GIRS - GENETIC INFORMATION RETRIEVAL SYSTEM - UMA PROPOSTA EVOLUTIVA PARA SISTEMAS DE RECUPERAÇÃO DE INFORMAÇÕES

- (-)
-

A recuperação de informação é uma subárea da ciência da computação que estuda o armazenamento e recuperação automática de documentos, que são objetos de dados, geralmente textos.O objetivo preliminar de um sistema de recuperação de informaçção é recuperar todos os documentos que são relevantes a uma solicitação do usuário com uma quantidade mínima de documentos não-relevantes.

Este artigo apresenta um sistema de recuperação de informações (SRI), o GIRS - Genetic Information Retrieval System. O GIRS foi projetado utilizando-se algoritmos genéticos, onde a informação sobre a relevância dos documentos recuperados é atualizada a cada iteração através do uso de algoritmos genéticos e com o feedback de relevância fornecido pelos usuários deste sistema. O feedback dos usuários, indicando quais documentos recuperados foram relevantes para sua consulta, também ajuda a melhorar a adaptação dos termos de busca em relação aos documentos, visando uma evolução na relevância das futuras consultas.

Palavras-chaves: Algoritmos Genéticos, Recuperação de Informação

1. Introdução

Desde a década de 40 a recuperação de informação tem sido tema de pesquisas, a conceituação de Recuperação de Informação, definida inicialmente por Calvin Mooers (Mooers, 1951 apud Saracevic, 1995), vem dada, de uma forma eminentemente funcional e não descritiva: *"Recuperação de informação é o nome do processo ou método onde um possível usuário de informação pode converter sua necessidade de informação numa lista real de citações de documentos armazenados que contenham informações úteis a ele . . . recuperação de informação abarca os aspectos intelectuais da descrição da informação e a sua especificação para busca, assim como também quaisquer sistemas, técnicas ou máquinas que sejam empregadas para efetuar a operação"*

O crescimento do volume de informação nas corporações sob a forma de documentos internos, normas, resoluções, atas, comunicações, etc, suscitou o desenvolvimento de técnicas de recuperação de informação para responder às necessidades dos usuários destas bibliotecas, tradicionais ou digitais. A ferramenta mais importante para auxiliar o processo de recuperação é denominada índice, que é uma coleção de termos que indicam o local onde a informação desejada pode ser localizada (Frakes, 1992). Estes termos devem ser organizados de forma a facilitar sua busca.

Atualmente já não se pode falar em crescimento do volume de publicações mas em uma verdadeira explosão. As bibliotecas digitais, que são publicações armazenadas e manipuladas eletronicamente, aparecem como um paradigma para melhorar a busca e apresentação de informações desejadas. Neste contexto são estudadas técnicas de digitalização de objetos originados de fontes heterogêneas, técnicas de armazenamento, processos de busca, recuperação e apresentação de forma amigável das informações. A indexação ainda é a principal ferramenta para recuperação de informação.

A crescente complexidade dos objetos armazenados e o grande volume de dados exigem processos de recuperação cada vez mais sofisticados. Diante deste quadro, recuperação de informação apresenta a cada dia, novos desafios e se configura como uma área de significância maior.

As corporações utilizam mecanismos de busca como o Google, Yahoo e Alta Vista nas suas bibliotecas digitais, onde os documentos apresentados como resultado das consultas, podem não apresentar uma relevância considerável para o usuário corporativo que efetivou a consulta. Isto tende a diluir as informações verdadeiramente relevantes em meio à grande quantidade de documentos recuperados.

Neste artigo apresentamos uma proposta de um sistema de recuperação de informações (SRI), o *GIRS* – Genetic Information Retrieval System onde a informação sobre a relevância dos documentos recuperados é atualizada a cada iteração através do uso de algoritmos genéticos e com o *feedback* de relevância fornecido pelos usuários deste sistema. O *feedback* dos usuários, indicando quais documentos recuperados foram relevantes para sua consulta, também ajuda a melhorar a adaptação dos termos de busca em relação aos documentos, visando uma evolução na relevância das futuras consultas.

