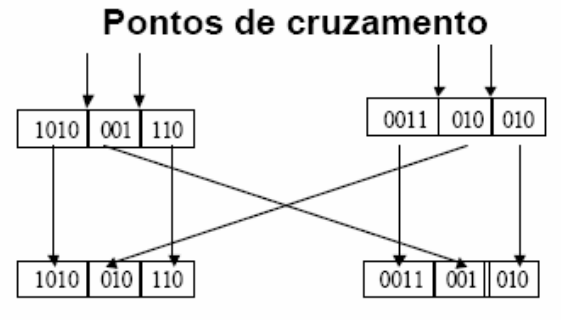

Figura 6 - Esquema com dois pontos de cruzamento
Fonte: adaptado de Santos (2001)

3.3.4.1.3. Cruzamento uniforme

O cruzamento uniforme (*uniform crossover*) é significativamente diferente dos outros dois cruzamentos apresentados anteriormente. Conforme ilustrado na figura 7, primeiramente é criada uma máscara de cruzamento de forma aleatória; posteriormente, cada gene do descendente é criado, copiando-se o gene correspondente de um dos pais, que é escolhido de acordo com a máscara de cruzamento, de modo que, se certo bit da máscara de cruzamento for 1, o gene correspondente será copiado do primeiro pai; se certo bit da máscara de cruzamento for 0, será copiado do segundo pai, conforme figura 7.

O processo é repetido com os pais trocados, para produzir o segundo descendente. Uma nova máscara de cruzamento é criada para cada par de pais.

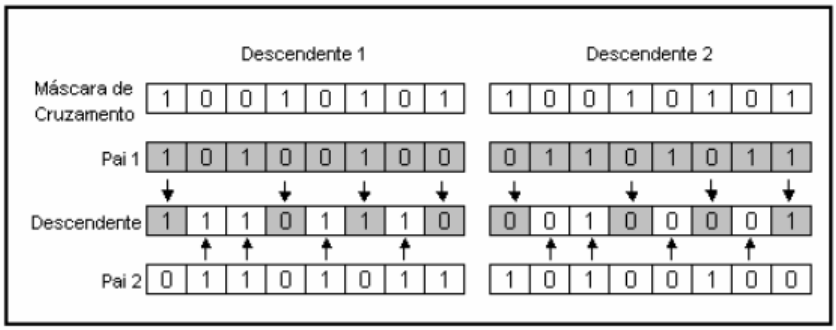

Figura 7 - Cruzamento uniforme
Fonte: adaptado de Santos (2001)

3.3.4.2. Operador Mutação

A mutação é vista como o operador responsável pela introdução e manutenção da diversidade genética na população. Ela trabalha alterando arbitrariamente, logo após o cruzamento, um ou mais componentes de uma estrutura escolhida entre a descendência, fornecendo dessa forma meios para a introdução de novos elementos na população. O operador de mutação é aplicado aos indivíduos com uma probabilidade dada por uma taxa de mutação. (HOLLAND, 1992)

Quando se usa uma representação binária, um bit é escolhido aleatóriamente e substituído por seu complemento (um 0 é substituído por 1 e vice-versa). Este operador é responsável pela introdução de um novo material genético na população de cromossomos, tal como acontece com as espécies na natureza. A Figura 8 ilustra o processo de mutação em um indivíduo.

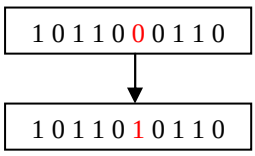

Figura 8 – Esquema de mutação
Fonte: elaborado pelo autor

3.3.5. Critérios de Parada

Diferentes critérios podem ser utilizados para terminar a execução de um algoritmo genético. Como exemplo, podem ser citados os seguintes (LOBO, 2005):

• após um dado número de gerações, ou seja, um total de ciclos de evolução de um algoritmo genético;

• quando a aptidão média ou do melhor indivíduo não melhorar;

• quando as aptidões dos indivíduos de uma população se tornarem parecidas;

• ao conhecer a resposta máxima da função-objetivo;

• no caso de perda de diversidade da população.

3.4 Considerações

Os algoritmos genéticos possibilitam uma diversidade nas respostas obtidas para um mesmo problema, o que sugere que podem ser uma excelente ferramenta para auxiliar a área da recuperação da informação.

Contudo, é muito importante a escolha da forma de representação dos indivíduos da população e dos parâmetros a serem utilizados pelo AG, tais como tamanho da população, número de gerações, taxa de mutação e, os critérios de parada.

O próximo capítulo apresenta alguns aspectos referentes a trabalhos relacionados com os conteúdos desta dissertação.

4. TRABALHOS RELACIONADOS

Neste capítulo são apresentados alguns trabalhos relacionados e que de alguma forma contribuíram com a definição do conteúdo desta dissertação.

Ferneda (2009) apresenta uma forma de aplicação dos algoritmos genéticos em sistemas de recuperação da informação onde o “código genético” é composto das possíveis representações de um mesmo documento. Conforme o autor, “as buscas realizadas pelos usuários são consideradas o “meio ambiente” no qual os documentos estão inseridos, e nesse ambiente as diversas representações de um mesmo documento competem entre si na busca de uma descrição mais adequada para o documento.

A representação dos documentos proposta no trabalho é composta por um conjunto de termos de indexação ou palavras-chave, no qual um gene binário de valor “1” representa a presença de um determinado termo de indexação na representação do documento, o valor “0” representa a sua ausência. Analogamente, a expressão de busca também é composta de um “código genético” onde os indivíduos criados com os termos de busca serão comparados com os indivíduos gerados pelos documentos para obtenção de suas avaliações. As operações genéticas são realizadas sobre os indivíduos de cada documento, o que pode ter um custo computacional muito grande.

Através da observação deste autor, foi utilizada a representação binária na codificação dos indivíduos no Sistema GIRS.

Cazella (2006), em sua tese de doutorado, apresenta uma proposta para modelar e incluir a relevância da opinião do usuário no processo de recomendação colaborativa, que

conforme o autor, “apresenta uma abordagem de Sistemas de Recomendação para recomendar itens baseando-se em informação adicional − definida como relevância da opinião do usuário − além das típicas informações utilizadas na grande maioria dos Sistemas de Recomendação”. A inclusão da relevância fornecida pelos usuários ajuda para que a recomendação consiga identificar qual a importância de um determinado item recomendado frente à relevância de opinião dos recomendadores.

Este autor nos forneceu a idéia de incluir o *feedback* de relevância dos usuários no sistema GIRS, para que a cada nova busca o sistema apresente melhores resultados em termos de relevância.

Pathak (2000) e Radwan (2006) sugeriram modelos onde ocorre a adaptação de diversas funções de *fitness* dos indivíduos, através do uso de algoritmos genéticos, diferentemente dos modelos clássicos apresentados pela literatura. Segundo Pathak, essa adaptação das funções consiste em uma combinação ponderada dos escores produzidos pelas funções individuais o que melhora o desempenho da recuperação em relação à obtida usando uma única função. Esta pontuação geral é usada para classificar e recuperar os documentos.

A análise destes autores nos levou a propor uma função de *fitness* diferente das sugeridas pela literatura.

Gomes (2001) apresentou um sistema de busca de informações personalizadas (FIDUS: Uma Ferramenta para Busca de Informações Personalizadas na Web) que segundo o autor, “provê informações relevantes a um usuário, segundo o seu perfil de consumidor de informação. Essa solução é baseada no algoritmo TF-Seno, uma variação do TF-IDF que permite uma maior precisão de busca de informações perfiladas”. O algoritmo TF-Seno serve para a determinação dos pesos dos termos em relação às páginas Web.

A proposta do *Fidus* é que os usuários possam efetuar buscas de páginas que estão estruturadas por categorias e subcategorias onde, para cada categoria são definidos os critérios de busca de documentos relevantes, possibilitando ao usuário refinar sua busca dentro desta estrutura. Este sistema utiliza–se da métrica da precisão média para avaliação da relevância dos resultados sugeridos.

Gottschalg-Duque (2005) em sua Tese de Doutorado, apresentou o SiRILiCO-Uma Proposta para um Sistema de Recuperação de Informação baseado em Teorias da Lingüística Computacional e Ontologia. Em seus experimentos utiliza textos eletrônicos, disponibilizados em língua portuguesa, e faz a indexação por meio da aplicação de teorias de lingüística computacional e utilização de ontologia.

A hipótese proposta pelo autor é “que é possível desenvolver e implementar um sistema de recuperação de informação totalmente baseado em teorias lingüísticas, teorias de lingüística computacional e ontologia”. Os resultados de precisão e revocação obtidos pelo SiRILiCO sugerem que não só é viável a hipótese defendida como também é muito promissora, afirma Gottschalg-Duque(2005).

Devido a Gomes (2001) e Gottschalg-Duque (2005) utilizarem-se da métrica de precisão média para realizar a avaliação de seus sistemas, também foi utilizado esta metodologia na avaliação do Sistema GIRS.

4.1 Considerações

Muitos são os esforços despendidos pelos pesquisadores na área da Recuperação da Informação, porém ainda existem questões a serem resolvidas no tocante a melhora da relevância dos documentos recuperados pelos mais diversos sistemas propostos.

Os trabalhos analisados apresentam grandes avanços neste sentido, mas novos enfoques e métodos precisam ser testados. Com o objetivo de propor um modelo de recuperação da informação com base nos algoritmos genéticos, o Sistema GIRS é apresentado no próximo capítulo.

5. GIRS (GENETIC INFORMATION RETRIEVAL SYSTEM)

Este capítulo apresenta a lógica do modelo proposto: uma ferramenta de recuperação de informação baseada nos algoritmos genéticos.

A utilização do GIRS implica em duas fases distintas: a primeira fase consiste na Configuração do Sistema, onde são efetuadas as operações que possibilitarão o uso do sistema e a segunda fase que consta da Utilização pelo usuário.

Esse capítulo está organizado da seguinte forma: a seção 5.1 apresenta a metodologia utilizada nesta pesquisa e a seguir são relatados os procedimentos metodológicos utilizados para a realização dos objetivos propostos. A seção 5.2 apresenta a fase de configuração do sistema. Na seção seguinte (5.3) serão detalhados os passos referentes ao sistema de consulta dos usuários, detalhando em especial a implementação do algoritmo genético utilizado. Na seção 5.4 são explanadas algumas considerações sobre este capítulo. A modelagem do sistema pode ser encontrada no Anexo A desta dissertação na página 88.

5.1 Métodos de Pesquisa

Um método científico configura-se como “o conjunto das atividades sistemáticas e racionais que, com maior segurança e economia, permite alcançar o objetivo - conhecimentos válidos e verdadeiros -, traçando o caminho a ser seguido, detectando erros e auxiliando as decisões do pesquisador” (LAKATOS, 2005).

5.1.1 Natureza da Pesquisa

Com relação à natureza da pesquisa, ela gerou conhecimentos novos e úteis para o avanço da ciência e para aplicação prática, dirigidos à solução do problema da recuperação de informações.

5.1.2 Abordagem do Problema da Pesquisa

No que diz respeito à forma da abordagem do problema, essa pesquisa teve uma abordagem qualitativa visando mensurar a qualidade dos sistemas de recuperação da informação disponíveis.

A abordagem qualitativa é evidenciada na fase de avaliação e validação do modelo de recuperação de informações baseada nos algoritmos genéticos onde a relevância dos documentos recuperados evolua a cada nova interação com o usuário.

Além do aspecto qualitativo, a pesquisa também teve sua vertente quantitativa, de caráter interpretativo, pois este projeto, através da revisão da literatura, possibilitou conhecer as características de diversos sistemas de recuperação de informações.

Essas duas abordagens (qualitativa e quantitativa) se complementam para atingir os objetivos desse projeto, permitindo análises interpretativas e construtivas, devido à complexidade do assunto pesquisado sob diferentes ângulos e conteúdos.

5.1.3 Caracterização dos Objetivos da Pesquisa

Do ponto de vista de seus objetivos, a pesquisa realizada pode ser considerada exploratória pois, investigou a evolução da relevância na recuperação de informações com a utilização de algoritmos genéticos e o *feedback* implícito de relevância obtido dos usuários.

