

## Aplicando Algoritmos Genéticos na Recuperação de Informação

*Applying Genetic Algorithms in Information Retrieval*

por [Edberto Fernalda](#)

**Resumo:** O desenvolvimento dos Algoritmos Genéticos é baseado na teoria evolucionista de Darwin e nas descobertas sobre a reprodução humana e a genética. Este artigo apresenta uma forma de aplicação dos algoritmos genéticos em sistemas recuperação de informação na qual as possíveis representações de um mesmo documento são consideradas um tipo de “código genético” deste documento. As buscas realizadas pelos usuários são consideradas o “meio ambiente” no qual os documentos estão inseridos. Nesse ambiente as diversas representações de um mesmo documento competem entre si na busca de uma descrição mais adequada para o documento. Diversos experimentos têm apresentado resultados prometedores na aplicação de algoritmos genéticos na recuperação de informação na Web.

**Palavras-chave:** Recuperação da informação; Algoritmos genéticos; Sistema evolutivo.

**Abstract:** The development of Genetic Algorithms is based on Darwin's evolutionary theory, human reproduction and genetics concepts. This article presents a way of applying genetic algorithms in information retrieval systems where possible representations of a document can be viewed as a kind of "genetic code". Searches conducted by users are seen as the "environment" where the documents are inserted. In this environment, the representations of a document compete with each other in finding a more appropriate description for the document. Experiments have shown promising results in the application of genetic algorithms in Web information retrieval.

**Keywords:** Information retrieval, Genetic algorithm, Evolutionary system

### Introdução

Desde o surgimento dos primeiros computadores eletrônicos diversos métodos computacionais foram propostos na tentativa de gerenciar o constante e acelerado aumento das informações científicas, principalmente a partir da Segunda Guerra Mundial. A maioria desses métodos se caracteriza pela utilização de modelos matemáticos, criados a partir de hipóteses restritas e aplicados de forma absoluta nos sistemas de recuperação de informação.

Em tais sistemas cada documento é geralmente representado por um conjunto de termos de indexação ou palavras-chave, associados ou não a um número que indica o grau de relevância do termo para a descrição do conteúdo do documento. [Fernalda](#) (2003, cap.4) apresenta detalhadamente diversos modelos matemáticos (“*quantitativos*”) e argumenta que o processo de representação e recuperação de informação é inerentemente impreciso, sendo sua modelagem matemática possível apenas por meio de simplificações teóricas e da adequação de conceitos subjetivos tais como o próprio conceito de “*informação*” e o conceito de “*relevância*”. Estas simplificações geram limitações qualitativas que podem ser notadas nas atuais ferramentas de busca na Web.

Tendo em vista as limitações e o esgotamento dos modelos matemáticos, surgem novas abordagens para o problema do tratamento e recuperação da informação. Essas abordagens buscam reduzir o caráter impositivo e absoluto dos modelos matemáticos, atribuindo ao processo de recuperação de informação características evolutivas, tal como percebidas nos sistemas naturais.

Diversas obras apresentam os sistemas evolutivos como alternativas aos tradicionais sistemas computacionais. Segundo [Johnson](#) (2003), esse tipo de sistema apresenta um comportamento “*emergente*” (“*botton-up*”), permitindo solucionar problemas relativamente complexos com o auxílio de regras simples, inspiradas em mecanismos da natureza. [Bentlet](#) (2002) apresenta diversos modelos computacionais inspirados em processos biológicos, tais como as Redes Neurais e os Algoritmos Genéticos. [Cordón](#) (2003) analisam a aplicação da Computação Evolutiva no desenvolvimento de sistemas de recuperação de informação.

Através de um exemplo simplificado, este artigo apresenta e avalia a utilização dos algoritmos genéticos na representação de documentos de um sistema de informação. A aplicação dos conceitos de [Algoritmos Genéticos](#) permite o desenvolvimento de sistemas evolutivos, nos quais os usuários, através de suas buscas, são elementos efetivamente participantes do processo de representação dos documentos do corpus do

sistema.

### Algoritmos Genéticos

Sabe-se hoje que todos os organismos vivos são constituídos de células, cada qual com o seu conjunto de cromossomos. Os cromossomos são cadeias de DNA (*ácido desoxirribonucléico*) que carregam as informações necessárias para a geração de outros seres vivos da mesma espécie.

