UNIVERSIDADE FEDERAL DE SÃO CARLOS

CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO

EM CIÊNCIA DA COMPUTAÇÃO

“ExtraWeb: um sumarizador de documentos Web baseado em etiquetas HTML e ontologia”

Patrick Pedreira Silva

SÃO CARLOS
Julho/2006

Ficha catalográfica elaborada pelo DePT da Biblioteca Comunitária da UFSCar

S586es

Silva, Patrick Pedreira.
ExtraWeb: um sumarizador de documentos Web baseado em etiquetas HTML e ontologia / Patrick Pedreira Silva. -- São Carlos : UFSCar, 2006.
158 p.

Dissertação (Mestrado) -- Universidade Federal de São Carlos, 2006.

1. Inteligência artificial. 2. Processamento da linguagem natural. 3. Sumarização automática. I. Título.

CDD: 006.3 (20ª)

Universidade Federal de São Carlos

Centro de Ciências Exatas e de Tecnologia

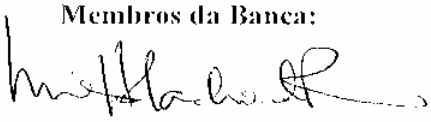
Programa de Pós-Graduação em Ciência da Computação

"ExtraWeb: um sumarizador de documentos Web baseado em etiquetas HTML e ontologia"

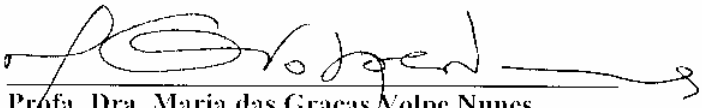
PATRICK PEDREIRA SILVA

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de São Carlos, como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação.

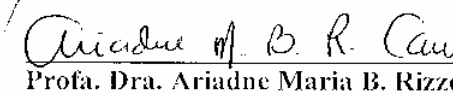
Membros da Banca:



Profa. Dra. Lúcia Helena Machado Rino
(Orientadora – DC/UFSCar)



Profa. Dra. Maria das Graças Volpe Nunes
(ICMC/USP)



Profa. Dra. Ariadne Maria B. Rizzoni de Carvalho
(UNICAMP)

São Carlos
Julho/2006

AGRADECIMENTOS

A Deus por me acompanhar sempre e me dar a força necessária para superar as dificuldades.

Aos meus pais, Gilberto e Vera, e minha irmã, Yanna, pelo apoio incondicional e estímulo.

A professora Lucia, pela orientação e auxílio.

A Flávia pela compreensão, apoio, amor e carinho.

A Elaine, Flávio e Lucas por terem me recebido e tratado como um membro da família.

Ao meu amigo Wilson pela amizade, bate-papos e força.

Aos meus colegas de LIAA e NILC pelo auxílio.

A CAPES pelo suporte financeiro a este trabalho.

RESUMO

Esta dissertação propõe um sumarizador de documentos *Web* baseado em etiquetas HTML e conhecimento ontológico, derivado de outras duas abordagens independentes: uma que contempla somente etiquetas HTML e outra, somente conhecimento ontológico. As três abordagens foram implementadas e avaliadas, indicando que a composição desses dois tipos de conhecimento tem um bom potencial descritivo de documentos *Web*. O protótipo resultante é denominado ExtraWeb.

O ExtraWeb explora a estrutura de marcação de documentos em português e informações de nível semântico usando a ontologia do Yahoo em português, enriquecida com vocabulário extraído de um thesaurus, Diadorim, e da Wikipédia. Em uma tarefa simulada por internautas, de busca de documentos, o ExtraWeb obteve um grau de utilidade próximo ao do Google, evidenciando seu potencial para indicar, por meio de extratos, a relevância de documentos recuperados na *Web*. Esse foco é de grande interesse atualmente, pois os extratos podem ser particularmente úteis como substitutos das descrições atuais das ferramentas de busca ou, mesmo, como substitutos dos documentos correspondentes completos. No primeiro caso, as descrições nem sempre contemplam as informações mais relevantes dos documentos; no segundo, sua leitura implica um esforço considerável por parte do internauta. Em ambos os casos, extratos podem otimizar essa tarefa, se comprovada sua utilidade para a indicação da relevância dos documentos. Assim, o ExtraWeb tem potencial para ser um acessório das ferramentas de busca, para melhorar a forma como os resultados são apresentados, muito embora sua escalabilidade e implantação em um ambiente real ainda não tenham sido exploradas.

ABSTRACT

This dissertation presents an automatic summarizer of Web documents based on both HTML tags and ontological knowledge. It has been derived from two independent approaches: one that focuses solely upon HTML tags, and another that focuses only on ontological knowledge. The three approaches were implemented and assessed, indicating that associating both knowledge types have a promising descriptive power for Web documents. The resulting prototype has been named ExtraWeb.

The ExtraWeb system explores the HTML structure of Web documents in Portuguese and semantic information using the Yahoo ontology in Portuguese. This has been enriched with additional terms extracted from both a thesaurus, Diadorim and the Wikipedia. In a simulated Web search, ExtraWeb achieved a similar utility degree to Google one, showing its potential to signal through extracts the relevance of the retrieved documents. This has been an important issue recently. Extracts may be particularly useful as surrogates of the current descriptions provided by the existing search engines. They may even substitute the corresponding source documents. In the former case, those descriptions do not necessarily convey relevant content of the documents; in the latter, reading full documents demands a substantial overhead of Web users. In both cases, extracts may improve the search task, provided that they actually signal relevant content. So, ExtraWeb is a potential plug-in of search engines, to improve their descriptions. However, its scability and insertion in a real setting have not yet been explored.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1 Introdução

A sumarização de textos é uma atividade rotineira na vida das pessoas. O propósito dos sumários é destacar as informações mais importantes de textos-fonte para produzir versões condensadas dos mesmos, visando determinado usuário e/ou tarefa (MANI; MAYBURY, 1999). Por essa razão, eles podem ser utilizados, por exemplo, em substituição à leitura completa do documento correspondente, minimizando o tempo de leitura despendido pelo leitor. A elaboração automática de sumários é estudada desde 1958 (LUNH, 1958), quando foram propostos os primeiros sistemas de Sumarização Automática (SA). Na atualidade, com o crescimento espetacular da Internet, em que uma quantidade cada vez maior de informações é disponibilizada *on-line* (LYMAN; VARIAN, 2003; WITTEN et al., 1994), a tecnologia de sumarização automática vem se tornando indispensável para lidar com o grande volume de informações, propiciando o desenvolvimento e o interesse cada vez maior por pesquisas nessa área.

Sendo as *webpages* um dos mais expressivos repositórios de informações, é natural que ferramentas de sumarização automática sejam desenvolvidas para sintetizá-las, buscando minimizar o esforço dos usuários e, assim, permitindo que estes apreendam rápida e eficazmente somente as informações mais relevantes. Uma das tarefas mais rotineiras desenvolvidas na Internet envolve a interação entre internautas e sistemas de Recuperação de Informação (RI), aqui denominados mecanismos de busca. Esses mecanismos indexam imensas coleções, acarretando a recuperação de milhares de documentos considerados relevantes, o que exige que o usuário tome uma decisão, selecionando, dentre os documentos retornados, aqueles que realmente atendam sua necessidade de informação. Na tomada de decisão, a forma como o documento é apresentado ao usuário é extremamente importante. A qualidade das descrições fornecidas pelo mecanismo de busca é crucial para o sucesso da busca, mas muitas

vezes, estas não são coerentes ou são confusas. Assim, a única maneira de verificar a relevância do documento apresentado é navegando pelo *link* indicado pelo mecanismo de busca. Isso aumenta o *overhead cognitivo* do usuário, que é definido como o esforço adicional que o internauta deve fazer para acessar o documento-fonte e verificar seu conteúdo (CONKLIN, 1987). Uma das formas de minimizar esse esforço seria considerar que é de responsabilidade do sistema fornecer melhores descrições dos documentos recuperados, para que estas permitissem ao usuário selecioná-los. Neste trabalho, propomos descrições de *webpages*, na forma de sumários gerados automaticamente, para fornecer aos usuários uma visão geral do conteúdo dos documentos. A questão principal, nesse caso, é fornecer ao usuário indícios claros sobre o conteúdo principal dos documentos e, assim, permitir que este faça sua seleção de documentos sem recorrer a eles diretamente. O foco deste trabalho é investigar técnicas de SA que possam contribuir para a melhoria da apresentação de resultados de mecanismos de busca. O cenário considerado é o de recuperação de *webpages*: dada uma consulta, o mecanismo de busca recupera as *webpages* que se relacionam a ela e mostra descrições curtas de seu conteúdo. O processo de busca do Google (BRIN; PAGE, 1998) é um bom exemplo disso: ele reproduz um trecho relativo à página recuperada – a descrição – juntamente com seu *link*. Propomos somente melhorar essa descrição, com os sumários das *webpages*. Cabe salientar que, apesar de a Internet se caracterizar pela publicação de documentos em diversos formatos, a nossa proposta de SA limita-se ao processamento de documentos *Web* cujo conteúdo esteja exclusivamente formatado por etiquetas HTML.

Atualmente, uma das principais abordagens utilizadas na SA é a profunda. Ela se baseia em modelos lingüísticos e/ou discursivos para simular a sumarização humana em um processo de interpretação e condensação do texto, para produção do sumário. Um

bom sistema automático que siga essa abordagem deve fazer uso intensivo de regras gramaticais e de habilidades de inferência lógica e conhecimento de mundo (SAMPSON, 1987).

Apesar da grande quantidade de técnicas e algoritmos de sumarização propostos na Lingüística Computacional, a sumarização de *webpages* requer métodos ligeiramente diferentes, devido à natureza da estrutura e do conteúdo de seus documentos (BERGER; MITTAL, 2000). Em vez de textos puros, isto é, textos sem nenhum tipo de formatação, com uma estrutura de discurso bem definida, os documentos geralmente possuem um misto de frases, *links*, gráficos e comandos de formatação que dificultam a modelagem dos processos de identificação, seleção e estruturação de tópicos para a SA. Sua codificação em HTML (*HyperText Markup Language*), contudo, poderia fornecer importantes indícios para o processo de sumarização. Assim, propomos explorar os elementos potenciais indicados pelo código HTML para distinguir as informações relevantes das supérfluas, a fim de considerar as primeiras e descartar as segundas, na elaboração de sumários. Esses elementos relacionam-se tanto à macro quanto à micro-estrutura de um documento, ou seja, é possível considerar etiquetas HTML estruturais ou etiquetas que, simplesmente, indicam alterações de estilo introduzidas pelo autor para ressaltar informações mais pontuais. Ao mesmo tempo, propomos enriquecer o modelo de SA com informações semânticas, fazendo uso de conhecimento ontológico, aliando o uso de informações estruturais provenientes das etiquetas HTML com informações semânticas oriundas de uma ontologia para detectar, por exemplo, os tópicos principais de um documento, a fim de guiar a composição dos sumários.

Ao contemplarmos o processamento semântico, uma questão relevante a considerar seria o uso da XML (*eXtensible Markup Language*), como meta-linguagem natural de etiquetação semântica de documentos. Excluímos seu uso, no entanto,

porque, diferentemente da linguagem HTML, que provê um conjunto de etiquetas e uma semântica bem definidos e padronizados por normas internacionais, a XML é de uso livre. Assim é que documentos anotados em XML podem incorporar etiquetas definidas pelo próprio usuário, dificultando a modelagem pretendida. Por exemplo, é possível encontrar em um documento XML uma etiqueta <ENDEREÇO>, que pode indicar tanto um endereço residencial quanto um endereço de e-mail. Embora essa flexibilidade da XML seja uma vantagem, por permitir que os usuários definam as suas próprias etiquetas de acordo com suas necessidades, ela resulta em versões muito diversificadas (e, muitas vezes, incompatíveis) de conjuntos de etiquetas. Além de ser de difícil modelagem, o uso livre de etiquetas em XML se torna um sério fator limitante para o processamento automático, já que o sistema não consegue prever toda a semântica de etiquetas que possam ser encontradas nos documentos. Dessa forma, optamos pela utilização da HTML, que é uma linguagem bem formalizada.