Assim, para atender os objetivos propostos, o problema foi abordado com um estudo exploratório de cunho quanti-qualitativo, embasado tanto na literatura como na experimentação de alguns sistemas de recuperação de informações.

5.2 Configuração do sistema

Esta seção apresenta as configurações do sistema que consistem em 4 operações principais que possibilitarão o uso do sistema:

• Cadastramento das *Stopwords*: a primeira das operações necessárias é o cadastramento das *stopwords,* que são as palavras com pouco significado a serem extraídas dos documentos antes da indexação. As *stopwords* (da língua em que os documentos estão escritos) são relacionadas pelo administrador do sistema através da interface de configuração do sistema conforme a Figura 9.

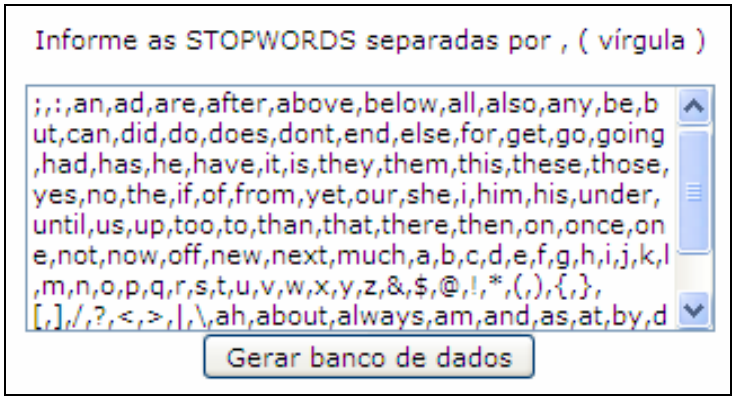

Figura 9 – Interface de Cadastro das *Stopwords*
Fonte: elaborado pelo autor

• Catalogação dos Documentos: a catalogação dos documentos consiste no armazenamento dos documentos num arquivo contendo a identificação, o título e o local de armazenamento de cada documento. Durante este processo, os documentos não sofrem nenhum tipo de modificação, apenas são criadas as referências dos documentos na base de dados gerada pelo sistema. A Figura 10 apresenta um exemplo da catalogação de dois documentos de bases diferentes que estão armazenados em locais distintos.

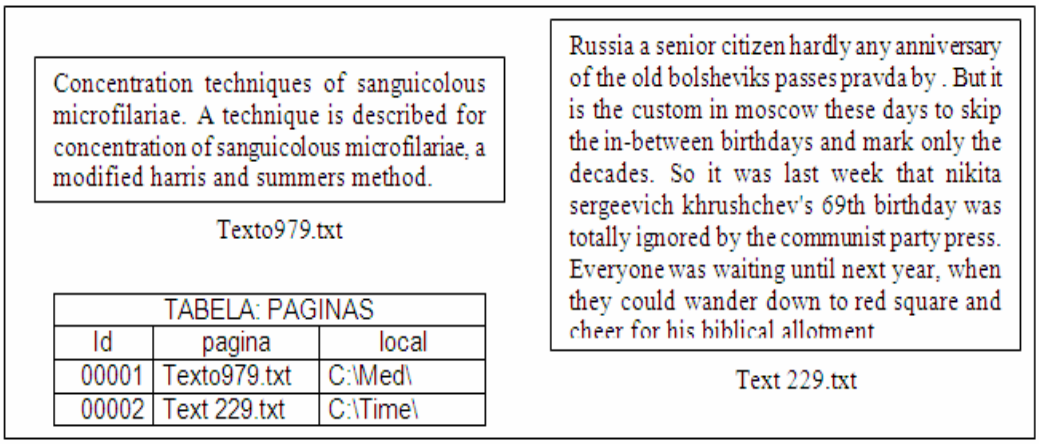

TABELA: PAGINAS		
Id	pagina	local
00001	Texto979.txt	C:\Med\
00002	Text 229.txt	C:\Time\

Figura 10 – Exemplo de Catalogação dos Documentos
Fonte: elaborado pelo autor

• Indexação dos termos: a próxima etapa consiste na indexação dos termos, onde será realizada a análise do texto existente em cada um dos documentos catalogados, eliminando as *stopwords* neles existentes, restando, apenas, os termos que podem ter certa importância no contexto do documento. Esta etapa pode ser visualizada na Figura 11 onde o texto do exemplo é apresentado em sua forma original e após a extração das *stopwords.* Após a eliminação das *stopwords*, o sistema armazena todos os termos em um arquivo, conforme Tabela: Palavras da Figura 11.

• Cálculo da Frequência dos Termos: para cada termo restante no documento, o Sistema efetua o cálculo do valor da frequência (F) do termo em relação à quantidade total de termos do documento, gerando um peso que pode ser definido como um indicador de importância da palavra em relação ao documento. Este indicador serve para a avaliação da relevância no momento da recuperação, ou seja, na consulta do usuário.

O valor da freqüência (F) é obtido conforme a equação 7, pela razão entre a quantidade de vezes que o termo aparece no documento (QR) e o valor resultante da subtração entre a quantidade total de palavras extraídas (TP) e da quantidade de *stopword* existente no documento (QS).

$$F = QR/(TP - QS) \tag{7}$$

Após estes cálculos, os termos são inseridos na base de dados do sistema a fim de formarem o arquivo invertido de termos dos documentos, com seus pesos respectivos, conforme a Tabela: Palavras_Paginas da Figura 11.

Concentration techniques of sanguicolous microfilariae. Techniques are described for concentration of sanguicolous microfilariae, a modified harris and summers method.

Texto979.txt original

Concentration techniques sanguicolous microfilariae techniques described concentration sanguicolous microfilariae modified harris summers method

Texto979.txt sem as *stopwords*

TABELA: PALAVRAS

Id	palavra
00001	Concentration
00002	Techniques
00003	Microfilariae
00004	Sanguicolous

TABELA: PALAVRAS_PAGINAS

palavra_id	pagina_id	frequencia
00001	00001	0.15384615
00002	00001	0,07519232
00002	00006	0,00123765
00003	00001	0,07519232

Figura 11 – Exemplo de Indexação dos Documentos
Fonte: elaborado pelo autor

A interface de configuração do Sistema pode ser visualizada na Figura 12. Estas operações são gerenciadas pelo administrador do sistema e devem ser atualizadas periodicamente.

Figura 12 – Interface de configuração do sistema
Fonte: elaborado pelo autor

5.2.1 Resumo das Etapas

O quadro resumo dos passos desta etapa pode ser visualizado a seguir.

Configuração do sistema
– Cadastro das *stopwords*. – Catalogação dos documentos. – Para cada documento catalogado: o Análise do texto. o Eliminação das *stopwords* existentes nos documentos. o Calcular a quantidade de *stopwords* existente no documento (QS). o Calcular a quantidade de palavras existente no documento (TP). o Calcular a quantidade de vezes que cada termo aparece no documento (QR), agrupando-os. – Indexação a partir dos dados armazenados anteriormente, repetindo a sequência de eventos abaixo para cada termo encontrado: o Inserir termo no índice. o Vincular termo ao documento. o Calcular o índice de freqüência do termo no documento, realizado pela fórmula: • $F = QR/(TP - QS)$. o Atribuir o valor do peso (F) como um indicador de importância da palavra em relação ao documento.

Quadro 1: Configuração do Sistema

Fonte: elaborado pelo autor

Após as etapas descritas acima, o sistema estará apto para que os usuários possam efetuar suas consultas.

5.3 Consultas dos usuários

O sistema de consulta ou de busca utiliza uma interface similar àquelas utilizadas nos motores de busca atuais (Google, Yahoo e Altavista), com a opção de o usuário escolher o tipo de busca que desejar, se Direta pelos termos escolhidos, ou, busca pelo GIRS (*Genetic Information Retrieval System*), onde o usuário informa os termos a serem pesquisados e os submete ao tipo de consulta escolhida, conforme ilustrado na Figura 13.

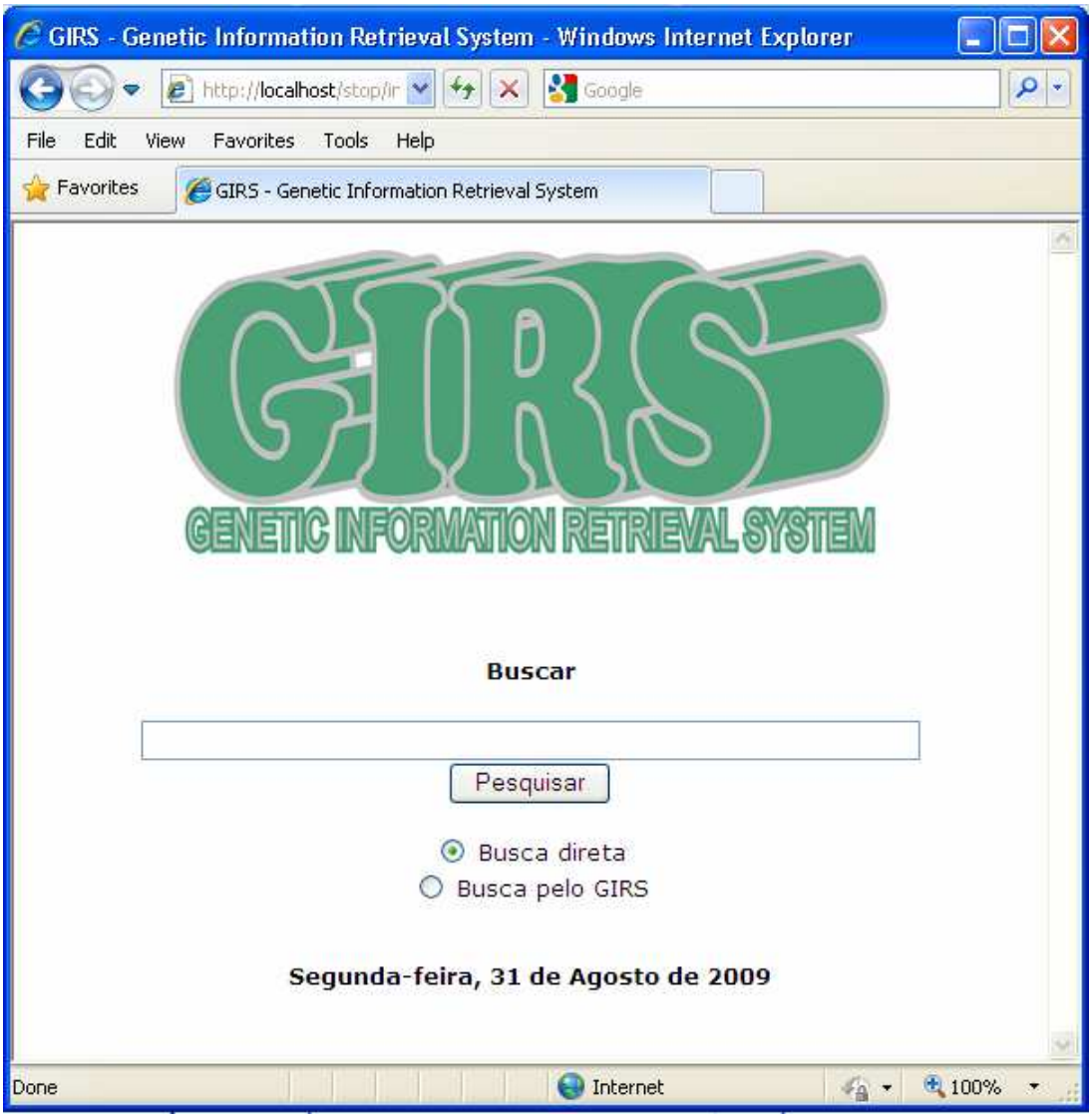

Figura 13 – Interface de Consultas
Fonte: elaborado pelo autor

5.3.1 Busca Direta

Ao optar pela busca direta e, após o usuário digitar os termos de seu interesse (podem ser utilizados termos simples, a combinação de diversos termos juntos ou termos compostos), submete sua consulta ao sistema, onde através destes termos o sistema fará a busca na base de dados dos documentos indexados pelo sistema e apresentará os 20 melhores documentos ordenados em ordem decrescente pela frequência (F) dos termos solicitados na consulta, o resultado da busca pode ser visualizado na Figura 14.

