Universidade Federal do Amazonas
Instituto de Ciências Exatas
Departamento de Ciência da Computação
Programa de Pós-Graduação em Informática

Identificando o Tópico de Páginas *Web*

Márcia Sampaio Lima

Manaus - Amazonas

Abril de 2009

Márcia Sampaio Lima

Identificando o Tópico de Páginas *Web*

Dissertação apresentada ao Programa de Pós-Graduação em Informática do Departamento de Ciência da Computação da Universidade Federal do Amazonas, como requisito parcial para obtenção do Título de Mestre em Informática. Área de concentração: Recuperação de Informação.

Orientador: Prof. Dr. João Marcos Bastos Cavalcanti, Ph.D.

Márcia Sampaio Lima

Identificando o Tópico de Páginas *Web*

Dissertação apresentada ao Programa de Pós-Graduação em Informática do Departamento de Ciência da Computação da Universidade Federal do Amazonas, como requisito parcial para obtenção do Título de Mestre em Informática. Área de concentração: Recuperação de Informação.

Banca Examinadora

Prof. Dr. João Marcos Bastos Cavalvanti, Ph.D. - Orientador.
Departamento de Ciência da Computação – UFAM/PPGI

Prof. Dr. Altigran Soares da Silva.
Departamento de Ciência da Computação – UFAM/PPGI

Prof. Dr. Edleno Silva de Moura.
Departamento de Ciência da Computação – UFAM/PPGI

Prof. Dr. José Palazzo Moreira de Oliveira.
Instituto de Informática da Universidade Federal do Rio Grande do Sul - UFRGS

Manaus - Amazonas

Abril de 2009

Agradecimentos

À minha amada filha, Annabel Lima, que me ensinou a ser persistente e forte logo em seus primeiros dias de vida.

Ao meu marido, Leonardo Lima, que me ensinou as primeiras linhas de código, em 1995, e agora me apóia na defesa desta dissertação.

À minha família, pelo apoio, carinho e compreensão.

A Deus, por permitir que eu realize mais esse sonho.

Ao meu orientador, prof. João Marcos Cavalcanti, pelo incentivo, apoio e dedicação.

Aos professores da UFAM, que me ensinaram, apoiaram e orientaram nesta etapa.

A todos que contribuíram de alguma forma para a realização deste trabalho.

Resumo

Evidências textuais e estruturais que podem ser extraídas dos documentos *web* são frequentemente usadas na busca pela melhoria da qualidade dos resultados obtidos pelos diversos sistemas de recuperação de informação (RI). O tópico de uma página *web* é uma evidência textual que possui uma vasta aplicabilidade nesses sistemas, podendo servir como uma nova fonte de evidência para melhorar *ranking* de páginas *web*, melhorar sistemas de classificação e filtragem destas páginas, entre outros.

O presente trabalho tem por objetivo estudar, desenvolver e avaliar um método para identificar automaticamente o tópico de páginas *web* através da combinação de diferentes fontes de evidências. Definimos o tópico de uma página como sendo um conjunto de, no máximo, cinco termos distintos relacionadas ao assunto principal da página. Em linhas gerais, o método de identificação de tópicos proposto nesta dissertação, está dividido em quatro fases distintas: (1) identificação dos possíveis termos descritores de uma página *web*, fazendo uso de múltiplas fontes de evidências; (2) utilização de um algoritmo genético na combinação das fontes de evidências usadas; (3) definição dos três melhores termos descritores da página; e (4) utilização da estrutura hierárquica de um diretório abrangente e popular da w*eb* com o objetivo de identificar o tópico da referida página.

Os resultados obtidos nos experimentos realizados para avaliar o método proposto foram os seguintes: (1) alto grau de importância do uso da concatenação do texto de âncora de *links* na descoberta dos termos descritores de uma página *web*; (2) boa avaliação da eficiência do método proposto na identificação de tópicos de páginas *web*: ***0.9129***, em uma escala de zero a um; e (3) boa avaliação da utilização de parte do método proposto na classificação automática de páginas *web* na estrutura hierárquica do diretório *Google*, atingindo ***88%±0.11*** de acertos das páginas classificadas.

Os experimentos realizados demonstram que o modelo proposto é útil na identificação do tópico de uma página *web* e também na classificação de páginas na estrutura hierárquica do diretório *Google*.

Palavras-chaves: Tópico de Páginas *Web*, Algoritmos Genéticos, Múltiplas Fontes de Evidências, Diretórios *Web*.

Abstract

Textual and structural sources of evidences extracted from web pages are frequently used to improve the results of Information Retrieval (IR) systems. The main topic of a web page is a textual source of evidence that has a wide applicability in IR systems. It can be used as a new source of evidence to improve ranking results, page classification, filtering, among other applications.

In this work, we propose to study, develop and evaluate a method to identify the main topic of a web page using a combination of different sources of evidences. We define the main topic of a web page as a set of, at most, five distinct keywords related to the main subject of the page. In general, the proposed method, is divided in four distinct phases: (1) identification of the keywords that describe the web page content, using multiple sources of evidences; (2) use of a genetic algorithm to combine the sources of evidences; (3) definition of the three better keywords of the page; and (4) use of a web directory to identify the page main topic.

The results of the experiments show that: (1) the best source of evidence used to describe the keywords of a web page is the content link; (2) the proposed method is efficient to identify the main topic of a web page: 0.9129, in a scale of zero to one; and (3) the proposed method is also efficient to automatic classify web pages within the Google directory, reaching 88%±0.11 of precision in the classification task.

Keywords: Topic of Web Page, Genetic Algorithm, Multiple Sources of Evidences, Web Directories.

Sumário

Introdução **12**

Trabalhos Relacionados **16**

Conceitos Básicos **20**

3.1 Definições Iniciais 20
Termos Descritores de uma Página 20
Tópico de uma Página 20
Texto de Âncora de *Link* 21
***Term Frequency* × *Inverse Document Frequency* (*TF*×*IDF*)** 22
Fontes de Evidências 22
Diretório *Web* 23
3.2 *N-grams* 23
3.3 Algoritmos Genéticos 24
3.3.1 Codificação dos Cromossomos 25
3.3.2 Operadores Genéticos 25
3.3.3 Função Objetivo 28
3.3.4 Validação Cruzada 32

Identificação dos Tópicos de Páginas *Web* **33**

4.1 O Método Proposto 33
4.1.1 Identificação dos Termos Candidatos a Descritores da Página 35
Fase 1: Geração de Termos Compostos 37
Fase 2: Eliminação de *Stopwords* 38
Fase 3: Eliminação de Termos não Alfabéticos 38
Fase 4: Cálculo do *TF*×*IDF* relativo 38
Fase 5: Execução dos Pontos de Corte 38
4.1.2 Utilização de AG na identificação dos Pesos das Evidências 39
Fase 1: Planejamento dos Cromossomos 40
Fase 2: Definição da Função Objetivo 41
Fase 3: Definição dos Parâmetros Genéticos 43
Fase 4: Processo de Evolução 43
4.1.3 Identificação dos Descritores de uma Página *Web* 44
4.1.4 Utilização de um diretório *Web* 45

Experimentos e Discussão dos Resultados **50**

5.1 Experimento 1 – Obtenção dos Pesos das Evidências 50
5.2 Experimento 2 - Avaliação do Método 54
5.3 Experimento 3 - Utilização do método para classificação de páginas no diretório *Google*. 65

Conclusões e Trabalhos Futuros **69**

6.1 Conclusão 69
6.2 Trabalhos Futuros 70

Referências Bibliográficas 72

Lista de Figuras

Figura 1 - Ilustração do objetivo do método proposto nesta dissertação. 14

Figura 2 – Exemplos de termos descritores da página http://www.ufam.edu.br..... 20

Figura 3 – Tópico da página http://www.ufam.edu.br 21

Figura 4 - Lista do texto de âncora de 30 apontadores que referenciam a página http://www.ufam.edu.br 21

Figura 5 - Exemplo de codificação de uma cromossomo com 5 genes e alfabeto binário..... 25

Figura 6 - Descrição da operação de cruzamento..... 26

Figura 7 - Ocorrência de mutação no gene 2 do cromossomo codificado no alfabeto binário. .. 27

Figura 8 - Etapas do método de identificação do tópico de uma página web *p*. 33

Figura 9 - Subfases da etapa de identificação dos termos candidatos a descritores de uma página *web p*. 37

Figura 10 - Representação do cromossomo utilizado na solução do subproblema de identificação dos pesos das evidências utilizadas no processo de determinação do tópico de uma página *web*..... 41

Figura 11 - Algoritmo que representa o processo evolutivo do AG utilizado na solução do subproblema de descoberta de pesos das evidências..... 44

Figura 12 - Etapas do processo de otimização da lista de categorias, objetivando a definição do tópico de uma página *web*. 46

Figura 13 - Lista das 20 primeiras categorias associadas às 20 primeiras respostas obtidas a partir da submissão dos termos *Stem Cells*, *Stem* e *Cells* ao serviço de busca em diretório. 48

Figura 14 - Texto da evidência *Texto de Âncora* da página *web* que discorre sobre *Donald Knuth.* 53

Figura 15 – Exemplo de página submetida ao método de identificação de tópicos de páginas *web.* Esta página discorre sobre *Barack Obama.* 55

Figura 16 – Exemplo de página submetida ao método de identificação de tópicos de páginas *web.* Esta página discorre sobre *Climate Change.* 55

Figura 17 - Lista de termos utilizados para obtenção das páginas *web* que compõem a base de teste. 56

Figura 18 – Exemplo de página *web* utilizada nos experimentos do método proposto e que discorre sobre *Acne* enfatizando a definição do problema, os tipos de acnes existentes e o grupo de pessoas mais acometidas pela doença. 57

Figura 19 - Exemplo de página *web* utilizada nos experimentos do método proposto e que discorre sobre *Acne* enfatizando a definição do problema, a sua forma de tratamento e propondo a compra de um produto, desenvolvido pela empresa dona do *site*, utilizado no tratamento da doença. 58

Figura 20 – Procedimento utilizado na definição da qualidade de um tópico para uma determinada página *web* p_1 62

Figura 21 - Página pessoal de Donald Knuth. http://www-cs-faculty.stanford.edu/~knuth..... 64

Figura 22 - Lista de categorias utilizadas na realização do experimento 3. 66

Lista de Tabelas

Tabela 1 – Configuração genética dos 10 melhores resultados obtidos com a execução do AG, cujo objetivo era maximizar o valor da função objetivo. 51

Tabela 2 - Relação de termos utilizados para obter as páginas teste e seus respectivos tópicos, gerados pelo sistema de RI aqui proposto. 60

Tabela 3 – Tópicos obtidos pela submissão de seis páginas *web* ao método proposto por Rafiei e Mendelzon e ao método de identificação automática de tópicos proposto nesta dissertação. ... 65

Tabela 4 – Resultados obtidos com a aplicação prática do método proposto na classificação de páginas *web* no diretório *Google*. 68

Capítulo 1

Introdução

A *World Wide Web* (*WWW*) é considerada um repositório universal do conhecimento e da cultura humana que viabiliza o compartilhamento de idéias e de informações numa proporção jamais vista. Seu grande sucesso é decorrente da facilidade com que usuários publicam suas páginas, pois nenhum conhecimento técnico profundo é exigido destes [3].

Contudo, associados a essa imensa quantidade e variedade de informações disponíveis surgem problemas característicos a esse contexto como, por exemplo, encontrar informações úteis para usuários, classificar documentos e filtrar informações [3].

Dentre os vários ramos de atuação, a área de recuperação de informação atua na pesquisa e no desenvolvimento de técnicas eficientes para resolver os problemas acima citados, assim surgiram as máquinas de busca, os sistemas de classificação automática de documentos *web* e os sistemas de filtragem de informações que devem funcionar de forma eficiente para atender as necessidades de seus usuários.

As máquinas de busca na *web* permitem que usuários expressem, através de um conjunto de palavras, sua necessidade de informação, para que em seguida um possível conjunto de documentos relacionados à suas necessidades seja devolvido como resposta a solicitação feita [3]. Os sistemas de classificação automática de documentos *web* viabilizam a classificação dos documentos em diversas categorias pré-definidas de acordo com o assunto abordado por estes, como por exemplo: biologia, informática, geografia, etc [19]. Já os sistemas de filtragem permitem que somente as informações de interesse de um usuário em particular cheguem a ele [3].

Com a finalidade de melhorar a qualidade dos resultados obtidos pelos variados sistemas de RI, por exemplo: (1) o *ranking* das máquinas de busca; (2) a classificação

de documentos; (3) a filtragem de informações; (4) a exibição de propagandas; e (5) a recomendações automáticas de informações, diversas evidências textuais e estruturais úteis que podem ser extraídas da *web* são frequentemente utilizadas. Tanto as evidências textuais - como o próprio texto do documento *web*, o resumo de um documento (*snippet*), a concatenação dos textos de âncora dos *links* que apontam para um documento, o tópico do documento – quanto às evidências estruturais - como o valor de *PageRank* [5], de *Hypertext Induced Topic Search (HITS)* [13], o nível *da Universal Resource Locator (*URL) – podem ser utilizadas isoladamente ou podem ser combinadas provendo uma coleção útil de informações acerca de um determinado documento *web*.

A crescente diversidade de informações textuais disponíveis na Internet motiva muitos pesquisadores a concentrarem seus estudos na análise e no processamento de textos oriundos de documentos *web*. Uma tarefa importante no processamento destes textos é a descoberta automática de seus tópicos, ou seja, o principal assunto sobre o qual o documento *web* discorre. A identificação do tópico de uma página *web* possui uma vasta aplicabilidade nos sistemas de RI em geral, podendo servir como uma nova fonte de evidência para melhorar *ranking* de páginas *web*, classificação e filtragem destas páginas e na recomendação automáticas de informações. Ainda pode ser utilizado em sistemas de exibição automática de propagandas contextuais e na validação de páginas comerciais e pessoais, informando como estas páginas são conhecidas na *web*.

Este trabalho tem como objetivo o desenvolvimento de um método utilizado para identificar automaticamente o tópico de uma página *web.* Definimos o tópico de uma página *web p* como um conjunto de, no máximo, cinco palavras distintas que estão associadas ao assunto principal de *p.* O método aqui exposto propõe a utilização de diversas fontes de evidências textuais, extraídas de *p*, que combinadas darão origem a uma lista de termos associados ao assunto desta, e que posteriormente será utilizada na definição de seu tópico. A combinação das evidências utilizadas deve ser feita automaticamente, evitando a subjetividade de opiniões humanas. Para atingir esse objetivo faz-se necessário o uso de algoritmos genéticos (AG).

Algoritmos Genéticos são métodos de busca e otimização inspirados nos conceitos da teoria de seleção natural das espécies. Os sistemas desenvolvidos a partir deste princípio são utilizados para procurar soluções de problemas complexos ou que possuam espaço de busca muito grande e que não podem ser resolvidos por técnicas

tradicionais, como por exemplo, a força bruta onde todas as possíveis soluções devem ser avaliadas [14].

A utilização de algoritmo genético no processo de identificação dos tópicos de páginas *web* visa à descoberta dos pesos a serem associados a cada uma das fontes evidências utilizadas no método. Os pesos serão aplicados em uma equação linear, que definirá o grau de importância de um termo *t* para a definição do assunto de um documento *D*. O espaço de busca do problema em questão compreende todos os números reais entre 0 e 1, por isso, utilizaremos a técnica de AG.

Na etapa final do método, onde será gerado o tópico de *p*, será utilizado um serviço de busca em diretório. A este serviço serão submetidos os três melhores termos associados ao assunto de *p*, e dele serão obtidas as melhores categorias relacionadas às primeiras vinte respostas desta busca, estas categorias passarão por um processo de otimização e originarão o tópico de *p*. A Figura 1 ilustra o método proposto nesta dissertação, no qual uma determinada página *web p* é a ele submetida e, através do uso de diversas fontes de evidências e de recursos providos pela própria *web*, o tópico de *p* é identificado.

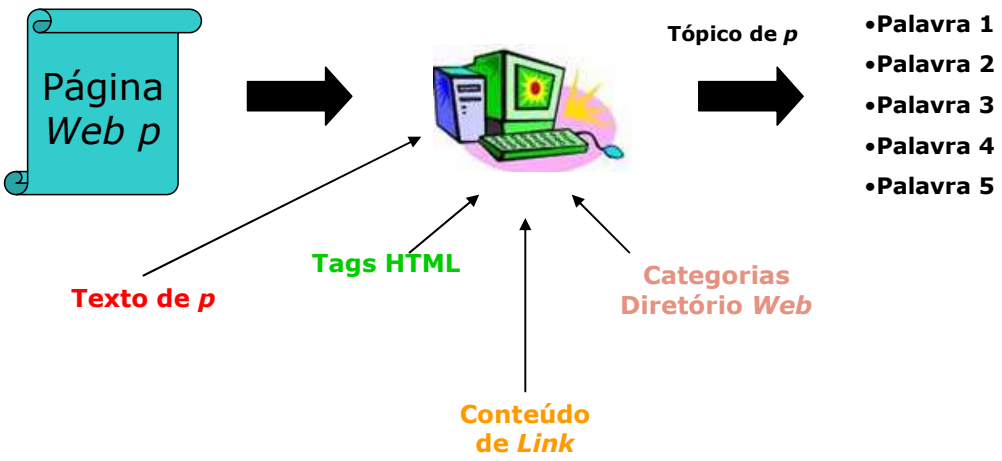

Figura 1 - Ilustração do objetivo do método proposto nesta dissertação.