O trabalho aqui apresentado consiste na implementação de três modelos de sumarização: um modelo que faz uso de etiquetas HTML; outro que utiliza somente conhecimento ontológico para detectar os tópicos principais dos documentos e o último, que foi construído segundo uma metodologia mista utilizando tanto etiquetas HTML quanto conhecimento ontológico, dando origem ao sumarizador extrativo ExtraWeb que é a maior contribuição deste trabalho.

Com base no exposto anteriormente, diante da necessidade/utilidade de melhorar as descrições fornecidas por mecanismos de busca e com o intuito de coletar subsídios para a proposta dos modelos citados, no próximo capítulo apresentaremos uma análise das descrições geradas por mecanismos de busca por meio de um estudo de casos do Google, visando verificar as fragilidades da apresentação de resultados desses mecanismos. No capítulo 3, descreveremos alguns trabalhos correlatos de SA de

documentos *Web* bem como trabalhos relacionados à detecção automática, baseada em ontologia, de conceitos subjacentes a documentos. No Capítulo 4, apresentaremos duas implementações preliminares de sumarizadores baseados em etiquetas HTML e ontologia. No Capítulo 5, apresentaremos o sumarizador extrativo ExtraWeb. Por fim, no Capítulo 6 apresentaremos algumas considerações finais.

2 Estudo de casos: descrições do Google

Com o objetivo de verificar as principais deficiências da apresentação de resultados dos mecanismos de busca, elaboramos um estudo de casos para colecionar subsídios para a definição de um modelo de SA que privilegiasse descrições mais informativas para o usuário.

Esse estudo baseou-se em um dos principais mecanismos de busca existentes - o Google - mais particularmente, em sua forma de apresentação dos resultados. Em nenhum sentido questiona-se aqui a eficácia de identificação de documentos relevantes, ou seja, a própria eficácia de busca, já que o foco deste trabalho está exclusivamente na SA dos documentos, considerados previamente recuperados.

Diferenciamos aqui dois componentes importantes para esse estudo: os documentos, que são constituídos por tudo o que é visível ao se acessar o *link* exibido pelo Google, e a janela de descrição, que é a resposta visível dada ao usuário, após o Google recuperar o documento. Essa janela é composta do *link* para o documento e de sua própria descrição textual ou frasal (GRIESBAUM, 2004). Embora outros trabalhos (INKTOMI, 2003; SHERMAN, 2002) não façam essa diferenciação, para o nosso trabalho ela é interessante, já que nosso objetivo é verificar a qualidade das descrições e não propriamente dos documentos recuperados.

Geralmente, num primeiro passo, os usuários decidem, lendo as descrições, quais documentos devem ser acessados, para verificar quais atenderão às suas necessidades de informação. Somente aqueles cujas descrições pareçam ser relevantes terão alguma chance de serem selecionados. Assim, as descrições são o ponto de partida para as escolhas dos usuários, que atribuem relevância aos documentos de acordo com a forma como estes são descritos e não de acordo com o seu conteúdo real. Se a qualidade das estimativas de relevância com base nas descrições corresponder à qualidade real dos

documentos, não existirá problema. Ao contrário, uma visão geral permitirá que os usuários diferenciem rapidamente os bons e os maus resultados. Contudo, existirá um problema quando a descrição do documento apontada pelo mecanismo de busca não levar a um documento relevante. O usuário só descobrirá isso ao acessar o próprio documento e analisar seu conteúdo. Neste caso, o *overhead cognitivo* será maior. Em outros casos, o usuário pode também deixar de verificar documentos relevantes, com base no julgamento feito pela descrição. Em ambos sua busca será prejudicada.

Para verificar em que medida os usuários podem ficar satisfeitos com as respostas do Google, sem grande *overhead cognitivo*, é que o estudo relatado a seguir foi efetuado. O Google foi escolhido para gerar a coleção de dados sob análise por ser o mais utilizado dos mecanismos de busca e ser uma ferramenta de alta qualidade (SULLIVAN, 2004).

2.1 Descrições de documentos pelo Google

A criação de descrições pelo Google é completamente automatizada e leva em consideração os termos da consulta apresentada pelo usuário, assim como o conteúdo da página e as referências a ela na *Web* (http://www.google.com/about.html). As descrições exibidas são constituídas de excertos retirados das páginas recuperadas, em vez de seus resumos estáticos, isto é, de resumos criados manualmente, quando existirem. São priorizados os segmentos textuais cujos termos correspondam àqueles usados na consulta, isto é, sejam exatamente os mesmos fornecidos pelo usuário, aparecendo, inclusive, na mesma ordem definida na consulta. Preferencialmente, o Google exibe os segmentos textuais contidos na descrição construída manualmente pelo autor do documento. Neste caso, ela é indicada pela etiqueta HTML <META NAME="Description">. O critério para essa seleção dos segmentos textuais é dependente da consulta: se os termos pesquisados pelo usuário estiverem incluídos no

conteúdo da etiqueta <META NAME="Description">, então esse conteúdo será exibido. Caso os termos pesquisados não se encontrem neste conteúdo, a janela exibida será composta por segmentos textuais contidos no corpo da página, conforme descrito anteriormente.

Essas descrições têm como objetivos principais prover o usuário da informação que ele está procurando, ilustrar o conteúdo que será encontrado no documento apontado e indicar outros termos que possam ser utilizados em buscas subseqüentes (http://www.googleguide.com/results_page.html). Em algumas situações, o Google não é capaz de associar uma descrição a um *link* recuperado, deixando o usuário sem nenhum tipo de informação prévia sobre o conteúdo do documento. Isso ocorre quando o documento não foi completamente indexado, o que significa que o mecanismo não foi capaz de recuperar segmentos textuais relacionadas à consulta. Sem esse tipo de informação, somente é exibido o *link* do documento. Essa situação não foi considerada nesse estudo, já que nesses casos não seria possível avaliar as descrições.

2.2 Análise de discurso das descrições do Google

Pelo exposto vê-se que as descrições produzidas pelo Google não são necessariamente textuais e nem correspondem a segmentos explicitamente apontados no documento-fonte, já que o conteúdo da etiqueta <META NAME="Description"> não é visível no documento para os leitores. A análise relatada nesta seção visa avaliar sua qualidade, considerando somente seu grau de informatividade, já que o julgamento da qualidade textual seria claramente prejudicado.

2.2.1 Caracterização da análise das descrições

Nossa análise do discurso busca responder, assim, às seguintes perguntas:

1. A resposta dada pelo mecanismo de busca é suficientemente descritiva?
2. A descrição apresentada reflete o conteúdo do documento-fonte?

A resposta afirmativa à questão (1) pode implicar uma das seguintes perspectivas:

a. Descartar a necessidade de o usuário buscar o documento-fonte. Neste caso, o usuário terá uma idéia geral do documento, ficando satisfeito com a descrição. Nesta situação a necessidade de informação do usuário é atendida pela descrição;
b. Recuperar o documento-fonte com a certeza de que a descrição espelha bem seu conteúdo. Neste caso, a etapa de crítica do documento-fonte é descartada e, assim, o *overhead cognitivo* do usuário é evitado.

A resposta à questão (2) exige a comparação da descrição com o conteúdo do documento-fonte.

Como essas perguntas indicam, avaliar a qualidade de uma descrição fornecida por um mecanismo de busca é um processo subjetivo e está ligado à percepção dos usuários, a qual pode variar muito, tornando o julgamento difícil. Objetivos distintos de busca também podem interferir no julgamento de qualidade. Além disso, usuários distintos, com necessidades de informação variadas, também têm percepções diferentes da qualidade das respostas.

A fim de delimitar a análise das respostas do Google e torná-la consistente, definimos duas variáveis: *keywords* e conceitos. As *keywords* têm correspondentes explícitas com aquelas definidas pelos autores e são indicadas em HTML, pela etiqueta <META NAME="Keywords">, nos próprios documentos[1]. Não só palavras simples podem ser indicadas por essa meta-etiqueta, mas também frases. Considera-se que elas refletem os tópicos principais dos documentos correspondentes (RAGGETT et al.,

[1] Daí o uso do próprio termo em inglês

1999). Os conceitos correspondem às palavras (excluindo as *stopwords* (CINTRA et al., 2001)) presentes nas descrições geradas pelo mecanismo de busca tendo, desta forma, natureza distinta das *keywords*.

O método escolhido para avaliarmos as descrições do Google consiste em verificar se as *keywords* de um determinado documento estão contempladas em sua descrição. Em outras palavras, pretendemos verificar se os conceitos apresentados nas descrições correspondem às *keywords*. Para esta avaliação, desconsideramos variações morfológicas entre a lista de *keywords* e os conceitos presentes nas descrições, pois o conteúdo é mais importante do que a forma das palavras. Desse modo, adotamos simplesmente as formas canônicas das palavras. Estabelecemos a seguinte hipótese:

- quanto mais *keywords* de um documento uma descrição contiver, maior será sua informatividade.

Quanto mais informativa uma descrição, melhor ela atenderá às necessidades de informação do usuário, ajudando-o a decidir com maior grau de certeza a relevância do documento correspondente, em relação à sua consulta.

2.2.2 Medidas de desempenho consideradas

Para julgar a qualidade das descrições, somente consideramos os primeiros 20 resultados para cada consulta já que, geralmente, a maior parte dos usuários visualiza até duas páginas de resultados de uma busca (INKTOMI, 2003; JANSEN et al., 2001). Griesbaum (2004) sugere que sejam utilizadas consultas distintas para a verificação de resultados, já que este é o procedimento usado pela TREC *(Text REtrieval Conference*). Seguindo essa diretriz, montamos um corpus de análise a partir da coleta de resultados provenientes de 27 consultas (conforme descrito na Seção 2.2.3), reproduzindo o mesmo número usado por Griesbaum (2004) para avaliar respostas de mecanismos de

busca alemães. Com este corpus foi possível calcular a precisão e a cobertura do Google na simulação de recuperação de documentos.

Cobertura e precisão são medidas de desempenho padrão da recuperação de informação (CLEVERDON et al., 1966). Seja Kdesc o número de *keywords* da descrição, Nc o número total de conceitos da descrição e Kdoc o número total de *keywords* do documento. A precisão e cobertura são definidas, respectivamente, por P= Kdesc/Nc e C=Kdesc/Kdoc. P indica a proporção de conceitos relevantes de uma janela; C indica a proporção de *keywords* da janela que são representativas do documento (CRAVEN, 2003). Adicionalmente, calculamos também a *f-measure* (F), como medida de eficiência do Google. Quanto mais próxima de 1, melhor será a informatividade das janelas analisadas.

2.2.3 Descrição do corpus de análise

As consultas submetidas ao Google foram construídas manualmente. De acordo com Wolfram et al. (2001) e Spink et al. (2000), consultas típicas são geralmente curtas (na maioria dos casos de 1 a 3 palavras) e possuem poucas restrições de uso de operadores booleanos. Para tentar ser o mais objetivo possível e simular da melhor maneira o contexto de uma pesquisa real, construímos consultas típicas segundo essa mesma definição, de forma aleatória em relação a seus domínios, ou seja, o número de domínios ou de consultas por domínio não foi considerado relevante.

Com o corpus, foi possível analisar dois tipos de dados. Primeiro, os referentes às janelas de descrição dos documentos geradas com base nas consultas. Segundo, os referentes às *keywords* indicadas pela meta-etiqueta <META NAME="Keywords">. Aproximadamente 2 a 8 janelas dos primeiros documentos *Web* recuperados foram coletadas, considerando a mesma ordem apontada pelo Google. Estes valores

correspondem ao número de documentos, entre os 20 primeiros, que traziam em seu código uma lista de *keywords* explicitamente indicada pela meta-etiqueta <META NAME="Keywords">. No total foram recolhidas 124 janelas (descrições e *links*) referentes a 27 consultas distintas (Tabela 1). Na média, 4 ½ janelas por consulta foram recuperadas, cujas descrições são, em média, de 22 palavras, das quais 14 eram conceitos.