Figura 14 – Interface de Resposta para Busca Direta
Fonte: elaborado pelo autor

5.3.2 Busca pelo GIRS (Genetic Information Retrieval System)

Nesta opção o usuário digita os termos e submete sua consulta ao sistema, onde através destes termos o sistema inicia todo o processo de consulta.

A seguir, serão explicados como foram implementados os passos do algoritmo genético proposto.

5.3.2.1 Cromossomo ou indivíduo

Em algoritmos genéticos clássicos, um cromossomo é representado por uma sequência de bits. A proposta aqui apresentada é a utilização de um cromossomo composto por dois vetores com tamanho igual ao número total de documentos indexados na base de dados. Um dos vetores com valor binário na forma {0, 1}, de modo que as posições deste vetor que apresentam valor = 1 indicam a presença do documento correspondente no resultado da busca, enquanto que o valor = 0 indica a sua ausência, que chamaremos de Vetor de Presença. O outro vetor indica o peso do termo em relação ao documento, onde cada posição deste vetor refere-se a um documento, os valores variam entre 0 e 0,001, devido a ser esta a grandeza dos valores dos pesos calculados pelo sistema, que chamaremos de Vetor dos Pesos. A Figura 15 apresenta um exemplo de indivíduo.

Indivíduo

Vetor de Presença	1	0	0	0	1	1	0	1	1	0
Vetor dos Pesos	0,00013	0	0	0	0,00087	0,00002	0	0,00095	0,00021	0

Figura 15 – Exemplo de Indivíduo
Fonte: elaborado pelo autor

5.3.2.2 Inicialização da população

A cada consulta do usuário, o sistema gera uma população inicial de 20 indivíduos aleatoriamente. Cada cromossomo inicial tem seus genes gerados randomicamente com

valores 0 ou 1 para o Vetor de Presença e caso o valor para esta posição seja igual a 1, a posição correspondente no Vetor dos Pesos recebe um valor randômico entre 0 e 0,001 indicando o peso dos termos para o documento correspondente (conforme detalhado na seção anterior).

5.3.2.3 Avaliação dos Indivíduos

Para a avaliação dos indivíduos da população inicial é criado um indivíduo de referência que utiliza como parâmetros os termos da consulta do usuário e os pesos correspondentes armazenados na base de dados. Este indivíduo é originado através de uma consulta com os termos solicitados pelo usuário, similar ao sistema de busca direta[1], e cujo resultado é armazenado em um indivíduo idêntico aos da população inicial, onde o Vetor de Presença indica os documentos relevantes para a consulta em questão de modo que as posições deste vetor que apresentam valor = 1 indicam a presença do documento correspondente no resultado da busca, enquanto que o valor = 0 indica a sua ausência. O Vetor dos Pesos indica o peso do termo em relação ao documento, onde cada posição deste vetor refere-se a um documento, os valores dos pesos de referência são aqueles armazenados na base de dados. A Figura 16 apresenta um exemplo deste indivíduo de referência.

Indivíduo de Referência

Vetor de Presença	1	0	0	0	1	1	0	1	1	0
Vetor dos Pesos	0,00013	0	0	0	0,00087	0,00002	0	0,00095	0,00021	0

Figura 16 – Exemplo de Indivíduo de Referência
Fonte: elaborado pelo autor

Cada indivíduo da população inicial passa a ser comparado com o indivíduo de referência, e o valor de avaliação será calculado pela equação 8, elaborada pelo autor:

$$\text{Avaliação} = \sum(Pc - |Fp - Fr|) \quad (8)$$

[1] Sistema de Busca Direta efetua uma consulta na base de dados com os termos solicitados pelo usuário, conforme seção 5.2.1. da página 49.

Pc = Total de posições coincidentes entre os documentos escolhidos para a resposta (valores iguais a 1) do indivíduo avaliado e o de referência;

|Fp – Fr| = Para as posições coincidentes do item anterior, subtrai-se os valores dos pesos correspondentes onde, o módulo do resultado é subtraído da avaliação, onde Fp = Peso do indivíduo avaliado e Fr = peso do indivíduo de referência, valores estes calculados para os mesmos índices dos vetores.

A equação 8 tem por objetivo apurar o valor de avaliação de forma que os indivíduos que apresentam a maior quantidade de posições coincidentes (valores iguais a 1 no indivíduo de referência e no indivíduo avaliado) e a menor diferença entre os pesos atribuídos recebam uma melhor avaliação, enquanto que os indivíduos que apresentam uma menor quantidade de posições coincidentes e uma maior diferença dos pesos recebam uma avaliação menor.

A comparação é feita em todas as posições dos vetores e, para todos os indivíduos da população. O valor de avaliação final de cada indivíduo é armazenado em outro vetor para ser utilizado posteriormente pelo operador de seleção.

A figura 17 apresenta um exemplo do cálculo da avaliação utilizado. O exemplo utiliza dois indivíduos e vetores com tamanhos reduzidos apenas para um melhor entendimento.

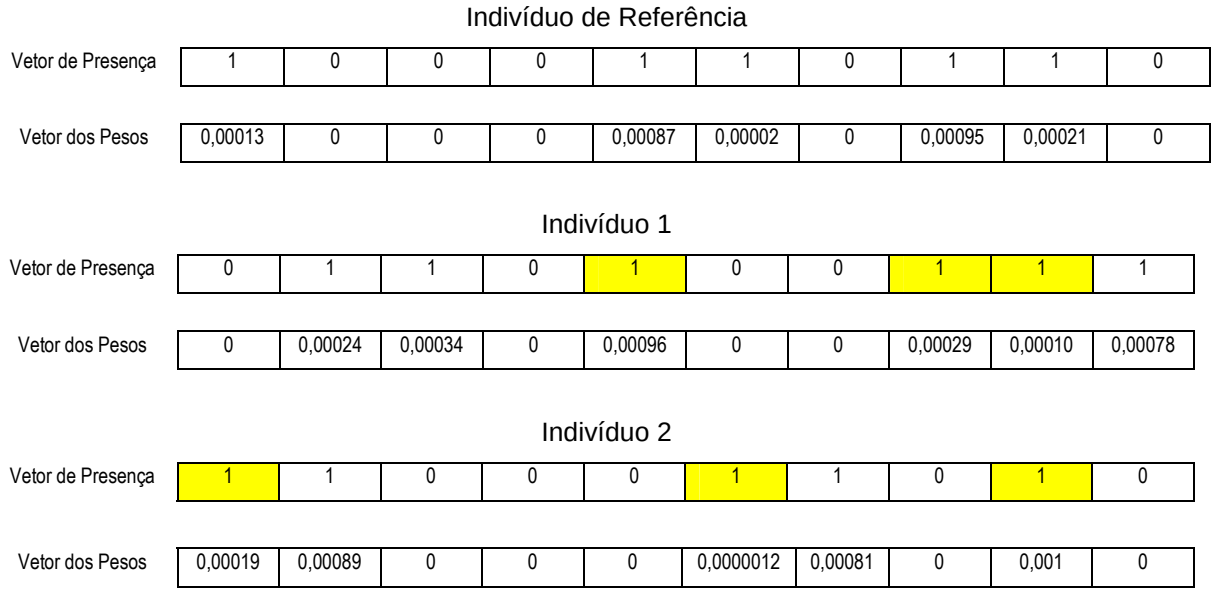

Figura 17 – Exemplo do cálculo da avaliação
Fonte: elaborado pelo autor

Avaliação Indivíduo 1:

Aval = [1 – |(0,00087 – 0,00096)| + 1 – (0,00095 – 0,00029) + 1 – (0,00021 – 0,00010)]
= 1 – 0,00009 + 1 – 0,00066 + 1 – 0,00011 = 2,99914

Avaliação Indivíduo 2:

Aval = [1 - |(0,00013 – 0,00019)| + 1 - (0,00002 – 0,0000012) + 1 – |(0,00021 - 0,001)|]
= 1 – 0,00006 + 1 – 0,0000188 + 1 – 0,00079 = 2,99913

Embora muito próximos os valores de avaliação dos indivíduos do exemplo, o Indivíduo 1 apresenta a melhor avaliação entre os dois.

5.3.2.4 Seleção e Cruzamento (*Crossover*)

O método de seleção utilizado é misto, onde se utiliza o método da roleta e o método elitista. No método da roleta os indivíduos de uma geração (ou população) são escolhidos para fazer parte da próxima geração, através de um sorteio de roleta. Já no método elitista os piores indivíduos da nova geração são substituídos pelos melhores indivíduos da população antiga.

Criamos uma roleta (virtual) na qual cada cromossomo recebe uma porção proporcional à sua avaliação (a soma das porções é igual a 100%). Dessa forma, para indivíduos com alta avaliação, é dada uma porção maior da roleta, enquanto aos indivíduos de aptidão mais baixa é dada uma porção relativamente menor.

A roleta é “girada” sendo então escolhido um indivíduo da população. A roleta é girada novamente sendo escolhido o outro indivíduo da população para efetuar o cruzamento com o primeiro indivíduo escolhido. A operação de *crossover* consiste em gerar novos cromossomos a partir destes dois cromossomos originais, chamados pai e mãe. Estes novos elementos devem manter as características dos pais.

O cruzamento em um ponto foi o utilizado, e o ponto de quebra do cromossomo é

escolhido de forma aleatória. A partir do ponto de quebra dos vetores se realiza a troca de material cromossômico entre os dois indivíduos escolhidos gerando dois novos indivíduos. Cada um dos dois descendentes recebe informação genética de cada um dos pais.

A Figura 18 apresenta os dois indivíduos que irão sofrer o cruzamento, com o ponto de cruzamento sorteado. A figura 19 mostra os novos indivíduos gerados após o cruzamento efetuado.

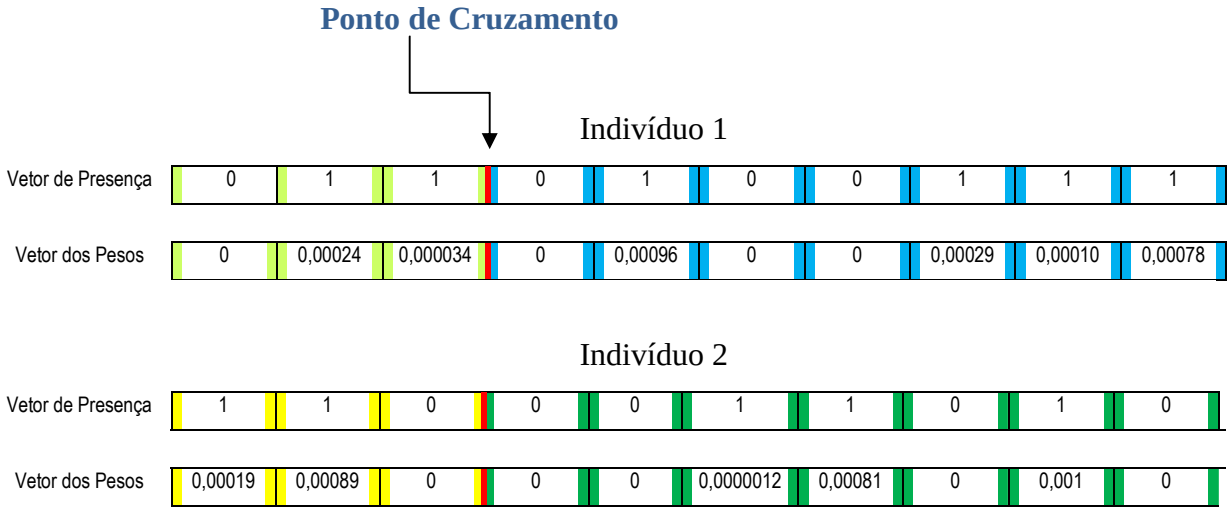

Figura 18 – Indivíduos que irão sofrer o cruzamento
Fonte: elaborado pelo autor

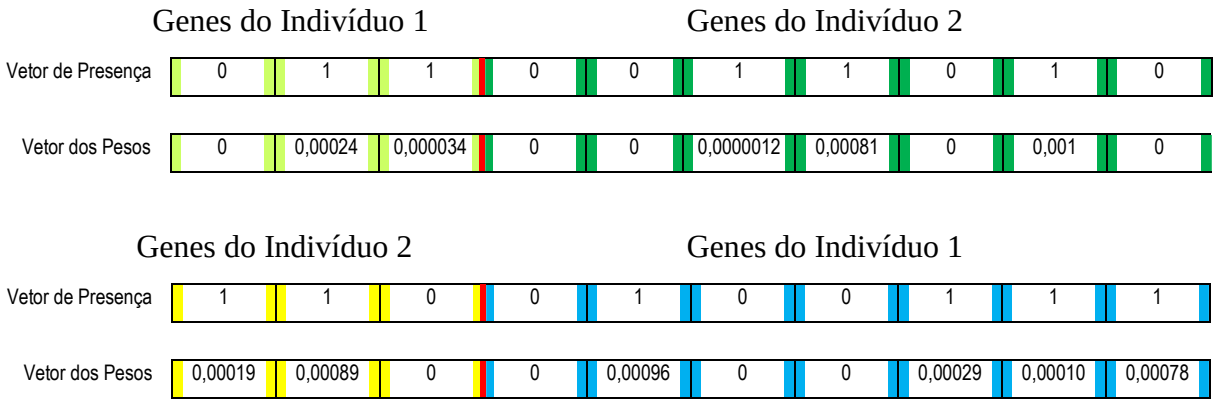

Figura 19 – Novos Indivíduos gerados após o cruzamento
Fonte: elaborado pelo autor

Este processo se repete o número de vezes igual à metade do tamanho da população gerando assim uma nova população com o mesmo tamanho da inicial. Após ser gerada esta nova população, os 20 (tamanho da população inicial) melhores indivíduos entre estas duas

populações (população inicial e a nova população gerada) serão selecionados para a próxima iteração do algoritmo genético.