As principais contribuições deste trabalho são: (1) o desenvolvimento de um método simples e eficiente para se determinar o tópico de uma página *web*, visando sua utilização como nova fonte de evidência nos sistemas de RI em geral; e (2) a avaliação, quanto à importância, das diversas fontes de evidências utilizadas no método.

Os resultados experimentais, gerados pelo método proposto, serão avaliados por 10 especialistas, que julgarão a importância de cada uma das cinco palavras que compõem o tópico de uma página *p*.

Esta dissertação é constituída de seis capítulos. No Capítulo 2 são apresentados os trabalhos relacionados utilizados como fonte de estudo para o desenvolvimento desta dissertação. No Capítulo 3 são apresentados os principais conceitos necessários para o entendimento deste trabalho. O Capítulo 4 descreve o método proposto de identificação de tópicos de páginas *web*, que utiliza a combinação de diferentes fontes de evidências. No Capítulo 5 são apresentados os experimentos realizados e discutidos seus respectivos resultados. Finalmente, no Capítulo 6 são apresentadas as conclusões e as direções para trabalhos futuros.

Capítulo 2

Trabalhos Relacionados

O problema de explorar diversas fontes de evidências na descoberta de termos descritores de páginas *web* foi explorado nos trabalhos de [15], [20] e [24]. Nos três trabalhos citados as diversas evidências eram combinadas, de diferentes formas, sempre com o objetivo de identificar os melhores termos descritores de um conjunto de páginas ou de apenas uma página *web*. A descoberta dos melhores termos descritores de um conjunto de páginas pode ser útil na solução de problemas de agrupamento de páginas, por exemplo. Já a descoberta dos melhores termos descritores de uma única página possui aplicabilidade em problemas de classificação, de filtragem, de *ranking* e também na definição do tópico principal desta página.

O objetivo de Liu e Chin, em [15], é a geração automática de um sumário onde, dado um tópico, seus subtópicos e suas definições equivalentes devem ser listados hierarquicamente. O sistema recebe um conjunto de termos, que representa o tópico da página, em seguida, este tópico é submetido ao serviço de uma máquina de busca e dela são obtidas as primeiras 100 páginas *web* retornadas. Do conteúdo textual dessas páginas são extraídos os dados das fontes de evidências utilizadas: freqüência com que os termos aparecem destacados pelas *tags* html (*HyperText Markup Language*) *<h1>,<h2>,<h3>,<h4>,<bold>* e freqüência dos termos no texto da página. O texto de âncora dos *links* é uma fonte de evidência utilizada com restrição, ele é utilizado para descobrir definições de subtópicos, no segundo nível de estrutura de *links*, somente se o texto de âncora do *link* possuir grafia igual ao subtópico procurado. A heurística utilizada para remover palavras irrelevantes do conjunto de tópicos e subtópicos é feita com base na freqüência dos termos: um termo só é freqüente se aparecer em mais de dois documentos. Outra característica deste trabalho é que não são atribuídos pesos diferentes às diversas fontes de evidências utilizadas. No Capítulo 5 veremos que experimentos, feitos com o método proposto nesta dissertação, demonstram que existem

fontes de evidências que são mais importantes na definição do tópico das páginas *web* se comparados a outras, logo estas não podem ser igualadas em relação aos seus respectivos grau de importância. No método de identificação de tópicos, aqui proposto, diversas fontes de informação serão utilizadas como evidências e a cada uma delas será atribuído um peso específico.

Rafiei e Mendelzon, em [20], propõem uma nova forma de se determinar automaticamente os principais termos de uma página *web* baseada nos valores de relevância destes termos para aquela página. Para isso, foram generalizadas duas técnicas bastante utilizadas de cálculo de *ranking* de páginas *web*, que são: (1) *PageRank* [5]; e (2) *Hubs and Authorits* [13], gerando dois algoritmos que determinam os termos relevantes de páginas *web*: (1) algoritmo para computação dos termos relevantes de uma página com um nível de propagação de influência – generalização do *PageRank* [5]; e (2) algoritmo para computação dos termos relevantes de uma página com dois níveis de propagação de influência – generalização do *Hubs and Authorits* [13]. No modelo de propagação de influência em um nível, a reputação de um termo *t* em uma página *p* é expressa pela seguinte fórmula:

$$R^n(p,t) = \begin{cases} \dfrac{d}{N_t} + (1-d)\sum\limits_{q->p} \dfrac{R^{n-1}(q,t)}{O(q)} & \textit{se t aparece em p} \\ (1-d)\sum\limits_{q->p} \dfrac{R^{n-1}(q,t)}{O(q)} & \textit{caso contrário} \end{cases} \quad (2.1)$$

onde *d* é a probabilidade de um surfista randômico, procurando por um termo *t* nas páginas da *web*, pular para uma página escolhida randomicamente entre as que contêm o termo *t*, (*1-d*) a probabilidade de ele seguir um *link* de saída da página *p*, *Nt* é o número de páginas na *web* que contém o termo *t*, *q→p* representa um *link* da página *q* para a página *p* e *O(q)* é o número de *links* de saída de *p*. Já no modelo de propagação de influência em dois níveis, a reputação de um termo *t* em uma página *p* é expressa utilizando duas métricas, a de *Hub* e de Autoridade, da seguinte forma:

$$H^n(p,t) = \begin{cases} \frac{d}{2N_t} + (1-d)\sum_{q->p} \frac{A^{n-1}(q,t)}{I(q)} & \textit{se t aparece em p} \\ (1-d)\sum_{q->p} \frac{A^{n-1}(q,t)}{I(q)} & \textit{caso contrário} \end{cases} \quad (2.2)$$

$$A^n(p,t) = \begin{cases} \frac{d}{2N_t} + (1-d)\sum_{q->p} \frac{H^{n-1}(q,t)}{O(q)} & \textit{se t aparece em p} \\ (1-d)\sum_{q->p} \frac{H^{n-1}(q,t)}{O(q)} & \textit{caso contrário} \end{cases} \quad (2.3)$$

onde *I(q)* representa o número de *links* de entrada da página *q*. Com esses algoritmos Rafiei e Mendelzon tentam inferir os melhores termos descritores de uma página *p*. O objetivo de ambos, na descoberta dos principais termos descritores de *p*, é utilização desta informação na validação de páginas comerciais e pessoais, informando como estas páginas são conhecidas na *web*.

Rafiei e Mendelzon não exploram as *tags* html, a freqüência dos termos e nem a concatenação dos textos de âncora de *links* como fonte de evidência. No Capítulo 5 veremos a importância dessas evidências na identificação dos termos relevantes de uma página *web*, todas foram utilizadas no método de identificação de tópicos aqui proposto.

Zeng e outros, em [25], reformularam o problema de *clustering* para um problema de criação de *ranking* de frases relevantes obtidas de um conjunto de páginas *web*. Zeng tem o objetivo de agrupar, *on-line*, páginas obtidas a partir de uma pesquisa submetida a uma máquina de busca, e para isso, é necessário descobrir os termos que melhor descrevem cada grupo de páginas. Por exemplo, quando o termo "*Jaguar*" for submetido à máquina de busca, o sistema de RI proposto em [25] deverá dividir as páginas, obtidas como respostas, em cinco grupos: *Jaguar Cars*, *Panthera Onca*, *Mac OS*, *Big Cats*, *Clubs* e *Others*. O nome e a quantidade de grupos são definidos pelo sistema proposto. Com essa finalidade, são utilizados os dados oriundos dos resumos e dos títulos das 200 primeiras páginas obtidas como resposta a uma consulta feita a uma máquina de busca. Zeng ainda aplica a técnica de *n-grams* (técnica utilizada para gerar subseqüências a partir de seqüências maiores), elimina as *stopwords* (termos muito comuns - artigos, preposições e pronomes - em textos) e utiliza o valor de *TF×IDF* (*Term Frequency* × *Inverse Document Frequency*) dos termos e seus respectivos

tamanhos como fonte de evidência para identificar o termo que melhor descreve o conteúdo de um conjunto de páginas. Porém, o texto de âncora dos *links* das páginas não é utilizado como fonte de evidência.

Tuin, Abdullah e Kong, em [24], propõem o uso de ontologias na identificação automática de tópicos de páginas *web*. A idéia principal do método é a exploração da estrutura hierárquica dos conceitos pertencentes a uma ontologia na descoberta do tópico de uma página *web p*. Através do mapeamento feito entre as palavras-chaves extraídas do texto da página e os conceitos da ontologia, um conceito específico é obtido e definido com sendo o tópico de *p*. Contudo, [24] explica que problemas na limitação do número de conceitos presentes nas ontologias forçou-os a enriquecer cada um deles com novos conceitos provindos do *WordNet* (http://wordnet.princeton.edu/), pois quanto maior o número de palavras-chaves, extraídas de *p*, mapeadas nos conceitos da ontologia melhor é o desempenho da técnica de identificação de tópico proposta. Para evitar problemas relacionados à limitação de conceitos, o método proposto nesta dissertação usará a estrutura hierárquica de um diretório abrangente e popular da *web*.

As técnicas que utilizam estruturas hierárquicas para identificação de tópico são usualmente aplicadas aos sistemas de classificação de páginas [23], [8]. Verificaremos no Capítulo 5 que o método aqui proposto também possui aplicação prática na classificação de páginas *web* dentro da hierarquia do diretório *Google*.

Capítulo 3

Conceitos Básicos

Neste capítulo são expostos todos os conceitos necessários para a compreensão do método de identificação de tópicos de páginas *web* aqui proposto e dos resultados obtidos através dos experimentos efetuados para avaliar a eficiência do mesmo.

3.1 Definições Iniciais

Termos Descritores de uma Página

Os termos descritores de uma página *web* constituem um conjunto de palavras-chaves que estão associadas ao tópico principal desta página. A utilização de termos descritores é uma forma de definir o tipo de conteúdo presente em um documento, sendo úteis para caracterizar o assunto do documento. A Figura 2 lista, sem ordem de relevância, 20 termos descritores, identificados manualmente, referentes à página http://www.ufam.edu.br.

1. graduação
2. faculdade
3. federal do amazonas
4. campus universitário
5. mestrado
6. doutorado
7. aluno
8. professor
9. amazonas
10. reitor
11. universidade do amazonas
12. ufam
13. universidade federal do amazonas
14. universidade federal
15. universidade
16. ensino
17. educação
18. pesquisa
19. pós-graduação
20. especialização

Figura 2 – Exemplos de termos descritores da página http://www.ufam.edu.br

Tópico de uma Página

O tópico de uma página *web p* será aqui representado, por um conjunto de, no máximo, cinco termos descritores distintos de *p*. É importante observar que estes cinco termos podem não ocorrer no texto original da página. A Figura 3 demonstra o tópico da página http://www.ufam.edu.br.

1. universidade federal do amazonas
2. ufam
3. ensino superior
4. educação
5. amazonas

Figura 3 – Tópico da página http://www.ufam.edu.br

Texto de Âncora de *Link*

O texto de âncora de um *link* corresponde ao texto de um *hyperlink* que pode ser lido e "clicado" pelo usuário. Esta informação geralmente representa uma descrição simples e objetiva do conteúdo da página para a qual o *link* faz referência. A concatenação de vários textos de apontadores que referenciam uma página *p* constitui uma fonte de evidência. Os textos de âncora des *links* é uma fonte de informação originária de autores distintos da página referenciada, portanto a descrição do documento referenciado se baseia em um ponto de vista externo ao autor da página, tornando-se uma importante fonte de informação.

Outra vantagem na utilização da concatenação dos textos de âncora dos *links* como fonte de informação é a descrição textual obtida para páginas de caráter não textual (páginas que cujo conteúdo seja apresentado por imagens, aplicações, gráficos e sons). Na Figura 4 estão listados o texto de âncora de *link* de 30 apontadores que referenciam a página http://www.ufam.edu.br.

1. universidade federal do amazonas
2. universidade federal do amazonas
3. federal do amazonas
4. minha facul
5. universidade federal do amazonas's website
6. universidade federal do amazonas
7. universidade federal do amazonas
8. universidade federal do amazonas
9. universidade federal do amazonas
10. universidade federal do amazonas
11. universidade federal do amazonas
12. ufam
13. ufam
14. ufam
15. federal do amazonas
16. universidade federal do amazonas - ufam
17. ufam
18. ufam
19. ufam
20. ufam
21. universidade federal do amazonas
22. universidade federal do amazonas
23. universidade federal do amazonas
24. universidade federal do amazonas
25. ufam
26. universidade federal do amazonas
27. universidade do amazonas
28. ufam
29. ufam
30. universidade federal do amazonas

Figura 4 - Lista do texto de âncora de 30 apontadores que referenciam a página http://www.ufam.edu.br

***Term Frequency* × *Inverse Document Frequency* (*TF*×*IDF*)**

O valor de *TF*×*IDF* tem por objetivo identificar a importância de um termo para uma coleção, ele é amplamente utilizado como fonte de informação nos sistemas de RI. O valor de *TF* (*Term Frequency*) de um termo qualquer *t* presente em um documento *D* indica a freqüência com que *t* ocorre em *D*. Já o valor de *IDF* (*Inverse Document Frequency*) de um termo qualquer *t*, expressa a importância de *t* para esta coleção e é dado por $\frac{\log N}{Nt}$, onde *N* corresponde à quantidade de documentos presente na coleção e *Nt* corresponde a quantidade de documentos da coleção em que o termo *t* aparece, logo temos:

$$TF \times IDF(t, D) = TF(t, D) \times \frac{\log N}{N_t} \quad (3.1)$$

Fontes de Evidências

As fontes de evidências servem como fonte de informações (dados) para variados sistemas de RI e podem ser extraídas tanto dos textos das páginas *web* quanto da estrutura de *link* das mesmas. Em [15], as *tags* html (<*h1*>, ... , <*h4*>,<*b*>) e a freqüência com que frases ocorrem na página são utilizadas como fonte de evidência para descobrir uma lista de sub-tópicos relacionados a um tópico específico. Em [22], é utilizada a combinação de três fontes de evidências distintas com o objetivo de melhorar a qualidade da ordenação de documentos em máquinas de buscas, são elas: (1) o conteúdo textual dos documentos; (2) o valor de reputação dos documentos; e (3) a concatenação dos textos de âncora de *links* que referenciam cada documento.

No processo de identificação dos tópicos de páginas *web* será utilizada a combinação de diferentes fontes de evidências, como: (1) o conteúdo textual da própria página *p*, explorando as *tags* html; (2) concatenação dos textos de âncora de *links* que referenciam *p*; e (3) valores de *TF*×*IDF* dos termos presentes em *p*. A justificativa da utilização de tais fontes e a combinação, aplicada a elas, serão discutidos no Capítulo 4.

Diretório *Web*

Os diretórios *web*, também chamados de catálogos e de *yellow pages*, são taxonomias hierárquicas utilizadas para classificar o conhecimento humano [3]. Eles se caracterizam pela categorização e pela organização em tópicos de suas páginas. Normalmente, a classificação de novas páginas na estrutura hierárquica de um diretório é feita manualmente. Os principais diretórios existentes são o diretório *Google* (http://dir.google.com/) e o diretório *Yahoo!* (http://br.yahoo.com/info/diretorio.html).

O diretório G*oogle* utiliza os *links* e a categorização do *Open Directory Project* (ODP) [18], porém a tecnologia de pesquisa é a da própria *Google* (http://www.google.com). O ODP [18] é um grande diretório público gerenciado pela *Netscape*, é mantido por um grupo de editores voluntários de todo o mundo que avaliam os *sites* para a sua inclusão na estrutura hierárquica [2]. Já o diretório *Yahoo!*, possui seu próprio diretório, onde a classificação das páginas também é feita manualmente [21].

3.2 *N-grams*

De acordo com [7], *n-grams* é uma subseqüência de *n* itens obtidos de uma seqüência maior. Tais itens podem ser representados por letras, sílabas ou palavras de um texto, por exemplo. Neste trabalho os itens utilizados pra formar as subseqüências são palavras e as subseqüências formadas serão chamadas de termos. O valor de *n* representa a quantidade máxima de palavras utilizadas para compor um termo. Utilizaremos *n* igual a quatro, assim, termos compostos por uma, duas, três e quatro palavras serão utilizados no processo de identificação de tópico. Como exemplo, o texto “Universidade Federal do Amazonas” originaria os seguintes *n-grams*:

- *Uni-grams*: Universidade, Federal, do, Amazonas
- *Bi-grams*: Universidade Federal, Federal do, do Amazonas
- *Tri-grams*: Universidade Federal do, Federal do Amazonas
- *Quad-grams*: Universidade Federal do Amazonas

Os *n-grams* serão utilizados com o propósito de obter do texto de uma página *web* os possíveis termos descritores desta. No exemplo acima, 10 termos se tornariam candidatos a descritores.

A técnica *n-grams* é frequentemente utilizada nos sistemas da área de RI [15], [25] e [1]. Ela será aplicada no sistema de identificação de tópicos de páginas *web* com o objetivo de gerar novos termos para compor o conjunto de possíveis descritores de uma página, a aplicação desta técnica é justifica, pois alguns termos isolados não possuem sentido completo.

3.3 Algoritmos Genéticos

Algoritmos Genéticos são métodos de busca e otimização inspirados nos conceitos da teoria de seleção natural das espécies. Os sistemas desenvolvidos a partir deste princípio são utilizados para procurar soluções de problemas complexos ou que possuam espaço de busca muito grande. Estes algoritmos são baseados nos processos genéticos (hereditariedade, mutação, seleção natural e cruzamento) de organismos biológicos para procurar soluções ótimas ou aproximadas do problema [14].