Um cromossomo é formado por genes que são responsáveis pelas características individuais dos organismos, tais como altura, cor dos olhos, a cor dos cabelos, etc. Na reprodução cada um dos pais passa metade de seus cromossomos aos filhos. Durante esse processo os cromossomos podem sofrer recombinações, em um processo denominado *crossover*, tendo como consequência uma diversificação nas características do indivíduo.

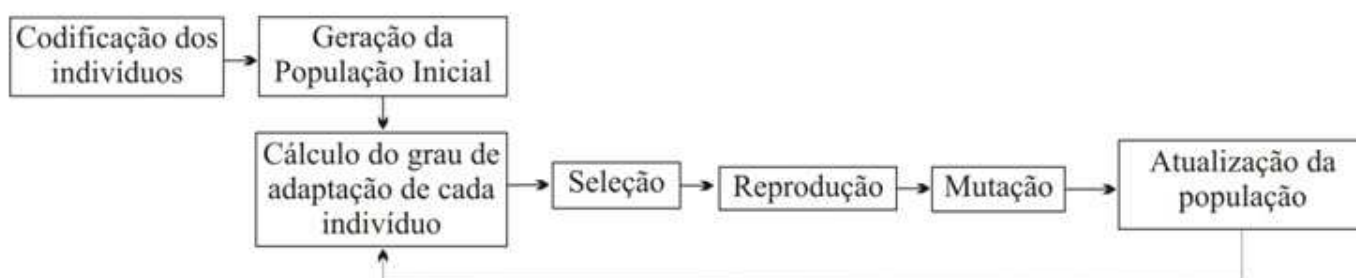
Sobre a inerente casualidade do processo de reprodução age a seleção natural, que seleciona os indivíduos cujas características os fazem mais adaptados ao meio ambiente onde vivem. Esses indivíduos possuem maiores chances de sobreviver e de se reproduzir, transmitindo assim seu material genético às gerações futuras.

A primeira tentativa de representar matematicamente a teoria da evolução das espécies foi apresentada no livro The Genetical Theory of Natural Selection de Fisher (1930). Segundo o autor, a aprendizagem e a evolução são formas de adaptação que se diferem apenas na escala de tempo. A evolução, em vez de ser o processo de uma vida, é o processo de várias gerações.

Em meados da década de 60, John Holland (1975), juntamente com seus alunos da Universidade de Michigan, desenvolveu diversas pesquisas com o objetivo de estudar o fenômeno da adaptação como ocorre na natureza e desenvolver modelos que pudessem ser utilizados em sistemas computacionais.

Os Algoritmos Genéticos (Mitchell, 2002) são técnicas que simulam o processo de evolução natural em uma população de possíveis soluções para um determinado problema. A cada iteração do algoritmo (“*geração*”), um novo conjunto de estruturas é criado através da troca de informações entre estruturas selecionadas da geração anterior. O resultado tende a ser um aumento da adaptação dos indivíduos ao meio ambiente, podendo acarretar também um aumento da aptidão de toda a população a cada nova geração, aproximando-se de uma solução ótima para o problema em questão. A estrutura funcional de um algoritmo genético está representada na Figura 1.

Figura 1 Seqüência de execução de um algoritmo genético



Um algoritmo genético é estruturado de forma que as informações referentes a um determinado sistema possam ser codificadas de maneira análoga aos cromossomos biológicos. Busca-se implementar um processo repetitivo que se assemelhe ao processo de evolução natural.

A partir dos anos 80 os Algoritmos Genéticos receberam um grande impulso em diversas áreas científicas devido principalmente à versatilidade e aos excelentes resultados apresentados. A popularização dos computadores e o aparecimento de sistemas cada vez mais rápidos e potentes também ajudaram muito o seu desenvolvimento.

### Algoritmos Genéticos na Recuperação de Informação

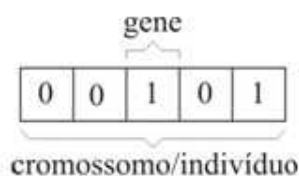
A aplicação dos algoritmos genéticos em sistemas de informação representa uma nova forma de pensar o processo de recuperação de informação na qual as representações dos documentos são alteradas de acordo com a necessidade de informação da comunidade de usuários, manifestada através de suas buscas.