5.3.2.5 Mutação

A mutação é uma operação que modifica um dos genes do cromossomo. Na implementação proposta, são escolhidos aleatoriamente os indivíduos e os genes que serão modificados, sendo utilizado uma taxa de mutação de 1% dos indivíduos e 0,1% dos seus genes. Este valor não deve ser muito grande, para não perder a similaridade com o indivíduo inicial. Também não pode ser muito pequeno, pois dificultará o processo de mutação, podendo levar à ocorrência de situações em que esta operação gera um cromossomo idêntico ao original, sem que a mesma tenha qualquer efeito na solução do problema.

Caso o valor do gene a ser mutado seja igual a 1, os valores nos dois vetores para esta posição recebem valor igual a 0. Já para o caso do valor ser igual a 0, o valor do Vetor de Presença para esta posição recebe valor igual a 1 e para o Vetor dos Pesos, na posição correspondente recebe um valor randômico entre 0 e 0,001.

O cromossomo mutado será considerado um novo cromossomo, estando sujeito as mesmas regras de sobrevivência dos demais cromossomos. A figura 20 apresenta um indivíduo que vai sofrer mutação, com os campos sorteados para permutação. A figura 21 mostra este mesmo indivíduo, depois da mutação.

Vetor de Presença	0	1	1	0	1	0	0	1	1	1
Vetor dos Pesos	0	0,00024	0,000034	0	0,00096	0	0	0,00029	0,00010	0,00078

Figura 20 – Indivíduo que irá sofrer mutação
Fonte: elaborado pelo autor

Vetor de Presença	0	1	1	0	0	0	1	1	1	1
Vetor dos Pesos	0	0,00024	0,000034	0	0	0	0,00016	0,00029	0,00010	0,00078

Figura 21 – Indivíduo depois da mutação
Fonte: elaborado pelo autor

5.3.2.6 Critério de Parada

Para o sistema proposto, são utilizados os seguintes critérios de parada para o algoritmo genético:

• quando a avaliação do melhor indivíduo não melhorar após 3 gerações, ou seja, quando não houver a criação de melhores indivíduos nas novas populações, visando à economia de tempo de resposta, ou;

• após 30 gerações, ou seja, um total de 30 ciclos de evolução do algoritmo genético, este valor foi adotado porque se verificou[2] que um número maior de gerações não representa uma melhora significativa no resultado, apenas retarda ainda mais o tempo de resposta estimado em torno de 18 a 20 segundos.

5.3.2.7 Apresentação do resultado final

Após a conclusão das operações do algoritmo genético, o indivíduo que tiver a melhor avaliação será escolhido pelo GIRS como sendo a resposta para a consulta em questão.

O indivíduo escolhido será então ordenado em ordem decrescente dos valores do Vetor dos Pesos e por sua vez apresentado ao usuário onde os documentos que possuem maior valor no Vetor dos Pesos são apresentados em primeiro lugar e assim sucessivamente em ordem decrescente. A Figura 22 apresenta um indivíduo escolhido como resposta com os pesos correspondentes, para um melhor entendimento.

---

[2] O número de 30 gerações foi o escolhido após serem testadas diversas possibilidades, sendo que um número maior de iterações não apresentava uma melhora significativa nos resultados, por outro lado, por se tratar de um motor de busca, o tempo de execução deve ser o mais breve possível.

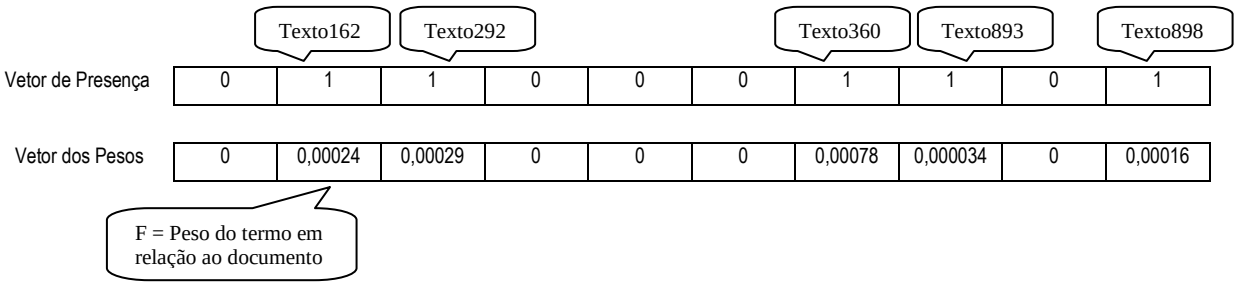

Figura 22 – Indivíduo escolhido como resposta
Fonte: elaborado pelo autor

O resultado final a ser apresentado para o usuário, conforme o exemplo apresentado na Figura 22, pode ser visualizado na Figura 23 que mostra a interface de resposta proposta.

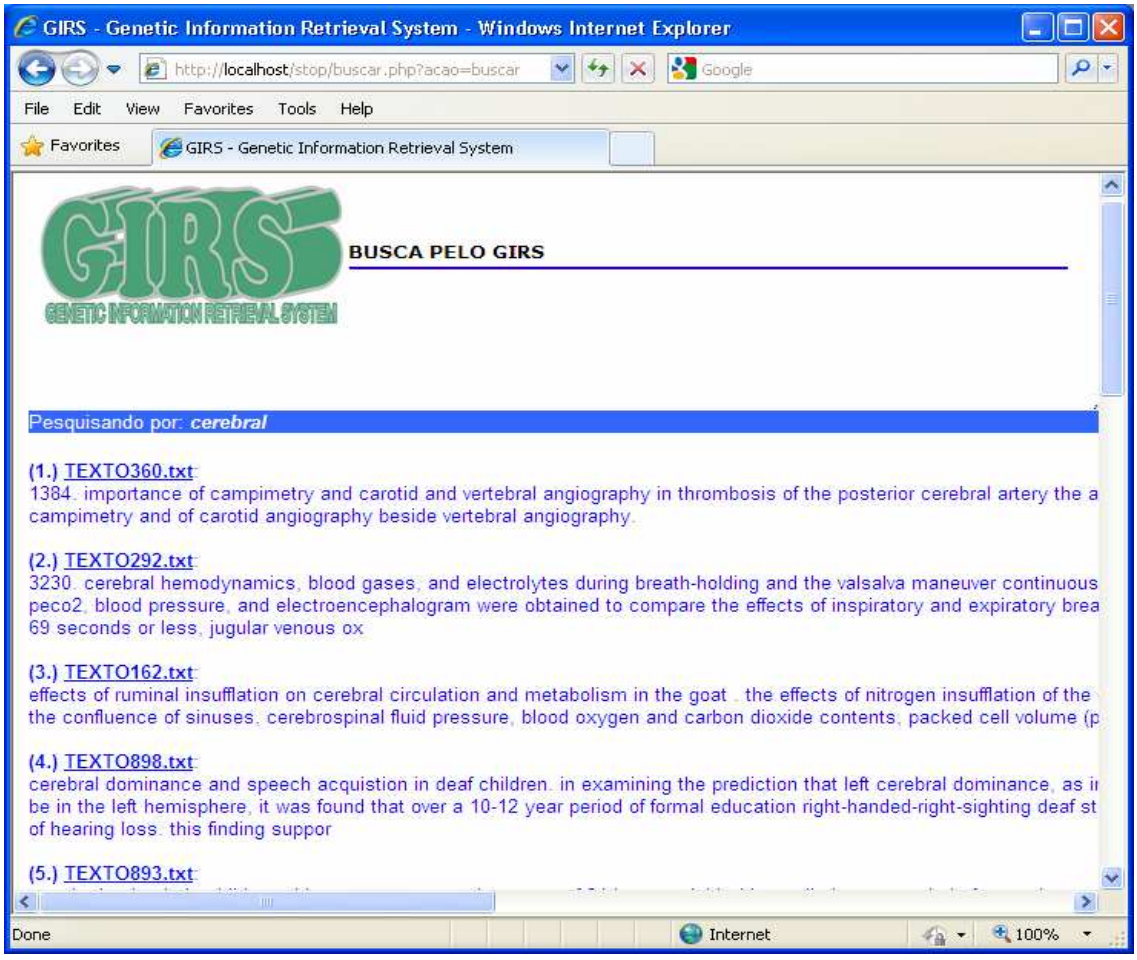

Figura 23 – Interface de Resposta para Busca pelo GIRS
Fonte: elaborado pelo autor

5.3.2.8 Visualização dos documentos

Para abrir a visualização dos documentos escolhidos pelo usuário, basta clicar duas vezes no título do documento na interface de resposta (Figura 23) e o sistema abre uma nova janela com todo o conteúdo referente ao documento escolhido, conforme Figura 24.

Ao abrir o documento escolhido pelo usuário, o Sistema atualiza o valor da freqüência (F) dos termos da consulta em relação ao documento visualizado, somando um valor constante de 0,001 ao valor anteriormente armazenado. Esta atualização serve para melhorar a relevância do documento quando das próximas consultas dos usuários, para os mesmos termos.