Para tanto, deve-se adequar o problema a ser resolvido aos requisitos exigidos por um AG, dessa forma três fases distintas de adequação do problema devem ser destacadas:

1. Codificação de cada possível solução do problema em uma estrutura chamada cromossomo;
2. Definição das configurações genéticas utilizadas: taxa de mutação, taxa de cruzamento, método de seleção, tamanho da população e número de gerações usadas no processo de evolução do AG.
3. Definição da função objetivo, que tem por finalidade avaliar o grau de adequação de cada cromossomo como solução do problema.

A utilização de algoritmo genético no processo de identificação dos tópicos de páginas *web*, proposto nesta dissertação, visa à descoberta dos pesos a serem associados a cada uma das seis evidências utilizadas no método (Seção 4.1.1). Os pesos serão aplicados na equação linear mostrada na Equação 4.1, que define o grau de importância de um termo *t* para um documento *D*. O processo de descoberta dos pesos deve ser automático, evitando as suposições humanas. Os AGs são boas técnicas utilizadas para atacar problemas de busca com espaço de busca intratavelmente grandes e que não podem ser resolvidos por técnicas tradicionais, como por exemplo a força bruta onde

todas as possíveis soluções devem ser avaliadas[14]. Por isso, é justificada a escolha de AGs na solução do problema de identificação de pesos cujo espaço de busca corresponde a todos os números reais compreendidos entre 0 e 1.

As subseções seguintes discorrem com mais detalhes cada uma das três fases de adequação de um problema para a utilização de AGs.

3.3.1 Codificação dos Cromossomos

A codificação dos cromossomos é fundamental na modelagem o algoritmo genético, ela consiste em uma maneira de traduzir a informação do problema a ser resolvido em uma forma viável a ser tratada pelo AG [14].

Os cromossomos representam possíveis soluções do problema e podem ser vistos como um ponto, do espaço de busca, candidato a solução [17]. Eles devem ser codificados de acordo com as características do problema a ser resolvido. Cada cromossomo é composto por vários genes e cada gene representa um aspecto distinto da solução.

Os cromossomos possuem diferentes formas de serem representados, entre elas: binária, inteira ou real. A essa representação se dá o nome de alfabeto do AG [14]. De acordo com a classe de problema que se deseja resolver pode-se usar qualquer um dos tipos.

A Figura 5 ilustra a codificação de um cromossomo com 5 genes cujo alfabeto utilizado é o binário.

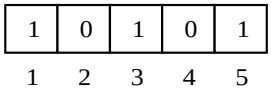

Figura 5 - Exemplo de codificação de uma cromossomo com 5 genes e alfabeto binário.

3.3.2 Operadores Genéticos

Os operadores genéticos mais conhecidos e utilizados nos algoritmos genéticos são os de seleção, cruzamento (*crossover*) e de mutação [17].

- **Seleção**

Este operador seleciona cromossomos da população para a realização da reprodução. Quanto maior a sua aptidão maior é a chance dele ser escolhido para reprodução [17].

O método de seleção de pais deve ser semelhante ao mecanismo de seleção natural que atua sobre as espécies biológicas, em que pais mais aptos geram mais filhos e pais menos aptos geram menos filhos [14]. Conseqüentemente deve-se privilegiar os indivíduos mais aptos, sem desprezar completamente os de aptidão inferior, pois, estes podem ter características genéticas que sejam favoráveis à criação de um indivíduo que representa a melhor solução para o problema. Por outro lado, se apenas os melhores indivíduos se reproduzirem ocorrerá um efeito chamado de convergência genética.

A convergência genética ocorre quando a população se compõe por indivíduos cada vez mais semelhantes, acarretando a falta de diversidade o que impede a evolução satisfatória da população [14].

O método da Roleta é uma maneira de selecionar indivíduos, onde cada indivíduo possui uma fatia da roleta proporcional à sua adaptação. A cada giro da roleta um indivíduo é selecionado, tendo maior chance aqueles que possuem as maiores fatias, sem deixar de lado a diversidade dos menos adaptados [9]. Porém, outras formas de seleção podem ser aplicadas dependendo do problema a ser tratado.

- **Cruzamento (*Crossover*)**

Operador genético que cria novos indivíduos através da combinação das características de outros dois indivíduos [16]. Este processo é ilustrado na Figura 6, onde a solução está codificada no alfabeto binário.

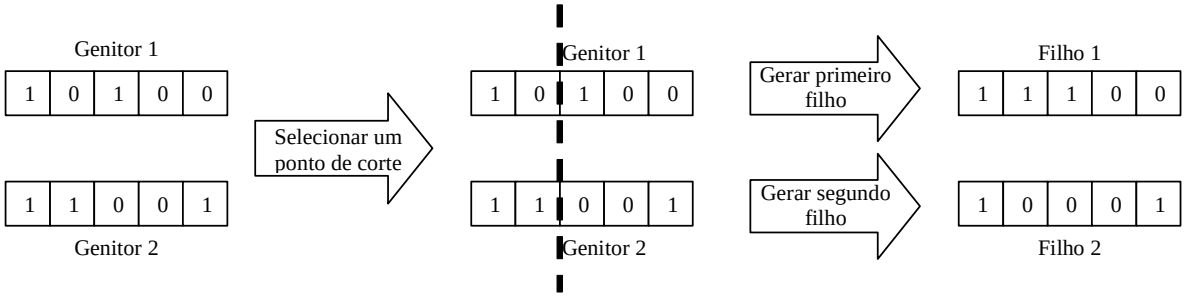

Figura 6 - Descrição da operação de cruzamento.

O funcionamento do operador genético de cruzamento consiste em: (1) selecionar os cromossomos genitores; (2) escolher, aleatoriamente, o ponto onde ocorrerá o corte para a realização do cruzamento; (3) separar as características genéticas dos cromossomos genitores em duas partes (uma a esquerda e outra a direita do ponto de corte); (4) gerar o primeiro filho, que será composto pela parte esquerda do primeiro "pai" e pela parte direita do segundo "pai"; e (5) gerar o segundo filho, que será composto pela parte direita do primeiro "pai" e pela parte esquerda do segundo "pai" [14]. A realização do cruzamento garante a troca de informações genéticas entre diferentes e possíveis soluções.

- **Mutação**

É um operador unário que cria novos indivíduos através da modificação aleatória dos valores contidos em um ou mais genes de um cromossomo [16].

Ao operador de mutação é associada uma probabilidade baixa de ocorrência, caso contrário o funcionamento do AG se parecerá com uma técnica chamada *random walk*, na qual a solução é determinada de forma aleatória [14]. Quando a probabilidade atinge um gene em questão, então seu valor é aleatoriamente alterado por outro pertencente ao domínio válido [14]. A mutação garante a diversidade das características dos indivíduos da população e permite que sejam introduzidas características que não estavam presentes em nenhum dos indivíduos [16]. O valor da probabilidade de ocorrência de mutação (taxa de mutação) é definido como parâmetro do AG.

A Figura 7 demonstra graficamente a ocorrência de mutação no gene 2 do cromossomo, cuja solução está codificada no alfabeto binário.

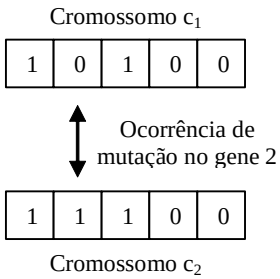

Figura 7 - Ocorrência de mutação no gene 2 do cromossomo codificado no alfabeto binário.

3.3.3 Função Objetivo

A função objetivo tem por finalidade determinar a qualidade de um indivíduo (cromossomo) como solução do problema, isto é, ela retorna um valor numérico que reflete quão bons os parâmetros representados no cromossomo resolvem o problema em questão [14]. A função objetivo deve refletir os objetivos a serem alcançados na resolução do problema.

Jarvelin e Kelalainen, em [11], propõem a utilização de duas novas métricas como forma de avaliar o desempenho de um sistema cuja resposta seja uma lista ordenada onde cada elemento desta lista possui um grau de importância, assim elementos com maior grau de importância devem aparecer à frente de elementos de menor importância. Nas métricas propostas em [11], o julgamento de relevância de um elemento da resposta não é binário (é importante ou não é importante), ele propõe a existência um grau de relevância que é atribuído a cada possível elemento da resposta. O sistema ideal é aquele cujos elementos de resposta estejam ordenados decrescentemente por valor de relevância.

As métricas propostas foram: (1) *Directed Cumulated Gain – CG*; e (2) *Discounted Cumulated Gain – DCG* [11]. O funcionamento destas métricas será explanado a seguir:

1. *Directed Cumulated Gain* – CG

Nesta métrica a lista de documentos (*L'*) retornada pelo sistema a ser avaliado é transformada em uma lista de ganho de valores apenas substituindo os identificadores (*IDs*) dos documentos de *L'* por seus valores de relevância. Assumindo que o valor de relevância de cada documento varia entre 0 e 3 (3 denota o maior grau de relevância e 0 nenhum grau de relevância), será, então, gerado um vetor (*G*) de tamanho *n* (mesmo tamanho da lista *L'*), onde cada elemento *i* do vetor conterá os valores 0, 1, 2 ou 3 correspondente ao respectivo valor de relevância do documento *i* da lista *L'*.

Por exemplo:

L' = (121; 145; 32; 6; 6789; 98; 56; 43; 2; 111; ....)

G' = (3; 2; 3; 0; 0; 1; 2; 2; 3; 0; ....)

O ganho cumulativo da posição *i* de *G'* é calculado somando-se os primeiros *k* elementos de *G'*, onde $1 \leq k \leq i$. Definindo *G[i]* como sendo a posição *i* no vetor *G*, a Equação 3.1 demonstra o cálculo utilizado para computar o vetor de ganho cumulativo (*CG*).

$$CG[i] = \begin{cases} G[i] & se\ i = 1 \\ CG[i-1] + G[i] & se\ i > 1 \end{cases} \quad (3.1)$$

Por exemplo, de *G'* obtém-se:

CG'=(3;5;8;8;8;9;11;13;16;16;....)

O ganho cumulativo da sétima posição do *ranking* é 11, *CG'[7] = 1*.

Dado que o valor de *CG[i]* é obtido pela soma dos primeiros *i* elementos do vetor *G*, os valores dos primeiros *n* elementos de *CG* tendem a crescer. A partir da posição *n*, os valores dos elementos de *CG* convergem para um valor fixo, significando que apenas os *n* primeiros documentos da lista *L'* (resposta do sistema a ser avaliado) possuíam algum grau de relevância (1 a 3, por exemplo), já os documentos retornados após a posição *n* do *ranking* não foram considerados relevantes no resposta.

2. *Discounted Cumulated Gain – DCG*

Outra métrica proposta é a *DCG*. Jarvelin e Kelalainen afirmam que quanto maior a posição de *ranking* de um determinado documento *d*, menos valioso *d* é para o usuário, pois a possibilidade deste último acessar *d* é pequena, devido ao esforço e tempo gastos, além do acumulo de informação já obtido com o acesso a documentos já examinados. Por isso, com o objetivo de reduzir gradativamente o valor de importância de um documento à medida que seu *ranking* aumenta, uma função de desconto é utilizada no cálculo do vetor de ganho cumulativo, passando a se chamar vetor de ganho cumulativo com desconto (*DCG*). A forma de desconto proposta por Jarvelin e Kekalainen, é a divisão do valor de importância do documento pelo *log* de seu *ranking* (Equação 3.2).

Definindo *b* como sendo a base do logaritmo, a Equação 3.2 demonstra a fórmula utilizada para computar os elementos do vetor de ganho cumulativo com desconto (*DCG*).

$$DCG[i] = \begin{cases} CG[i], & se\ i < b \\ DCG[i-1] + \dfrac{G[i]}{\log_b^i}, & se\ i \geq b. \end{cases} \quad (3.2)$$

Por exemplo, sendo *b* = 2, de *G'* e *CG'* obtém-se:

DCG'= (3; 5; 6.89; 6.89; 6.89; 7.28; 7.99; 8.66; 9.61; 9.61;...)

A habilidade de uma consulta retornar documentos de maior relevância mais próximos ao topo da lista de resultados pode ser avaliada por ambos as métricas: *CG* e *DCG*.

Os vetores de ganho cumulativo (*CG*) e de ganho cumulativo com desconto (*DCG*) também podem ser comparados com os seus vetores ideais. Em um vetor ideal todos os documentos com maior relevância estão à frente dos documentos de menor importância.

O vetor ideal é obtido da seguinte forma: sendo *k, l* e *m* documentos com valores de relevância 1, 2 e 3 respectivamente para uma consulta qualquer, o vetor ideal (*BV*) é obtido preenchendo os seus elementos de índice 1 ate *m* com o valor 3, os elementos de índice *m+1* até *m+l* com valor 2, os elementos de índice *m+l+1* até *m+l+k* com valor 1, e o restante dos elementos devem ser preenchidos com o valor 0, conforme Equação 3.3.

$$BV[i] = \begin{cases} 3, & se\ i \leq m, \\ 2, & se\ m < i \leq m + l, \\ 1, & se\ m + l < i \leq m + l + k, \\ 0, & caso\ contrário. \end{cases} \quad (3.3)$$

Por exemplo, o vetor ideal correspondente a *G'* é:

I' = (3; 3; 3; 2; 2; 2; 1; 0; 0; 0; ....)

Baseado no exemplo acima, pode-se obter o vetor *CG* ideal e o vetor *DCG* ideal, como mostra a Equação 3.4 e a Equação 3.5, respectivamente:

$$CG_I[i] = \begin{cases} G_I[i] & se\ i = 1 \\ CG_I[i-1] + G_I[i] & se\ i > 1 \end{cases} \quad (3.4)$$

$$DCG_I[i] = \begin{cases} CG_I[i], & se\ i < b \\ DCG_I[i-1] + \dfrac{G_I[i]}{\log_b^i}, & se\ i \geq b. \end{cases} \quad (3.5)$$

Por exemplo, de *G'* obtém-se os seguintes vetores ideais:

CG $_I$***'*** = (3; 6; 9; 11; 13; 15; 16; 16; 16; 16; ....)

DCG $_I$***'*** = (3; 6; 7.89; 8.89; 9.75; 10.52; 10.88; 10.88; 10.88; 10.88;...)

Em [11] é proposta a normalização dos vetores *CG* e *DCG*, pois desta forma podemos acompanhar o desempenho do sistema de RI a cada posição do vetor normalizado, onde o valor normalizado compreende a faixa [0,1]. O valor 1 representa o desempenho ideal do sistema de RI. Para normalizar os vetores *CG* e *DCG* basta dividi-los por seus correspondentes vetores ideais CG_I e DCG_I, como mostra a Equação 3.6.

$$norm - vect(V, I) = (v_1 / i_1, v_2 / i_2, ..., v_k / i_k) \quad (3.6)$$

Por exemplo, baseado em *DCG'* e DCG_I', obtém-se o vetor normalizado *nDCG*:

$$nDCG' = norm - vect(DCG', DCG_I')$$

nDCG' = (1; 0.83; 0.87; 0.77; 0.70; 0.69; 0.73; 0.79; 0.88; 0.88;...)

Um dado sistema de RI pode ser avaliado, quanto a sua eficiência, analisando-se a média de suas respostas quando um conjunto de consultas teste é submetido ao sistema. A média dos elementos de um vetor *(n)(D)CG*, acima de uma determinada posição que representa o *ranking* de resposta, sumariza o desempenho do sistema para uma dada consulta. O cálculo da média das *k* primeiras posições de um vetor *V*, que pode representar os vetores *(n)(D)CG*, é apresentado pela Equação 3.7:

$$avg - pos(V, k) = k^{-1} * \sum_{i=1...k} V[i] \quad (3.7)$$

A métrica utilizada para avaliar o desempenho do sistema de detecção de tópicos de páginas *web* descrito neste trabalho foi a *nDCG*. A utilização desta métrica se justifica, pois a mesma não se baseia em julgamentos binários de relevância (é importante ou não é importante), como a precisão e a revocação, além disso, é necessário que haja um *ranking* na lista de termos descritores da página, onde termos que melhor descrevem o seu conteúdo estejam no topo da lista e termos de menor importância apareçam no final desta lista, logo o julgamento de relevância de um termo para uma página não poderia ser binário, pois apenas os termos de maior relevância serão utilizados na detecção de seu tópico.

3.3.4 Validação Cruzada

A validação cruzada é uma técnica frequentemente utilizada em aprendizado de máquinas (*machine learning*). Ela consiste em dividir um conjunto de dados D em n subconjunto D_i. Com essa divisão um dado algoritmo pode ser executar n vezes, cada vez usando um conjunto de treino diferente $D - D_i$ e o teste do algoritmo pode ser feito com o subconjunto D_i [4].

O conjunto de treino ainda é subdividido em um subconjunto de dados para validação (D_v) e um subconjunto de dados para estimação ($D - D_i - D_v$). A idéia é utilizar o conjunto de treinamento para avaliar o desempenho dos indivíduos candidatos à solução do problema e assim, escolher o melhor. O subconjunto de treinamento permite selecionar o indivíduo e o subconjunto de validação permite validar o indivíduo escolhido como solução. Já com o conjunto de testes é verificada a generalização do modelo.