[Gordon](#) (1988) e [Blair](#) (1990, p.254) apresentam um modelo no qual cada documento é representado por um conjunto de “*cromossomos*” binários. Segundo Blair, a inerente indeterminação da representação de um documento pode ser interpretada como um tipo de variabilidade genética que permite aos documentos se adaptarem progressivamente ao meio ambiente. Entendendo-se por “meio ambiente” o conjunto das buscas realizadas pelos usuários do sistema de informação.

### Codificação dos Indivíduos

O ponto de partida para a utilização de um algoritmo genético consiste em definir uma representação adequada dos indivíduos (*soluções*) envolvidos no problema de maneira que o algoritmo possa operá-los. No algoritmo proposto por [Holland](#) (1975), cada cromossomo é representado por uma cadeia binária de tamanho fixo, onde cada gene pode assumir o valor “0” (*zero*) ou o valor “1” (*um*), como exemplificado na Figura 2.

Figura 2 Representação de um cromossomo de genes binários



Em um sistema de recuperação de informação os documentos são geralmente representados por um conjunto de termos de indexação ou palavras-chave. A representação de um documento pode ser visto como o seu “código genético” 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. A Figura 3 apresenta um exemplo de um documento representado por apenas três de cinco possíveis termos: “*algoritmos genéticos*”, “*Recuperação de informação*” e “*Web*”.

Figura 3 Representação de um documento através de um cromossomo binário



### População Inicial

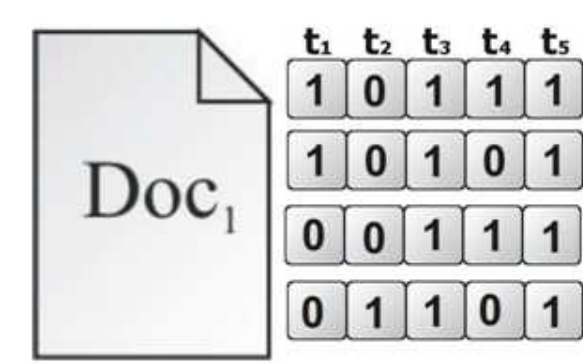
A tarefa de associar termos de indexação aos documentos de um sistema de informação pode ser efetuada por profissionais da informação, como bibliotecários, ou por especialistas da área de conhecimento do corpus documental do sistema. Porém, mesmo utilizando uma equipe de profissionais qualificados e uma política de indexação consistente, a subjetividade desse processo pode levar a situações em que um mesmo documento pode ser representado de diferentes formas.

Um algoritmo genético utiliza essa variação de forma pró-ativa, adequando progressivamente as representações de um mesmo documento às necessidades dos usuários do sistema, observadas através de suas buscas. As diversas representações de um mesmo documento, a população inicial, pode também ser conseguida através da geração automática de indivíduos, obedecendo a certas condições pré-estabelecidas.

O número de indivíduos da população inicial pode afetar o desempenho e a eficiência dos algoritmos genéticos. Populações muito pequenas podem perder a diversidade necessária para convergir para uma boa solução do problema que se deseja resolver. Por outro lado, se a população tiver um número excessivo de indivíduos o algoritmo poderá perder grande parte de sua eficiência. [Blair](#) (1990, p.256) sugere a utilização de

aproximadamente dez indivíduos na representação de um mesmo documento. O exemplo da Figura 4 apresenta o documento *Doc1* representado por quatro indivíduos (*cromossomos*) contendo cinco termos de indexação (*genes*): *t1*, *t2*, *t3*, *t4* e *t5*.

Figura 4. “Código genético” de um documento (Doc1).



A população inicial deve ser composta por um conjunto de indivíduos razoavelmente plausível para a solução do problema em questão. Através da competição esses indivíduos serão continuamente modificados, tornando-se progressivamente mais efetivos na identificação de documentos relevantes.