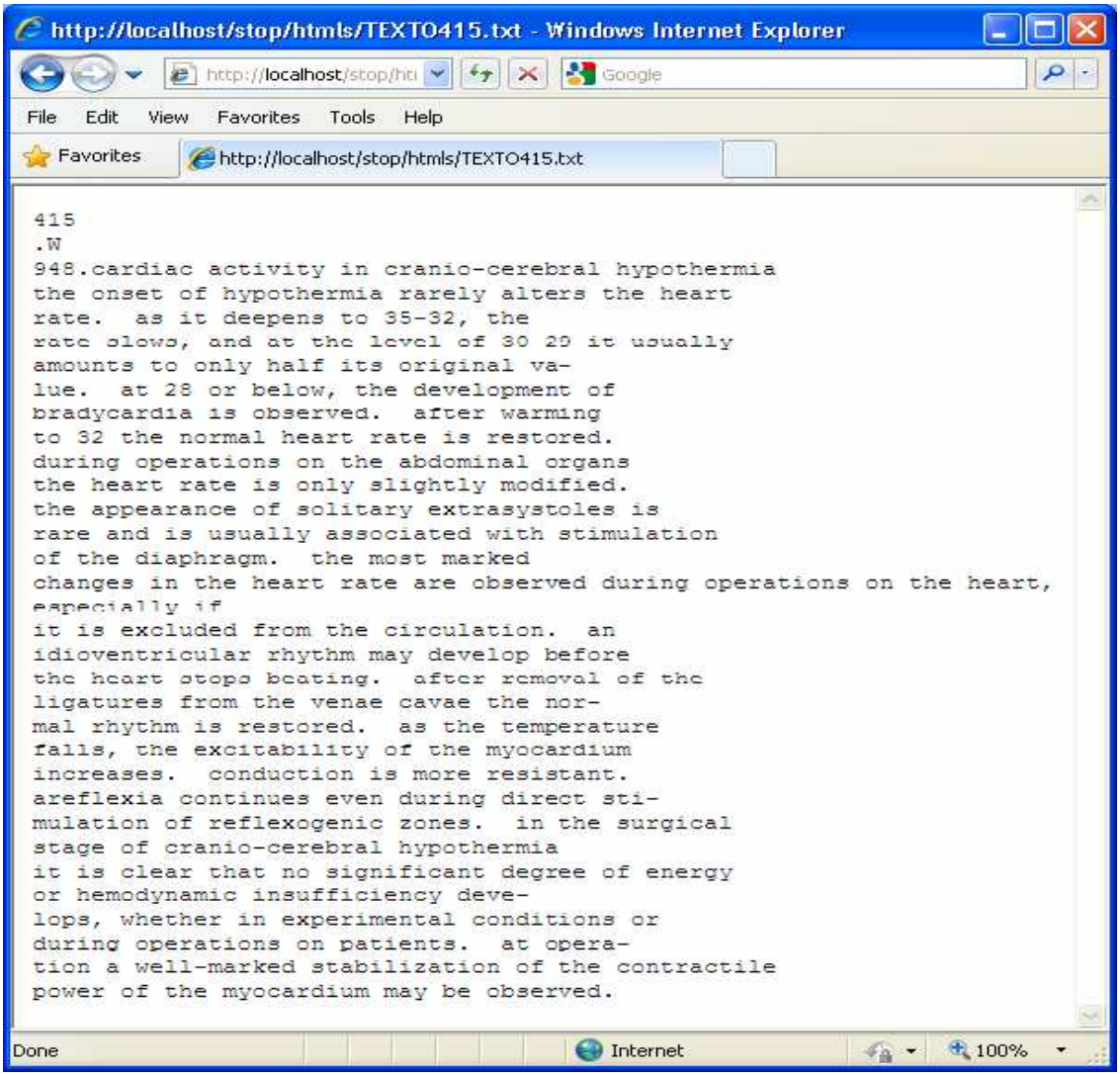

Figura 24 – Interface de Visualização dos Documentos
Fonte: elaborado pelo autor

5.4 Considerações

O Sistema GIRS através de uma interface simples (similar aos buscadores tradicionais), procura apresentar de forma objetiva o resultado das consultas dos usuários, tanto pela busca direta, quanto pelo sistema de algoritmos genéticos.

Os parâmetros do algoritmo genético implementado, tais como tamanho da população, número de gerações, taxa de mutação e, os critérios de parada, foram ajustados após serem testadas várias possibilidades, sendo os valores adotados os que apresentaram a melhor resposta para o caso em questão.

6. METODOLOGIAS DE AVALIAÇÃO DE SISTEMAS DE RECUPERAÇÃO DA INFORMAÇÃO

Os sistemas de recuperação da informação precisam ser avaliados com a finalidade de aferir o quanto o sistema atende a necessidade de seu usuário final, para todos os usuários reais e potenciais na comunidade. (TAGUE-SUTCLIFFE, 1996)

A avaliação de um sistema de recuperação da informação pode ser apurada sem a participação dos usuários, mas segundo Rodrigues (2008), a última palavra em termos de desempenho de um sistema de recuperação da informação só pode ser dada após alguns usuários reais ou potenciais tiverem usado o sistema em um experimento controlado de recuperação da informação.

Muitas vezes, segundo Rodrigues (2008), efetuar este tipo de avaliação com usuários reais pode ser uma tarefa difícil de controlar. Por esse motivo, a determinação de quais os documentos são relevantes para determinada consulta é resultado da avaliação das bases de documentos por usuários reais, gerando assim coleções de testes. Após a consolidação destas coleções de testes, elas podem ser utilizadas como padrões para avaliações de sistemas de recuperação de informação, sem a necessidade de consultar os usuários novamente.

Estas coleções são compostas pelos documentos, um conjunto de consultas e os julgamentos de relevância.

Conforme Rodrigues (2008), para se fazer uma avaliação é necessária uma coleção de testes, composta de documentos, consultas e níveis de relevância para as consultas

apresentadas, além de uma metodologia estatística que determine se as diferenças observadas no desempenho entre os sistemas analisados são estatisticamente significantes.

Em sistemas de recuperação de informação textual, as medidas mais utilizadas são precisão (*precision*) e revocação (*recall*).

Precisão: é a razão entre o número de documentos relevantes retornados em relação ao número de documentos recuperados, conforme equação:

$\text{Pr}ecisão = \frac{Ra}{A}$, onde Ra = documentos relevantes retornados e A = documentos retornados.

Revocação (R: é a razão entre o número de documentos relevantes retornados em relação a todos os documentos relevantes da base de dados, conforme equação:

$\text{Re}vocação = \frac{Ra}{R}$, onde Ra = documentos relevantes retornados e R = documentos relevantes na base de dados.

Conforme Rijsbergen (1979) outras medidas utilizadas em RI são a Medida F e o Fallout. A Medida-F é a média harmônica da precisão e da revocação, enquanto que *Fallout* é a razão entre o número de documentos não relevantes recuperados e o número total de documentos não relevantes da base de dados, conforme equações a seguir:

$\text{Medida}\, F = \frac{2PR}{P+R}$, onde P = precisão e R = revocação.

$Fallout = \frac{NRa}{NR}$, onde NRa = número de documentos não relevantes recuperados e NR = número total de documentos não relevantes da base de dados.

A decisão de quais as medidas utilizarem numa avaliação depende da aplicação, havendo sempre discussões sobre a confiabilidade de tais medidas, afirma Silva(2003).

A TREC (*Text Retrieval Conference*) é uma conferência anual organizada pelo NIST (*National Institute of Standards and Technology*), fornece a infra-estrutura necessária para a avaliação em larga escala de metodologias de recuperação de texto. Consiste numa série de workshops de propostas onde o objetivo é comparar as diversas técnicas utilizadas pelos grupos participantes. Para cada tarefa há uma base de documentos com cerca de 2 gigabytes de texto (entre um milhão e um milhão e meio de documentos) e algumas consultas que estabelecem o que é a informação procurada e o que constitui um documento relevante. A TREC (*Text Retrieval Conference)*[3] sugere três medidas diferentes de avaliação: precisão em níveis específicos de revocação, precisão em pontos específicos dos documentos recuperados e precisão média dos documentos recuperados. (RODRIGUES, 2008)

a) Na precisão em níveis específicos de revocação, escolhe-se o número de níveis de revocação, por exemplo, 10 níveis: {0.1, 0.2, 0.3, ..., 1.0}. Os níveis correspondem a usuários que estão satisfeitos ao encontrarem 10%, 20%, 30%, ..., 100% dos documentos relevantes.

Para cada um dos níveis a precisão correspondente é calculada, por exemplo, se desejamos encontrar a precisão no nível 0.5 calculamos a razão entre o número de documentos relevantes recuperados e o número de documentos recuperados quando o total de documentos relevantes recuperados corresponde a 50% do total de documentos relevantes existentes na base de documentos.

b) O método do número fixo de pontos na lista de documentos recuperados pode atender às necessidades mais específicas dos usuários, por exemplo: os 5, 10, 30, 100 e 1000 mais relevantes documentos recuperados. Estes pontos correspondem a usuários que procuram 5, 10, 30, 100 ou 1000 documentos por consulta, o que obriga o sistema a recuperar a quantidade de documentos solicitada, independente de haver esta quantidade de documentos relevantes para a consulta em questão.

O problema com esse tipo de método é que a precisão estará restrita a pequenas faixas, ou pontos fixos, da lista de documentos recuperados. Por exemplo, se para uma consulta existem 50 documentos relevantes na base de documentos, a precisão no ponto fixo 100 será

[3] As informações relativas às conferências da TREC, bem como as coleções de testes, podem ser encontradas em http://trec.nist.gov/

de no máximo 0.5, e no ponto fixo 1000 será de no máximo 0.05, o que pode não representar a qualidade da avaliação de desempenho de um sistema de busca que esteja conseguindo recuperar todos os 50 documentos relevantes.

c) O método da precisão média é calculada tomando-se a média aritmética das precisões de cada uma das consultas efetuadas para a avaliação. Por exemplo, suponhamos que uma determinada avaliação consista de três consultas, uma com 80 documentos relevantes, outra com 10 documentos relevantes e a terceira com 30 documentos relevantes. O sistema recuperou 24 documentos relevantes entre os documentos recuperados para a primeira consulta, 6 documentos relevantes entre os documentos recuperados para a segunda consulta e 15 documentos relevantes entre os documentos recuperados para a terceira consulta. Então a precisão média será:

$$precisão\_média = \frac{(24/80) + (6/10) + (15/30)}{3} = 0{,}47$$

Diante disso, podemos dizer que a precisão média é uma medida que privilegia sistemas que recuperam primeiramente os documentos relevantes.

6.1 Metodologia de avaliação do Sistema GIRS

Nesta seção será explicada a metodologia utilizada para avaliação do Sistema GIRS.

Para avaliação dos resultados do Sistema GIRS foi utilizada a métrica de precisão média[4] que, segundo Rodrigues (2008), é uma das métricas mais utilizada em recuperação da informação, e utilizada para avaliação de diversos sistemas de recuperação de informação como em GOMES (2001), GOTTSCHALG-DUQUE (2005) e NASCIMENTO (2004).

[4] O método da precisão média é calculada tomando-se a média aritmética das precisões de cada uma das consultas efetuadas para a avaliação.

As coleções de testes utilizadas foram a *Time collection* e a *Medline collection*. A *Time collection* consiste em uma coleção com 423 artigos selecionados da revista Time versando sobre assuntos diversos, já a *Medline collection* armazena 1033 resumos da área médica extraídos da *National Library of Medicine*. Estas coleções podem ser encontradas no Departamento de Ciência da Computação da Universidade de Glasgow[5].

Estas coleções foram escolhidas com a finalidade de testar o Sistema com uma base de documentos de uma área específica, neste caso da medicina (*medline collection*), e uma outra base com documentos sobre assuntos diversos e de áreas diferentes (*time collection*).

Neste trabalho utilizamos as coleções de documentos referidas acima mas desconsideramos as consultas e os julgamentos de relevância por elas sugeridas, uma vez que o sistema GIRS tem a finalidade de recuperação de documentos através da busca por palavras chaves e não como as consultas que compõem as coleções escolhidas, ou seja, através de sentenças.

Desta forma, se faz necessária a comparação dos resultados obtidos pelo Sistema com o resultado obtido em um sistema de busca de documentos tradicional e desenvolvido por uma empresa com boa credibilidade.

A ferramenta escolhida para servir de base de comparação foi o *Google Desktop Search*[6]. O *Google Desktop Search* é um aplicativo de busca no desktop que permite encontrar termos em arquivos, mensagens de e-mails, páginas webs visitadas, entre outros documentos, e apresenta como uma das opções de classificação pela relevância dos termos. O aplicativo cria um índice do conteúdo e o armazena no próprio computador para posteriormente encontrar a informação solicidada.

O resultado de cada consulta efetuada pelo Sistema GIRS foi comparado com o resultado obtido pela consulta ao *Google Desktop Search*. Através desta comparação se efetuou os cálculos de precisão do Sistema GIRS em relação ao *Google Desktop Search.*

[5] As coleções de testes utilizadas podem ser encontradas em http://ir.dcs.gla.ac.uk/resources/test_collections/

[6] O *Google Desktop Search* tem seu desenvolvimento sob a responsabilidade da *Google Inc.*, com sede na cidade de Mountain View, Califórnia, USA. Para download acesse http://desktop.google.com/

Para efeito de cálculos, foi tomado por base os 20 primeiros documentos sugeridos por ambos os Sistemas, uma vez que, conforme Rodrigues(2008), quanto maior a quantidade de documentos recuperados, maior será o número de documentos não relevantes recuperados, tornando desta forma, a busca mais dispersa.