Neste trabalho, aplicaremos a técnica de validação cruzada no treinamento, validação e teste do algoritmo genético utilizado para determinar a combinação das diferentes fontes de evidências usadas na descoberta do tópico de uma página *web*.

Capítulo 4

Identificação dos Tópicos de Páginas *Web*

Neste capítulo detalhamos o método proposto para identificar automaticamente o tópico de páginas *web*. Serão discutidas as etapas e as decisões tomadas durante todo o processo, as vantagens do método e suas limitações.

4.1 O Método Proposto

Como citado anteriormente, este trabalho tem o objetivo de desenvolver um sistema de RI para identificar o tópico de uma página *web*. Durante a fase de revisão da literatura, observou-se que as idéias exploradas em outros trabalhos poderiam ser combinadas viabilizando a construção de um algoritmo para identificação de tópicos. O tópico de uma página pode ser utilizado como uma nova fonte de evidência para melhorar o desempenho de vários outros sistemas de RI. A Figura 8 demonstra graficamente as etapas do método utilizado no processo de identificação do tópico de uma página w*eb p*.

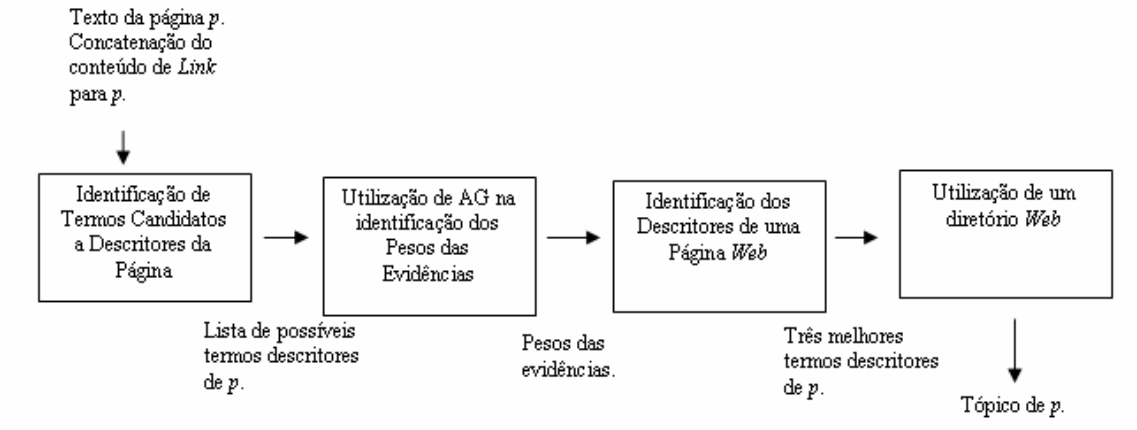

Figura 8 - Etapas do método de identificação do tópico de uma página web *p*.

Em linhas gerais, o método utilizado está dividido em quatro fases distintas:

(1) Identificação dos termos candidatos a descritores da página *web p*.

(2) Utilização de um algoritmo genético com o objetivo de identificar os pesos das fontes de evidências usadas na primeira fase.

(3) Identificação dos três melhores termos descritores de *p*.

(4) Utilização da estrutura hierárquica de um diretório abrangente e popular da *web* objetivando a identificação do tópico de *p*.

Para que o método seja desenvolvido, testado e avaliado foi necessário a criação de uma base de referência. O objetivo desta base é servir como um repositório de documentos que possam ser utilizados nos experimentos do processo de identificação do tópico de uma página *web*. Para compô-la foram coletadas da *WWW* 100 páginas, aleatoriamente, de assuntos diversos. Quatro restrições foram obedecidas durante a fase de coleta de tais páginas: (1) o idioma das páginas deve ser o inglês; (2) todas as páginas devem possuir conteúdo html; (3) as páginas devem ser referenciadas por, no mínimo, 5 outras página da *web*; e (4) páginas do tipo portal, páginas iniciais de *sites* de jornais e *blogs* não podem fazer parte da base de referência.

A restrição ao idioma inglês é necessária, pois pesquisas feitas no diretório *Google* comprovaram que existe um maior número de páginas classificadas neste idioma se compararmos ao número de páginas classificadas em outros idiomas. Como exemplo, em [6] pode-se verificar que existem 25.398 páginas classificadas no idioma português (data de acesso: 24/02/2009), no entanto, a partir das informações disponíveis em [18] verificamos que existem cerca de 2.657.234 páginas escritas em inglês classificadas no diretório (data de acesso: 24/02/2009). A escassez de informações no diretório utilizado é observada como um ponto negativo, podendo influenciando negativamente na avaliação do método.

A segunda restrição foi feita, pois o método utiliza o conteúdo destacado por *tags* html na descoberta dos termos descritores da página *web*.

Páginas do tipo portal, páginas iniciais de *sites* de jornais e *blogs* não foram coletadas, pois muitas vezes, por apresentarem notícias diversificadas, nem mesmo o ser humano consegue definir o assunto predominante destas, como exemplo citamos o portal da UOL(http://www.uol.com.br), do Terra (http://www.terra.com.br), o *site* do

jornal Acrítica (http://www.acritica.com.br) e o *blog* do ator Agnaldo Silva (http://bloglog.globo.com/aguinaldosilva/). Evitamos a utilização deste tipo de página em nossos experimentos, porém sugerimos como trabalho futuro o estudo da eficiência do método aqui proposto quando páginas com essas características são submetidas a ele.

Outra restrição imposta foi quanto a número de páginas que referenciam *p*. O número mínimo estipulado foi de cinco páginas, sendo obtido, no máximo, os dados de 30 *links* de entrada. O número mínimo cinco foi estipulado com o objetivo de garantir que o campo *TextoAncora*, da tupla que representa uma página *web* na base de referência, não seja vazio.

Na base de referência, cada página *web p* é representada por uma tupla <*TextoHtml*,*TextoAncora*,*ArqDescritores*>, onde *TextoHtml* corresponde ao conteúdo textual de *p* com *tags* de marcações, *TextoAncora* representa a concatenação dos textos de âncora dos *links* que referenciam *p* e *ArqDescritores* corresponde a um arquivo que contém os termos descritores pré-avaliados de *p*.

Para cada página foram identificados, manualmente, os termos que a descrevem. A estes foram associados valores de relevância que variam de 0 a 3, onde 0 indica nenhum grau de importância e 3 indica o maior grau de importância do termo para descrever o assunto da página. É importante observar que os termos descritores da página podem não pertencer ao texto da mesma. Assim, cada página pertencente à base de referência possui um arquivo de descritores pré-avaliados. Os dados presentes nestes arquivos serão utilizados posteriormente na avaliação do método proposto.

Após a obtenção da base de referência pode-se experimentar e avaliar o método de identificação dos tópicos aqui proposto. O processo de identificação do tópico de uma página *web* está dividido em quatro etapas distintas (Figura 8) a serem explicadas a seguir:

4.1.1 Identificação dos Termos Candidatos a Descritores da Página

O objetivo desta etapa é determinar um conjunto de termos candidatos a descritores de uma página *web p*. É importante observar que estes termos podem não ocorrer no texto da página em questão.

Como dado de entrada a esta etapa são fornecidas a página *web p* e a concatenação do conteúdo dos *links* de entrada de *p*. Como dado de saída é gerado um conjunto *T* de possíveis termos descritores de *p*, que servirá como entrada para segunda etapa do processo de identificação do tópico de *p*.

Na identificação dos termos candidatos a descritores de *p*, múltiplas fontes de evidências foram utilizadas. Na hipótese de tornar o método mais preciso e eficiente, a utilização de múltiplas evidências, como o texto completo de *p*, as *tags* de marcação html e a exploração da estrutura de *link* da página é justificada. Na linguagem de marcação html existem várias formas de se destacar palavras e/ou frases. Tendo em mente que o usuário desenvolvedor da página tende a destacar suas principais frases e palavras, e ainda que estas possam estar associadas ao assunto principal da página, as seguintes *tags* html foram selecionadas como fonte de evidência: *<h1>*, *<h2>*, *<bold>*, ** e *<title>*. Em [15], Liu e Chin utilizam a freqüência com que os termos aparecem destacados pelas *tags* html *<h1>*,*<h2>*,*<h3>*,*<h4>*,*<bold>* com o objetivo de gerar automaticamente um sumário onde, dado um tópico, seus subtópicos e suas definições equivalentes devem ser listados hierarquicamente.

Além das *tags* html, também foi utilizado como fonte de evidência a concatenação dos textos de âncora de *links* que referenciam a página *p* da qual se deseja extrair o tópico, explorando desta forma a estrutura de *links* de *p*. Os documentos que referenciam *p* não precisam obrigatoriamente pertencer a base de referência, eles são identificados com o auxilio de *Application Program Interface* (APIs) disponíveis em diversas máquinas de busca presentes na *web*, e em seguida o texto de âncora presente no apontador adequado é extraído. Damos destaque especial para a concatenação dos textos de âncora de *links* que referenciam *p*, pois nenhum dos trabalhos citados utiliza esta informação como fonte de evidência. Os experimentos da Seção 5.1 confirmam a importância desta fonte para o método. Além disso, a concatenação dos textos de âncora permite a aplicação do método em páginas de caráter não textual.

Outra fonte de evidência utilizada é o valor de *TF×IDF* para cada termo *t* candidato a descritor da página, conforme Equação 3.1. O objetivo da utilização desta fonte de evidência é identificar a importância de um termo *t* como descritor de uma página (Seção 3.1). Em [25], Zeng utiliza o valor de *TF×IDF* com o objetivo de criar um *ranking* de frases relevantes obtidas de vários conjuntos de páginas *web* e assim descobrir os termos que melhor descrevem cada grupo de páginas.

O processo de identificação dos possíveis termos descritores de *p* é subdividido em outras cinco fases, como mostra a Figura 9. Na primeira subfase é gerado um conjunto inicial *T'* contendo todos os termos candidatos a possíveis descritores de *p*, nas subfases seguintes o conjunto *T'* passa por mudanças (exclusão de termos) gerando conjuntos intermediários que ao final do processo darão origem ao conjunto definitivo *T* de termos candidatos a descritores de *p*. A seguir, será explicada cada subfase.

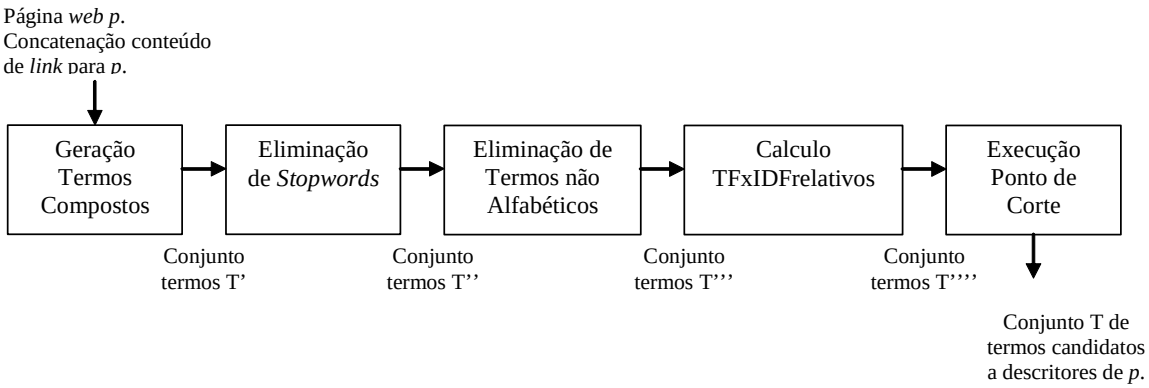

Figura 9 - Subfases da etapa de identificação dos termos candidatos a descritores de uma página *web p*.

Fase 1: Geração de Termos Compostos

Utilizando-se da técnica de *n-grams* (Seção 3.2) foram gerados novos termos para compor os possíveis descritores da página, a aplicação da técnica é justifica, pois alguns termos isolados não possuem sentido completo.

O valor de *n* utilizado no método aqui exposto é quatro, o mesmo usado em [15]. Assim, as palavras compostas por um, dois, três e quatro termos, podem fazer parte do novo conjunto de possíveis descritores. Uma restrição imposta durante a etapa de geração de termos compostos é que os novos termos gerados não podem começar ou terminar com *stopwords*, porém as *stopwords* podem aparecer no meio da composição do termo, como exemplo o termo: "Universidade do Amazonas", onde a *stopword* "do" aparece sua na composição.

Fase 2: Eliminação de *Stopwords*

A eliminação de termos muito comuns como artigos, preposições, pronomes e algumas palavras como *e-mail*, *everything* e *few* se faz necessário com a finalidade de minimizar o tamanho do conjunto de possíveis termos descritores da página. Analisando tais termos verificamos que estes não possuem significados relevantes que possam influenciar no resultado final do método, ou seja, na definição do tópico da página.

Como o objetivo da eliminação de *stopwords* é o de prover uma diminuição no número de termos descritores, a lista de termos classificados como *stopwords* deve ser extensiva a: alguns verbos, advérbios e adjetivos além dos artigos, preposições e conjunções [3].

Fase 3: Eliminação de Termos não Alfabéticos

Com o mesmo propósito justificado pela eliminação das *stopwords*, termos não alfabéticos como números, datas, códigos de endereçamento postal, números de telefone e endereços eletrônicos são eliminados.

Fase 4: Cálculo do *TF*×*IDF* relativo

O valor de *TF* (*Term Frequency*) de um termo qualquer *t* presente em um documento indica a freqüência com que *t* ocorre no documento, assim para cada termo candidato a descritor da página *web p* foi computada a sua freqüência em *p*.

Para cálculo dos valores de *IDF*, a coleção utilizada foi a *web*. Com auxilio da *API* de uma máquina de busca da *Google* foi obtido o valor aproximado do total de documentos presentes na coleção e o valor aproximado do total de documentos que contêm cada termo *t* candidato a descritor de *p*. Por este motivo o valor de *IDF* utilizado não é o real e sim o relativo à base indexada pela máquina de busca utilizada. Todos os valores de *IDF* necessários durante o processo de definição de tópicos foram pré-calculados e armazenados em memória.

Fase 5: Execução dos Pontos de Corte

Ainda com a finalidade de diminuir o tamanho do conjunto de possíveis termos descritores, dois pontos de corte foram estabelecidos: (1) eliminação de termos cuja

freqüência no documento é igual a 0; e (2) eliminação de termos cujo valor de *IDF* relativo seja superior a 16, pois, realizando experimentos e analisando os cálculos de *IDF*, percebemos que os extremos, ou seja, termos com altos valores e termos com valores muito baixos de *IDF*, representam um conjunto de termos que podem ser descartados do conjunto de possíveis descritores para a página, pois se tratam de termos extremamente raros e termos muito comuns (artigos, preposições, conjunções), respectivamente.

4.1.2 Utilização de AG na identificação dos Pesos das Evidências

O objetivo desta etapa é determinar os pesos a serem associados a cada uma das seis evidências identificadas na fase anterior. Os pesos serão utilizados na equação linear mostrada na Equação 4.1. O processo de descoberta dos pesos deve ser automático, evitando as suposições humanas, para isso fez-se necessário a utilização de algoritmos genéticos.

Como dado de entrada a esta etapa é fornecido um conjunto de páginas *web*, pertencentes à base se referência, e seus respectivos candidatos a termos descritores. Como dado de saída é gerado o peso das seis evidências identificadas na etapa anterior. Os pesos servirão como entrada para terceira etapa do processo de identificação do tópico de uma página *web*.

A Equação 4.1 mostra a equação linear utilizada para computar a importância *Imp* de um termo *t* para a definição do tópico de um documento *D*. *D* representa uma página *web*.

$$\mathrm{Im}p(t,D) = \alpha * eH1(t) + \beta * eH2(t) + \gamma * eBold(t) + \delta * eTitle(t) + \varepsilon * eTfIdf(t) + \xi * eTextAncora(t) \quad (4.1)$$

Imp(t,D) representa a importância do termo *t* para o documento *D*, *eH1* corresponde a freqüência com que o termo *t* ocorre entre as *tags* <*h1*> e </*h1*> no documento *D*, *eH2* corresponde a freqüência com que o termo *t* ocorre destacado pela *tag* <*h2*> no documento *D*, *eBold* corresponde a freqüência com que o termo *t* ocorre destacado pelas *tags* <*Bold*> e <*Strong*> no documento *D*, *eTitle* corresponde a freqüência com que o termo *t* ocorre entre as *tags* <*Title*> e </*Title*> em *D*, *eTfIDf* corresponde ao valor de *TF*×*IDF* relativo do termo *t* e *eTextAncora* corresponde a freqüência com que o termo *t*

ocorre na concatenação dos textos de âncora dos *links* que apontam para *D*. Já $\alpha, \beta, \gamma, \delta, \varepsilon, \xi$ representam os pesos associados as evidências *H1*, *H2*, *Bold*, *Title*, *TF×IDF* relativo e *Texto de Âncora* respectivamente e seus valores serão obtidos através do uso de algoritmos genéticos.

O uso de algoritmos genéticos nesta etapa do processo de identificação do tópico de uma página visa aperfeiçoar a escolha dos pesos associados a cada evidência utilizada, pois o espaço de solução do subproblema em questão é muito grande (conjunto dos números reais). Logo, temos um problema de busca no qual um AG específico será utilizado para solucioná-lo.