### Cálculo do grau de adaptação (*fitness*)

Para a população inicial e a cada nova geração é calculado o grau de adaptação (*fitness*) de cada indivíduo. Esse cálculo é feito através de uma função de adaptação (*função de fitness*) que deve ser definida tendo em vista o tipo de problema a ser resolvido. A função de *fitness* deve refletir a qualidade de cada indivíduo em solucionar o problema. Na literatura sobre o tema podem ser encontradas diversas propostas para a esta função (Radwan, 2006). Uma função de *fitness* bastante utilizada é o *Coeficiente de Similaridade de Jaccard* (van Rijsbergen, 1979). Esta função calcula o valor da similaridade entre duas seqüências binárias e é definida como o número de posições com valor “1” em ambas as seqüências, dividido pelo número de posições com valor “1” em pelo menos uma das seqüências.

$$\text{fitness} = \frac{\text{Quantidade de posições com "1" em ambas as seqüências}}{\text{Quantidade de posições com "1" em pelo menos uma das seqüências}}$$

Da mesma forma dos documentos, as buscas dos usuários podem ser também representadas através de uma seqüência binária, como por exemplo, 01110. Esta seqüência representa uma expressão de busca contendo os termos *t2*, *t3* e *t4*. Supondo que, após a execução da busca, o documento *Doc1* tenha sido considerado relevante pelo usuário, este documento apresentará os valores de *fitness* descritos na Tabela 1.

Tabela 1. Cálculo do grau de adaptação (*fitness*) após uma busca.

		t <sub>1</sub>	t <sub>2</sub>	t <sub>3</sub>	t <sub>4</sub>	t <sub>5</sub>	
Expressão de busca		0	1	1	1	0	
<div>Doc<sub>1</sub></div>	1	1	0	1	1	1	<i>fitness</i> 0.4
	2	1	0	1	0	1	0.2
	3	0	0	1	1	1	0.5
	4	0	1	1	0	1	0.5
<i>fitness</i> do documento							0.4

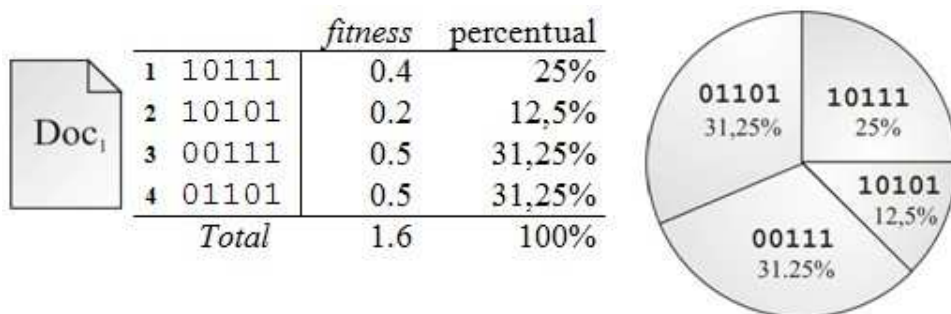
Estes cálculos são feitos para todos os documentos recuperados e considerados relevantes pelo usuário. O *fitness* dos documentos, calculado pela média aritmética do *fitness* de cada indivíduo, pode ser utilizado no ordenamento dos documentos resultantes das buscas futuras.

## Reprodução

De acordo com a teoria de Darwin, os indivíduos mais adaptados (com maior *fitness*) ao meio ambiente têm maior chance de se reproduzirem. Para simular a seleção natural, um algoritmo genético pode utilizar alguns métodos para selecionar aleatoriamente os indivíduos que deverão se reproduzir. Um dos métodos mais utilizados é chamado de “*Roleta*” (Roulette Wheel).

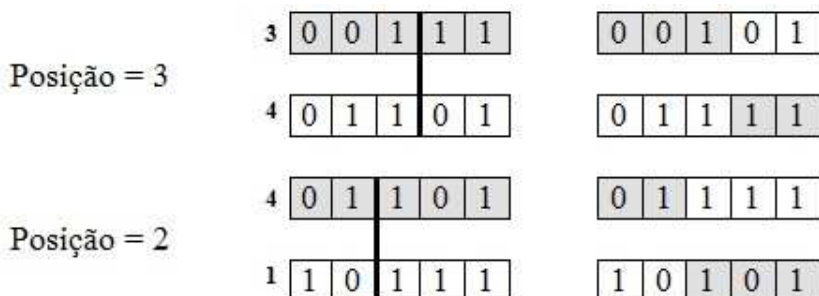
No método da Roleta o *fitness* de cada indivíduo é utilizado para construir uma “roleta” que fornecerá a base para o processo de seleção. Para cada indivíduo é calculado o percentual do *fitness* em relação à soma geral do *fitness* de todos os indivíduos. Dessa forma cada indivíduo terá chance de reprodução proporcional ao seu *fitness*, como exemplificado na Tabela 2.