O artigo é dividido em seis seções, esta introdução, na seção dois o foco é na metodologia de avaliação de um SRI, a seção três descreve os modelos clássicos de SRI. A partir da seção

quatro, são apresentados os pilares do sistema: retro-alimentação de relevância, algoritmos genéticos e o sistema *GIRS* propriamente dito.

2. Sistemas de Recuperação de Informação

A recuperação de informação é uma subárea da ciência da computação que estuda o armazenamento e recuperação automática de documentos, que são objetos de dados, geralmente textos. "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)

Um sistema de Recuperação de Informação (SRI) pode ser estruturado conforme a Figura 1 (Gey,1992). Conforme a figura 1, compõe um SRI: documentos e necessidades do usuário. A partir das necessidades do usuário é formulada uma consulta que dispara o processo de recuperação que, à partir das estruturas de dados e da consulta formulada, recupera uma lista de documentos considerados relevantes.

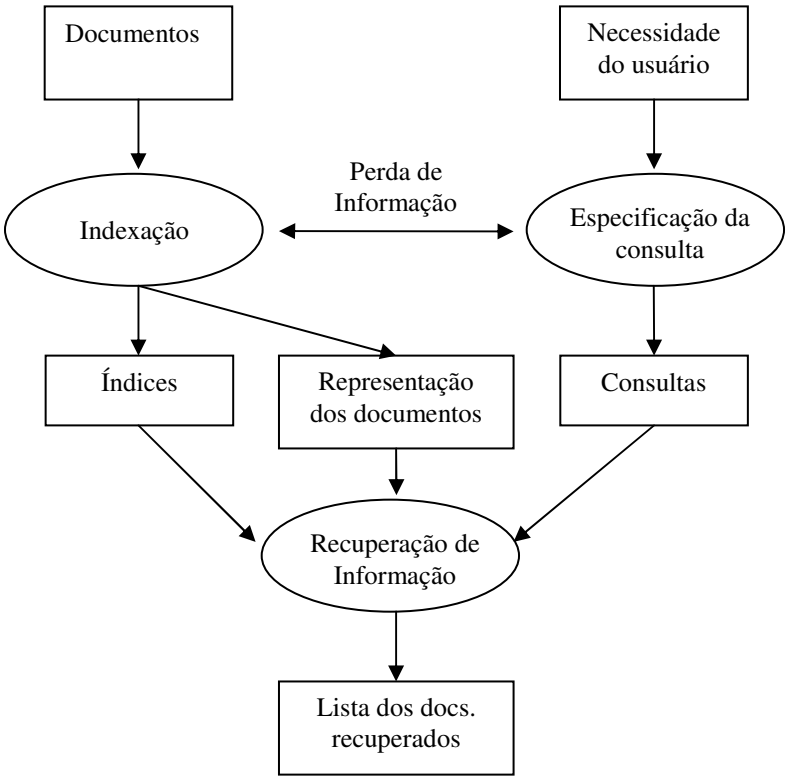

Figura 1 – Componentes de um SRI

O processo de indexação envolve a criação de estruturas de dados associados à parte textual dos documentos, por exemplo, as estruturas de arranjos de sufixos (*PAT arrays*) e arquivos invertidos, discutidas em (Frakes, 1992). Estas estruturas podem conter dados sobre características dos termos na coleção de documentos, tais como a freqüência de cada termo em um documento.

A especificação de uma consulta geralmente é uma tarefa difícil, freqüentemente há uma distância semântica entre a real necessidade do usuário e o que ele expressa na consulta formulada. Essa distância é gerada pelo limitado conhecimento do usuário sobre o universo de pesquisa e pelo formalismo da linguagem de consulta.