Objetivando o uso de um AG na descoberta dos pesos das evidências, deve-se adequar o subproblema, de descoberta de pesos, aos requisitos de um AG, dessa forma quatro fases distintas de adequação do problema são destacadas:

Fase 1: Planejamento dos Cromossomos

Como visto na Seção 3.3.1, os cromossomos representam possíveis soluções do problema e devem ser modelados de acordo com as características do problema a ser resolvido. Dentre os possíveis tipos de representação para os cromossomos, a utilizada neste trabalho será a real, ou seja, cada gene pertencente ao cromossomo representará um número real compreendido entre a faixa 0 e 1, estes valores corresponderão aos valores de $\alpha, \beta, \gamma, \delta, \varepsilon, \xi$, que representam, respectivamente, os pesos atribuídos as evidências *H1, H2, Bold, Title, TF×IDF* relativo e *Texto de Âncora*, aplicados na fórmula da Equação 4.1 que tem o objetivo de identificar os melhores descritores de uma página *web*.

A Figura 10 demonstra graficamente o cromossomo utilizado na resolução do subproblema. Cada cromossomo será composto por seis genes que representam os pesos atribuídos a cada uma das seis evidências utilizada no processo de identificação de tópicos. Assim, o primeiro, segundo, terceiro, quarto, quinto e o sexto gene representarão os pesos atribuídos às evidências *H1, H2, Bold, Title, TF×IDF* relativo e ao *Texto de Âncora* respectivamente.

H1	***H2***	***Bold***	***Title***	***TF×IDF***	***Texto Âncora***
1	2	3	4	5	6

Figura 10 - Representação do cromossomo utilizado na solução do subproblema de identificação dos pesos das evidências utilizadas no processo de determinação do tópico de uma página *web*.

Fase 2: Definição da Função Objetivo

A função objetivo tem por finalidade avaliar o grau de adequação de cada indivíduo (cromossomo) como solução do problema, sendo associado a cada cromossomo um valor de aptidão gerado por esta função. A função objetivo, aqui utilizada, reflete corretamente o objetivo do subproblema em questão, ou seja, o de encontrar os melhores pesos para as seis evidências utilizadas na definição os termos descritores de uma página, através da maximização de seu valor.

Utilizamos uma função objetivo que se baseia na métrica proposta por Jarvelin e Kekalainen [11]. Eles propõem duas novas métricas para computar o ganho cumulativo dos resultados retornados em uma lista ordenada de respostas: (1) *Cumulated Gain*; e (2) *Discounted Cumulated Gain*, ambas expostas na Seção 3.3.3. A métrica utilizada como função objetivo do subproblema é a *Normalized Discounted Cumulated Gain* (*nDCG*), pois, com a normalização, os valores resultantes estarão compreendidos entre 0 e 1, onde 1 representa o desempenho ideal do sistema avaliado e 0 representa o extremo oposto.

O cálculo do valor de *nDCG* para uma determinada página *web* p_1, pertencente a base de referência, dada uma possível solução s_1 (cromossomo) do problema, é feito da seguinte forma:

1. Do arquivo de descritores pré-avaliados de p_1, obtém-se uma lista com os termos descritores de p_1 e seus respectivos valores de relevância.
2. Em seguida, um vetor ideal V_i, de termos descritores de p_1 é gerado. Em V_i os termos aparecem em ordem decrescente de relevância, ou seja, os termos de maior relevância aparecem em primeiro seguidos dos demais. V_i é uma ordenação decrescente do arquivo de descritores pré-avaliados de p_1 e é ideal, pois retrata um cenário de um sistema ideal onde todos os termos de maior relevância são listados a frente dos termos de menor relevância.

3. O próximo passo é calcular para cada termo *t* pertencente a p_1 o valor de $Imp(t, p_1)$ (conforme Equação 4.1), ou seja, a importância de *t* para p_1. A partir deste cálculo gera-se o vetor R_1 de resposta contendo os termos candidatos a descritores de p_1 ordenado decrescentemente conforme valor de $Imp(t, p_1)$. Os pesos associados a cada uma das evidências utilizadas no cálculo de $Imp(t, p_1)$ são representados pelos genes do cromossomo s_1.

4. A partir de R_1, gera-se um vetor G_1', intitulado por Jarvelin e Kekalainen de Vetor de Ganho. Cada elemento $G_1'[i]$ representa o valor de relevância do termo $R_1[i]$, obtido a partir do arquivo de descritores pré-avaliados de p_1 e que pode variar de 0 (menor grau de relevância) a 3 (maior grau de relevância).

5. Em seguida, os vetores de ganho cumulativo CG_1', de ganho cumulativo ideal CG_{1I}' e de ganho cumulativo normalizado nCG_1'são calculados conforme Equações 3.1, 3.4 e 3.6 respectivamente.

6. O próximo passo é efetuar o cálculo dos vetores de ganho cumulativo com desconto DCG_1' e de ganho cumulativo com desconta ideal DCG_{1I}', conforme Equações 3.2 e 3.5, respectivamente. A ambos é aplicado um fator de desconto log_2 (proposto em [11]). Finalmente o vetor de ganho cumulativo com desconto normalizado $nDCG_1'$ é gerado pela divisão de cada elemento do DCG_1' por cada elemento do DCG_{1I}' (Equação 3.6). A sumarização da eficiência da solução s_1 como resposta do problema é feita a partir do vetor $nDCG_1'$, bastando calcular a média (conforme Equação 3.7) dos elementos das três primeiras posições de $nDCG_1'$. O valor da média informa o quão apto é s_1 para a solução do subproblema de descoberta dos pesos das evidências, quanto mais próximo do valor 1 melhor é a solução. A média calculada se restringe as três primeiras posições do verto $nDCG_1'$, pois serão os três melhores termos descritores, de cada página, que serão utilizados na quarta etapa do processo de identificação de tópico de páginas *web*.

Fase 3: Definição dos Parâmetros Genéticos

Além de planejar os cromossomos e definir a função objetivo, algumas configurações genéticas devem ser definidas objetivando a melhora do desempenho do AG. Na solução do subproblema de definição dos pesos das evidências foram utilizados diversos parâmetros de configuração no algoritmo genético, dando origem à execução de vários experimentos discutidos na Seção 5.1. Abaixo estão expostas as configurações genéticas utilizadas no AG:

- Taxa de *crossover*: 60% e 75% dos cromossomos.
- Taxa de mutação: 2% e 3% dos genes.
- Tamanho da população: foram feitos treinamentos com população de 120, 220, 320, 520, 720 e 1000 indivíduos.
- Número de geração: 10 e 20 gerações.
- Seleção de indivíduos: Método da Roleta (Seção 3.3.2).
- Criação da população: a população inicial foi criada com valores aleatórios.

Fase 4: Processo de Evolução

O algoritmo da Figura 11 descreve o processo evolutivo do AG utilizado. Para cada possível solução do problema (cromossomo) é calculado o valor da média dos três primeiros elementos do vetor *nDCG'* gerado para cada uma das páginas de treino, valor chamado de `MediaLocal`. A soma das médias locais é, em seguida, dividida pelo número de páginas de treino e o valor obtido é associado ao cromossomo como sendo seu valor de aptidão. O cálculo da média dos elementos de *nDCG'* acima de uma determinada posição do vetor é utilizada com o objetivo de sumarizar o desempenho do sistema avaliado.

Ao final do ciclo de execução de treinos e validações do AG, o cromossomo com maior valor de aptidão corresponde a melhor solução do subproblema de identificação dos pesos das evidências e os valores, por ele representado, serão aplicados a fórmula da Equação 4.1.

```
1. Algoritmo EvoluçãoAG
2.   Para cada um dos s[i] cromossomos que compõem a população faça:
3.      Para cada uma das 60 páginas de treino faça:
4.          Aplica-se o peso representado pelo cromossomo s[i] as evidências.
5.          Gera-se o vetor nDCG_i', conforme descrito em 3.3.3.
6.          MediaLocal = a média dos três primeiros elementos de nDCG_i'.
7.          MediaGeral = MediaGeral + MediaLocal
8.      MediaTotal = MediaGeral / 60
9.      Valor de aptidão do cromossomo s[i] = MediaTotal
10. Fim Algoritmo
```

Figura 11 - Algoritmo que representa o processo evolutivo do AG utilizado na solução do subproblema de descoberta de pesos das evidências.

4.1.3 Identificação dos Descritores de uma Página *Web*

O objetivo desta terceira etapa do método proposto é identificar os melhores termos descritores de uma página *web p*. É importante observar que estes termos podem não ocorrer no texto de *p*.

Como dado de entrada a esta etapa são fornecidos os pesos de cada uma das seis evidências utilizadas, a lista de termos candidatos a descritores da página e os valores das seis evidências computados para cada termo candidato. Como dado de saída são identificados os três melhores termos descritores de *p*.

Os valores das seis evidências utilizados para cada termo candidato a descritor de *p*, foram computados na primeira etapa do processo de identificação do tópico e correspondem respectivamente a: (1) freqüência com que o termo aparece destacado pela *tag* de marcação <*h1*>; (2) freqüência com que o termo aparece destacado pela *tag* de marcação <*h2*>; (3) freqüência com que o termo aparece destacado pelas *tags* de marcação <*b*> e <*Strong*>; (4) freqüência com que o termo aparece entre as *tags* <*title*> e <*/title*>; (5) o valor de *TF×IDF relativos* do termos; e (6) a freqüência com que o termo aparece na concatenação dos textos de âncora de *links* que referenciam *p*.

Tendo acesso aos valores das evidências e aos seus respectivos pesos, obtidos na etapa anterior, pode-se então utilizar a fórmula da Equação 4.1. Para cada termo *t* candidato a descritor de *p* é calculado o valor de sua importância no documento: *Imp(t,p)*. Em seguida, é feita uma ordenação desses valores e os três termos com maior valor são definidos como sendo os melhores termos descritores de *p*. Esta etapa é executada para todas as páginas *web* pertencentes na base de referência.

4.1.4 Utilização de um diretório *Web*

O objetivo da última etapa de identificação do tópico de uma página *web p* é, finalmente, determinar o tópico de *p*. É importante salientar que o tópico de *p* será definido por uma lista ordenada que contenha, no máximo, cinco palavras distintas que não pertencem necessariamente ao conteúdo textual de *p*.

Como dado de entrada a esta etapa são fornecidos os três melhores descritores da página *p*. Como dado de saída é definido o tópico de *p*, finalizando o processo de identificação automática do tópico de uma página *web*.

Para determinar o tópico de *p* é utilizada a estrutura hierárquica de um diretório abrangente e popular da *web*. Inicialmente, os três melhores termos descritores da página *p*, obtidos na etapa anterior, são submetidos ao serviço de busca em diretório, em seguida as vinte primeiras categorias associadas às vinte primeiras respostas retornadas pela submissão são obtidas e passam por um processo de otimização para que, por fim, as cinco melhores categorias que descrevam o conteúdo de *p* sejam associadas ao seu tópico.

O serviço de diretório utilizado nos exemplos e experimentos aqui citados foi o da máquina de busca da *Google*. O diretório *Google* consiste em uma coleção de *links* organizados por tópicos em uma estrutura hierárquica de categorias. Seus *links* e sua categorização são oriundos do *Open Directory Project* (ODP) [18], porém a tecnologia de pesquisa é a da própria *Google* [2]. O ODP [18] é um grande diretório público gerenciado pela *Netscape*, é mantido por um grupo de editores voluntários de todo o mundo que avalia manualmente os sites para a inclusão no diretório. As páginas da *web* selecionadas por esses editores são classificadas em diversas categorias que compõem a hierarquia do diretório [2].

Como resposta a uma busca, o diretório *Google* retorna uma lista ordenada de página referente ao assunto buscado e a categoria a qual estas últimas pertencem. Estas categorias são acessíveis através de APIs [10] fornecidas pelo provedor do serviço de busca. As cinco melhores categorias que descrevem o conteúdo da página serão associadas ao seu tópico.

O processo de otimização das listas de categorias visa gerar um *ranking* com os termos presentes nessas listas. Este *ranking* é feito em seis etapas distintas conforme mostra a Figura 12. Todas as seis etapas de otimização da lista de categorias foram

testadas e seus resultados foram analisados utilizando páginas *web* pertencentes a uma base de referência diferente da base utilizada para testar e avaliar o método de identificação de tópicos de páginas, assim podemos afirmar que as etapas de otimização das categorias são aplicáveis a qualquer conjunto de páginas *web* obtendo resultados válidos.

Na primeira etapa de otimização da lista é feita a eliminação da estrutura hierárquica das categorias, isto é, a relação pai/filho utilizada pela taxonomia do diretório deixa de ser relevante e os termos que antes compunham a categoria passam a ser tratados de forma individualizada, formando uma lista *L* de termos candidatos a tópicos da página. O descarte da hierarquia se justifica, pois, conforme descrito em [12], a relação pai/filho utilizada pela taxonomia dos diretórios disponíveis pode ser de dois tipos: (1) "é um"; ou (2) "é parte de", o que impede a afirmação de que o tópico da página corresponde a cada uma das subcategorias presentes na hierarquia de categorias que melhor descreve o conteúdo da página. Como exemplo, suponha que a seguinte hierarquia de categorias foi identificada como sendo a hierarquia que melhor descreve o termo "*John Paul II*" (Papa):`/top/Society/Religion_and_Spirituality/Christianity/Denominations/Catholicism/Popes/J/John_Paul_II`, pode-se afirmar que uma página que discorre sobre "*John Paul II*" está relacionada ao tópico *Christianity*, porém não faz sentido afirmar que esta mesma página esteja relacionada ao tópico *Top* ou *J*.

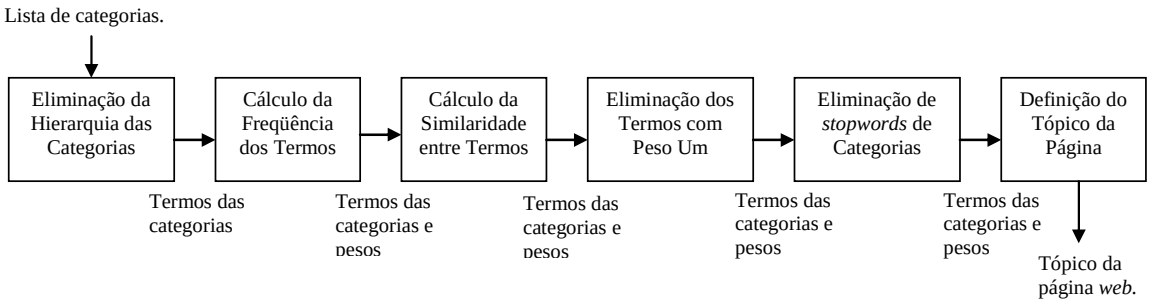

Figura 12 - Etapas do processo de otimização da lista de categorias, objetivando a definição do tópico de uma página *web*.

Na segunda etapa de otimização, é computada a freqüência com que cada termo *t* ocorre na lista *L*, essa freqüência é atribuída a *t* com sendo o seu peso inicial. Nesta etapa os termos redundantes também são eliminados da lista *L*. O cálculo da freqüência

dos termos se justifica, pois verificamos que alguns dos termos mais comuns de *L* estão associados ao tópico da página.

A terceira etapa consiste em computar a similaridade entre os termos presentes em *L* e os três termos descritores da página que foram utilizados na submissão ao sistema de busca em diretório. Os termos presentes em *L*, cuja grafia coincida com ou contenha termos descritores da página, possuem seus pesos multiplicados por 100. O valor 100 é fator multiplicador utilizado para aumentar o peso do termo candidato a tópico. Também verificamos que a coincidência na grafia nos leva a melhores candidatos.

Na próxima etapa de otimização, são eliminados os termos de *L* que possuem peso igual a 1, o que acarreta na diminuição da quantidade de termos candidatos a tópicos de *p*. A eliminação destes termos é feita, pois a maioria dos termos que possuem peso 1, após a execução da terceira etapa de otimização, são termos considerados irrelevantes.

Na quinta etapa é feita a eliminação dos termos presentes na lista de *stopwords* de categorias. Termos considerados *stopwords* de categorias são aqueles de sentido amplo utilizados apenas para compor a hierarquia das categorias do diretório, que isolados não possuem sentido completo e sob as quais estão muitas outras subcategorias específicas, por isso não são usados na definição do tópico da página, como por exemplo: *top, issue, reference, by_region*. A lista de *stopwords* de categorias foi feita analisando a estrutura hierárquica do diretório *Google*.

Na última etapa de otimização, os cinco termos com maior peso são obtidos e definidos como sendo o tópico da página *p*.

Como exemplo do processo de otimização da lista de categorias, suponha uma página *web* p_1 que discorra sobre o assunto *Stem Cells*, suponha ainda que os três melhores termos que descrevem a página são: *Stem Cells*, *Stem* e *Cells*. A lista de categorias formada pelas 20 primeiras respostas, ao submetermos os termos que descrevem p_1 ao sistema de busca em diretório, é mostrada pela Figura 13.