Tabela 1. Consultas utilizadas para a recuperação de janelas

Consulta	Total de janelas por consulta
1. redes AND neurais AND artificiais	2
2. informações AND viagem AND "porto seguro" AND bahia	5
3. bibliografia AND "glauber rocha"	5
4. tutorial AND de AND prolog	3
5. Roberto AND Carlos AND cantor NOT jogador	4
6. jogo AND fifa NOT www.fifa.com AND EA AND Sports	5
7. laço AND programação AND repetição	2
8. dificuldades AND terceira AND idade AND internet	5
9. curso AND de AND inglês AND gratuito	4
10. receitas AND doces AND salgados AND carnes AND massas	2
11. sorriso AND monalisa NOT filme NOT columbia NOT dvd NOT cd	4
12. musical AND o AND fantasma AND da AND opera	3
13. livraria AND virtual	4
14. recursos AND gratuitos AND para AND webmasters	4
15. astronomia AND "Sistema Solar" AND planetas	5
16. download AND musicas AND mp3	7
17. informações AND divórcio NOT europa	3
18. poluição AND do AND ar	6
19. dinossauros	6
20. doenças AND respiratórias	5
21. namoro AND virtual	6
22. medicina AND alternativa NOT livro NOT livros	4
23. "Elba Ramalho" AND cantora	5
24. animais AND ameaçados AND de AND extinção	6
25. cartões AND virtuais	7
26. mp3	8
27. receita AND federal	4
Total	**124 janelas**

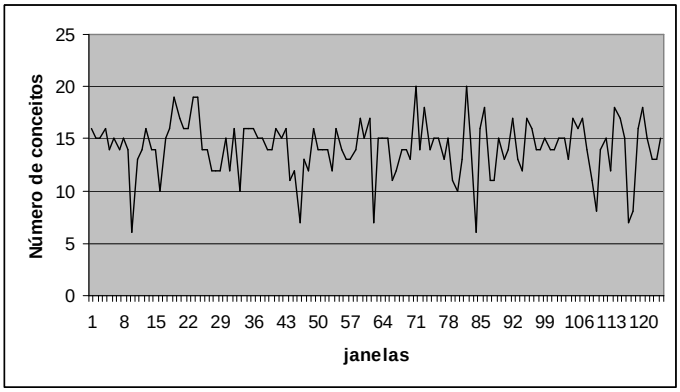


Figura 1. Distribuição dos conceitos nas janelas

A distribuição do número de conceitos em cada uma das janelas é apresentada na Figura 1. Essa distribuição no corpus é importante porque reflete a natureza heterogênea dos documentos encontrados na Internet, a qual, associada ao próprio mecanismo de recuperação do Google, resulta em janelas cujas descrições possuem diferentes números de conceitos. Essa diversidade é importante já que, desta forma, os nossos dados de análise refletem as diferentes descrições geradas pelos mecanismos de busca durante o processo de interação com o usuário e, conseqüentemente, diferentes graus de qualidade, já que descrições com características muito similares poderiam, de alguma forma, corromper a análise dos dados.

O número médio de *keywords* de cada documento foi de aproximadamente 20, porém, houve uma grande variação com relação a este número, como pode ser visto na Figura 2. Alguns documentos apresentavam somente uma *keyword*, enquanto outros tinham 138. Contudo, a maioria dos documentos analisados (101) tinha menos de 31 *keywords*.

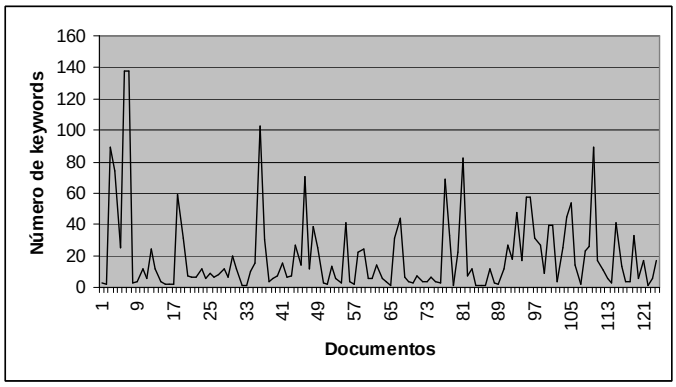


Figura 2. Distribuição das keywords nos documentos

2.3 Síntese da análise de corpus

Em média, as janelas cobriram cerca de 29% das *keywords* de um documento. Contudo, a variação de cobertura entre as diversas janelas foi relativamente grande, já que algumas descrições (36) não possuíam *keywords* enquanto outras (17) possuíam 100% delas. Esses resultados indicam que os usuários do Google têm à sua disposição descrições deficientes dos documentos de uma coleção, em relação à reprodução de suas *keywords*. Um outro dado relevante foi a ausência de *keywords* em 36 janelas, o que corresponde a 29% do total analisado. Isso mostra que um número significativo de janelas não conseguiu cobrir o conteúdo considerado relevante pelo autor. As outras 88 janelas continham pelo menos uma *keyword*. A distribuição da cobertura das janelas recuperadas para todas as consultas é apresentada na Figura 3.

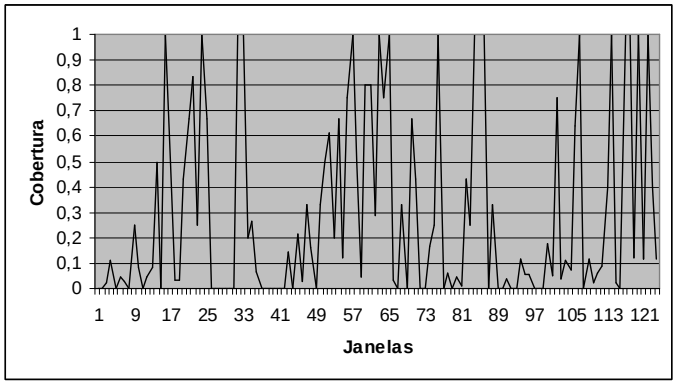


Figura 3. Distribuição da medida de cobertura nas janelas

A precisão da coleção recuperada variou entre 0 e 0.8, sendo seu valor médio bastante baixo (0.15). Esse baixo valor tem implicação direta no *overhead cognitivo* do usuário: sua decisão sobre a relevância do documento para sua consulta não pode ser tomada somente a partir do conteúdo das descrições. Da mesma forma que a cobertura, um número considerável de janelas (cerca de 29%) teve um valor de precisão nulo. A distribuição dos valores de precisão das janelas é vista na Figura 4.

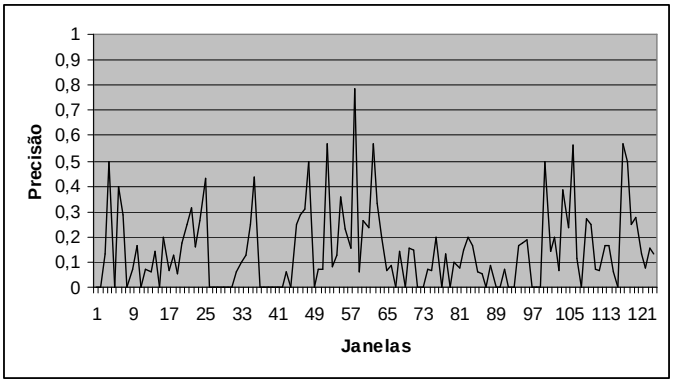


Figura 4. Distribuição da medida de precisão nas janelas

Esses valores médios de precisão e cobertura levam, certamente, a um baixo valor da *f-measure* médio da coleção, de 15% (variando de 0 a 0,73, aproximadamente), o qual evidencia a baixa eficácia do Google na geração de suas descrições.

A Tabela 2 traz os valores da *f-measure* considerando as janelas agrupadas por consulta. Esses resultados mostram que a qualidade de uma janela é muito mais dependente dos termos utilizados pelo usuário para formular a sua consulta do que da sua especificidade e estrutura (com uso de operadores booleanos, por exemplo), além de ser dependente do processo de casamento de padrão entre termos e documentos realizado pelo mecanismo de busca. A estrutura das consultas parece não ter influenciado diretamente a qualidade das descrições, já que consultas com mesma estrutura (mesma quantidade de termos e operadores booleanos) obtiveram valores de *f-measure* bem diferentes, por exemplo, as consultas 11 e 20 (Tabela 2). Casos similares,

em termos de especificidade dos termos da consulta, podem ser vistos comparando os valores da *f-measure* (Tabela 2) para consultas semelhantes, que obtiveram resultados bem diferentes, por exemplo, as consultas 3 e 27, ou ainda as consultas 2 e 14.

Tabela 2. Resultados agrupados por consulta

Consulta	*F-measure*
1. download AND musicas AND mp3	0,44
2. mp3	0,35
3. jogo AND fifa NOT www.fifa.com AND EA AND Sports	0,34
4. astronomia AND "Sistema Solar" AND planetas	0,28
5. laço AND programação AND repetição	0,26
6. recursos AND gratuitos AND para AND webmasters	0,24
7. livraria AND virtual	0,24
8. cartões AND virtuais	0,23
9. animais AND ameaçados AND de AND extinção	0,22
10. curso AND de AND inglês AND gratuito	0,22
11. doenças AND respiratórias	0,20
12. receita AND federal	0,19
13. Roberto AND Carlos AND cantor NOT jogador	0,17
14. dinossauros	0,12
15. poluição AND do AND ar	0,11
16. tutorial AND de AND prolog	0,10
17. informações AND divórcio NOT europa	0,09
18. informações AND viagem AND "porto seguro" AND bahia	0,07
19. bibliografia AND "glauber rocha"	0,07
20. namoro AND virtual	0,06
21. receitas AND doces AND salgados AND carnes AND massas	0,06
22. medicina AND alternativa NOT livro NOT livros	0,05
23. "Elba Ramalho" AND cantora	0,03
24. musical AND o AND fantasma AND da AND opera	0,03
25. dificuldades AND terceira AND idade AND internet	-
26. redes AND neurais AND artificiais	-
27. sorriso AND monalisa NOT filme NOT columbia NOT dvd NOT cd	-

Esses resultados corroboram aqueles obtidos por Jansen (2001), que investigou o efeito da estrutura das consultas nos resultados recuperados por sistemas de busca na *Web*. Seu experimento envolveu 15 consultas simples (sem qualquer recurso avançado, como o uso de operadores booleanos ou pesquisa por frase exata) submetidas a cinco mecanismos de busca: Alta Vista[2], Excite[3], FAST Search[4], Infoseek[5] e Northern Light[6]. Posteriormente, essas 15 consultas foram modificadas utilizando os operadores

[2] www.altavista.com
[3] www.excite.com
[4] www.alltheweb.com
[5] infoseek.go.com
[6] www.northernlight.com

permitidos por cada um dos mecanismos de busca, dando origem a 210 novas consultas que foram submetidas a cada um desses mecanismos. Comparando os resultados, Jansen concluiu que modificar a estrutura da consulta aparentemente tem um impacto muito pequeno nos resultados recuperados, pois elas trazem em média apenas 2,7 resultados diferentes dos que haviam sido apresentados pelas consultas simples.

Cabe destacar que esses resultados não levam em consideração a qualidade da consulta. Muitas vezes, resultados considerados ruins podem decorrer de consultas mal formuladas, que podem resultar na incapacidade de recuperação de material relevante. Essa situação torna-se ainda mais evidente quando as consultas não são suficientes para caracterizar de maneira precisa o contexto de busca desejado.

Pelo próprio caráter geral da Internet, supõe-se que qualquer internauta, com qualquer grau de proficiência, seja um usuário competente. Mas como o nosso estudo teve como foco o usuário típico, que geralmente não faz nenhum tipo de refinamento de sua consulta a fim de garantir sua qualidade (SPINK et al., 2000; WOLFRAM et al., 2001), esses aspectos não foram aqui considerados. Assim, uma descrição produzida automaticamente, quando não informativa, foi considerada ruim, na maioria dos casos. Considerando os resultados descritos anteriormente, na próxima seção, apresentamos algumas considerações sobre a análise de corpus e a solução proposta para lidar com os problemas apontados.