Tabela 2. Formação da “roleta” com as probabilidades de reprodução.



A roleta é “*girada*” quatro vezes a fim de selecionar dois casais de indivíduos para reprodução. Para cada casal utiliza-se uma posição aleatória para a troca de “material genético”. Supondo-se que para o documento *Doc1* foram escolhidos os casais de cromossomos 3-4 e 4-1, e as posições 3 e 2, respectivamente, a reprodução será executada conforme o exemplo da Figura 5.

Figura 5. Representação do processo de *crossover*.



Após a reprodução, o documento *Doc1* será representado por quatro novos cromossomos, conforme a Figura 6.

Figura 6. Representação do documento *Doc1* após a reprodução.

	$t_1$	$t_2$	$t_3$	$t_4$	$t_5$	<i>fitness</i>
1	0	0	1	0	1	0.4
2	0	1	1	1	1	0.2
3	0	1	1	1	1	0.5
4	1	0	1	0	1	0.5

## Mutação

A capacidade dos algoritmos genéticos provém da diversidade dos indivíduos. As mutações ajudam a prevenir



a estagnação das populações, ajudando a preservar esta diversidade através das gerações. Para uma adequada simulação do processo natural, a simulação computacional da mutação inclui um parâmetro ao sistema: a “*probabilidade de mutação*”. Este parâmetro regula a frequência que as mutações serão efetuadas.

Após a reprodução, e observada a frequência de mutação, será selecionado aleatoriamente um ou mais indivíduos que deverão sofrer uma modificação compulsória em seus cromossomos. Para cada indivíduo será escolhida aleatoriamente a posição (*gene*) onde esta mutação será efetuada. Utilizando ainda o documento Doc1 como exemplo, e supondo terem sido escolhidos os cromossomos 4 e 1 e os respectivos genes 3 e 4, a mutação será processada da forma como representada na Figura 7.

Figura 7. Representação do processo de mutação.

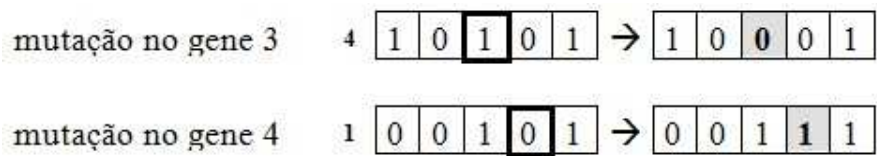
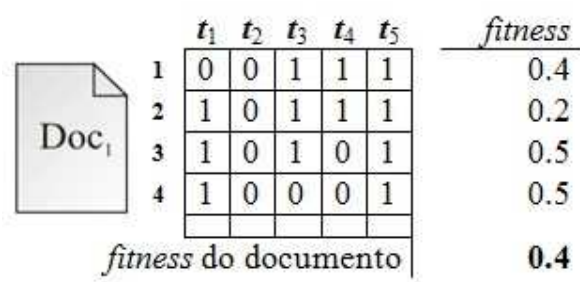
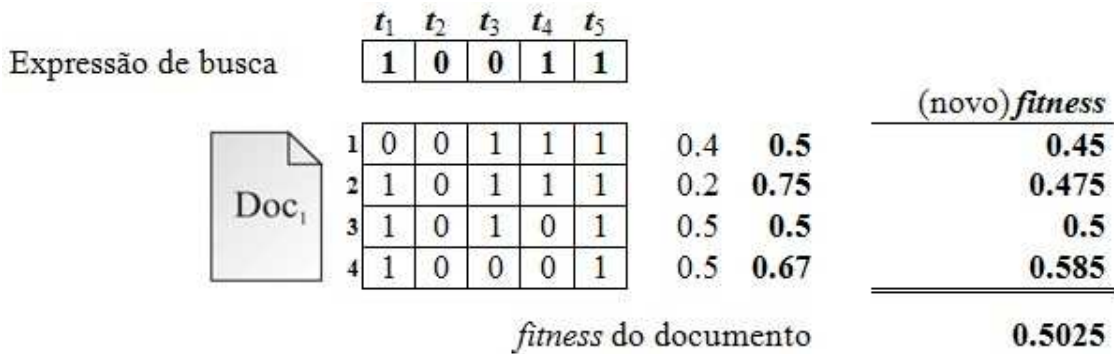


Figura 8. Representação do documento Doc1 após o processo de mutação.