O processo de recuperação consiste na geração de uma lista de documentos recuperados para responder a consulta formulada pelo usuário. Os índices construídos para uma coleção de documentos são usados para acelerar esta tarefa. Além disso, a lista de documentos recuperados é classificada em ordem decrescente de um grau de similaridade entre o documento e a consulta. A grande questão é alinhar a similaridade com a relevância desejada pelo usuário.

2.1 Avaliação do sistema de Recuperação de Informação proposto

Os sistemas de recuperação de informação podem ser avaliados através de consultas que fazem parte de uma coleção de referência. No sistema proposto é utilizado uma coleção de documentos oriundos do cotidiano da universidade com cerca de dez mil documentos entre atas, resoluções e normas. Nesta coleção há um conjunto de consultas e para cada consulta é fornecido um conjunto ideal de documentos resposta, criado por especialistas nos temas envolvidos.

O SRI classifica os documentos recuperados para cada consulta, de acordo com uma ordem de relevância gerando um vetor resultado. Avalia-se o SRI através da comparação das respostas geradas por este sistema e o conjunto ideal de respostas. Para isso, o vetor resultado é examinado e comparado com o conjunto ideal, obtendo-se dois índices de avaliação: precisão e revocação. A precisão e revocação são medidas baseadas na noção de documentos relevantes de acordo com uma determinada necessidade de informação.

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) Os 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)

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 à sua busca e submete essa informação ao sistema. (FAGUNDES, 2007)

Conforme a figura 2, R é o conjunto de todos os objetos relevantes contídos na base de dados, normalmente desconhecido pelo usuário, A é o conjunto de todos os objetos retornados pela consulta e Ra é o conjunto de todos os objetos relevantes retornados pela consulta.

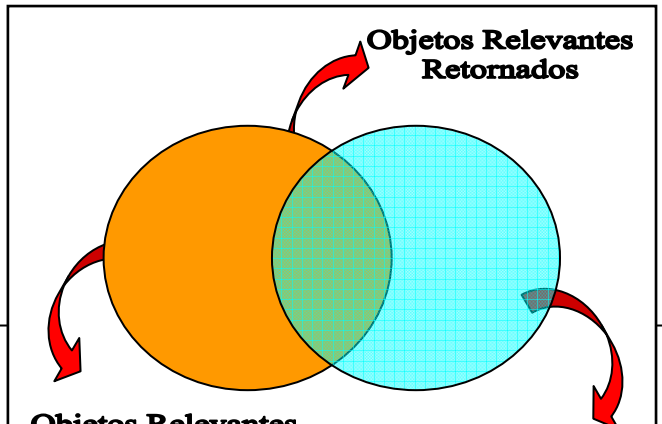

Figura 2 – Conjuntos de documentos usados na estimativa de precisão e revocação

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

Revocação: é a razão entre o número de objetos relevantes retornados (intersecção entre os conjuntos R e A) em relação a todos os objetos relevantes da base de dados (conjunto R), conforme equação 1:

$$\text{Revocação: } \frac{Ra}{R} \qquad (1)$$

Precisão: é a razão entre o número de objetos relevantes retornados (intersecção entre os conjuntos R e A) em relação ao número de objetos recuperados (conjunto A), conforme equaçao 2:

$$\text{Precisão: } \frac{Ra}{A} \qquad (2)$$

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

Um dos grandes problemas encontrados na implementação das métricas definidas acima é determinar a quantidade de documentos relevantes a ser utilizada (conjunto R) no cálculo da recall. Neste contexto surge uma contradição: Uma vez conhecendo-se a quantidade de documentos relevantes, certamente houve uma forma de apurar esta contagem. Por causa desta contradição é que utilizamos um conjunto previamente analisado por especialistas humanos.

3. Modelos clássicos

Os modelos clássicos, utilizados no processo de recuperação de informação (booleano, vetorial e probabilístico) apresentam estratégias de busca de documentos relevantes para uma consulta (*query*).