2.4 Considerações sobre a análise de corpus

Uma das formas de melhorar a resposta dos mecanismos de busca, fazendo com que eles superem os problemas apontados, é produzir resultados mais claros e coerentes; outra é melhorar a consulta através de ferramentas de refinamento mais sofisticadas.

Nosso objetivo, nesse trabalho, foi o de melhorar as descrições recuperadas pelo mecanismo de busca.

O Google mostra diferentes descrições para um mesmo documento dependendo da forma como ele é recuperado. Essas descrições dependem diretamente do termo de consulta utilizado pelo usuário. Como existe um número grande de termos de consulta que são usados para recuperar um mesmo documento, o número de possíveis descrições geradas para esse documento é igualmente considerável. Assim, nem sempre as descrições geradas pelo mecanismo de busca trarão os principais tópicos de um documento, já que a necessidade de um casamento entre a consulta do usuário e o conteúdo do documento (*keyword in context*[7]) pode trazer nas descrições tópicos secundários ou até mesmo irrelevantes. Para a geração da descrição, o mecanismo de busca não considera as *keywords* ou tópicos ressaltados pelo autor do documento em sua estrutura HTML. Estas, quando aparecem na descrição, são provenientes dos termos da consulta ou fazem parte dos segmentos textuais recuperados referentes a cada um deles. Isso ajuda a explicar a baixa presença de *keywords* (cerca de 29%, segundo nossa investigação) nas janelas que descrevem os documentos. Outro fator limitante à presença de um número maior de *keywords* é o tamanho destinado às descrições no Google: são aproximadamente 70 caracteres para o título de uma página e cerca de 150 caracteres (ou 22 palavras) para o texto da descrição, conforme os dados da nossa coleção.

Como já mencionado, o número de respostas exibidas por qualquer mecanismo de busca é muito grande se comparado ao número que realmente interessa ao usuário. Além disso, o número médio de respostas que um usuário qualquer considera relevantes não ultrapassa uma página (ou seja, aproximadamente 10 descrições produzidas

[7] Representação onde os termos da consulta do usuário são mostrados em destaque (geralmente em negrito) circundados por palavras que procuram recuperar o contexto destes termos no documento-fonte.

automaticamente). A análise descrita neste capítulo confirma esses dados: o conteúdo das janelas cobre pouco o conteúdo dos documentos e a precisão da descrição dos documentos também é questionável, levando ao baixo desempenho do mecanismo de busca. Se considerada fixa a consulta, mesmo que o usuário não tenha condições de garantir sua qualidade, o mecanismo de busca deveria gerar descrições de documentos que refletissem o conteúdo adequado. O que os resultados evidenciam é que sequer uma boa correspondência com a lista de tópicos relevantes produzida pelo autor do documento é garantida.

Esse estudo de casos evidenciou a necessidade de melhorar a apresentação de resultados de mecanismos de busca, considerando métodos que possam gerar descrições mais informativas.

A nossa proposta para lidar com esta questão é focar a composição de descrições, não na consulta do usuário, mas em termos recuperados da estrutura HTML, que podem não fazer parte do texto visível, mas que contêm informações importantes e úteis. Exemplos deste tipo de texto são os textos incluídos nas meta-etiquetas para descrição e palavras-chave. A estrutura HTML é rica e fornece as palavras-chave designadas pelo autor para descrever seu conteúdo ou assunto; além disso, propomos realizar uma análise conceitual, por meio de uma ontologia, para determinar os tópicos principais sobre os quais a página versa, a fim de preservá-los durante a geração dos sumários.

Nesse contexto, a nossa proposta de usar sumários como descrições de páginas se apresenta como uma alternativa para auxiliar os usuários destes mecanismos. No próximo capítulo, apresentaremos alguns trabalhos correlatos que fornecem indícios sobre como a estrutura HTML e o processamento semântico, visando à detecção de conceitos, podem colaborar com a SA de documentos *Web*.

3 Utilização da marcação HTML e conhecimento ontológico para o processamento de documentos

Conforme acusamos no estudo de casos do capítulo anterior, a apresentação de resultados de buscas, utilizando descrições que tenham como foco apenas os termos da consulta dos usuários, gera descrições que não são suficientemente informativas, aumentando o esforço do usuário em sua seleção de documentos relevantes. Sendo assim, justifica-se o desenvolvimento de ferramentas que diminuam esse esforço durante a seleção de documentos. A nossa proposta para lidar com esse problema foi a utilização da estrutura HTML e de processamento semântico para detecção de tópicos em documentos *Web*, a fim de produzir sumários que preservem estes tópicos.

A seguir, apresentaremos o formalismo HTML e alguns trabalhos correlatos à nossa proposta de pesquisa, que consideram a estrutura HTML durante o processo de sumarização, bem como trabalhos que fazem uso de conhecimento ontológico para processamento semântico de documentos.

3.1 O formalismo HTML

A HTML é uma linguagem que permite a representação semi-estruturada de dados, mas ela é, sobretudo, orientada à apresentação, isto é, ao modo como as informações são visualizadas. Essas informações podem ser de diversos tipos: textos, gráficos, vídeos, sons e ainda outros itens multimídia. Sua finalidade básica é a criação, para a *Web*, de documentos hipertexto que nada mais são do que uma coleção de caracteres e etiquetas de marcação. Esta linguagem foi utilizada na *Web* devido às suas características, como por exemplo: portabilidade (o código-fonte é escrito usando apenas a tabela ASCII[8] e pode funcionar em qualquer sistema ou plataforma),

8 ASCII (American Standard Code for Information Interchange) é um conjunto de códigos usados para o computador representar números, letras, pontuação e outros caracteres

flexibilidade (com a HTML, pode-se obter a formatação de textos constituídos de forma organizada, por intermédio de agrupamento de etiquetas), tamanho de código reduzido (seu código em comparação a outras linguagens de marcação é pequeno, possuindo em torno de 100 etiquetas, e ocupa pouco espaço, o que é altamente favorável para a transferência de dados através da *Web*).

Grande parte de suas etiquetas permite explorar esse modo usando fontes ou estilos que ressaltam visualmente o segmento mais relevante (por exemplo, o uso de negrito). Adicionalmente, é possível também usar marcadores HTML para estruturar o texto de modo a fazê-lo mais proeminente. O uso de informações de *layout* é explorado com esse fim, conforme veremos na próxima seção, citando algumas propostas específicas do uso de HTML em sistemas de SA.

A especificação 4.01 do HTML possui cerca de 100 etiquetas, das quais algumas foram consideradas como potencialmente úteis para processos de sumarização automática, conforme veremos no Capítulo 5. Essas etiquetas estão hierarquicamente organizadas segundo a estrutura de um documento conforme pode ser visto na Figura 5.

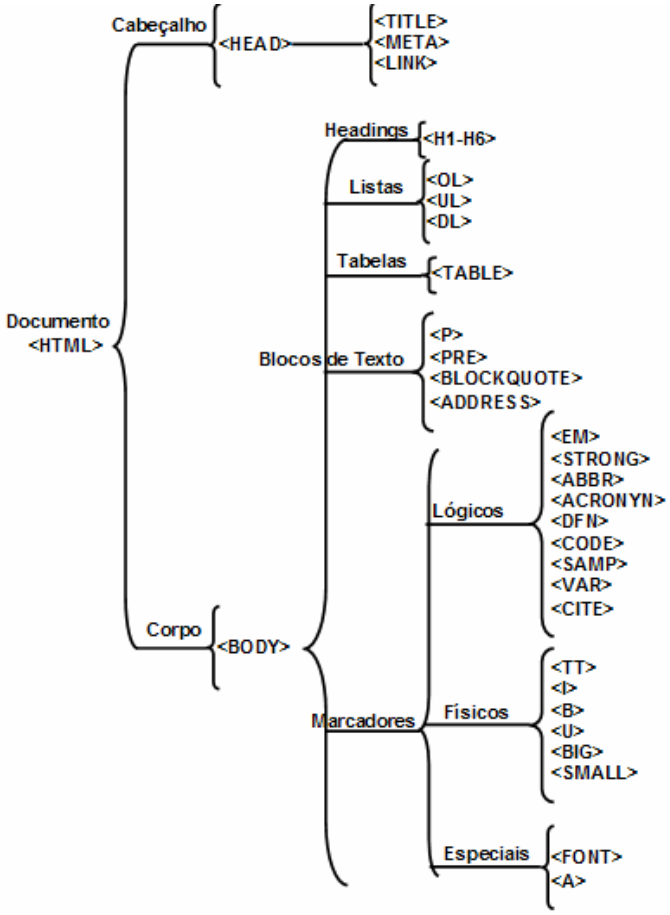


Figura 5. Organização hierárquica das etiquetas HTML

A sumarização automática de documentos *Web* é interessante por dois motivos principais:

a) a maior parte dos documentos *Web* estão em formato HTML (WOODRUFF et al., 1996);

b) o código HTML é rico em informações que podem ser exploradas para tarefas de sumarização automática.

Valer-se da estrutura HTML no processo de SA de documentos *Web* é uma estratégia interessante, porém insuficiente em muitos casos, já que a estrutura textual, isto é, o conteúdo dos documentos, é igualmente relevante. Desta forma, aliar o uso de informações provenientes das etiquetas HTML a recursos que permitam detectar a conectividade semântica entre elementos textuais dos documentos *Web*, identificando,

por exemplo, seus tópicos principais para guiar a composição dos sumários, pode trazer ganhos para processos de sumarização, proporcionando melhoria dos resultados dos mecanismos de busca. Diante deste fato, nas próximas seções, apresentaremos alguns trabalhos que exploram o potencial da estrutura de marcação HTML em tarefas de sumarização bem como trabalhos que fazem uso de conhecimento ontológico para identificação de tópicos em documentos.

3.1.1 Sistemas de SA que usam o potencial de marcação HTML

Nesta seção, descrevemos as principais características de alguns sistemas de SA de documentos *Web* relacionados à nossa proposta.

3.1.1.1 O Sistema SweSum

O *SweSum* (DALIANIS, 2000) é um sumarizador automático de textos jornalísticos desenvolvido inicialmente para o sueco, mas hoje disponível também para outros idiomas, como o norueguês, dinamarquês, espanhol, inglês, francês, alemão, persa e grego. Ele faz uso de heurísticas de extração baseadas em informações estatísticas e lingüísticas, gerando sumários de textos jornalísticos em formato HTML. O processo de sumarização se dá em três passos principais (Figura 6): pré-processamento e extração de palavras-chave, ranqueamento de sentenças e composição do extrato.

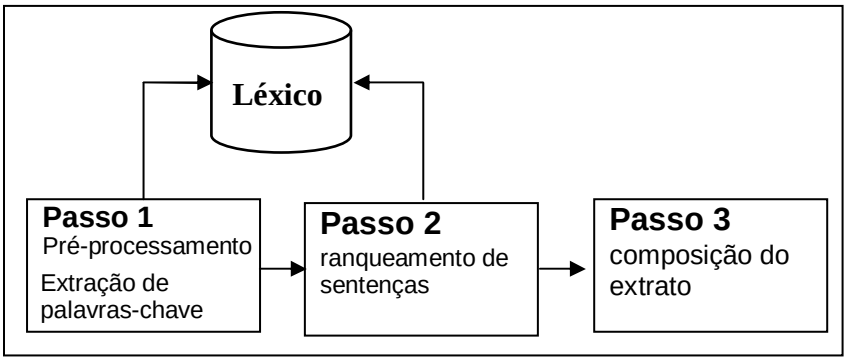


Figura 6. O processo de sumarização no SweSum

No primeiro passo, as sentenças do documento são delimitadas considerando os símbolos de pontuação usuais e a etiqueta
 que indica quebras de linha. Além das sentenças delimitadas, o sistema armazena a posição (linha) de cada uma delas dentro do código-fonte HTML. Após essa delimitação, as palavras-chave do documento são extraídas. São consideradas palavras-chave os substantivos, adjetivos, advérbios ou quaisquer palavras definidas pelo usuário do sistema. Para cada palavra-chave é computada sua freqüência no documento e, para isso, é usado um léxico. A freqüência das palavras-chave é usada posteriormente no processo de ranqueamento de sentenças.