Como resultado da primeira etapa de otimização desta lista de categorias, no qual a hierarquia entre as categorias é desprezada, obtêm-se os seguintes termos: `top, biotechnology, science, biology, stem_cells, institutions, issues, science_and_technology, society, associations, journals, stem_cell_research , directories, products_and_services, research_centers, health, medicine, medical_specialties, hematology, blood_and_bone_marrow_transplantation, cell_biology, publications e living_systems.`

```
1.  /top/society/issues/science_and_technology/biotechnology/stem_cell_research
2.  /top/science/biology/biotechnology/stem_cells/associations
3.  /top/science/biology/biotechnology/stem_cells
4.  /top/science/biology/biotechnology/stem_cells/journals
5.  /top/science/biology/biotechnology/stem_cells/institutions
6.  /top/science/biology/biotechnology/stem_cells
7.  /top/science/biology/institutions/research_centers
8.  /top/science/biology/biotechnology/stem_cells/associations
9.  /top/science/biology/biotechnology/stem_cells/institutions
10. /top/society/issues/science_and_technology/biotechnology/stem_cell_research
11. /top/society/issues/science_and_technology/biotechnology/stem_cell_research
12. /top/science/biology/biotechnology/stem_cells/directories
13. /top/science/biology/biotechnology/stem_cells/journals
14. /top/science/biology/biotechnology/stem_cells/institutions
15. /top/science/biology/biotechnology/stem_cells/products_and_services
16. /top/science/biology/biotechnology/stem_cells/associations
17. /top/science/biology/biotechnology/stem_cells/directories
18. /top/health/medicine/medical_specialties/hematology/blood_and_bone_marrow_transpl
    antation
19. /top/science/biology/cell_biology/publications/journals
20. /top/science/biology/products_and_services/living_systems
```

Figura 13 - Lista das 20 primeiras categorias associadas às 20 primeiras respostas obtidas a partir da submissão dos termos *Stem Cells, Stem* e *Cells* ao serviço de busca em diretório.

Na segunda etapa, é associado a cada termo *t* acima citado um peso que corresponde à freqüência com que *t* aparece na lista de categorias, além disso, os termos duplicados são eliminados de *L*, sendo assim temos como resultado: *top 20, biotechnology 16, science 16, biology 16, stem_cells 13, institutions 4, issues 3, science_and_technology 3, society 3, associations 3, journals 3, stem_cell_research 3, directories 2, products_and_services 2, research_centers 1, health 1, medicine 1, medical_specialties 1, hematology 1, blood_and_bone_marrow_transplantation 1, cell_biology 1, publications 1 e living_systems 1.*

Na terceira etapa os termos candidatos a tópicos de p_1 que contenham ou que sejam iguais aos termos descritores de p_1 são enfatizados, eles possuem seu peso multiplicado por 100. Logo, como resultado temos: *stem_cells 1300000000, stem_cell_research 30000, living_systems 10000, top 20, biotechnology 16, science 16, biology 16, institutions 4, issues 3, science_and_technology 3, society 3, associations 3, journals 3, directories 2, products_and_services 2, research_centers 1, health 1, medicine 1, medical_specialties 1, hematology 1, blood_and_bone_marrow_transplantation 1 e cell_biology 1, publications 1.*

A próxima etapa consiste em eliminar os termos com valor de peso igual a 1, obtemos o seguinte resultado*: stem_cells 1300000000, stem_cell_research 30000, living_systems 10000, top 20, biotechnology 16, science 16, biology 16, institutions 4, issues 3, science_and_technology 3, society 3, associations 3, journals 3, directories 2 e products_and_services 2.*

Na quinta etapa é feita a eliminação das *stopwords* de categorias, resultando a seguinte lista de termos*: stem_cells 1300000000, stem_cell_research 30000,*

living_systems 10000, biotechnology 16, science 16, biology 16, institutions 4, science_and_technology 3, associations 3, journals 3 e products_and_services 2.

Na sexta e última etapa, obtemos os cinco termos de maior valor e determinamos o tópico da página p_1. Sendo assim o tópico de p_1 está relacionado aos seguintes termos: *stem cells, stem cell research, living systems, biotechnology e science.*

Capítulo 5

Experimentos e Discussão dos Resultados

Neste capítulo são expostos os experimentos realizados com o método de identificação de tópico de páginas *web* proposto. São discutidos os objetivos de cada experimento, as configurações utilizadas, as suas formas de execução, os resultados obtidos e um exemplo de aplicação prática de classificação de páginas em um diretório da *web*.

5.1 Experimento 1 – Obtenção dos Pesos das Evidências

Este experimento foi realizado com o objetivo de determinar, automaticamente, os melhores pesos das seis fontes de evidências utilizadas no processo de identificação de tópico de páginas *web*. Os pesos, aqui definidos, foram aplicados na Equação 4.1 que quantifica a importância de um termo *t* para uma página *p*. Esta equação é fundamental na escolha dos três melhores termos descritores de *p*, que são utilizados na descoberta de seu tópico, por isso as suposições humanas na definição dos pesos não poderiam ser cogitadas.

A Seção 4.1.2, enumera os vários parâmetros genéticos utilizados durante a execução do AG que definirá os pesos das evidências. O uso desta variedade de configurações tem como objetivo obter o melhor resultado para a solução do problema, ou seja, maximizar o valor da função objetivo utilizada.

A função objetivo utilizada para avaliar o grau de aptidão de cada possível solução (cromossomo) foi a *Normalized Discounted Cumulated Gain (nDCG)* explicada na Seção 4.1.2.

As páginas *web* utilizadas neste experimento pertencem à base de referência criada (Seção 4.1). Objetivando a exploração exaustiva desta coleção, optou-se por

utilizar a técnica de validação cruzada (Seção 3.3.4) nos experimentos realizados, logo a base de referência foi dividida em cinco subconjuntos, cada um contendo 20 páginas. Do total de páginas que compunham a base 60% foram utilizadas na fase de treino do AG, 20% foram utilizadas para a fase de validação dos experimentos e os outros 20% restantes foram utilizados para a fase de teste (discutido na Seção 5.2).

Tamanho da População	Taxa de *Crossover*	Taxa de Mutação	Número de Gerações	Valor Função Objetivo
520	***60%***	***2%***	***10***	***0.7801***
720	60%	3%	10	0.7466
720	60%	2%	10	0.7387
320	75%	3%	20	0.7379
320	60%	2%	10	0.7366
720	75%	3%	20	0.7366
1000	60%	3%	10	0.7315
220	75%	3%	10	0.7200
320	75%	3%	10	0.7192
120	60%	2%	10	0.7161

Tabela 1 – Configuração genética dos 10 melhores resultados obtidos com a execução do AG, cujo objetivo era maximizar o valor da função objetivo.

Utilizando todas as combinações geradas pelas configurações genéticas citadas na Seção 4.1.2, foram executados 48 experimentos distintos com a finalidade de identificar os melhores pesos das evidências utilizadas na identificação de tópicos. A melhor solução do problema corresponde àquela de maior valor gerado pela função objetivo. O processo evolutivo do AG está descrito na Seção 4.1.2. A Tabela 1 apresenta as configurações genéticas e os valores gerados pela da função objetivo dos 10 melhores resultados obtidos com a execução do AG, as demais soluções obtiveram valores inferior a 0.7161. A solução de maior aptidão foi obtida com as seguintes configurações: a taxa de *crossover* foi de 60% dos cromossomos, a taxa de mutação foi de 2% dos genes, tamanho da população foi de 520 cromossomos e foram utilizadas 10 gerações (Tabela 1). Esta configuração obteve o valor de aptidão da solução igual a 0.7801, valor máximo encontrado, e os pesos das evidências correspondentes a esta solução são os seguintes:

1. Peso da evidência *H1*: 0.033
2. Peso da evidência *H2*: 0.212
3. Peso da evidência *Bold*: 0.842
4. Peso da evidência *Title*: 0.820
5. Peso da evidência *TF×IDF* relativo: 0.744
6. Peso da evidência *Texto de Âncora*: 0.880

Logo a Equação 4.1 pode ser redefinida da seguinte forma:

$$\mathrm{Im}p(t,D) = 0.033 * eH1(t) + 0.212 * eH2(t) + 0.842 * eBold(t) + 0.820 * eTitle(t) + 0.744 * eTfIdf(t) + 0.880 * eTextAncora(t) \quad (5.1)$$

onde *Imp(t,D)* representa a importância do termo *t* para definição do tópico de um documento *D*, *eH1* é a freqüência com que o termo *t* ocorre entre as *tags* <*h1*> e </*h1*> em *D*, *eH2* corresponde a freqüência com que o termo *t* ocorre destacado pela *tag* <*h2*> em *D*, *eBold* corresponde a freqüência com que *t* ocorre destacado pelas *tags* <*Bold*> e <*Strong*> em *D*, *eTitle* corresponde a freqüência com *t* ocorre entre as *tags* <*Title*> e </*Title*> em *D*, *eTfIDf* corresponde ao valor de *TF×IDF* relativo de *t* e *eTextAncora* corresponde a freqüência com que *t* ocorre na concatenação dos conteúdos de *link* das páginas que apontam para *D*. Esses valores são obtidos através do processamento do conteúdo de cada página pertencente a base de referência.

Pelos resultados obtidos concluímos que a evidência de maior peso é a *Texto de Âncora*, isto significa que as informações obtidas a partir desta fonte são as que melhor descrevem o conteúdo de uma página *p*. Esta conclusão é justificada, pois muitas vezes o conteúdo dos apontadores que referenciam *p* são descrições simples e objetivas de *p*. Para exemplificar, selecionamos uma página p_1 pertencente à base de referência e que discorre sobre o cientista da computação *Donald Knuth* (Figura 21). Em seguida obtivemos o conteúdo da evidência *Texto de Âncora* que contém a concatenação do texto de 30 apontadores que referenciam p_1, conforme mostra a Figura 14.

Pela Figura 14, observamos que 26.66% do conteúdo da evidência *Texto de Âncora* de p_1 é composto pelo termo "`donald knuth`", 23.33% é composto pelo termo "`donald e. knuth`", 13.33% é composto pelo termo "`don knuth`", 6.66% é composto pelo termo "`professor donald knuth`" e pelo termo "`donald knuth's web site`", 10.00% é

composto pelo termo “knuth”, 3.33% é composto pelo termo “donald knuth homepage”, outros 3.33% é composto pelo termo “d. e. knuth”, e os últimos 3.33% é composto pelo termo “amazing person donald e. knuth”, . Todos eles são bons termos descritores do conteúdo de p_1.

1. donald knuth
2. donald knuth
3. donald knuth
4. professor donald knuth
5. donald knuth's web site
6. donald knuth
7. knuth
8. donald knuth's web site
9. knuth
10. knuth
11. donald e. knuth, stanford university
12. d.e. knuth
13. amazing person donald e. knuth
14. donald e. knuth
15. donald e. knuth
16. donald knuth
17. donald knuth
18. donald e. knuth
19. donald knuth
20. donald e. knuth
21. donald e. knuth
22. donald e. knuth
23. donald knuth homepage
24. donald e. knuth
25. don knuth
26. don knuth
27. donald knuth
28. don knuth
29. don knuth
30. professor donald knuth

Figura 14 - Texto da evidência *Texto de Âncora* da página *web* que discorre sobre *Donald Knuth*.

Abaixo podemos verificar os fragmentos de texto, obtido de p_1, utilizado para computar conteúdo das evidências *Bold*, *H1*, *H2* e *Title* da página:

- Fragmento de texto para computar a evidência *Bold* : donald e. knuth
- Fragmento de texto para computar a evidência *H1*: frequently asked questions infrequently asked questions recent news computer musings known errors in my books important message to all users of tex help wanted diamond signs preprints of recent papers curriculum vitae; pipe organ downloadable graphics downloadable programs expecting a check from me? did you borrow a video from me?.
- Fragmento de texto para computar a evidência *H2*: stanford computer science home page.
- Fragmento de texto para computar a evidência *Title*: don knuth's home page.

Note que as evidências *Bold* e *Title* possuem pesos semelhantes ao peso atribuído à evidência *Texto de Âncora*, justificado também pela importância dos termos presentes nestas evidências para identificação do tópico de uma página. O exemplo da página p_1, que discorre sobre *Donald Knuth*, endossa o resultado obtido.

Os baixos pesos das evidências *H1* e *H2* são explicados devido à baixa freqüência das *tags* <*h1*> e <*h2*> entre as páginas que compõem a base de referência, 30% e 25% respectivamente, se compararmos com a porcentagem de páginas que contém as *tags* <*b*> ou <*strong*>: 75%. As evidências *Bold*, *Title* e *TF×IDF* relativo foram classificadas, pelo AG, como a segunda, a terceira e a quarta evidências mais importantes, respectivamente.

A definição correta dos pesos das evidências utilizadas é de grande importância para o bom funcionamento do método de identificação de tópico de páginas *web*, pois eles serão utilizados na definição dos melhores termos descritores de uma página *p*, que servirão como semente na identificação do tópico de *p*.

5.2 Experimento 2 - Avaliação do Método

O segundo experimento executado teve a finalidade de testar a eficiência do sistema de identificação de tópicos de páginas *web*, cujo método foi descrito no Capítulo 4.

Durante a execução deste experimento foram utilizadas 20 páginas da *web*, obtidas sob certos critérios. Páginas de *blogs*, jornais e portais não foram submetidas ao método, devido à dificuldade encontrada, até mesmo por seres humanos, na identificação do tópico principal de cada uma delas. Neste experimento foram utilizadas páginas com as seguintes características: (1) discorrem acerca de um determinado assunto, (2) possuem conteúdo html; e (3) são escritas em inglês. As Figuras 15 e 16 ilustram as características das páginas utilizadas.

A obtenção das 20 páginas que compõem a base de teste, utilizada para avaliar a eficiência do método, procedeu-se da seguinte forma: (1) foram selecionados manualmente dezoito termos, aleatórios, distintos e em inglês, que caracterizassem o tópico de vinte páginas diferentes; (2) cada termo foi submetido ao serviço da máquina de busca *Google*; (3) para cada termo, foram avaliadas, manualmente, as dez primeiras respostas da máquina de busca; (4) a página que melhor descrevesse o assunto buscado foi então selecionada para compor a base de teste. A Figura 17 mostra a lista de termos utilizada para obter as páginas que compõem a base de teste. É importante destacar que as páginas da base de teste pertencem a diferentes *sites*, assim o método também é

avaliado quanto a sua eficácia quando páginas que apresentam diferentes características estruturais são submetidas a ele.

Figura 15 – Exemplo de página submetida ao método de identificação de tópicos de páginas *web*. Esta página discorre sobre *Barack Obama*.

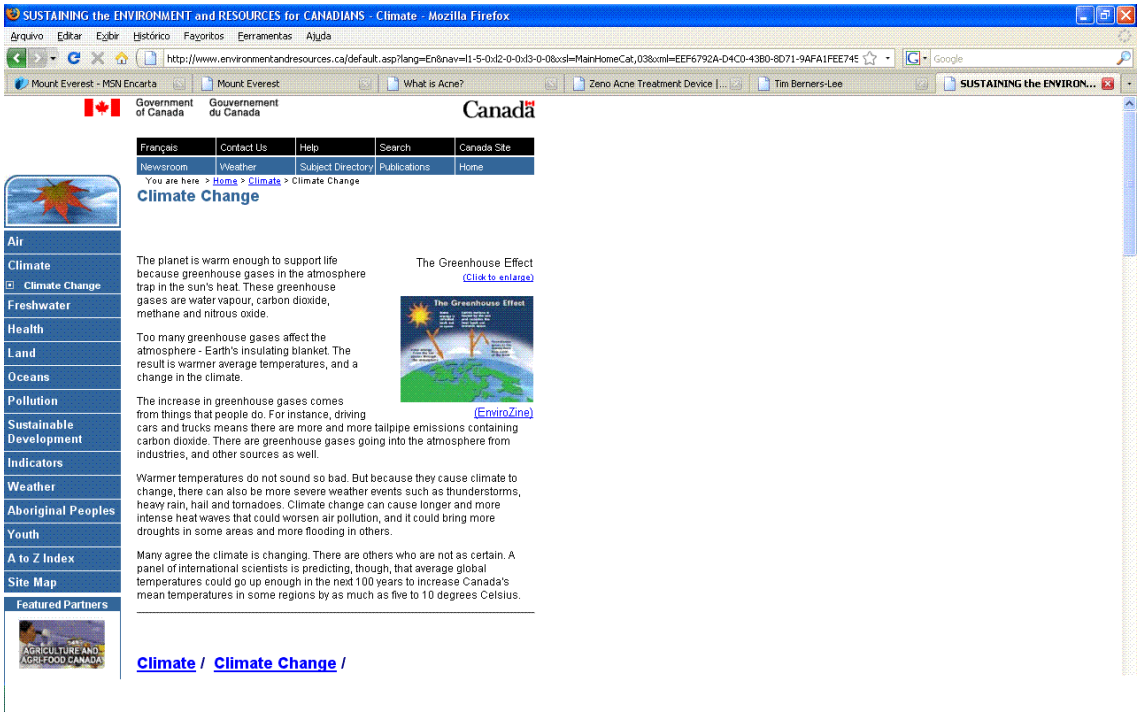

Figura 16 – Exemplo de página submetida ao método de identificação de tópicos de páginas *web*. Esta página discorre sobre *Climate Change*.