Estes modelos consideram que cada documento é descrito por um conjunto de palavras chaves, chamadas termos de indexação. Associa-se a cada termo de indexação t_i em um documento d_j um peso $w_{ij} > 0$, que quantifica a correlação entre os termos e o documento.

Além dos modelos clássicos, modelos muito mais avançados de recuperação de informação tem sido propostos ao longo dos anos, dentre estes, destacam-se modelos baseados em bases de conhecimento Biwas *apud* Cardoso (2000), lógica *fuzzi* Bookstein *apud* Cardoso (2000) e redes neurais Kwok *apud* Cardoso (2000).

3.1 Modelo Booleano

Dada uma consulta Q e um conjunto de documentos considerados relevantes para a Q, o índice atribuído aos documentos deve indicar qual documento é mais relevante que outro, estabelecendo uma ordem de relevância. Esses índices são calculados com base na comparação entre a consulta e os documentos.

No modelo booleano os documentos recuperados são aqueles que contêm os termos que satisfazem a expressão lógica da consulta. Uma consulta é considerada como uma expressão booleana convencional formada com os conectivos lógicos *AND*, *OR* e *NOT*.

Uma maneira direta de implementar o modelo booleano seria Salton(1989): assuma a existência de uma lista invertida na qual cada entrada corresponde a um termo de indexação, ademais, a entrada t_i aponta para uma lista de documentos nos quais o termo t_i ocorre. O conjunto de documentos recuperados pode ser obtido pela interseção das listas invertidas de documentos, dos termos que aparecem na consulta. Assim, somente documentos cujos termos de indexação satisfazem a consulta booleana são recuperados.

Os principais problemas do modelo booleano são a ausência de ordem na resposta, e as respostas podem ser nulas ou muito grandes. As vantagens desse modelo são a facilidade de implementação, e a expressividade completa das expressões.

3.2 Modelo vetorial

O modelo vetorial, representa documentos e consultas como vetores de termos. Termos são ocorrências únicas nos documentos. Os documentos devolvidos como resultado para uma consulta são representados similarmente, ou seja, o vetor resultado para uma consulta é montado através de um cálculo de similaridade.

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 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, através da seguinte fórmula Salton (1988):

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

Figura 3 – Cálculo de Similaridade

Os pesos quantificam a relevância de cada termo para as consultas (*Wiq*) e para os documentos (W_{id}) no espaço vetorial. 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 documento frequency*) é dado por:

$$idf_i = \log \frac{N}{n_i}$$

Figura 4 – Cálculo da Frequência Inversa

Este valor é usado para calcular o peso, utilizando a seguinte fórmula: $W_{id} = freq(_{ti},d) * idf_i$, 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.

3.3 Modelo probabilístico

O modelo probabilístico descreve documentos considerando pesos binários que representam a presença ou ausência de termos. O vetor resultado gerado pelo modelo tem como base o cálculo da probabilidade de que um documento seja relevante para uma consulta. A principal ferramenta matemática do modelo probabilístico é o teorema de Bayes Van(1979).

O modelo probabilístico é baseado no princípio probabilístico de ordenação *Probability Ranking Principle*, que estabelece que este modelo pode ser usado de forma ótima. Este princípio é baseado na hipótese de que a relevância de um documento para uma determinada consulta é independente de outros documentos. O princípio é o seguinte:

"Se a resposta de um sistema de recuperação de referência a cada requisição, é uma ordem de documentos classificada de forma decrescente pela probabilidade de relevância para o usuário que submeteu a requisição, onde as probabilidades são estimadas com a melhor precisão com base nos dados disponíveis, então a efetividade geral do sistema para o seu usuário, será a melhor que pode ser obtida com base naqueles dados".

O modelo probabilístico considera um processo iterativo de estimativas da probabilidade de relevância.

Devem ser calculados: $P(+R_q|d)$ a probabilidade de que um documento d seja relevante para uma consulta q e $P(-R_q|d)$ a probabilidade de que um documento d não seja relevante para uma consulta q.