Fecha-se assim um ciclo da evolução dos documentos, exemplificado através do documento Doc1. Assim como o Doc1, todos os documentos do corpus do sistema terão o seu “código genético” modificado em função da expressão de busca do usuário. Posteriormente, em uma nova busca, expressa pela sequência 10011 , por exemplo, o documento Doc1 terá os valores de *fitness* apresentados na Tabela 3.

Tabela 3. *Fitness* do documento Doc1 após uma segunda busca.



O novo valor do *fitness* de cada cromossomo é calculado através da média aritmética entre o *fitness* da busca anterior e o *fitness* da busca atual. A nova configuração do documento Doc1 é apresentada na Figura 9.

Figura 9. Representação do documento Doc1 após uma segunda busca.

		$t_1$	$t_2$	$t_3$	$t_4$	$t_5$	<i>fitness</i>
	1	0	0	1	1	1	0.45
	2	1	0	1	1	1	0.475
	3	1	0	1	0	1	0.5
	4	1	0	0	0	1	0.585
	<i>fitness do documento</i>						<u>0,5025</u>

Verifica-se um aumento do valor do *fitness* do documento *Doc1*, configurando-se, portanto, uma evolução do documento ao seu "*meio ambiente*", isto é, aos interesses dos usuários do sistema. Para efeito didático, o exemplo utilizado para ilustrar o funcionamento dos algoritmos genéticos foi bastante simplificado. Os algoritmos genéticos possuem diversos parâmetros e funções que podem variar dependendo do tipo de aplicação a que se destinam. Essa variabilidade faz com que os algoritmos genéticos se configurem como um campo experimental regido apenas por uma idéia genérica, permitindo uma grande diversidade nas formas de implementação.

[Gordon](#) (1988) implementou um pequeno sistema contendo 18 documentos, cada um contendo 17 diferentes descrições fornecidas por usuários. O sistema de Gordon obteve resultados expressivos. Após 40 gerações, as descrições dos documentos estavam cerca de 19% mais aptas para identificar documentos relevantes.

[Vrajitoru](#) (1998) apresenta uma proposta de algoritmo genético na qual cada documento é representado por um único cromossomo não-binário e define um operador de *crossover* específico, denominado "dissociated crossover", que obteve melhor desempenho do que outras formas de reprodução. Apesar de utilizar funções e cálculos relativamente simples, os algoritmos genéticos exigem um processamento exaustivo e sua aplicação em grandes bases documentais torna-se dependente do modelo de algoritmo e dos recursos computacionais utilizados.

### Conclusão

A utilização dos algoritmos genéticos na recuperação de informação apresenta-se como uma possibilidade, uma proposição para futuras implementações de sistemas com características evolutivas. Sua aplicação rompe com a rigidez dos modelos puramente matemáticos, reconhecendo a inerente indeterminação do processo de representação do conteúdo dos documentos.

Os trabalhos práticos disponíveis na literatura apresentam apenas testes utilizando pequenos protótipos de sistemas, não determinando sua aplicabilidade em sistemas reais ([Gordon](#), 1988; [Vrajitoru](#), 2000). Apesar da característica evolutiva representar uma forma inovadora de abordar o problema da recuperação de informação, introduz diversos questionamentos relacionados aos efeitos de sua inerente imprevisibilidade, quando utilizado em situações reais.

No atual contexto da Web, cuja dinamicidade muitas vezes não permite uma indexação adequada dos documentos a serem disponibilizados, os algoritmos genéticos poderiam representar uma alternativa, ao permitir que as representações dos documentos se configurem adequadamente ao longo de um período, de acordo com a recuperação desses documentos por grupos de usuários com interesses comuns. Diversos projetos ([Martín-Bautista](#), 1999; [Chen](#), 2001; [Sobrinho; Girardi](#), 2003) buscam incorporar as idéias evolutivas dos algoritmos genéticos ao contexto heterogêneo, complexo e dinâmico da Web.