Dois termos, utilizados na escolha das páginas de teste foram repetidos, são eles: *Acne* e *Mount Everest* (Figura 17). A intenção, na repetição dos termos citados, é verificar o desempenho do sistema quando duas páginas que possuem estruturas distintas, porém discorrem sobre o mesmo assunto, são submetidas a ele. A Tabela 2 mostra o tópico, gerado pelo sistema, associado a cada termo utilizado na obtenção das páginas que compõem a base de teste. Para o termo *Mount Everest* os dois tópicos gerados foram iguais, porém para o termo *Acne* os tópicos gerados foram diferentes. Analisando o conteúdo das duas páginas que discorrem sobre *Acne*, verificamos que ambas possuem objetivos diferentes: (1) a página representada pelo termo *Acne* de número 10 (Figura 17) discorre acerca do assunto através da definição do problema dermatológico, identificação os tipos de acnes existentes e identificação do grupo de pessoas mais acometidas pela doença (Figura 18); (2) a página representada pelo termo *Acne* de número 11 (Figura 17) discorre acerca do assunto conceituando o problema dermatológico, destacando a sua forma de tratamento e propondo a compra de um produto, desenvolvido pela empresa dona do *site*, utilizado no tratamento da doença (Figura 19). Por isso, a diferença existente no tópico das páginas é justificada. Já as páginas representadas pelos termos *Mount Everest*, de número 15 e 20 da Figura 17, discorrem sobre as características deste, como sua a localização, altura e origem de seu nome.

```
1. Bill Clinton          11.   Acne
2. Bermuda Triangle      12.   Washington, dc
3. Diabetes              13.   Barack Obama
4. Types of Diabetes     14.   Sahara
5. Greenpeace            15.   Mount Everest
6. Mysql                 16.   Egypt
7. Insulin               17.   Hillary Clinton
8. European Union        18.   Texas
9. Coffee                19.   Panama Canal
10.  Acne                20.   Mount Everest
```

Figura 17 - Lista de termos utilizados para obtenção das páginas *web* que compõem a base de teste.

O próximo passo foi a submissão das páginas de teste ao método exposto e, como resultado, obteve-se o tópico destas. A Tabela 2 mostra cada termo utilizado para obtenção das páginas *web* que compõem a base de teste e o tópico resultante da submissão destas ao sistema de identificação de tópicos.

Figura 18 – Exemplo de página *web* utilizada nos experimentos do método proposto e que discorre sobre *Acne* enfatizando a definição do problema, os tipos de acnes existentes e o grupo de pessoas mais acometidas pela doença.

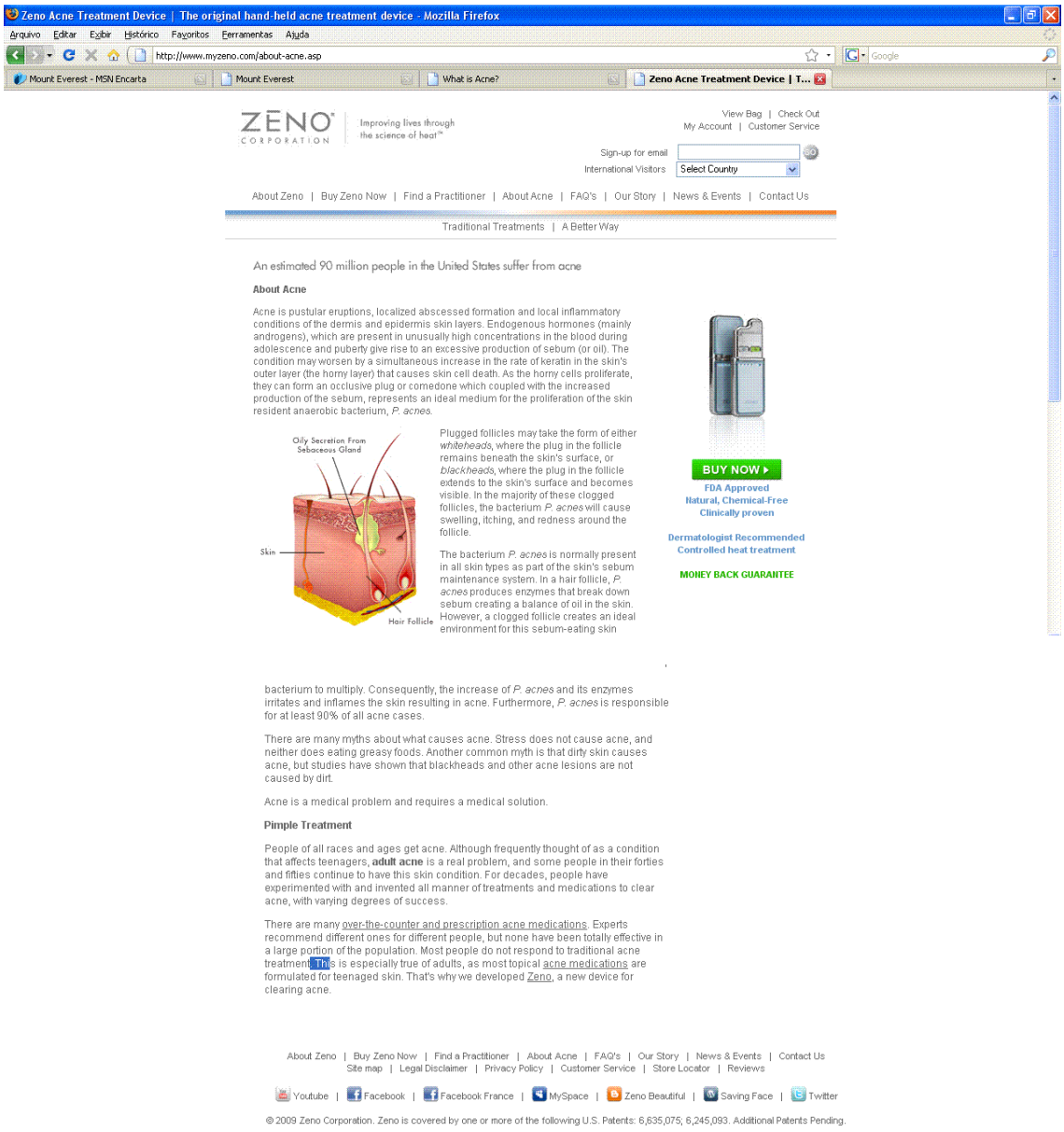

Figura 19 - Exemplo de página *web* utilizada nos experimentos do método proposto e que discorre sobre *Acne* enfatizando a definição do problema, a sua forma de tratamento e propondo a compra de um produto, desenvolvido pela empresa dona do *site*, utilizado no tratamento da doença.

Termo	Tópico	Termo	Tópico
1.Bill Clinton	Clinton, William jefferson Clinton foundation North america United states Presidents	11.Acne	Acne Health Shopping Conditions and diseases Skin disorders

2.Bermuda Triangle	Bermuda triangle Paranormal	12.Washington, dc	Washington, dc United states North america Government Embassies and consulates
3.Diabetes	Diabetes Diabetes mellitus Gestational diabetes Diabetic retinopathy health	13.Barack Obama	Obama, barack Obama, barrack [d-il] North america United states Candidates and campaigns
4.Types of Diabetes	Diabetes Gestational diabetes Health Conditions and diseases Endocrine disorders	14.Sahara	Western sahara Saharan dust Sahara Deserts África
5.Greenpeace	Greenpeace Environment Society and culture Oceania politics	15.Mount Everest	Mount everest Mountaineering Everest Recreation Climbing
6.Mysql	Mysql Computers Languages Software Programming	16.Egypt	Science Archaeology Social sciences Egypt Pyramids
7.Insulin	Insulin Insulin resistance Health Conditions and diseases Endocrine disorders	17.Hillary Clinton	Clinton, hillary rodham Clinton, hillary United states North america Candidates and campaigns
8.European Union	European union European commission European studies Monetary union - euro Europe	18.Texas	Texas United States North america Localities Business and economy
9.Coffee	coffee coffee and tea	19.Panama Canal	Panama canal Panama

	coffee makers Food Drink		Canals Panama city Central America
10.Acne	Dermatology Acne Health Medical specialties Medicine	20.Mount Everest	Mount everest Mountaineering Everest Recreation Climbing

Tabela 2 - Relação de termos utilizados para obter as páginas teste e seus respectivos tópicos, gerados pelo sistema de RI aqui proposto.

Para avaliar a eficiência do método foram utilizados os dados oriundos do julgamento de dez especialistas que analisaram os resultados gerados pelo sistema de identificação de tópicos de páginas *web*. Cada especialista analisou o resultado do sistema aplicado a dez páginas *web* distintas, todas provindas da base de teste. A distribuição das páginas de testes e seus respectivos tópicos entre os especialistas, para fins de avaliação, foi feita de tal forma que o tópico de uma determinada página fosse analisado cinco vezes, logo cada página teve seu tópico analisado por cinco especialistas. Cada especialista recebeu o endereço eletrônico das dez páginas, para que os mesmos pudessem ler e compreender seus conteúdos, e, associado a cada endereço, uma lista de cinco termos, originada pela execução do sistema, e que representa o tópico da página *web.* Os especialistas deveriam destacar os termos pertencentes ao tópico da página que fossem julgados como relevante para descrever o assunto da mesma.

A sumarização dos dados, provindos da análise feita pelos especialistas, gerou para cada termo t_i ($1 \leq i \leq 5$) que compõe o tópico de uma determinada página p um valor de relevância que indica o quão importante t_i é para a definição do assunto de p. O valor de relevância gerado para o termo varia de 0 a 5, onde o valor 0 indica que nenhum especialista julgou o termo como relevante para definir o assunto da página e 5 indica que todos os especialistas julgaram o termo como relevante para definir o assunto da página.

Portanto, para cada uma das 20 páginas de teste foram obtidos: (1) o tópico da página p, representado por uma lista ordenada de termos, gerado pelo sistema de identificação de tópicos; e (2) os valores de relevância, julgado pelos especialistas, para

cada termo t_i ($1 \leq i \leq 5$) que compõe o tópico de p. Estas informações foram aplicadas a métrica *Normalized Discounted Cumulated Gain* (*nDCG*) (Seção 4.1.2) utilizada na avaliação do desempenho do método proposto.

Objetivando avaliar a qualidade do tópico da página p_1, gerado pelo sistema aqui proposto, transformamos a lista L_1' de termos que compõem o tópico de p_1 em uma lista G_1' de ganho de valores. Cada elemento $G_1[i]$' de G_1' representa o valor de relevância do termo $L_1[i]$', gerado a partir dos julgamentos dos especialistas. O valor de relevância atribuído a cada termo pode variar de 0 a 5.

Por exemplo, em uma página que discorre sobre *Bill Clinton*:

$\boldsymbol{L_1'}$ = (`Clinton, William Jefferson;`
`Clinton foundation;`
`North America;`
`United states;`
`Presidents`)

$\boldsymbol{G_1'}$ = (`5;1;5;5;5`)

A partir de G_1', é gerado um vetor ideal I_1', onde os termos com maior relevância são listados a frente dos termos de menor relevância. Logo:

$\boldsymbol{L_{1I}'}$ = (`Clinton, William Jefferson;`
`Presidents;`
`North America;`
`United states;`
`Clinton foundation`)

$\boldsymbol{I_1'}$ = (`5;5;5;5;1`)

O trecho de código da Figura 20 demonstra uma função que contém os próximos passos a serem executados na quantificação da qualidade do tópico L_1' para a página p_1. Recebendo como dado de entrada a lista de ganho de valores G_1' e o vetor ideal I_1', gerados a partir de L_1' e dos julgamentos feitos pelos especialistas, as linhas 1,2 e 3 da função demonstram que devem ser calculados os vetores de ganho cumulativo CG_1', de ganho cumulativo ideal CG_{1I}' e de ganho cumulativo normalizado nCG_1'. Estes vetores são obtidos conforme as Equações 3.1, 3.4 e 3.6, respectivamente. Nas linhas 5 e 6 da função são efetuados os cálculos dos vetores de ganho cumulativo com desconto DCG_1'

e de ganho cumulativo com desconto ideal DCG_{1I}', conforme as Equações 3.2 e 3.5, respectivamente. O fator de desconto utilizado na determinação dos valores dos elementos de DCG_1' e DCG_{1I}' foi log_2. Na linha 7, finalmente, o vetor de ganho cumulativo com desconto normalizado $nDCG_1$' é gerado, conforme descrito na Equação 3.6. Na linha 8, a sumarização da eficiência da qualidade do tópico L_1' para a página p_1 é feita a partir do vetor $nDCG_1$', bastando calcular a média (Equação 3.7) dos 5 elementos de $nDCG_1$', chamaremos aqui de `mediaLocal`. O valor da `mediaLocal` informa o quão bom foi a resposta do sistema (L_1'), quanto mais próximo do valor 1 melhor é a solução.

```
1.   Função QualidadeTópico(G_i', I_i')
2.      Gera-se o vetor CG_i'.
3.      Gera-se o vetor CG_1I'.
4.      Gera-se o vetor nCG_1'.
5.      Gera-se o vetor DCG_1'.
6.      Gera-se o vetor DCG_1I'.
7.      Gera-se o vetor nDCG_1I'.
8.      MediaLocal = a média dos cinco primeiros elementos de nDCG_iI'.
9.      Retorne (MediaLocal)
10.  Fim Função
```

Figura 20 – Procedimento utilizado na definição da qualidade de um tópico para uma determinada página *web* p_1 .

A função da Figura 20 é executada para cada uma das 20 páginas de teste. Em seguida, uma média geral (chamada de `mediaGlobal`) é obtida dividindo-se a soma de todos os valores de retorno da função (`mediaLocal`) por 20, então quantificamos o desempenho do sistemas. Com base na análise dos especialistas o valor obtido e que qualifica o sistema proposto, quando analisados os cinco elemento do vetor *nDCG*, foi ***0.9129***. Este valor significa uma avaliação bastante satisfatória do sistema, já que o valor máximo a ser atingido é 1.0, equivalente ao sistema ideal. Logo, concluímos que a escolha das diversas fontes de evidências utilizadas e a combinação, a elas aplicadas, foram essenciais para o bom resultado obtido nos experimentos do processo de identificação de tópicos aqui proposto.

A Tabela 3 demonstra a comparação entre os resultados obtidos na identificação dos tópicos de seis páginas *web* quando estas foram submetidas ao método proposto por Rafiei e Mendelzon, em [20], e ao método de identificação automática de tópicos proposto nesta dissertação. A coluna ***Endereço da Página Web*** (Tabela 3) identifica o endereço *web* das seis páginas utilizadas nos experimentos feitos por Rafiei e

Mendelzon em [20]. A coluna ***Descrição da Página*** contém uma descrição simplificada do conteúdo de cada uma das seis páginas analisadas. A coluna ***Técnica de Rafiei e Mendelzon*** identifica os termos descritores, identificados em [20], utilizados para definir o tópico das seis páginas analisadas. A coluna ***Tópico identificado pelo nosso método*** identifica os cinco termos descritores utilizados para definir o tópico das seis páginas, quando estas foram submetidas ao método proposto nesta dissertação.

Os resultados presentes na Tabela 3 são meramente ilustrativos, nos impedindo de determinar qual é o método mais adequado para identificação de tópicos de páginas *web*, pois: (1) a amostra e a variedade de páginas comparadas foram pequenas; e (2) devido a possíveis mudanças acorridas no conteúdo textual destas páginas.

Analisando os resultados obtidos para a página http://java.sun.com (Tabela 3) verificamos que o método proposto por Rafiei e Mendelzon identifica o termo `Microsoft` como sendo um bom descritor do assunto desta, no entanto sabemos que esta afirmação não é válida. Porém, os cinco termos identificados pelo nosso método são bons identificadores do assunto desta página.

Na análise da página http://www-cs-faculty.stanford.edu/~knuth (Figura 21) observamos que dos seis termos identificados por Rafiei e Mendelzon como sendo bons descritores desta (`Don Knuth; Donald E. Knuth; Tex; Dilbert Zone; Látex; e ACM`), apenas os três primeiros estão realmente relacionados ao seu assunto. Porém, o nosso método identifica cinco termos que caracterizam o assunto da página pessoal de Donald Knuth: `Knuth, Donald; Computers; Pioneers; History; e Software.`

A interseção entre os termos que compõem os tópicos das páginas http://www.cafeaulait.org/javafaq.html e http://www.w3.org/People/Berners-Lee/ nos deixa afirmar que o desempenho do nosso sistema foi satisfatório, ou seja, o método proposto atingiu o seu objetivo: determinar, através de um conjunto de cinco termos, o assunto principal destas páginas.

A maior diferença existente entre os termos que compõem os tópicos avaliados se concentra nas páginas http://www.eff.org/ e http://www.cdt.org/, acreditamos que uma provável mudança do conteúdo textual das mesmas tenha influenciado no resultado, já que ambos os métodos utilizam o conteúdo textual das páginas como fonte de evidência. Porém, baseado nas propostas dos dois *sites* (privar pelos direitos dos usuários da *web*) concluímos que nosso método gerou resultados que definem o conteúdo destas páginas.