O documento d é considerado relevante para a consulta q se $P(+R_q|d) > P(-R_q|d)$, e o vetor resultado é decidido com base num fator $W_{d|q}$, definido por:

$$W_{d|q} = \frac{P(+R_q \mid d)}{P(-R_q \mid d)}$$

Figura 5 – Cálculo do Fator W

Este fator minimiza a média do erro probabilístico. Através do teorema de Bayes e estimativas de relevância baseadas nos termos da consulta, pode-se chegar a seguinte equação:

$$sim(d,q) = W_{d|q} = \sum_{i=1}^{t} x_i \times W_{qi}$$

Figura 6 – Cálculo da Similaridade Probabilística

Onde:

· $x_i \in \{0, 1\}$;

· $W_{qi} = \log r_{qi} (1\text{-}s_{qi}) / s_{qi}(1\text{-}r_{qi})$;

· r_{qi} é a probabilidade de que um termo de indexação *i* ocorra no documento, dado que o documento é relevante para a consulta *q*; e

· s_{qi} é a probabilidade de que um termo de indexação *i* ocorra no documento, dado que o documento não é relevante para a consulta *q*.

O modelo probabilístico tem como vantagem, além do bom desempenho prático, o princípio probabilístico de ordenação, que uma vez garantido, resulta em um comportamento ótimo do método. Entretanto, a desvantagem é que este comportamento depende da precisão das estimativas de probabilidade. Além disso, o método não explora a freqüência do termo no documento e ignora o problema de filtragem de informação.

4. Relevance Feedback ou Retro-alimentação de Relevância

Uma forma de aumentar a eficiência de um sistema de recuperação de informação é reduzir as diferenças lingüísticas, sociais ou culturais existentes entre usuários e indexadores, incorporando as decisões de ambos na forma de representar os documentos. (FERNEDA, 2003)

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.

5. Algoritmos Genéticos

Um **algoritmo genético** (**AG**) é uma técnica de procura utilizada na ciência da computação para achar soluções aproximadas em problemas de otimização e busca. Algoritmos genéticos são uma classe particular de algoritmos evolutivos que usam técnicas inspiradas pela biologia evolutiva como hereditariedade, mutação, seleção natural e recombinação (ou *crossing over*).

Algoritmos genéticos são implementados como uma simulação de computador em que uma população de representações abstratas de solução é selecionada em busca de soluções melhores. A evolução geralmente se inicia a partir de um conjunto de soluções criado aleatoriamente e é realizada através de gerações. A cada geração, a adaptação de cada solução na população é avaliada, alguns indivíduos são selecionados para a próxima geração, e recombinados ou mutados para formar uma nova população. A nova população então é utilizada como entrada para a próxima iteração do algoritmo.

Algoritmos genéticos diferem dos algoritmos tradicionais de otimização em basicamente quatro aspectos:

- Se baseiam em uma codificação do conjunto das soluções possíveis, e não nos parâmetros da otimização em si;
- os resultados são apresentados como uma população de soluções e não como uma solução única;
- não necessitam de nenhum conhecimento derivado do problema, apenas de uma forma de avaliação do resultado;
- usam transições probabilísticas e não regras determinísticas.

6. O projeto GIRS – Genetic Information Retrieval System

O mecanismo de recuperação de informações ***GIRS*** pretende que a relevância dos documentos recuperados evolua a cada iteração através do uso de algoritmos genéticos e do *feedback* de relevância obtido dos usuários, o domínio é o conjunto de documentos legais gerados pelo cotidiano da universidade. O sistema proposto se divide em duas fases.

6.1 Primeira Fase: Preparação do sistema

Nesta fase são cadastradas as *stopwords*[1], que são as palavras a serem extraídas dos documentos antes da indexação. A seguir, ocorre a catalogação dos documentos, armazenando em um arquivo contendo a identificação, o título e o local de armazenamento de cada documento.