6.2 Considerações

As diversas formas de avaliação dos sistemas de recuperação da informação buscam medir, através das métricas específicas, a eficaçia dos sistemas avaliados comparando-os a bases de dados reconhecidas e amplamente testadas pela comunidade de pesquisadores. Porém, a avaliação destes sistemas nescessitam de uma melhor análise que só pode ser determinada após submetidos os sistemas a usuários reais.

Dentre as métricas sugeridas pela literatura, optou-se por utilizar a da precisão média, uma vez que é a mais utilizada para a avaliação de sistemas similares ao deste trabalho.

O próximo capítulo apresenta os resultados obtidos com as avaliações do Sistema GIRS.

7. RESULTADOS

Neste capítulo serão apresentados os resultados obtidos, as tabelas de valores e os gráficos de precisão e explanada a análise destes resultados.

O Sistema GIRS foi submetido a diversas consultas sobre as coleções de documentos, que se dividiram em dois grupos de termos a serem consultados, termos específicos da área médica (*arterial, pulmonary, health, heart, metabolism, cardiac.*) e termos genéricos da linguagem original em que os documentos estão escritos (*dinner, month, exercise, reported, advance, position*).

7.1 Consultas com termos da área médica

O primeiro conjunto de consultas foram feitas apenas utilizando termos específicos da área médica (*arterial, pulmonary, health, heart, metabolism, cardiac*), os termos foram escolhidos após verificar-se que a frequência destes termos nos documentos da base são significativos.

Os termos foram submetidos primeiramente aos documentos extraídos da *Medline collection*. A Tabela 3 apresenta a tabulação dos valores de precisão para cada uma das dez consultas efetuadas com os dez termos da área da medicina, bem como as precisões médias para cada termo.

Tabela 3 – Valores de precisão para termos da área da medicina para Medline Collection
Fonte: elaborado pelo autor

VALORES DE PRECISÃO											
Termos consultados	**Cons 1**	**Cons 2**	**Cons 3**	**Cons 4**	**Cons 5**	**Cons 6**	**Cons 7**	**Cons 8**	**Cons 9**	**Cons 10**	**Média**
Arterial	0,800	0,850	0,700	0,700	0,750	0,700	0,750	0,700	0,700	0,700	**0,735**
Syndrome	0,600	0,550	0,700	0,750	0,550	0,600	0,750	0,550	0,750	0,600	**0,640**
Pulmonary	0,300	0,600	0,550	0,500	0,650	0,500	0,250	0,600	0,650	0,650	**0,525**
Health	1,000	0,950	0,950	1,000	1,000	1,000	0,950	0,950	1,000	0,950	**0,975**
Heart	0,600	0,650	0,500	0,600	0,550	0,650	0,600	0,450	0,500	0,600	**0,570**
Diabetes	0,550	0,750	0,750	0,800	0,750	0,800	0,700	0,700	0,700	0,600	**0,710**
Patients	0,400	0,400	0,400	0,400	0,450	0,400	0,400	0,450	0,500	0,350	**0,415**
Metabolism	0,700	0,700	0,700	0,700	0,650	0,650	0,700	0,700	0,750	0,600	**0,685**
Cardiac	0,600	0,650	0,550	0,600	0,550	0,600	0,500	0,600	0,550	0,650	**0,585**
Plasma	0,550	0,550	0,650	0,600	0,650	0,750	0,600	0,600	0,600	0,700	**0,625**

Posteriormente, os termos foram submetidos ao conjunto dos documentos extraídos da *Medline collection + Time collection*. A Tabela 4 apresenta a tabulação dos valores de precisão para cada uma das dez consultas efetuadas com os dez termos da área da medicina, bem como as precisões médias para cada termo.

Tabela 4 – Valores de precisão para termos da área da medicina para Medline Collection+Time Collection
Fonte: elaborado pelo autor

VALORES DE PRECISÃO											
Termos consultados	**Cons 1**	**Cons 2**	**Cons 3**	**Cons 4**	**Cons 5**	**Cons 6**	**Cons 7**	**Cons 8**	**Cons 9**	**Cons 10**	**Média**
Arterial	0,400	0,450	0,450	0,500	0,450	0,400	0,450	0,500	0,450	0,400	**0,445**
Syndrome	0,550	0,500	0,500	0,600	0,550	0,600	0,450	0,550	0,600	0,600	**0,550**
Pulmonary	0,450	0,400	0,400	0,450	0,400	0,450	0,450	0,500	0,500	0,450	**0,445**
Health	0,900	0,950	0,900	0,850	0,850	0,900	0,950	1,000	0,900	0,850	**0,905**
Heart	0,350	0,300	0,350	0,350	0,300	0,450	0,400	0,350	0,400	0,350	**0,360**
Diabetes	0,600	0,650	0,750	0,700	0,700	0,650	0,550	0,600	0,600	0,700	**0,650**
Patients	0,350	0,400	0,450	0,400	0,500	0,350	0,400	0,500	0,500	0,400	**0,425**
Metabolism	0,500	0,450	0,400	0,400	0,500	0,450	0,450	0,500	0,550	0,550	**0,475**
Cardiac	0,600	0,650	0,500	0,650	0,550	0,600	0,600	0,600	0,500	0,600	**0,585**
Plasma	0,500	0,550	0,550	0,650	0,600	0,600	0,550	0,500	0,550	0,550	**0,560**

O gráfico da Figura 24 apresenta o valor de precisão média para os resultados obtidos pelo Sistema GIRS para o primeiro conjunto de consultas com os termos da área da medicina para os documentos da *Medline Collection* e da *Medline Collection + Time Collection*.

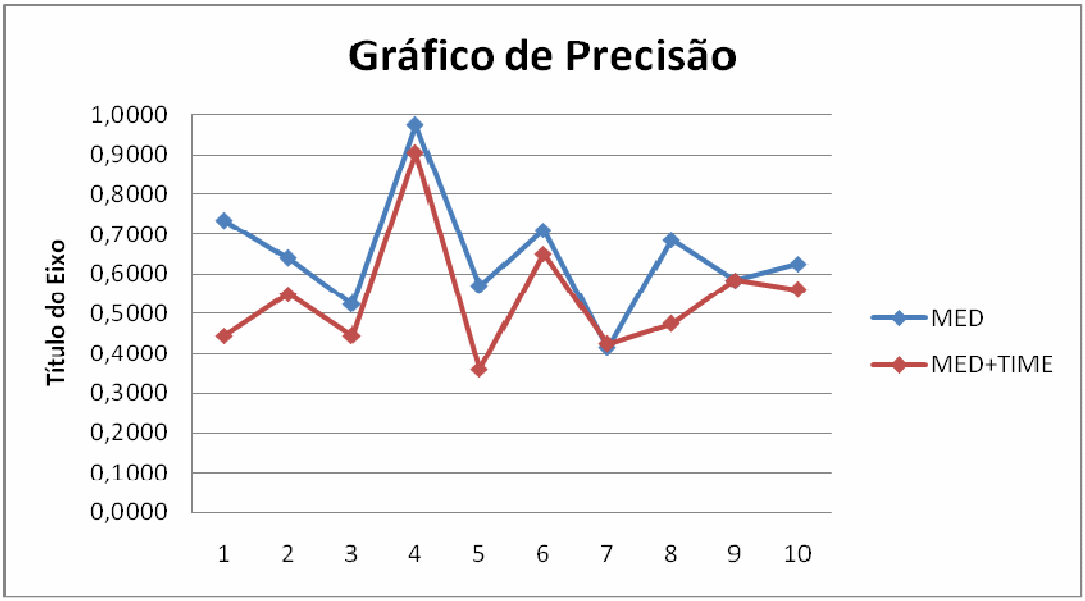

Figura 24 – Gráfico de Precisão Média das Consultas com Termos da Área da Medicina
Fonte: elaborado pelo autor

Os valores representados no gráfico da Figura 24 são as médias obtidas após a execução de dez consultas para cada termo com cada conjunto de documentos testado (*medline collection e medline collection + time collection).*

Pode-se verificar que para as consultas efetuadas apenas com os documentos da *Medline Collection*, a média de precisão para os dez termos é de 0,646 enquanto que para as consultas efetuadas com os documentos das duas coleções juntas, *Medline Collection + Time Collection*, a média de precisão caiu para 0,540.

7.2 Consultas com termos genéricos da linguagem

No segundo conjunto de consultas foram utilizados termos genéricos da linguagem original em que os documentos estão escritos (*dinner, month, exercise, reported, advance,*

position) os termos foram escolhidos da mesma forma que os anteriores, ou seja, após verificar-se que a frequência destes termos nos documentos da base são significativos.

Os termos selecionados foram submetidos primeiramente aos documentos da *Medline collection*. A Tabela 5 apresenta a tabulação dos valores de precisão para cada uma das dez consultas efetuadas com os dez termos genéricos, bem como as precisões médias para cada termo.

Tabela 5 – Valores de precisão para termos genéricos para Medline Collection
Fonte: elaborado pelo autor

VALORES DE PRECISÃO											
Termos consultados	**Cons 1**	**Cons 2**	**Cons 3**	**Cons 4**	**Cons 5**	**Cons 6**	**Cons 7**	**Cons 8**	**Cons 9**	**Cons 10**	**Média**
Reported	0,400	0,350	0,450	0,500	0,350	0,450	0,500	0,500	0,450	0,400	0,435
Buffer	0,950	1,000	1,000	0,900	0,950	0,850	1,000	0,850	0,900	0,950	0,935
Development	0,600	0,550	0,550	0,650	0,600	0,500	0,550	0,550	0,500	0,500	0,555
Dinner	0,950	1,000	1,000	1,000	0,950	0,950	0,900	1,000	1,000	1,000	0,975
Month	0,450	0,300	0,350	0,350	0,300	0,350	0,400	0,350	0,400	0,400	0,365
Exercise	0,910	0,850	0,850	0,900	0,830	0,850	0,910	0,830	0,860	0,910	0,870
Suspected	0,900	0,800	0,800	0,800	0,850	0,850	0,850	0,850	0,850	0,800	0,835
Advance	0,880	0,820	0,760	0,880	0,880	0,880	0,760	0,820	0,820	0,760	0,826
Position	0,600	0,600	0,550	0,550	0,600	0,550	0,650	0,550	0,650	0,650	0,595
Assembly	0,500	0,550	0,600	0,650	0,600	0,650	0,600	0,550	0,600	0,600	0,590

Posteriormente, os termos foram submetidos ao conjunto dos documentos extraídos da *Medline collection + Time collection.* A Tabela 6 apresenta a tabulação dos valores de precisão para cada uma das dez consultas efetuadas com os dez termos genéricos, bem como as precisões médias para cada termo.

Tabela 6 – Valores de precisão para termos genéricos para Medline Collection+Time Collection
Fonte: elaborado pelo autor

VALORES DE PRECISÃO											
Termos consultados	**Cons 1**	**Cons 2**	**Cons 3**	**Cons 4**	**Cons 5**	**Cons 6**	**Cons 7**	**Cons 8**	**Cons 9**	**Cons 10**	**Média**
Reported	0,250	0,300	0,450	0,300	0,300	0,400	0,300	0,300	0,300	0,300	0,320
Buffer	0,750	0,850	0,850	0,800	0,800	0,850	0,800	0,750	0,800	0,850	0,810
Development	0,500	0,550	0,450	0,450	0,450	0,450	0,550	0,300	0,400	0,350	0,445
Dinner	0,800	0,800	0,733	0,867	0,733	0,733	0,867	0,800	0,867	0,733	0,793
Month	0,400	0,350	0,350	0,250	0,400	0,400	0,350	0,400	0,350	0,350	0,360
Exercise	0,900	0,850	0,900	0,850	0,800	0,850	0,850	0,800	0,900	0,850	0,855
Suspected	0,750	0,800	0,700	0,800	0,850	0,800	0,800	0,750	0,800	0,800	0,785
Advance	0,700	0,700	0,700	0,700	0,650	0,650	0,700	0,700	0,750	0,600	0,685
Position	0,600	0,650	0,550	0,600	0,550	0,600	0,500	0,600	0,550	0,650	0,585
Assembly	0,500	0,500	0,550	0,650	0,550	0,650	0,550	0,550	0,500	0,600	0,560

O gráfico da Figura 25 apresenta os valores de precisão média para os resultados obtidos pelo Sistema GIRS para o segundo conjunto de consultas com os termos genéricos da linguagem original em que os documentos estão escritos, neste caso a língua inglesa, para os documentos da *Medline Collection* e da *Medline Collection + Time Collection*.