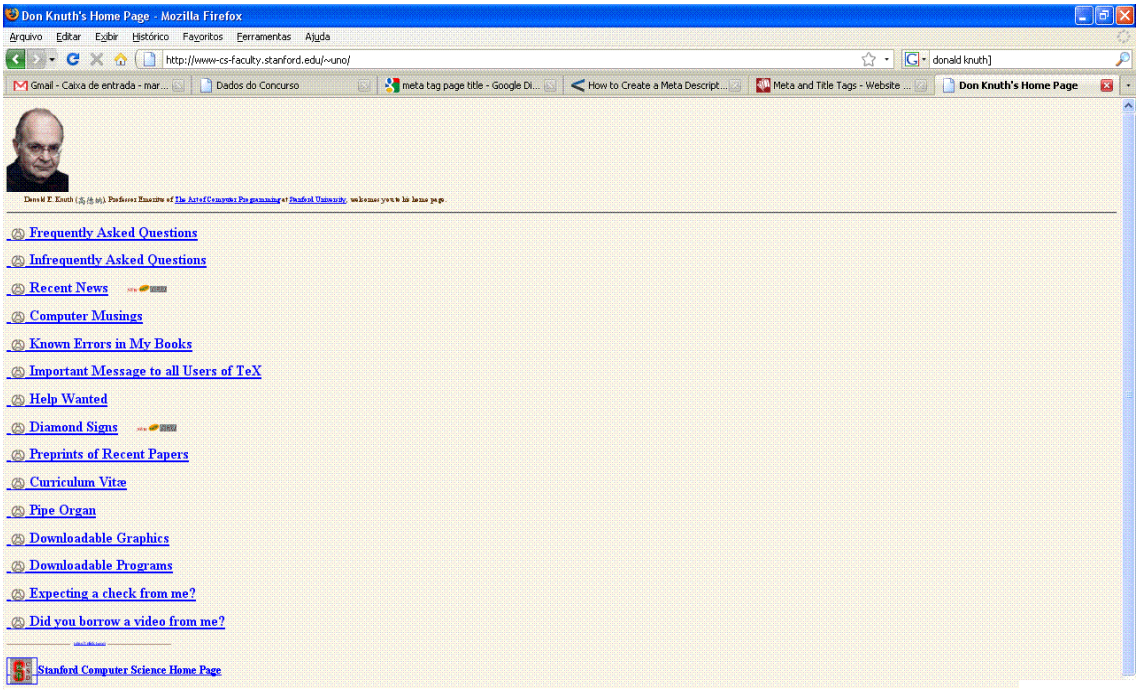

Figura 21 - Página pessoal de Donald Knuth. http://www-cs-faculty.stanford.edu/~knuth

Endereço da Página *Web*	Descrição da Página	Técnica de Rafiei e Mendelzon	Tópico identificado pelo nosso método
http://java.sun.com	Página inicial da java.sun utilizada como guia para desenvolvedores java.	1.Java 2.Software 3.Computer 4.Programming 5.Sun Development 6.Microsoft 7.Search	1.Java 2.Computers 3.Programming 4.Languages 5.Software
http://www.cafeaulait.org/javafaq.html	Página onde as perguntas freqüentes sobre java são respondidas.	1.Java 2.Programming 3.FAQ 4.Sun 5.Computer 6.Language 7.Tutorial 8.Java FAQ 9.Software	1.Java 2.Faqs Help and Tutorials 3.Faqs 4.Individual Group Faqs 5.computers
http://www.eff.org/	Página que discorre sobre a EFF que é uma organização defensora da liberdade civil dos direitos da população no mundo digital. Quando a individualidade destas pessoas, na rede mundial, está sob ataque, EFF se propõe a ser a primeira a defendê-las.	1.Anti-Censorship 2.Join the Blue Ribbon 3.Blue Ribbon Campaign 4.Eletronic Frontier 5.Fundation 6.Free Speech	1.Computers 2.Internet 3.Organizations 4.Intellectual Property 5.History

http://www.cdt.org/	Página que discorre sobre a CDT que é uma organização que promove os valores democráticos na era digital. Ela se utiliza de soluções práticas para melhorar a liberdade de expressão e a privacidade das pessoas que utilizam variadas formas de tecnologia que promovem a comunicação mundial.	1.Center for Democracy and Technology 2.Communications 3.Decency Act 4.Censorship 5.Free Speech 6.Blue Ribbon 7.Syllabus 8.Encryption	1.Center for Democracy 2.Computers 3.Privacy 4.Human rights and liberties 5.Organizations
http://www.w3.org/People/Berners-Lee/	Página pessoal de Tim Berners-Lee.	1.History of the Internet 2.Tim Berners-Lee 3.Internet History	1.Berners lee, tim 2.Internet 3.Computers 4.People 5.History
http://www-cs-faculty.stanford.edu/~knuth	Página pessoal de Donald Knuth.	1.Don Knuth 2.Donald E. Knuth 3.Tex 4.Dilbert Zone 5.Latex 6.ACM	1.Knuth, Donald 2.Computers 3.Pioneers 4.History 5.Software

Tabela 3 – Tópicos obtidos pela submissão de seis páginas *web* ao método proposto por Rafiei e Mendelzon e ao método de identificação automática de tópicos proposto nesta dissertação.

A identificação do tópico de uma página *web* possui uma vasta aplicabilidade nos sistemas de RI em geral, podendo servir como uma nova fonte de evidência para melhorar o *ranking* sistemas de busca de informações, aperfeiçoar sistemas de classificação e de filtragem de páginas. O tópico das páginas ainda pode ser utilizado em sistemas de exibição automática de propagandas contextuais e na validação de páginas comerciais e pessoais, informando como estas páginas são conhecidas na w*eb*.

5.3 Experimento 3 - Utilização do método para classificação de páginas no diretório *Google*.

Este experimento foi feito com objetivo de verificar a eficácia da aplicação de parte do método proposto na classificação de páginas *web* dentro da hierarquia de diretórios *Google*. Em sua execução foram utilizadas 50 páginas obtidas do diretório *Google*, através da análise manual de seus respectivos conteúdos. As páginas utilizadas neste

experimento são aquelas que: (1) discorrem acerca de um determinado assunto, (2) possuem conteúdo html; (3) são escritas em inglês; e (4) pertencem ao diretório *Google*.

Além das páginas *web* foram obtidas, também, as categorias a qual cada uma pertencia, com o objetivo de compará-las com os resultados retornados pelo sistema. A Figura 22 mostra uma lista com algumas das diversas categorias utilizadas na realização deste experimento.

- `/top/Computers/Data_Formats/Markup_Languages/HTML/`
- `/top/Society/Religion_and_Spirituality/Christianity/Church_History/`
- `/top/Society/Politics/Democracy/`
- `/top/Sports/Strength_Sports/Bodybuilding/Supplements/Anabolic_Steroids/`
- `/top/Society/Holidays/Halloween/History/`
- `/top/Health/Nutrition/Nutrients/Trans_Fats/`
- `/top/Arts/Music/Bands_and_Artists/L/Lennon,_John/`
- `/top/Science/Earth Sciences/Geology/Geologic_Hazards/Tsunami`
- `/top/Science/Astronomy/Solar_System/Planets/Mercury`

Figura 22 - Lista de categorias utilizadas na realização do experimento 3.

Uma limitação imposta a este experimento é que as páginas utilizadas para extrair os dados que compunham a evidência *Texto de Âncora* de uma determinada página *p* não podiam pertencer ao mesmo diretório *web* de *p*. Isso evita a influência positiva do conteúdo destes apontadores no método. Neste experimento o número mínimo de páginas utilizadas na concatenação dos textos de âncora de *links* que referenciam *p* foi 5 e o número máximo foi 30.

O próximo passo foi a submissão destas páginas ao método exposto, porém nem todas as fases que compõem o processo foram executadas. As três primeiras etapas do método foram executadas integralmente (Capítulo 4), contudo na última etapa não foi executado a subfase de otimização da lista de categorias. Desta etapa foi obtida apenas a primeira categoria associada à primeira resposta do sistema de busca em diretório *Google* quando os três melhores termos descritores das páginas foram submetidos a ele.

Para avaliar a eficiência da aplicação prática de parte do método proposto na classificação de páginas *web* no diretório *Google*, comparamos as categorias geradas como resposta do sistema às categorias de origem de cada página testada. O diretório *Google* utiliza o conceito de categorias relacionadas que são categorias que possuem páginas da *web* com conteúdo semelhante a páginas pertencentes à outra categoria do diretório[2]. Citamos o seguinte exemplo: a categoria `/top/Science/Earth_Sciences/Geology/Geologic_Hazards/Tsunami` mostra um *link* relacionado à categoria `/top/Kids_and_Teens/School_Time/Science/The_Earth/Geology/Tsunamis`, pois o conteúdo presente em ambas as categorias estão relacionados ao mesmo tema: `Tsunamis`. Portanto, assumimos que tanto as páginas classificadas em suas categorias de origem assim como as páginas classificadas em categorias relacionadas foram classificadas corretamente pelo sistema.

Para quantificar a eficiência do método na classificação de páginas *web*, o grupo inicial de 50 páginas de teste foi divido, aleatoriamente, em 5 grupos contendo 10 páginas cada. Para cada grupo foi calculado o valor de precisão do método (conforme Equação 44). Em seguida, foi calculado o desvio padrão destes valores de precisão. Como resultado, obtivemos precisão média de 88% com desvio padrão de 11%. Portanto, atingimos o resultado de 88%±0.11.

$$\text{Pr}\,ecisão = \frac{Número\ \ páginas\ classificadas\ corretamente}{Total\ \ de\ \ páginas\ no\ grupo} \tag{5.2}$$

Logo, conclui-se que o método também possui aplicação prática na classificação de páginas no diretório *Google*, a Tabela 4 sintetiza os resultados obtidos para cada um dos 5 grupos de páginas avaliados.

De acordo com Qi e Davison, em [19], a classificação automática de páginas *web* é uma tarefa essencial a diversos sistemas de RI, entre eles a manutenção dos diretórios *web* e a busca focada de informações.

	Grupo 1	**Grupo 2**	**Grupo 3**	**Grupo 4**	**Grupo 5**
Total de páginas classificadas corretamente	8	10	10	9	7
Total de páginas classificadas erroneamente	2	0	0	1	3
Valor de Precisão	80%	100%	100%	90%	70%

Tabela 4 – Resultados obtidos com a aplicação prática do método proposto na classificação de páginas *web* no diretório *Google*.

Capítulo 6

Conclusões e Trabalhos Futuros

6.1 Conclusão

Foi apresentado, neste trabalho, o desenvolvimento e a avaliação de um método utilizado para identificar automaticamente o tópico de uma página *web p*. O método aqui exposto propôs a utilização de diversas fontes de evidências textuais, extraídas de *p*, que combinadas deram origem a uma lista de termos associados ao assunto desta, e que foram utilizadas na definição de seu tópico. A combinação das evidências utilizadas foi feita automaticamente através do uso de algoritmos genéticos, evitando assim as suposições humanas. Na última etapa do método, onde é gerado o tópico de *p*, foi utilizado o serviço de busca em diretório da *Google*, a ele foram submetidos os três melhores termos descritores de *p*, e dele foram obtidas as categorias relacionadas às primeiras vinte respostas desta busca, estas categorias passaram por um processo de otimização e originaram o tópico de *p*.

Apenas páginas que discorressem acerca de um determinado assunto, que possuíssem conteúdo html e que fossem escritas em inglês foram submetidas ao método. A primeira exigência foi justificada, pois muitas vezes, por apresentarem notícias diversificadas, nem mesmo o ser humano consegue definir o assunto predominante de página com estas características. A segunda restrição foi feita, pois o método utiliza o conteúdo destacado por *tags* html na descoberta dos termos descritores de *p*. Já a terceira restrição foi aplicada, pois pesquisas feitas no diretório *Google* comprovaram que existe um maior número de páginas classificadas neste idioma se compararmos ao número de páginas classificadas em outros idiomas.

Os resultados obtidos nos experimentos realizados para avaliar o método proposto foram os seguintes: (1) alto grau de importância do uso da concatenação do texto de âncora de *links* na definição dos termos descritores de uma página *web p*; (2) boa avaliação da eficiência do método proposto em identificar o tópico de *p*: ***0.9129***, em uma escala de 0 a 1, onde 1 indica o comportamento ideal do sistema; e (3) boa avaliação da utilização de parte do método proposto na classificação automática de páginas *web* na estrutura hierárquica do diretório *Google*, atingindo 88%±0.11de acertos do total de páginas classificadas.

Os experimentos realizados demonstram que o modelo proposto é útil na identificação do tópico de uma página *web* e também na classificação de novas páginas na estrutura hierárquica do diretório *Google*.

6.2 Trabalhos Futuros

São sugeridos como trabalhos futuros:

- Submeter *blogs*, portais e páginas iniciais de *sites* de jornais ao método com a finalidade de experimentar e avaliar o desempenho do mesmo. Estudando e propondo a adaptação do mesmo caso os resultados obtidos não sejam satisfatórios.
- Realizar experimentos com o método proposto em outras coleções, com o objetivo de melhor embasar os resultado aqui obtidos. Assim como, realizar novos experimentos na classificação automática de páginas na estrutura hierárquica do diretório *Google*.
- Testar o método proposto com fontes de evidências adicionais, tais como: (1) o modelo de propagação de influência em um nível e o modelo de propagação de influência em dois níveis propostos por Rafiei e Mendelzo que tem o objetivo de identificar automaticamente os principais termos de uma página *web* baseada nos valores de relevância destes termos para aquela página, (2) as informações provindas de meta tags do html, por exemplo, a meta *tag Description* e a meta *tag Keywords* que descrevem de maneira sucinta o conteúdo da página e identificam as palavras-chaves associadas ao conteúdo da página, respectivamente.

- Avaliar a suscetibilidade da fonte de evidência *Texto de Âncora* a ruídos através da submissão de páginas populares e páginas não populares ao método proposto. Páginas populares são aquelas referenciadas por diversas outras páginas enquanto páginas não populares são páginas que possuem pouca referencia. A popularidade da página reflete a qualidade do conteúdo da evidencia *Texto de Âncora*, identificada como a principal evidência na identificação do tópico de uma página. Páginas mais populares possuem maior probabilidade de apresentar ruído nos textos de âncora de *links* que a referenciam.

Referências Bibliográficas

[1] Agyemang, M.; Barker, K.; Alhajj, R. *Mining Web Content Outliers using Structure Oriented Weighting Techniques and N-Grams*. SAC'05, Santa Fe, New Mexico, USA, 2005.

[2] Ajuda do diretório do *Google* na *web*. http://www.google.com/dirhelp.html. Data acesso: 22/03/2009.

[3] Baeza-Yates, R.; Ribeiro-Neto, B. *Modern Information Retrieval*. Addison Wesley, 1999.

[4] Blockeel, H. and Struyf, J. Efficient *Algorithms for Decision Tree Cross-validation*. The Journal of Machine Learning Research, 2003.

[5] Brin, S.; Page, L. *The anatomy of a large-scale hypertextual web search engine*. In Proceedings of the 7th International World Wide Web Conference, páginas 107-117, Brisbane, Australia. Elsevier Science, 1998.

[6] Categoria *World* do diretório da *Google*. http://directory.google.com/Top/World/. Data acesso: 24/02/2009

[7] Cavnar, W. e Trenkle, J. N-gram-based text categorization. In *Symposium On Document Analysis and Information Retrieval*, páginas 161–176, Universidade de Nevada - Las Vegas, 1994.

[8] Chakrabarti, S.; Dom, B.; Indyk, P. *Enhanced hypertext categorization using hyperlinks*. SIGMOD '98 Seattle. WA, USA.

[9] Goldberg, D. E. *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley, 1998.

[10] *Google SOAP Search API*. Disponível em http://code.google.com/apis/soapsearch/. Data acesso: 01/04/2008.

[11] Jarvelin, K.; Kekalainen, J. *Cumulated Gain-Based Evaluation of IR Techniques. ACM Transactions on Information Systems*, vol. 20, no. 4, páginas 422-446, 2002

[12] Kim, J.; Candan, K., CP/VC: *Concept Similarity Mining without Frequency Information from Domain Describing Taxonomies*. CIKM06, USA, 2006

[13] Kleinberg, J. M. *Authoritavite Sources in a Hyperlinked Enviroment*. Journal of the ACM (JACM), 1999.

[14] Linden, R. *Algoritmos Genéticos: Uma importante ferramenta da Inteligência Computacional*. Ed. Brasport. Rio de Janeiro, 2006.

[15] Liu, B.; Chin, C.; Ng, H.. *Mining* Topic-Specific Concepts and Definitions on the Web. WWW 2003, Hungary, 2003.

[16] Michalewicz, Z. Genetic Algorithms + Data Structures = Evolution Programs. Ed. Springer, 3a. ed., 1999.

[17] Mitchell, M. *An Introduction to Genetic Algorithms*. MIT Press, 1998.

[18] *Open Directory Project* http://dmoz.com/. Data acesso: 24/02/2009.

[19] Qi, X.; Davison, B. *Web Page Classification: Features and Algorithms*. ACM Computing Surveys, Vol. 42, No. 2, Article 12, 2009.

[20] Rafiei, D.; Mendelzon, A., *What is this Page Known for? Computing Web Page Reputations*. WWW9 Conference, Amsterdam, 2000.

[21] Recomendações para a submissão de *sites* ao diretório *Yahoo*!. http://help.yahoo.com/l/br/yahoo/dir/ctd/ctd-01.html. Data acesso: 17/03/2009.

[22] Silva, T.; Moura, E.; Cavalcanti, J.; Silva, A.; Carvalho, M.; Gonçalves, A. *An evolutionary approach for combining different sources of evidence in search engines*. Information Systems, 2009.

[23] Sun, A.; Lim, E.; Ng, W. *Performance Measurement Framework for Hierarchical Text Classification*. Journal of the American Society for Information Science and Technology (JASIST), 2003.

[24] Tiun, S.; Abdullah, R.; Kong, T. *Automatic Topic Identification Using Ontology Hierarchy*. Proceedings of the Second International Conference on Computational Linguistics and Intelligent Text Processing, 2001.

[25] Zeng, H.; He, Q.; Chen, Z.; Ma, W.; Ma, J. *Learning to Cluster Web Search Results*. SIGIR'04, Sheffield, South Yorkshire, UK, 2004.