No segundo passo, o sistema determina a importância das sentenças do texto e os seus respectivos ranques, atribuindo-lhes um escore. O escore de cada sentença depende de sua posição no texto e da freqüência das palavras que ela contém. Para isso são levados em consideração os seguintes critérios:

- Inclusão incondicional da primeira sentença do documento: é sempre incluída no sumário. A hipótese é que em textos jornalísticos a primeira sentença é sempre importante e deve fazer parte do sumário.
- Critério geográfico: as sentenças são ranqueadas de acordo com sua posição no documento. Por se tratar de documentos HTML, a posição de cada sentença corresponde à linha do código-fonte HTML onde se encontra a sentença. Considerando uma sentença que no código-fonte

ocorre após duas etiquetas quaisquer como <HTML> e <TITLE>, por exemplo, ela teria como posição o valor 3. O escore de posição é calculado conforme a fórmula seguinte:

$$Escore\ posição = \left(\frac{1}{linha\ da\ sentença} \right) * 10$$

De acordo com esta fórmula as sentenças que aparecem primeiro, ou seja, pertencem às primeiras linhas, apresentarão maior escore de posição.

- Premiação para a presença de valores numéricos: sempre que um número é encontrado em uma sentença, seu escore é acrescido de uma unidade. A hipótese é que os dados numéricos são importantes e as sentenças que trazem esses tipos de dados devem ser privilegiadas.
- Escore fixo para textos em negrito: o *SweSum* associa um escore de 100 para as sentenças que contêm textos em negrito; que são identificados no código HTML pela etiqueta <B>. A associação do valor 100 ao escore não é justificada pelo autor, intuitivamente isso garante que sentenças com textos em negrito tenham uma alta prioridade sobre as outras.
- Premiação de sentenças pela presença de palavras-chave: as sentenças são premiadas de acordo com freqüência das palavras-chave que elas contêm.

Os critérios definidos anteriormente são usados para determinar o escore de cada sentença. Ainda no segundo passo do processo de sumarização, um escore individual de cada palavra no documento é calculado e adicionado ao escore da sentença, conforme as fórmulas seguintes:

$$EscorePalavra = (FrequênciaPalavra) * ConstantePalavraChave$$

$$EscoreSentença = \sum EscorePalavra$$

em que *ConstantePalavraChave* tem o valor padrão de 0,3333. A escolha desse valor não é justificada no trabalho pelo autor.

Para evitar distorções no ranqueamento, como a introduzida por sentenças muito longas, todos os escores sentenciais são normalizados; para isso, o escore de cada sentença, calculado conforme a fórmula anterior, é multiplicado por um fator denominado ASL (*Average Sentence Length*) e dividido pelo número total de palavras na sentença.

$$EscoreSentença = \frac{ASL \ * \ EscoreSentença}{NumeroPalavrasNaSentença}$$

$$ASL = \frac{QuantidadePalavras}{QuantidadeLinhas}$$

em que *QuantidadePalavra* corresponde ao número total de palavras no documento e *QuantidadeLinhas* corresponde ao número de linhas do código-fonte.

O terceiro passo de composição do extrato, simplesmente seleciona as sentenças de maiores escores para compô-lo. O sistema também inclui no extrato todas as linhas de código HTML restantes a fim de manter a estrutura do documento no momento de apresentá-lo ao usuário.

O *SweSum* passou por um processo de avaliação automática (HASSEL, 2003) e manual. Na avaliação automática, o objetivo foi medir a precisão do sistema, verificando quantas das sentenças do extrato automático coincidem com aquelas contidas em um extrato ideal. Um corpus de extratos ideais de textos jornalísticos foi criado. Os extratos ideais foram manualmente construídos por especialistas que, após lerem os textos-fonte correspondentes, selecionaram as sentenças consideradas mais

relevantes para incluí-las nos extratos. Os extratos ideais foram construídos a partir de textos, com tamanhos que variam entre 5 e 500 linhas e com um número médio de 193 palavras. Na média, os extratos corresponderam a 37% do tamanho original dos textos-fonte.

A avaliação automática consistiu da comparação de 100 extratos gerados pelo sistema, sentença a sentença, com o respectivo extrato ideal, ou seja, para cada extrato obtido, checavam-se quantas de suas sentenças estavam presentes no extrato ideal correspondente. Os resultados obtidos indicaram uma precisão de 57.2% para o sistema.

Na avaliação manual, o objetivo foi medir a informatividade e a coerência dos extratos gerados pela ferramenta. Foi considerado um processo de redução gradual do tamanho dos extratos, aumentando a taxa de compressão, a fim de determinar até que ponto os documentos poderiam ser sumarizados mantendo a coerência e a informatividade. O objetivo foi, portanto, identificar até que valores se poderia aumentar a taxa de compressão sem prejudicar a coerência e a informatividade do extrato. Noventa extratos foram gerados usando o *SweSum*, usando taxas de compressão variando entre 10% e 90%. Para cada extrato, um juiz verificou se havia existido quebra de coerência ou perda de informatividade e anotou qual era a taxa de compressão quando isso ocorreu. Os resultados obtidos indicaram que, na média, os extratos mantinham-se coerentes com taxas de compressão de até 74% e a informatividade era mantida para taxas de compressão de até 69%. Usando taxas de compressão maiores, os extratos começam a apresentar problemas de coerência e informatividade. Esses resultados podem ser considerados ruins, sobretudo se considerarmos que a aplicação de extratos na apresentação de resultados de buscas exige altas taxas de compressão.

3.1.1.2 O Sistema WWW Site Summarization

O sistema *WWW Site Summarization (W3SS)* (ZHANG et al., 2004) é uma ferramenta de sumarização de *websites* que tem o mesmo objetivo adotado em nosso trabalho: o de fornecer ao usuário informação suficiente para que ele determine a relevância do site para seus interesses, sem necessitar visitá-lo antes. A abordagem proposta se baseia na aprendizagem de máquina e no processamento de língua natural para a construção de um conjunto de regras que posteriormente são usadas para a produção de sumários.

Os sumários gerados pelo sistema são compostos de três elementos principais: sentenças completas extraídas de parágrafos narrativos, palavras-chave e termos-chave. Todo o processo de seleção destes três elementos para composição do sumário foi feito com base em técnicas de aprendizagem supervisionada, que geraram regras que são aplicadas no momento de sumarizar um *website*. A geração das regras constitui um passo prévio à sumarização e esta visa garantir que somente aqueles elementos que possuíam determinadas características pudessem ser incluídos no sumário. As regras garantiram que as sentenças fossem extraídas somente daqueles parágrafos considerados narrativos. Desta forma, evitou-se que os sumários apresentassem palavras ou frases isoladas, sentenças muito curtas ou incoerentes. Um parágrafo foi considerado narrativo se o seu texto fosse coerente. Isto foi feito em dois estágios: primeiro, foram definidos alguns critérios para determinar se um parágrafo era longo o suficiente para fornecer as sentenças a serem incluídas no sumário; segundo, critérios adicionais foram definidos para classificar os parágrafos longos em narrativos ou não-narrativos. Somente os parágrafos narrativos foram usados na geração dos sumários. Essa classificação foi feita com base em técnicas de aprendizagem supervisionada, que geraram regras que garantiram a seleção de parágrafos com as características descritas.

Para geração das regras foi utilizado como corpus de treino 60.000 páginas de 60 sites do *DMOZ* (www.dmoz.org). Deste corpus foram extraídos 700 parágrafos (de 100 páginas diferentes) classificados manualmente e intuitivamente em curtos ou longos para servirem como dados de treinamento. Além dessa classificação manual, foram anotadas as seguintes características de cada parágrafo: número de caracteres do parágrafo (incluindo símbolos de pontuação), número de palavras, e número de caracteres nas palavras (excluindo símbolos de pontuação). A partir destes dados de treinamento foram geradas as regras (usando o programa C5.0) que permitiram a classificação dos parágrafos em curtos ou longos.

Uma vez definidas as regras que classificavam os parágrafos com relação ao seu tamanho, um procedimento similar foi usado para classificá-los em narrativos ou não-narrativos. Seguindo a mesma metodologia anterior, 3243 parágrafos longos (recuperados usando as regras anteriores) foram manualmente classificados em narrativos ou não-narrativos. As seguintes características de cada parágrafo também foram consideradas para a definição das regras: classes gramaticais das palavras, número de caracteres e de palavras no parágrafo.

A seleção das palavras-chave e termos-chave incluídos no sumário também foi baseada em regras. Para definição das regras, um conjunto de 5454 frases-chave foi extraído do corpus de treino *DMOZ*. Para cada palavra-chave e termo-chave, as seguintes características foram coletadas para formar o conjunto de treinamento: freqüência nos parágrafos narrativos, freqüência nos textos âncoras (aqueles delimitados pela etiqueta <A>), freqüência nos textos especiais (marcados em itálico (<I>) ou negrito (<B>)) e classe gramatical. Uma vez definidas as regras para seleção destes elementos, a sumarização é realizada em seis passos principais, como mostra a Figura 7.

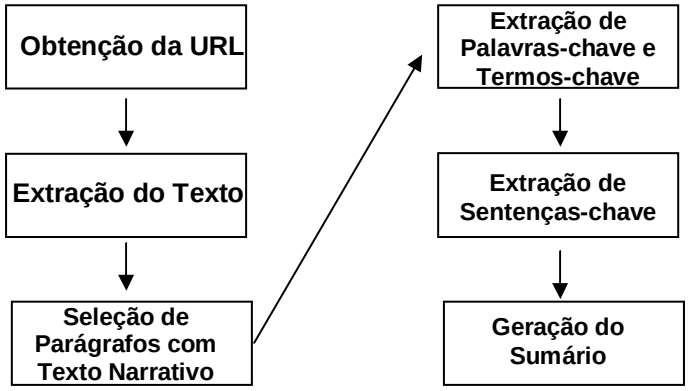


Figura 7. O processo de sumarização no W3SS

Primeiro, após a obtenção da URL fornecida pelo usuário contendo o endereço do *website* a ser sumarizado, para cada *website* o sistema coleta até 1.000 páginas a partir de sua página inicial. Segundo Zhang et al. (2004), este número garante um bom balanceamento com relação à velocidade de recuperação de páginas e informatividade dos sumários gerados. No segundo passo, somente o texto é extraído dessas páginas e particionado em parágrafos.

No terceiro passo, parágrafos são selecionados com base em seu tamanho e em suas características textuais. Somente os parágrafos narrativos são selecionados de acordo com as regras definidas na fase de treinamento.

O quarto passo da Figura 7 envolve a extração de palavras-chave e termos-chave dos documentos para inclusão no sumário. Palavras-chave são compostas por uma única palavra enquanto os termos-chave são compostos por duas palavras. A determinação das palavras e termos-chave é também baseada em regras, conforme relatamos anteriormente.

No quinto passo, é feita a extração de sentenças-chave contidas nos parágrafos narrativos. Uma vez identificadas as palavras-chave e termos-chave, as sentenças mais significativas para gerar os sumários são aquelas que contêm esses elementos. Para cada sentença é associado um fator de significância, que corresponde ao peso máximo

dos *clusters* identificados na sentença. Um *cluster C* é uma seqüência de palavras consecutivas em uma sentença, observando que: (1) a sentença inicia e termina em uma palavra-chave ou termo-chave; (2) menos de duas palavras (que não sejam palavras-chave ou termos-chave) devem separar duas palavras-chave ou termos-chave vizinhos dentro da sentença. O peso de cada *cluster* é computado somando o peso de todas as palavras-chave e termos-chave dentro do *cluster* e dividindo esta soma pelo número de palavras do *cluster*. O peso de uma palavra-chave ou termo-chave i é definido da seguinte maneira:

$$w_i = \frac{f_i}{\sum_{i=1}^{100} fi}$$

em que f_i é a freqüência do termo no *website*.

Finalmente, no último passo, o sumário é gerado consistindo de 25 palavras-chaves, 10 termos-chave e 5 sentenças-chave, selecionados com base nos critérios definidos anteriormente.