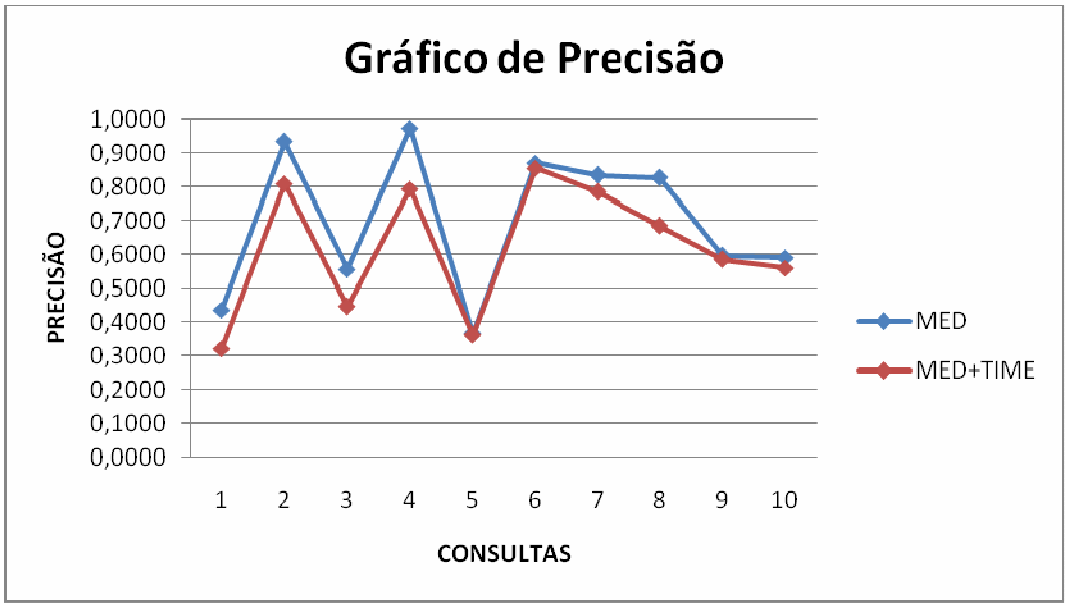

Figura 25 – Gráfico de Precisão Média das Consultas com Termos Genéricos
Fonte: elaborado pelo autor

Os valores representados no gráfico da Figura 25 são as médias obtidas após a execução de dez consultas para cada termo com cada conjunto de documentos testado.

Pode-se verificar que para as consultas efetuadas apenas com os documentos da *Medline Collection*, a média de precisão para os dez termos é de 0,698 enquanto que para as consultas efetuadas com os documentos das duas coleções juntas, *Medline Collection + Time Collection*, a média de precisão é de 0,620.

CONCLUSÃO

A cada dia que passa cresce a quantidade de documentos gerados e armazenados nas empresas e corporações. Com isso a necessidade de recuperação de informações nestes documentos torna-se cada vez mais importante e as dificuldades em encontrá-las também.

Diversos mecanismos de busca comerciais estão disponíveis no mercado, mas nem sempre estes produtos atendem as necessidades dos usuários, pois neste tipo de buscadores, serão recuperados todos os documentos que contenham os termos solicitados nas consultas dos usuários, resultando em uma quantidade muito grande de documentos e a consequente dificuldade de selecionar os documentos mais relevantes ao interesse do usuário. Cada sistema utiliza um algoritmo específico para ordenar os documentos que serão sugeridos ao usuário em resposta à consulta efetuada.

Tentando resolver este problema, os pesquisadores têm experimentado técnicas de inteligência artificial como redes neurais, sistemas *fuzzy*, sistemas multiagentes e algoritmos genéticos, em substituição de métodos puramente matemáticos, com a intenção de melhorar a relevância da recuperação da informação por meio de seus sistemas. A utilização dos AGs possibilita a obtenção de respostas diferentes para consultas efetuadas com os mesmos termos, o que dá ao sistema uma característica de diversificação das respostas oferecidas aos usuários.

Os algoritmos genéticos são baseados na teoria da evolução das espécies de Darwin, e sua utilização na recuperação da informação apresenta-se como uma alternativa para implementação de sistemas com características evolutivas.

Outra técnica importante que tem sido estudada no intuito de melhorar a relevância na

recuperação da informação é incluir nos sistemas o *feedback* dos usuários.

Em vista disto, esta dissertação propôs um sistema de recuperação de informações onde a relevância dos documentos recuperados evolua a cada iteração através do uso de algoritmos genéticos e com a ajuda direta dos usuários deste sistema através de um *feedback* implícito , visando uma melhora no tocante à relevância dos documentos recuperados, a cada nova consulta.

Analisando os resultados das avaliações efetuadas, verifica-se um bom nível de precisão média[7] quando o Sistema GIRS é submetido a documentos da base de documentos[8] de um domínio específico (no caso da área médica), tanto com termos específicos da área médica[9] como com termos genéricos da linguagem original em que os documentos estão escritos[10], no caso a língua inglesa.

Estes resultados são promissores na busca por melhores resultados e indicam que os algoritmos genéticos podem ser uma excelente técnica para auxiliar a área da recuperação da informação, devido a diversidade das respostas obtidas com a utilização desta técnica.

Os resultados analisados foram obtidos com coleções de testes, o que significa que para se ter um melhor dimensionamento da eficácia do sistema proposto, outros testes com usuários reais são necessários e os resultados apresentados por estes usuários deverão ser analisados. Um ponto importante a ser verificado é o tempo de execução da busca (em torno de 18 a 20 segundos), devido a ser inerente aos algoritmos genéticos o aumento no tempo de resposta com o aumento do tamanho dos indivíduos (quanto maior o número de documentos na base, maior serão os indivíduos).

---

[7] Quando comparados com os resultados do Google Desktop Search.

[8] As bases de documentos testadas foram a *Medline Collection* com 1033 documentos e a *Time Collection* com 423 documentos.

[9] Termos específicos da área médica (*arterial, pulmonary, health, heart, metabolism, cardiac*).

[10] Termos genéricos da linguagem original em que os documentos estão escritos (*dinner, month, exercise, reported, advance, position*).

Contudo, pode-se verificar que os sistemas de recuperação de informação com o auxílio dos algorítmos genéticos podem ser melhor explorados, com o intuito de alcançar uma melhora na relevância dos documentos recuperados.

Trabalhos Futuros

Esta dissertação permitiu que fossem identificados possíveis trabalhos de pesquisa a serem realizados, com o objetivo de melhorar este trabalho.

Como trabalhos futuros, podemos citar:

- a inclusão no sistema GIRS, de buscas em linguagem natural;

- criação de uma interface para que os usuários possam fornecer o *feedback* de relevância de forma explícita, indicando os documentos relevantes e os documentos não relevantes, ou ainda determinar o grau de relevância dos documentos, para que o sistema possa atualizar os pesos armazenados na base de dados;

- inclusão de um módulo indexador que efetue a indexação automática dos documentos, com a possibilidade de personalização de quais os tipos de documentos, os locais a serem verificados e a periodicidade da indexação, independente da atuação do administrador do sistema;

- pesquisar novas formas de codificação dos algoritmos genéticos com a finalidade de redução no tempo de execução das buscas, uma vez que os tempos obtidos variaram entre 18 e 20 segundos.

REFERÊNCIAS

AIRES, Rachel V. X. *Uma arquitetura lingüisticamente motivada para recuperação de informação de textos em português.* Prova de qualificação, USP, São Paulo, SP, 2006. Disponível em: http://www.linguateca.pt/documentos/QualificacaoRachelAires.pdf. Acesso em 01 jul 2009.

ALVES, A. C. ; SCHREIBER, J. N. C. ; FURTADO, J. C. ; KONZEN, A. A. ; MOLZ, R. F. . *GIRS - Genetic information Retrieval System*. In: ENEGEP, 2008, RIO DE JANEIRO. XXVIII Encontro Nacional de Engenharia de Produção, Rio de Janeiro: ABEPRO, 2008. v. 1.

ALVES, Antonio C. *Intrabusca - Um mecanismo de recuperação de informações corporativo*, Trabalho de Conclusão, UNISC, Santa Cruz do Sul, RS, Brasil, 2005.

BAEZA-YATES, Ricardo, RIBEIRO-NETO, Berthier. *Modern Information Retrieval.* New York: Addison-Wesley, 1999.

BARCELLOS, João Carlos Holland de. *Algoritmos genéticos adaptativos: um estudo comparativo*. Dissertação (Mestrado em Engenharia), Escola Politécnica, USP, São Paulo, 2000. Disponível em http://www.teses.usp.br/teses/disponiveis/3/3141/tde-05092001-141334/ Acesso em 10 ago 2009.

BUENO, R. et al. *Algoritmos Genéticos para Consultas por Similaridade Aproximadas.* Uberlândia, MG. Edições SBC, 2005. Disponível em: http://www.sbbd-sbes2005.ufu.br/arquivos/artigo-13-BuenoTraina.pdf. Acesso em 18 jul 2009.

CAZELLA, Sílvio César. *Aplicando a Relevância da Opinião de Usuários em Sistema de Recomendação para Pesquisadores.* Tese (Doutorado em Ciência da Computação), Instituto de Informática, UFRGS, Porto Alegre, 2006. Disponível em http://www.inf.unisinos.br/~cazella/papers/VersaoFinal2006TeseSilvioCazellahomologacao.pdf. Acesso em 15 set 2009.

CARDOSO, Olinda N. P. *Recuperação de Informação.* Infocomp Revista de Computação da UFLA, Lavras, v. 1, 2000. Disponível em: http://www.dcc.ufla.br/infocomp/artigos/v2.1/art07.pdf. Acesso em 28 ago 2009.

CARVALHO, André Ponce de Leon F. de. *Algoritmos Genéticos.* 2008. Disponível em: http://www.icmc.usp.br/~andre/research/genetic/. Acesso em 20 jan 2009.

CRESTANI, Fabio; van RIJSBERGEN, Cornelis J. *A model for adaptative information retrieval*. Journal of Intelligent Information Systems, v.8, 1997.

DANA, Vrajitoru. *Genetic Algorithms in Information Retrieval.* Université de Neuchâtel Institut Interfacultaire d'Informatique, Neuchâtel, Switzerland, 1997. Disponível em: http://www.cs.iusb.edu/~danav/papers/AidriEng.pdf. Acesso em 10 jan 2009.

DINIZ, Alexandre S. *Mecanismos e Estruturas de Busca Semântica*, UNIMONTES, Montes Claros, MG , Brasil, 2004. Disponível em http://www.ccet.unimontes.br/arquivos/monografias/16.pdf. Acesso em 21 set 2008.

FAGUNDES, Ricardo C. *Aplicação de Consultas Baseadas em Similaridade em Ambientes de Conhecimento Definidos por Tesauros.* Trabalho de Conclusão em Ciência da Computação – UNISC, Santa Cruz do Sul, RS, 2007.

FERNEDA, Edberto. *Recuperação da informação; Análise sobre a contribuição da ciência da computação para a ciência da informação.* Tese (Doutorado em Ciências da Comunicação) – Escola de Comunicação e Artes – USP, São Paulo, 2003. Disponível em: http://www.Teses.usp.br/teses/disponiveis/27/27143/tde-15032004-130230/publico/Tese.pdf. Acesso em 22 jul 2009.

FERNEDA, E., PINHEIRO, C.B.F. *Representação Dinâmica de Documentos em Bibliotecas Digitais*. In: 3º Simpósio Internacional de Bibliotecas Digitais, São Paulo, 2005. Disponível em http://bibliotecas-cruesp.usp.br/3sibd/docs/ferneda171.pdf. Acesso em 10 jan 2009.

FERRREIRA, A.C.P.L; BRAGA, A. P.; LUDERMIR, T.B. *Sistemas Inteligentes – Fundamentos e Aplicações.* São Paulo: Manole, 2003.