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 simples que podem ter certa importância no contexto do documento. Para estes termos é realizado o cálculo do valor da freqüência (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 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). Veja equação 3

$$F = QR/(TP - QS) \tag{3}$$

Ao final, 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. O quadro a seguir sistematiza o processo:

Preparação do sistema

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

6.2 Segunda Fase: Consultas dos usuários

A fase de consultas dos usuários se inicia quando o usuário informa os termos a serem pesquisados e os submete à consulta. Caso os termos solicitados pelo usuário constem no arquivo de termos indexados, o algoritmo genético (AG) gera aleatoriamente uma população inicial de M indivíduos, onde cada indivíduo é composto por N cromossomos com valores iguais a 1 ou 0, sendo N o número total de documentos do acervo, 1 indica a presença do documento referente àquele cromossomo para a resposta solicitada e 0 indica a ausência deste documento.

Cada indivíduo da população inicial é avaliado pela soma individual de cada cromossomo (documento) obtido através das métricas de revocação (*recall*) e precisão (*precision*), valores estes calculados tomando por base os pesos (F) para os termos da consulta, gerados anteriormente e armazenados na base de dados. A partir desta avaliação, o AG efetua a seleção dos indivíduos para o cruzamento e mutação através do método da roleta, onde os indivíduos mais aptos têm uma maior chance de se reproduzirem, gerando uma nova população com M indivíduos.

O sistema seleciona os M melhores indivíduos entre os da população inicial e os novos criados pelos cruzamentos e mutações, formando uma nova população inicial. Essa nova população segue os mesmos passos da população anterior e assim sucessivamente até o critério de parada a ser definido.

Após a evolução dos indivíduos modificados geneticamente, e selecionando o melhor indivíduo, os documentos que apresentam as melhores avaliações individuais serão apresentados ao usuário, sendo estes considerados mais relevantes para a consulta em questão.

Do usuário será obtida uma avaliação dos documentos que foram apresentados como resposta a sua consulta, onde será indicado quais entre os documentos apresentados são relevantes e quais não são relevantes para sua consulta, tal procedimento é conhecido como *relevance feedback.* A forma de obtenção deste *feedback*, é explícita (indicada pelo usuário) e implícita (através da verificação de acesso do usuário aos documentos sugeridos pelo *GIRS*).

Através deste *feedback*, caso o documento receba uma avaliação como relevante para os termos da consulta, os pesos armazenados para estes termos serão atualizados somando-se 0,1 ao valor existente, já no caso da avaliação ser como não relevante, os pesos sofrem um decréscimo de 0,1.Com a tabela devidamente atualizada, quando ocorrer uma próxima consulta, o sistema toma como base estes novos valores para os cálculos de avaliação do AG.

Com esta evolução, a expectativa é que o sistema com o passar do tempo, obtenha uma recuperação da informação com maior quantidade de documentos relevantes e menor quantidade de documentos não relevantes, baseando-se no *feedback* dos usuários, que no caso em questão, constituem um grupo bastante homogêneo. O quadro a seguir sistematiza o processo:

Consultas dos Usuários

- Usuário digita termos a serem consultados;
- Usuário submete sua consulta ao sistema;
- Caso os termos constem do arquivo de termos indexados:
    - Algoritmo Genético (AG):
        - Gera aleatoriamente a população inicial de M indivíduos (soluções) com N cromossomos (documentos) com valores 0 ou 1;
            - N = número total de documentos catalogados;
            - 0 = indica a ausência do documento na solução gerada;
            - 1 = indica a presença do documento na solução gerada;