### Referências Bibliográficas

- BENTLET, Peter J. *Biologia Digital: como a natureza está transformando nossa tecnologia e nossas vidas*. São Paulo: Berkeley Brasil, 2002. 320p.
- BLAIR, David C. *Language and representation in information retrieval*. Amsterdam: Elsevier, 1990.

CHEN, Yi-Shin; SHAHABI, Cyrus. Automatically Improving the Accuracy of User Profiles with Genetic Algorithm. IASTED International Conference on Artificial Intelligence and Soft Computing. Cancun, Mexico, Mai.2001

CORDÓN, Oscar; HERRERA-VIEDMA, Enrique.; LÓPEZ-PUJALTE, Cristina.; LUQUE, María; ZARCO, Carmen. A Review on the Application of Evolutionary Computation to Information Retrieval. International Journal of Approximate Reasoning, v.34, n.2-3, p.241-264, nov.2003.

FERNEDA, Edberto. Recuperação de Informação: estudo sobre a contribuição da Ciência da Computação para a Ciência da Informação. São Paulo, 2003. 147p. Tese (doutorado em Ciência da Informação). Escola de Comunicação e Artes, Universidade de São Paulo.

FISHER, Ronald. A. The Genetical Theory of Natural Selection. Oxford: Clarendon Press, 1930.

GORDON, Michael. Probabilistic and genetic algorithms for document retrieval. Communications of the ACM, v.31, n.10, p.1208-1218, out.1988.

HOLLAND, John H. Adaptation in Natural and Artificial Systems. University of Michigan Press: Ann Arbor, 1975.

JOHNSON, Steven. Emergência: a dinâmica de rede em formigas, cérebros, cidades e softwares. Rio de Janeiro: Jorge Zahar, 2003. 231p.

MARTÍN-BAUTISTA, Maria J.; MIRANDA, María-Amparo V.; LARSEN, Henrik L. A Fuzzy Genetic Algorithm Approach to an Adaptive Information Retrieval Agent. Journal of the American Society for Information Science Journal (JASIS) v.50, n.9, p.760-771. jul.1999.

MITCHELL, Melaine. An introduction to genetic algorithms. Cambridge: MIT Press, 2002. 209p.

RADWAN, Ahmed A. A.; LATEF, Bahgat A. Abdel; ALI, Abdel Mgeid A.; SADEK, Osman A. Using Genetic Algorithm to Improve Information Retrieval Systems. Proceedings of World Academy of Science, Engineering and Technology, v.17, dezembro 2006.

SOBRINHO, Antonio Carlos ; GIRARDI, Rosario . Uma Análise das Aplicações dos Algoritmos Genéticos em Sistemas de Acesso à Informação Personalizada. REIC: Revista Eletrônica de Iniciação Científica, v. III, n. IV, p. 1, 2003. Disponível em <<http://www.sbc.org.br/reic/edicoes/2003e4/tutoriais/AlgoritmosGeneticosEmSistemasDeAcessoAInformacaoPersonalizada.pdf>> Acessado em 4/10/2008.

VAN RIJSBERGEN, Cornelis J. Information retrieval. London: Butterworths, 1979. 152p.

VRAJITORU, Dana. Crossover improvement for the Genetic Algorithm in Information Retrieval. Information Processing and Management, v.34, n.4. p.405-415, 1998.

VRAJITORU, Dana. Large Population or Many Generations for Genetic Algorithms? Implications in Information Retrieval. In: Crestani, F., Pasi, G. (eds.): Soft Computing in Information Retrieval: Techniques and Applications. Heidelberg: Physica-Verlag, p.199-222. 2000.



**Sobre o autor / About the Author:**

Edberto Ferneda

[ferneda@ffclrp.usp.br](mailto:ferneda@ffclrp.usp.br)

Doutor em Ciência da Informação pelo ECA/USP; Mestre em Informática pela Universidade Federal da Paraíba;  
Professor do Curso de Ciências da Informação e Documentação da Universidade de São Paulo - USP - Ribeirão Preto.