O sistema W3SS foi avaliado manualmente com o objetivo de determinar quão informativos eram os seus sumários. Para isso, os sumários foram comparados aos sumários manuais do *DMOZ* e aos métodos de "visualização da página principal do site" (*home page browsing*) e "visita ao site com tempo limitado" (*time-limited site browsing)*, em uma tarefa que consistiu em responder a um conjunto de perguntas sobre um determinado site. Essas perguntas envolviam informações relativas ao site em questão, como por exemplo, nome da instituição, missão da instituição, tópicos principais do site etc. Para a realização do experimento, 20 sites foram considerados e 20 avaliadores foram responsáveis por responder às perguntas, tendo suas respostas posteriormente graduadas por um juiz, numa escala de 0 a 20, conforme elas estivessem

corretas ou não, em que 20 era o valor atribuído para o máximo de respostas corretas. Os resultados obtidos são mostrados na Tabela 3.

Tabela 3. Resultados da avaliação do W3SS

Sistema/ Método	Média de pontos
DMOZ	15.3
W3SS	15.0
time-limited site browsing	13.4
home page browsing	12.7

Apesar de apresentar um desempenho ligeiramente superior ao *W3SS*, o *DMOZ* apresenta uma variação maior entre as pontuações individuais obtidas pelos avaliadores, 11.0 (para a pior nota) e 19.2 (para a maior nota). A média de pontos do *W3SS* é mais constante neste aspecto. Para 11 sites a média de pontos do *W3SS* foi superior à obtida pelo *DMOZ*, para 8 sites foi inferior e para 1 site foi equivalente, o que mostra que a qualidade de seus sumários varia de site para site. De modo geral, seu desempenho também foi superior àqueles obtidos pelos métodos *home page browsing* e *time-limited site browsing.*

De acordo com Zhang et al. (2004), os resultados mostram que os sumários produzidos pelo sistema, embora não sejam coerentes, são tão informativos quanto aqueles produzidos por humanos. Assim, esses sumários automáticos podem facilmente ser convertidos em um texto coerente, por editores humanos, sem a necessidade de acessar o site. A performance dos métodos propostos depende da disponibilidade de conteúdo narrativo suficiente no *website* e da disponibilidade de informações explícitas, no conteúdo narrativo, que descrevam o site.

3.1.1.3 O Sistema InCommonSense

O sistema *InCommonSense* (PARIS; AMITAY, 2000) propõe uma abordagem baseada em contexto que fornece a base para construir o sumário de um site a partir de descrições manuais sobre o site em foco recuperadas de outros sites. O sistema se baseia no fato de que as pessoas descrevem, comentam e relatam outras páginas *Web* no contexto de seus documentos, ou seja, para gerar um sumário de um documento A, o sistema busca por documentos D_i que apontem para A. Posteriormente, recupera em D_i o texto (geralmente um parágrafo) visualmente próximo ao *link* que aponta para A, considerando-o como uma descrição candidata de A. Esse processo é visto na Figura 8.

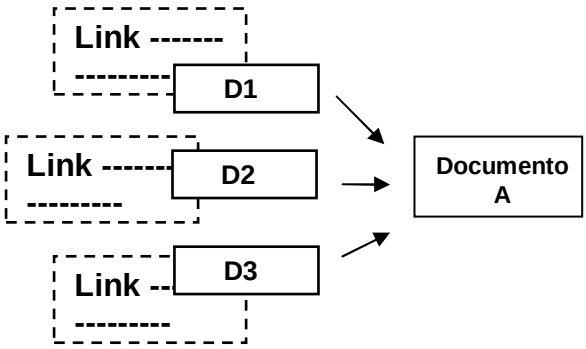


Figura 8. O processo de sumarização no InCommonSense

No *InCommonSense* as descrições candidatas são coletadas buscando-se em mecanismos de busca *links* para o documento que se deseja sumarizar (usando um consulta do tipo – "link: URL"). O sistema, então, recupera as páginas e as analisa para encontrar marcas que possam indicar parágrafos. Os parágrafos correspondem aos segmentos de texto, marcados no código-fonte pelas etiquetas HTML (<P>, <LI>, <HR>, <UL>, <DD>, <DT>, <IMG>, <H1-H6>, <BLOCKQUOTE>). Se o *link* da página analisada tiver um parágrafo associado, ambos são recuperados. Na sua versão atual, o *InCommonSense* recupera até 220 documentos relacionados à URL do

documento a ser sumarizado usando os mecanismos de busca Google, HotBot, AltaVista e Infoseek.

Após recuperar as descrições candidatas de uma página, o sistema deve decidir qual delas é a melhor para exibí-la como sumário. Com esta finalidade, foi desenvolvido um mecanismo de filtragem que analisa todas as descrições candidatas e escolhe a melhor com base em um conjunto de regras. As regras foram definidas anteriormente, usando a ferramenta C4.5 e um conjunto de dados de treinamento. Os dados de treinamento corresponderam a 252 descrições, classificadas manualmente em boas ou ruins, juntamente com um conjunto de informações sobre cada uma delas. As informações e as hipóteses para recolhê-las foram as seguintes:

1. Número de palavras: estima o tamanho de uma descrição. Boas descrições geralmente são curtas.

2. Número de vírgulas: indica enumeração e sentenças coordenadas que são construções recomendadas pelo *World Wide Web Consortium* para escrita de descrições.

3. Número de travessões: é geralmente uma indicação de que o texto é sobre o item que o precede, na maior parte dos casos um apontador para a página descrita no texto.

4. Número de exclamações: ajuda a identificar sentenças completas na descrição. Boas descrições são formadas por sentenças completas.

5. Número de pronomes pessoais: indica opiniões. Geralmente boas descrições devem incluir algum tipo de opinião sobre o site.

6. Número de acrônimos: acrônimos são muito comuns em descrições curtas, eles ajudam a detectar a brevidade do texto.

7. Freqüência da palavra "*about*": este termo ajuda a detectar trechos que discorrem sobre o site.

8. Freqüência de palavras que expressam opinião: ajuda a detectar opiniões e subjetividade nas descrições.

9. Texto inicia com travessão (sim, não): idem ao item 3.

10. Texto inicia com "*is*" (sim, não): se o texto do *link* é parte de uma sentença e é um sintagma nominal, é provável que uma descrição comece logo após este verbo.

11. Texto inicia com letra maiúscula (sim, não): quando as palavras estão em letras maiúsculas, observa-se que o texto geralmente refere-se a um título e não a uma descrição.

12. Texto inicia com ":" (sim, não): idem ao item 3.

13. Texto termina com interrogação (sim, não): idem ao item 4.

14. Texto termina com exclamação (sim, não): idem ao item 4.

O *InCommonSense* foi avaliado extrinsecamente em um contexto de busca de documentos com o objetivo de determinar a utilidade dos sumários produzidos. A avaliação envolveu a comparação entre os sumários produzidos pelo sistema e duas técnicas comumente usadas por mecanismos de busca. A primeira (*Alta Vista Style*) consistiu na extração das *top X* palavras do documento e a segunda (*Google Style*) consistiu em recuperar termos da consulta junto com algumas palavras à sua volta (contexto). Para essa avaliação foi montado um corpus de teste composto por cinco conjuntos de sumários, cada conjunto gerado para uma consulta distinta *("albert einstein","genome project","origami","wildlife on the web","world time zones"*). Cada conjunto foi composto por nove *links* de sites e seus respectivos sumários, representando uma página de resultados de busca, totalizando 135 sumários. Quatro questões referentes à utilidade das descrições foram elaboradas manualmente. A avaliação foi conduzida da seguinte maneira: ao juiz do experimento de avaliação, foi apresentada uma tarefa inicial de busca sobre certo tópico (indicado por uma consulta).

Um dos conjuntos de *links* e sumários referentes à consulta foi exibido para que o juiz pudesse escolher um dos *links*. Posteriormente, o juiz respondeu às questões graduando-as em uma escala de 0 a 7. No total, 738 juizes participaram do experimento realizando as tarefas propostas, tendo por base os sumários produzidos. Considerando a escala de 0 a 7 para a questão "*how easy it was to find the information needed*" o *InCommonSense* obteve uma média de 4.71, o *Google* 4.13 e A*lta Vista* 4.14. Em termos de utilidade do sistema, a saída textual do *InCommonSense* foi superior à saída dos sistemas comerciais testados.

A seguir, apresentaremos as principais características dos sistemas descritos anteriormente, sintetizando-as a fim de mostrar como elas contribuíram para a execução do nosso projeto de mestrado.

3.1.1.4 Características gerais no processamento de documentos *Web*

Todos os trabalhos descritos anteriormente apresentam características que contribuíram em maior ou menor grau para o desenvolvimento deste projeto de mestrado, já que eles se propõem exatamente a gerar sumários de documentos *Web*.

A primeira e mais evidente contribuição referiu-se à explicitação das etiquetas HTML que são sinalizadoras de informações relevantes e do tipo de informações que elas ressaltam. Neste aspecto, o *SweSum* (DALIANIS, 2000), o *W3SS* (ZHANG et al., 2004) e o *InCommonSense* (PARIS; AMITAY, 2000) remetem a um conjunto inicial de etiquetas HTML que foram consideradas para geração de nossas heurísticas. Cabe destacar que, além das etiquetas HTML utilizadas nos trabalhos citados, ampliamos este conjunto, incluindo outras etiquetas a fim de tornar nossa proposta mais abrangente. Essa estratégia de utilizar HTML para sumarizar documentos foi explorada levando à implementação do sumarizador HTMLSUMM (relatada no próximo capítulo).

Considerações sobre associação de pesos a etiquetas HTML para ponderar a relevância da informação, como proposta no *SweSum*, também foram adotadas em nossa proposta. Ainda com relação ao *SweSum*, observamos algumas limitações que procuramos solucionar ao definirmos nosso projeto. Um problema claro da abordagem proposta é que, ao apresentar um extrato ao usuário por meio de uma versão condensada do documento original, todas as etiquetas HTML associadas às sentenças não incluídas no extrato são eliminadas, gerando um código-fonte com problemas, já que a ausência de algumas etiquetas pode gerar um documento que tem um aspecto "deformado", dificultando inclusive a análise do documento pelo usuário. Outra limitação aparente é a utilização apenas de uma única etiqueta HTML (<B>) no processo de sumarização. Outras etiquetas com funções similares, responsáveis por ressaltar informações relevantes, poderiam enriquecer a abordagem. Esse mesmo problema pode ser notado nos outros sistemas, daí a nossa opção por considerar um número maior de etiquetas. Também não foi feita nenhuma consideração sobre *stopwords*, como por exemplo, a remoção destas palavras, já que elas podem interferir no escore das sentenças. Procuramos superar essas limitações nos modelos que propusemos nesta dissertação, propondo um conjunto maior de etiquetas, além de optarmos pela apresentação de extratos como textos simples e não como versões condensadas de código HTML. Além disso, também tivemos a preocupação de desconsiderar as *stopwords* durante a sumarização.

O sistema *IncommonSense*, em particular, apresenta algumas etiquetas (LI, DD, DT) que não foram representativas em um levantamento prévio que fizemos sobre quais as etiquetas são mais comumente encontradas em documentos *Web* com conteúdos em português, mas que foram consideradas em nossa proposta, já que se demonstraram úteis na determinação de conteúdo relevante para este sistema. O *IncommonSense*

propõe uma sistemática de avaliação extrínseca que foi adotada para avaliar o nosso sumarizador ExtraWeb (conforme descreveremos no Capítulo 5). Adotamos esta sistemática de avaliação, justamente por ela explicitar a utilidade de sumários em um contexto de busca real, utilizando-os em um processo de tomada de decisão sobre relevância de documentos, da mesma forma que ocorre na interação entre internautas e mecanismos de busca da Internet.