GOLDBERG, David E.. *Genetic Algorithms in Search, Optimization, and Machine Learning.* EUA: Addison-Wesley, 1989.

GOMES, Edeyson A.. *Fidus: Uma Ferramenta para Busca de Informações Personalizadas na Web*. Dissertação (Mestrado em Informática), UFPA, Paraíba, Brasil, 2001. Disponível em: http://www.edeyson.com.br/Arquivos/Fidus/Dissertacao_Edeyson.pdf. Acesso em 25 ago 2009.

GOTTSCHALG-DUQUE, C. A. *SiRILiCO-Uma Proposta para um Sistema de Recuperação de Informação baseado em Teorias da Lingüística Computacional e Ontologia*. Tese(Doutorado em Ciência da Informação), Escola de Ciência da Informação – UFMG, Belo Horizonte, 2005. Disponível em: http://www.bibliotecadigital.ufmg.br/dspace/bitstream/1843/EARM-7HBND8/1/doutorado_claudio_gottschalg_duque.pdf. Acesso em 15 ago 2009.

GEY, F. *Models in Information Retrieval*. Folders of Tutorial Presented at the 19th ACM Conference on Research and Development in Information Retrieval (SIGIR), 1992.

HOLLAND, J. H. *Adaptation in Natural and Artificial Systems*. The University of Michigan, USA: MIT Press, 1992.

KURAMOTO, Hélio. *Sintagmas nominais: uma nova proposta para a recuperação de informação*. Revista de Ciência da Informação, volume 3, fev. 2002. Disponível em: <http://dici.ibict.br/archive/00000060/01/Artigo_03.htm>. Acesso em 21 set 2008.

LAKATOS, Eva M.; MARCONI, Marina de A. *Fundamentos de metodologia científica.* São Paulo: Atlas, 2005.

LINDEN, Ricardo. *Algorítmos Genéticos: Uma importante ferramenta da Inteligência Computacional*. Rio de Janeiro: Brasport, 2006

LOBO, Eduardo Luiz Miranda. *Uma Solução do Problema de Horário Escolar via Algoritmo Genético Paralelo.* Dissertação de Mestrado em Modelagem Matemática e Computacional. CEFET-MG - Centro Federal de Educação Tecnológica de Minas Gerais. Belo Horizonte, 2005.

LOPES, Luciana de S.; LORENA, Luiz A. N. *Uma heurística baseada em algoritmos genéticos aplicada ao problema de cobertura de conjuntos.* Dissertação (Mestrado em Computação Aplicada). INPE, São José dos Campos, 1995. Disponível em: www.lac.inpe.br/~lorena/luciana/tese-luciana.pdf. Acesso em 25 mar 2009.

MILANI, Fábio; CAZELLA, Silvio César. *Um modelo para determinar a autoridade de usuários em Sistemas de Recomendação*. Canoas, RS. III Fórum de Inteligência Artificial. 2005. Disponível em http://www.inf.unisinos.br/~cazella/papers/forumia_Milani_Cazella.pdf. Acesso em 15 set 2009.

MIRANDA, Marcio Nunes. *Algoritmos Genéticos – Fundamentos e Aplicações.* UFRJ, 2008. Disponível em: http://www.gta.ufrj.br/~marcio/genetic.html. Acesso em 25 mar 2009.

MOURA, Gevilacio Aguiar Coêlho de. *Sistemas de busca da web: diretórios e mecanismos de busca*. 2001. Disponível em http://www.quatrocantos.com/tec_web/sist_busca/index.htm Acesso em 20 jun 2009.

NASCIMENTO, Luiz Antonio do. *Proposta de um Sistema de Recuperação de Informação para Extranet de Projeto.* Dissertação (Mestrado em Engenharia), Escola Politécnica, USP, São Paulo, 2004. Disponível em http://www.teses.usp.br/teses/disponiveis/3/3146/tde-13052004-140558/. Acesso em 20 jul 2009.

OCHI, Luis Satoru. *Algoritmos Genéticos: Origem E Evolução.* 1998. Disponível em http://www.sbmac.org.br/bol/bol-2/artigos/satoru/satoru.html. Acesso em 10 jan 2009.

PATHAK, P.; GORDON, M.; WEIGUO, F. *Effective information retrieval using genetic algorithms based matching functions adaptation.* Hawaii: IEEE, 2000. Disponível em: http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?tp=&arnumber=926653&isnumber=20043. Acesso em 15 jul 2009.

PACHECO, Marco Aurélio C. *Algoritmos Genéticos: Princípios e Aplicações.* Disponível em: http://www.ica.ele.pucrio.br/publicacoes/download/cnf_0111.pdf. Acesso em 20 jan 2009.

RADWAN, Ahmed A. A., et al. *Using Genetic Algorithm to Improve Information Retrieval Systems*. Proceedings of world academy of science, engineering and technology, volume 17, Barcelona, 2006.

RODRIGUES, Edmilson F. *Aprendizagem Estatística para Recuperação da Informação.* Dissertação (Mestrado em Informática), Departamento de Ciência da Computação, Universidade de Brasília, Brasília, DF, 2008. Disponível em: http://monografias.cic.unb.br/dspace/handle/123456789/155. Acesso em 04 out 2009.

ROMANHUKI, Emerson. *Aprendizagem de políticas de oferta de negociação entre agentes cognitivos*. Dissertação (Mestrado em Informática), PUC-PR, Curitiba, 2008.

SALTON, Gerard. *Automatic Text Processing: The Transformation, Analisys, and Retrieval of Information by Computer*, Addison Wesley, 1989. Disponível em: http://portal.acm.org/citation.cfm?id=77013. Acesso em 10 ago 2009.

SANTOS, Thiago L. V. De; BARROS, Santos F. de A. *Algoritmos Genéticos para Ordenamento em Sistemas de Busca na Web.* Recife, PE, 2001. Disponível em: http://www.cin.ufpe.br/~tg/2000-2/tlvls.doc. Acesso em 20 set 2009.

SILVA, Mário J.; MARTINS, Bruno; COSTA, Miguel. *Avaliação de Sistemas de Recuperação de Informação da Web em Português: Proposta Inicial à Comunidade.* Avalon'03, Encontro de Avaliação Conjunta de Sistemas de Processamento Computacional do Português, Lisboa, Portugal, 2003. Disponível em: http://www.linguateca.pt/aval_conjunta/acetatosAvalon/AvalonXLDB.pdf. Acesso em 04 out 2009.

SOBRINHO, Antonio C. C. Da S.; GIRARDI, Maria Del R. *Uma Análise das Aplicações dos Algoritmos Genéticos em Sistemas de Acesso à Informação Personalizada.* Revista Eletrônica de Iniciação Científica, Ano III, Volume III, Número IV, SBC, Porto Alegre, RS, 2003. Disponível em: http://www.sbc.org.br/reic/edicoes/2003e4/tutoriais/AlgoritmosGeneticosEmSistemasDeAcessoAInformacaoPersonalizada.pdf. Acesso em 15 mai 2009.

SOUZA, M.J.F., MACULAN, N., OCHI. L.S. *Uma heurística para o problema do horário escolar*. Proc. of the X CLAIO (X Latin-Ibero-American Conference on Operations Research and Systems), México, 2000.

SOUZA, Renato Rocha. *Sistemas de recuperação de informações e mecanismos de busca na web: panorama atual e tendências*. Perspectivas em Ciência da Informação, vol.11, nº 2, Belo Horizonte, 2006. Disponível em http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1413-99362006000200002&lng=en&nrm=iso. Acesso em 15 set 2009.

TAGUE-SUTCLIFFE, J. M. *Some Perspectives on the Evaluation of Information Retrieval Systems*. Journal of the American Society for Information Science, vol. 47, nº 1, New York, 1996. Disponível em: http://portal.acm.org/citation.cfm?id=231883. Acesso em 03 out 2009.

YEH, Jen-Yuan; LIN, Jung-Yi; KE, Hao-Ren; YANG, Wei-Pang. *Learning to Rank for Information Retrieval Using Genetic Programming*. In CIGIR-2007 - Learning to Rank for Information Retrieval, Amsterdam, 2007. Disponível em: http://jenyuan.yeh.googlepages.com/jyyeh-LR4IR07.pdf. Acesso em 10 jul 2009.

WAZLAWICK, Raul S. *Metodologia Científica para Ciência da Computação.* Rio de Janeiro: Elsevier, 2008.

WIVES, Leandro K.. *Um Estudo Sobre Técnicas de Recuperação de Informações com Ênfase em Informações Textuais*. Trabalho Individual (TI). UFRGS, Porto Alegre, RS, Brasil, 1997. Disponível em http://www.inf.ufrgs.br/~wives. Acesso em 10 out 2008.

WIVES, Leandro Krug. *Tecnologias de descoberta de conhecimento em textos aplicadas à inteligência competitiva*. 2002. Exame de qualificação (Programa de pós-graduação em computação) - UFRGS, Instituto de Informática, Porto Alegre.

ANEXO A

Modelagem do Sistema

O sistema foi desenvolvido utilizando a linguagem de programação PHP e HTML para acesso diretamente por um navegador de Internet (Firefox, Internet Explorer, Mozila) e tem os seguintes arquivos implementados:

a) index: a tela de inicialização do sistema apresenta as opções de informar as *stopwords* a serem removidas do documento, escolher os documentos que serão indexados na base de dados e ir para a tela de busca.

b) Tela_busca: é a tela de busca, onde são oferecidas duas opções de consulta, busca direta e busca pelo GIRS(Genetic Information Retrieval System).

c) Buscar: este arquivo gerencia toda a consulta do sistema. Se a busca for direta, é montada uma página HTML e apresentados os documentos que possuem os termos de busca, ordenando-os de forma decrescente do peso armazenado na base de dados. Caso a busca seja pelo GIRS, este arquivo faz o gerenciamento do algoritmo genético para posterior apresentação ao usuário.

d) config: sua finalidade é fazer a conexão com o banco de dados.

e) inicializaPopulacao: gera de forma randômica os valores atribuídos (0 ou 1) ao Vetor de Presença correspondente a cada um dos 20 indivíduos da população inicial e também de forma randômica gera pesos para as posições com valor igual a 1 que serão armazenados

no Vetor dos Pesos. Estes indivíduos têm o tamanho igual ao número de documentos indexados na base de dados.

f) avaliaPopulacao: efetua a avaliação da população através da seguinte fórmula:

Avaliação = $\sum$(Pc - |Fp - Fr|)

Pc = Total de posições coincidentes entre os documentos escolhidos para a resposta (valores iguais a 1) do indivíduo avaliado e o de referência;

|Fp – Fr| = Para as posições coincidentes do item anterior, subtrai-se os valores dos pesos correspondentes onde, o módulo do resultado é subtraído da avaliação, onde Fp = Peso do indivíduo avaliado e Fr = peso do indivíduo de referência, valores estes calculados para os mesmos índices dos vetores.

g) ordena: ordena os indivíduos da população por ordem decrescente de avaliação.

h) roleta: cria a roleta com as porções proporcionais a avaliação de cada indivíduo para o processo de seleção para o cruzamento.

i) cruzamento: seleciona os indivíduos (pai e mãe) utilizando-se da roleta criada, determina o ponto de cruzamento e efetua o cruzamento dos indivíduos selecionados, gerando uma nova população;

j) avaliaNovaPopulacao: efetua a avaliação da nova população de forma análoga a fórmula do item da letra “f”.

k) ordenaNova: ordena os indivíduos da nova população por ordem decrescente de avaliação.

l) melhores: seleciona os 20 melhores indivíduos das duas populações para formar a população inicial para a próxima iteração do algoritmo genético.

m) mutacao: seleciona os indivíduos e os genes que sofrerão a mutação e efetua tal operação.

n) apresenta: este arquivo monta uma página HTML e apresenta os vinte documentos com melhor avaliação extraídos do indivíduo escolhido como sendo o mais apto pelo algoritmo genético.

o) click: expande o conteúdo total do documento a ser visualizado, o qual foi escolhido pelo usuário do sistema. Neste arquivo ocorre a atualização do peso do arquivo em relação ao termo de busca selecionado, somando-se uma constante ao valor já armazenado.

p) funcoes: arquivo de funções auxiliares para a indexação, remoção das *stopwords*, limpeza das tags HTML e calculo do peso(F).