- Repete os passos a seguir até o critério de parada
  - Avaliação dos indivíduos:
    - Calcula para cada cromossomo (documento) com valor igual a 1 os valores de *precision* e *recall;*
    - Soma para cada indivíduo as avaliações de seus cromossomos;
  - Seleciona os indivíduos através do método da roleta:
    - Efetua o cruzamento dos indivíduos selecionados;
    - Efetua a mutação dos indivíduos selecionados;
  - Efetua avaliação dos novos indivíduos:
    - Calcula para cada cromossomo (documento) com valor igual a 1 os valores de *precision* e *recall;*
    - Soma para cada indivíduo as avaliações de seus cromossomos;
  - Seleciona os M melhores indivíduos entre os da população inicial e os novos criados pelos cruzamentos e mutações, formando uma nova população inicial.
  - Seleciona o melhor de todos os indivíduos;
- Ordena os documentos do indivíduo selecionado, pela ordem decrescente das avaliações individuais;
- Apresenta ao usuário os documentos ordenados, sendo estes considerados mais relevantes para a consulta em questão.

o Obtenção do *feedback* do usuário, implícita e explicitamente

- Se documento relevante
  - Atualiza os pesos armazenados para os termos da consulta somando-se 0,1 ao valor existente;
- Se documento não relevante
  - Atualiza os pesos armazenados para os termos da consulta subtraindo-se 0,1 ao valor existente;

– Caso os termos não constem do arquivo de termos indexados:

o O sistema apresenta uma resposta negativa ao usuário.

Conclusões

Neste artigo, foi apresentada uma visão geral de modelagem em sistemas de recuperação de informação, onde foram descritos os três modelos clássicos. A descrição dos modelos

clássicos embasou e justificou a apresentação de um modelo próprio, criado com a finalidade de aprimorar as pesquisas em um ambiente controlado.

O GIRS – Genetic Information Retrieval System é um ambiente de recuperação de informação em que a relevância dos documentos recuperados evolui a cada iteração através do uso de algoritmos genéticos e do *feedback* de relevância dado pelos usuários.

O domínio escolhido para avaliação é uma coleção de documentos provenientes do cotidiano da universidade com cerca de dez mil documentos entre atas, resoluções e normas. Nesta coleção há um conjunto de consultas e para cada consulta foi fornecido um conjunto ideal de documentos resposta, criado por especialistas nos temas envolvidos. Com isso, será possível avaliar com precisão a acurácia do nosso sistema.

No momento da redação deste artigo, os resultados da avaliação do sistema estavam sendo tabulados o que impediu a apresentação dos resultados finais, entretanto os resultados parciais demonstraram uma evolução na relevância dos documentos apresentados como resultado de uma consulta. Ressaltando que o conjunto de documentos relevantes, dado uma determinada consulta já estavam previamente definidos por uma equipe de especialistas e portanto a validação do sistema ocorre pela comparação simples entre documentos esperados versus documentos recuperados pelo nosso sistema.

Os resultados parciais apresentaram um percentual crescente de documentos relevantes recuperados pelo *GIRS* a cada interação demonstrando a atuação positiva do algoritmo genético implementado.

Após a tabulação total dos resultados da validação, o projeto será aplicado a uma biblioteca digital de uma indústria fumageira. Este biblioteca digital contém milhares de descrições de processos industriais aplicados a produção de tabaco.

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 15 out 2007.

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

CRESTANI, FABIO; VAN RIJSBERGEN, **CORNELIS J.** *A model for adaptative information retrieval*. Journal of Intelligent Information Systems, v.8, 1997.

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, E. *Recuperação da informação; Análise sobre a constribuição 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 2007.

FRAKES, W. B. & BAEZA-YATES, R. *Information Retrieval Data Structures & Algorithms*, Prentice Hall, 1992.

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

CARDOSO, O. N. P.. *Recuperação de Informação*. Infocomp Revista de Computação da Ufla, Lavras - MG, v. 1, p. 33-38, 2000.

SALTON, G. *Automatic Text Processing: TheTransformation, Analisys, and Retrieval of Information by Computer*, Addison Wesley, 1989.

SARACEVIC, T. *Evaluation of evaluation in information retrieval*. Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Special issue of SIGIR Forum, 138-146, 1995.

VAN RIJSBERGEN, C. J. *Information Retrieval*, Butterworths, 2ª edition, 1979.