De modo geral, todos esses trabalhos aplicam as informações providas pelas etiquetas HTML em conjunto com outras técnicas para realizar a SA de documentos. Isso evidencia que, apesar de sua utilidade, a linguagem HTML sozinha não é suficiente para garantir a geração de sumários de boa qualidade. Esse aspecto está relacionado, principalmente, à própria natureza da linguagem, já que ela é voltada sobretudo para a apresentação de dados e estruturação dos documentos. Essa característica faz com que seja necessário um processamento do conteúdo textual em um nível mais alto do que apenas no nível de estruturação da informação. Faz-se necessário que a informação também seja processada no seu nível semântico, a fim de agregar valor e fornecer subsídios que aprimorem processos de sumarização. Com o advento da *Web* Semântica (FENSEL et al., 2002) esta necessidade torna-se ainda mais evidente, já que muitas das limitações dos sistemas atuais residem na falta de tratamento semântico das informações contidas nos documentos *Web*. Esse fato levou-nos a propor uma estratégia de sumarização mista que explorasse tanto a estrutura quanto a semântica dos documentos, resultando na implementação do sumarizador ExtraWeb (relatada no Capítulo 5).

Outra característica evidenciada por essa revisão de trabalhos, refere-se à metodologia do desenvolvimento de sistemas de sumarização de documentos *Web*: ela é carregada de elementos empíricos, e, apesar das características gerais dos sistemas enumerados nesta seção, percebe-se que ainda não existe uma metodologia bem

sedimentada, que guie uma elaboração completamente padronizada, permitindo desenvolver sistemas de sumarização HTML com bons resultados. Em conseqüência disso, procuramos dar a nossa contribuição aprimorando e propondo uma metodologia, visando preencher este vácuo.

Na próxima seção, apresentaremos alguns trabalhos e evidências que ilustram como o processamento semântico de um texto pode ser feito com base em conhecimentos ontológicos e como isso pode representar ganhos para a SA.

3.2 Processamento semântico de conteúdo textual

O processamento semântico por meio de ontologias vem se tornando cada vez mais comum nos últimos anos, já que elas fornecem meios de representar e utilizar o conhecimento de mundo (ESTIVAL et al., 2004). Em aplicações que envolvam o processamento de conteúdo textual, esse conhecimento de mundo pode significar, por exemplo, entender sobre o que versa um texto. No caso específico da SA, as ontologias podem ser aplicadas para identificar em um texto-fonte seus tópicos principais a fim de subsidiar a seleção daquelas sentenças que são mais relevantes para expressá-los (LIN, 1995). Com o uso das ontologias é possível agregar valor à informação disponível, ou seja, associar um grau de utilidade às respostas geradas. Diante deste potencial, adotamos o uso de uma ontologia, como uma das nossas estratégias de sumarização, utilizando-a para a identificação de tópicos.

A acepção utilizada neste trabalho é aquela que denota uma ontologia como uma especificação explícita de uma conceitualização acerca de um domínio (GENESERETH; NILSSON, 1987; GRUBER, 1996). A conceitualização envolve, assim, a definição de uma coleção de conceitos que se assume existirem em um domínio, assim como os relacionamentos entre eles (GENESERETH; NILSSON, 1987). A conceitualização é expressa como uma representação do vocabulário

terminológico da ontologia e das relações entre esses termos (GRUBER, 1996). Guarino et al. (1994) indicam que a ontologia descreve uma taxonomia de palavras. Adicionalmente, Chandrasekaran et al. (1999) indicam que as ontologias são vocabulários de representação de conceitos, provendo termos potenciais para descreverem o conhecimento de um domínio.

3.2.1 A representação de conceitos ontológicos

Nesta dissertação, os conceitos ontológicos são utilizados para detectar os possíveis tópicos dos documentos, sendo utilizados como indicadores de informações relevantes. Um conjunto de conceitos é usado para formar a nossa base de dados ontológica. Dessa forma, nossa ontologia é formada por uma coleção de conceitos relacionados e estes, por sua vez, são representados por um conjunto de palavras. Os conceitos podem ser expressos através da língua natural, utilizando termos específicos, ou seja, palavras que, quando encontradas, indicam a presença do conceito (LOH, 2001). Desta forma, podem ser identificados através de técnicas que analisam o conteúdo textual dos documentos.

Com relação à sua estrutura, um conceito é composto de um identificador e de um conjunto de palavras que o descrevem (Figura 9), chamadas de descritores.

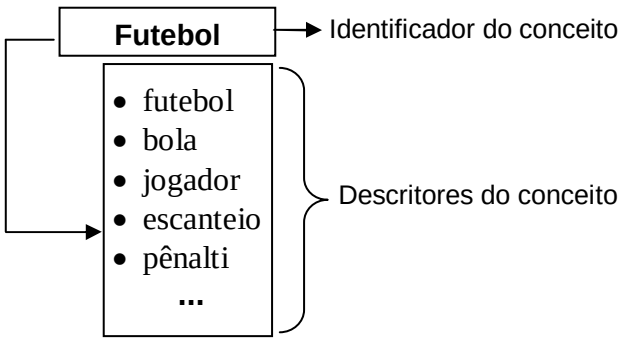


Figura 9. Estrutura de um conceito

O identificador é uma palavra da língua natural que dá a idéia geral do conceito (WIVES, 2004). Podem ser utilizados nomes de objetos, substantivos, verbos, etc., (“basquete”, “doença” e “jantar”, por exemplo).

Os descritores do conceito são palavras que sinalizam a presença do conceito. Para definir os descritores podem-se utilizar o próprio identificador e outras palavras relacionadas. Por exemplo, o conceito *Futebol* poderia ser expresso por palavras identificadoras como: futebol, pênalti, gol, escanteio, impedimento, etc.

As palavras descritoras podem estar relacionadas umas às outras de diferentes formas. As formas mais comuns de relacionamento são as sinonímias[9], hiponímias[10] e hiperonímias[11] (TIUN et al., 2001). A Figura 10 é um exemplo fictício de um conceito denominado *mandioca* cujos descritores foram definidos com base na relação de sinonímia.

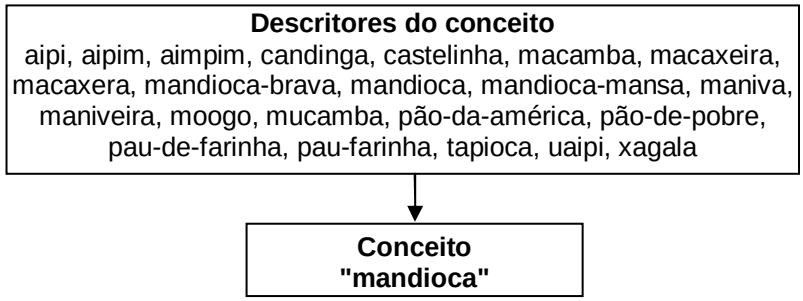


Figura 10. Descrição de um conceito

Além desses relacionamentos citados anteriormente, os descritores de conceitos também podem ser definidos por meio de variações morfológicas, tais como variações léxicas de gênero, número e grau, verbos e diferenças de grafia (ex.: acrobata/acróbata) (AGIRRE et al., 2000, 2001; LOH, 2001; WIVES, 2004).

[9] Relação de sentido entre dois vocábulos que têm significação muito próxima, permitindo que um seja escolhido pelo outro em alguns contextos, sem alterar o sentido literal da sentença como um todo.
[10] Relação existente entre uma palavra de sentido mais específico e outra de sentido mais genérico
[11] Relação estabelecida entre um vocábulo de sentido mais genérico e outro de sentido mais específico

A definição de uma base de dados ontológica requer a coleta e descrição de conceitos que se pretende representar. Uma vez definida ela pode ser usada, então, para a identificação de tópicos em documentos a fim de determinar aqueles que sejam os mais importantes para guiar, por exemplo, a seleção de sentenças durante a sumarização automática. Na próxima seção, descreveremos alguns trabalhos que utilizam bases ontológicas para a detecção de tópicos.

3.2.2 Trabalhos que fazem uso de ontologias

A identificação de conceitos baseada em ontologias para determinar tópicos de documentos é bastante explorada em diversos trabalhos, por exemplo, Lin (1995) adota essa estratégia para criar sumários de textos. Ele estende a contagem de palavras, como maneira de identificar os tópicos de um texto, para a contagem de conceitos, propondo um método para identificar automaticamente suas idéias centrais. Essa contagem é feita utilizando uma ontologia para realizar generalizações, como, por exemplo, inferir que um texto que traga as palavras *laptop* e *handheld* pode tratar do tópico *computadores portáteis*. Para isso, é definido um peso para cada conceito e esse peso representa a freqüência de ocorrência dos itens lexicais. Tanto conceitos quanto subconceitos de itens lexicais podem ser usados para determinar esse peso. Além do peso, é definido o grau de generalização do conceito – variável G - calculada para cada conceito C da ontologia da seguinte maneira:

$$G_C = \frac{\text{MAX}_C(\text{ maior peso entre todos os subconceitos de C})}{\text{SUM}_C(\text{soma dos pesos de todos os subconceitos de C})}$$

Essa expressão indica que, quanto maior o valor de G, mais o superconceito C reflete um único subconceito, ou seja, o subconceito de C de maior peso, tem

proporcionalmente um peso maior que os outros subconceitos. Opostamente, quanto menor o valor de G, maior será o equilíbrio entre os pesos dos subconceitos de C, e, portanto, C generaliza seus subconceitos. Nesta situação, se escolhêssemos um dos subconceitos de C como conceito principal, perderíamos informações, já que todos os subconceitos apresentam pesos similares e portanto são igualmente relevantes, devendo o superconceito ser escolhido como conceito principal. Por exemplo, suponha que, para certo texto, um conceito chamado "*empresas*" seja superconceito C dos seguintes subconceitos C_i (respectivos pesos entre parênteses): *Toshiba(0), NEC(1), Compaq(1), Apple(7) e IBM(1)*. O peso final do conceito "*empresas*" e o seu valor G serão, respectivamente, 10 (0+1+1+7+1) e 0.70 (7/10). Neste exemplo, poderíamos supor, com base no valor de G, que o subconceito *Apple* é o conceito principal do texto, pois ele é o conceito mais mencionado e que mais influi em G, cujo valor é 0.70. Segundo Lin (1995), o valor 0.68 (empiricamente determinado) serve como um limitante superior para G. Esse limite serve para indicar quais conceitos são considerados relevantes no documento. Assim, se o valor de G estiver abaixo de 0.68, ele é considerado um conceito importante dentro do texto.

Com o objetivo de testar esse modelo, um experimento foi realizado envolvendo a sumarização de artigos da *Business Week* (1993-1994). A coleção de testes foi formada por 50 artigos sobre *processamento de informação*, com tamanho médio de 750 palavras. As medidas utilizadas para verificar o desempenho foram cobertura e precisão e a WordNet (MILLER, 1995) foi utilizada como ontologia. Para cada texto foi obtido um *abstract* (7-8 sentenças) feito por um profissional. Adicionalmente, foram construídos manualmente seus extratos ideais (com 8 sentenças) pela extração e justaposição de sentenças que continham os conceitos principais mencionados nos *abstracts*. Os extratos ideais foram então comparados com aqueles gerados

automaticamente, cujas sentenças foram pontuadas e selecionadas considerando três variações: 1) a pontuação de uma sentença corresponde à soma dos pesos dos conceitos-pais das palavras presentes na sentença; 2) o peso de uma sentença corresponde à soma dos pesos dos conceitos na própria sentença; 3) similar à variação 1, mas somente considerando um conceito (o mais relevante) por sentença. Os resultados obtidos (considerando extratos automáticos de 8 sentenças) apresentaram respectivamente, os seguintes valores de precisão (P) e cobertura (R): variação 1 (P=0.37, R=0.32), variação 2 (P=0.34, R=0.30), variação 3 (P=0.33, R=0.28).

Usando uma metodologia distinta de sumarização, Wu e Liu (2003) utilizaram artigos do *The New York Times* e do *Wall Street Journal* como corpus para elaborar e construir a ontologia utilizada na sumarização dos documentos, que é codificada como uma estrutura em árvore onde cada nó representa um conceito. Todos os artigos possuem extratos ideais e estes, por sua vez, têm parágrafos como unidade básica. Parágrafos foram escolhidos como unidades básicas para possibilitar a comparação entre os extratos gerados por este método e os extratos ideais do corpus que usam essa granularidade. Um processo de mapeamento verifica a correspondência entre as palavras do texto e os conceitos ontológicos, atribuindo-lhes pesos. O peso de cada conceito ontológico em relação ao texto a ser sumarizado é calculado somando-se a freqüência das palavras que aparecem no documento e que correspondam ao conceito. Aqui, cada palavra tem correspondência unívoca com cada conceito da ontologia. Considerando a estrutura arbórea, quando o peso de um nó é incrementado, o incremento é propagado para seus ancestrais. Após rotular toda a árvore com os pesos, os nós de segundo nível da árvore (logo abaixo da raiz) com maiores pesos são considerados os tópicos principais do documento. Pelo fato de a ontologia estar organizada como uma árvore com uma única raiz principal, e de se propagarem os

pesos, a raiz da árvore (primeiro nível) receberia sempre o maior peso. Além disso, representaria um único conceito muito genérico. Desta forma, escolher os nós do segundo nível garante que diferentes conceitos com diferentes pesos sejam considerados. Posteriormente, os parágrafos que tiverem maior proximidade com esses conceitos são selecionados.

A seleção de parágrafos é feita pontuando-os de acordo com a presença de palavras que correspondem aos conceitos principais identificados: para cada palavra relacionada a um conceito identificado anteriormente, o parágrafo recebe uma quantidade de pontos relativa ao peso do conceito considerado. Por exemplo, suponha a existência de um conceito chamado "*cinema*" cujo peso seja 20. Um parágrafo obterá 20 pontos para cada palavra que ele contenha e que seja relacionada a este conceito, por exemplo, as palavras *filme* e *Spider-man*. Essa abordagem foi testada para verificar o seu potencial. As medidas utilizadas foram precisão e cobertura. Elas foram calculadas verificando se as sentenças presentes no extrato automático correspondiam àquelas de um extrato ideal. Para o experimento foi utilizado um corpus de 51 artigos (com extratos ideais), num total de 882 parágrafos (com 1 ou 2 sentenças), dos quais 133 faziam parte dos extratos ideais. Os resultados obtidos indicam que este método alcançou uma precisão que varia entre 0.24 (extratos com 10 parágrafos) e 0.70 (extratos com 1 parágrafo) e uma cobertura variando entre 0.94 (extratos com 10 parágrafos) e 0.70 (extratos com 1 parágrafo). Segundo Wu e Liu (2003), esses resultados mostram que a ontologia consegue captar informações relevantes para a tarefa de sumarização.

Além de aplicações na área de SA, como as citadas anteriormente, a identificação de conceitos em documentos mostrou-se útil em outras áreas de pesquisa, como relatamos a seguir.

Tiun et al. (2001) exploraram a ontologia do Yahoo[12] para determinação do tópico principal de um documento. Um conjunto de palavras-chave é extraído de sentenças significativas do documento e, posteriormente, é mapeado no conjunto de seus conceitos ontológicos. As sentenças significativas são indicadas por um módulo de extração, que se baseia nas etiquetas HTML do documento (por exemplo, extraindo palavras presentes nos textos-âncora dos *links*, palavras enfatizadas por itálico ou negrito ou marcadas como título do documento). As palavras-chave são, então, mapeadas em conceitos da ontologia. Essa correspondência é feita comparando-se cada uma delas a um conjunto de itens lexicais que estão associados a cada conceito da ontologia. Esse conjunto é construído juntamente com a ontologia; na medida em que os conceitos ontológicos são determinados, os itens lexicais que descrevem esses conceitos são associados. Inicialmente cada conceito tem associado a si um pequeno número de descritores que são extraídos diretamente do seu identificador. Por exemplo, um conceito ontológico do Yahoo denominado *Arts and Humanities* terá o item lexical *Arts* e o item lexical *Humanities* como descritores associados; este conjunto de descritores é estendido posteriormente com itens lexicais coletados da WordNet.

O mapeamento entre palavras-chave e conceitos visa definir o peso de cada conceito no documento. Esse peso é calculado pela soma da freqüência das palavras-chave, no texto, coincidentes com os itens lexicais que descrevem o conceito e com a forma como o mapeamento é realizado. Palavras-chave mapeadas alternativamente por meio de um vocabulário estendido, contribuirão com 50% da sua freqüência para o peso de um conceito. O peso acumulado final de um conceito será o seu peso somado ao peso acumulado total de seu(s) conceito(s) filho(s), multiplicado pelo número de palavras-chave mapeadas sem utilização do vocabulário estendido. Tiun et al. apontam, para um

[12] http://www.yahoo.com

conjunto de 202 documentos, uma precisão de 29.7% na determinação de seus tópicos principais.

Explorando a identificação de conceitos para a classificação de páginas Web, Mladenic e Grobelnik (1999) também utilizam como base a ontologia do Yahoo para o inglês. Eles têm como hipótese que documentos de texto podem ser caracterizados por um conjunto de palavras-chave que permitem indicar seu conteúdo. Os documentos são representados como um vetor de características e incluem, além de unigramas, seqüências de até cinco palavras (5-gramas), ou pentagramas. Inicialmente os conceitos da ontologia são descritos por meio de palavras extraídas das categorias definidas no Yahoo. Por exemplo, a categoria "*Machine Learning*" que, por sua vez, é uma subcategoria de "*Science*", está hierarquicamente subordinada do seguinte modo: *"Science: Computer Science: Artificial Intelligence: Machine Learning"*. Assim, o conceito *Machine Learning* tem associado a si, como descritores, as palavras-chave: *Science*, *Computer Science*, *Artificial Intelligence* e *Machine Learning.* Técnicas de aprendizado de máquina (*Naive-Bayes*) são utilizadas usando como dados de treinamento documentos previamente classificados segundo essa hierarquia. O problema de classificação é dividido em subproblemas: para cada conceito da ontologia existe um classificador associado, que visa reconhecer documentos relacionados àquele conceito. O resultado é um conjunto de classificadores independentes que, em conjunto, são usados para determinar o tópico principal de um documento pela associação de suas palavras-chave aos conceitos da ontologia. Isso é feito computando-se sua similaridade ou a probabilidade de a palavra-chave pertencer a uma determinada classe conceitual. Essa abordagem foi testada por meio de um experimento cujo objetivo foi verificar se um novo documento apresentado era corretamente classificado. Como medida de desempenho, foi realizada uma verificação simples, checando para quais classes

conceituais o documento era indicado com maior probabilidade. O conjunto de testes foi formado por 1.100 documentos, os resultados experimentais mostraram que, para cerca de 50% dos exemplos de testes, o conceito correto estava entre os três de maiores probabilidades apontados pelos classificadores.

De modo geral, os trabalhos apresentados demonstram o potencial de utilização de conhecimento ontológico para a detecção de tópicos em texto. Isso é extremamente relevante para nossa proposta uma vez que, detectar os tópicos principais de um documento a fim de preservá-los durante a sumarização, é fundamental. A motivação comum a todos eles é a de que a análise de palavras isoladas em um texto, sem nenhuma consideração sobre o relacionamento semântico entre elas pode ser um fator limitante no processamento de documentos. Desta forma, todos eles adotam como elemento fundamental uma ontologia. Outras similaridades podem ser apontadas como, por exemplo, processos de mapeamento que associam conceitos aos documentos verificando a correspondência entre as palavras do texto e a ontologia. Processos de propagação de peso entre os conceitos com o intuito de realizar generalizações também são pontos em comum entre os trabalhos citados. Todos esses aspectos foram observados e considerados para a elaboração da nossa proposta. A seguir, apresentamos uma síntese das características principais dos trabalhos descritos, já que eles apresentam estratégias e recursos que foram adaptados e/ou incorporados às nossas estratégias de sumarização, contribuindo para nosso projeto de mestrado.

3.2.3 Características gerais no processamento semântico de documentos

A primeira e mais evidente característica de todos os trabalhos citados que serviu de contribuição para nossa proposta referiu-se à explicitação da utilidade da ontologia para a detecção de tópicos em documentos *Web*. Os benefícios explicitados

nos levaram a optar pela inclusão de uma ontologia às nossas estratégias de sumarização já que, inicialmente, somente consideraríamos a estrutura do documento, focando as etiquetas HTML. Com a inclusão de uma ontologia foi possível explorar e processar o conteúdo textual de forma mais abrangente, já que as marcações HTML estão restritas a alguns trechos do documento. Essa estratégia foi explorada levando à implementação do sumarizador GEO (relatada no próximo capítulo). A opção pela utilização da ontologia do Yahoo em nossa proposta deveu-se, particularmente, aos trabalhos de Mladenic e Grobelnik (1999) e Tiun et al. (2001), já que essa ontologia é voltada ao processamento de documentos *Web* e está disponível em português.

Outra contribuição, particularmente do trabalho de Tiun et al. (2001), tem relação com a forma de descrever os conceitos ontológicos (processo de enriquecimento descrito no próximo capítulo). Adotamos as mesmas relações semânticas utilizadas por eles (sinonímia, hiperonímia, hiponímia). Lin (1995) e Tiun et al. (2001) enriquecem suas ontologias com fontes externas de conhecimento (WordNet, *thesaurus*, etc.), por isso também optamos por enriquecer a ontologia do Yahoo com informações coletadas do *thesaurus* Diadorim (GREGHI et al., 2002) e da versão em língua portuguesa da Wikipédia (http//pt.wikipedia.org).

Alguns processos cruciais, como por exemplo, o mapeamento do conteúdo textual em conceitos ontológicos, foram inspirados e utilizados de modo muito similar ao que é feito nesses trabalhos, especialmente em (TIUN et al., 2001; WU; LIU, 2003). Do mesmo modo que é feito nestes trabalhos, no nosso sistema verificamos a correspondência entre as palavras contidas no documento e aquelas definidas como descritores de conceitos na ontologia para detectar e pontuar conceitos ontológicos subjacentes ao texto. A propagação de pontos entre conceitos relacionados (LIN, 1995; TIUN et al., 2001; WU; LIU, 2003) também foi adotada por nós.

Processos de avaliação utilizados nesses trabalhos, especificamente aqueles que usaram métricas como precisão e cobertura (LIN, 1995; WU; LIU, 2003), também foram adotados por nós para avaliar a nossa estratégia de sumarização que faz uso de ontologia (sistema GEO).

No próximo capítulo, apresentaremos duas implementações preliminares de sumarizadores envolvendo etiquetas HTML e conhecimento ontológico para geração de sumários de documentos *Web*.

4 Implementações preliminares: os sumarizadores HTMLSUMM e GEO

Diante das evidências apontadas pelos trabalhos descritos no capítulo anterior sobre a utilidade das marcações HTML e do processamento semântico de conteúdo textual por meio de ontologias, desenvolvemos duas estratégias de sumarização que resultaram na implementação de dois sumarizadores: o HTMLSUMM e o GEO. O HTMLSUMM faz uso de informações provenientes da estrutura HTML para gerar sumários de documentos *Web*, já o GEO utiliza-se de conhecimento ontológico para a sumarização. Esses dois sistemas foram base para o desenvolvimento do sumarizador ExtraWeb que é a maior contribuição desta pesquisa e é descrito no capítulo seguinte. Neste capítulo descrevemos o HTMLSUMM e o GEO.

4.1 Sumarização baseada em etiquetas HTML

Nesta seção é descrito o sistema HTMLSUMM (HTML SUMMarizer). Ele faz uso do primeiro modelo que propomos nesta dissertação de mestrado: um modelo de sumarização baseado em etiquetas HTML.

4.1.1 O sistema de geração de extratos HTMLSUMM

O sumarizador HTMLSUMM é baseado em etiquetas HTML e segue um modelo que representa um documento como dois conjuntos distintos, de sentenças e palavras-chave, com um mapeamento associado entre esses dois conjuntos (FILATOVA; HATZIVASSILOGLOU, 2004). Essa representação é, então, usada para descrever a tarefa de seleção de sentenças para a composição de um extrato. A seguir são apresentadas a arquitetura e as principais características desse modelo.

4.1.1.1 Arquitetura do HTMLSUMM

A Figura 11 mostra a arquitetura do HTMLSUMM. Caracterizado como um sumarizador extrativo, ele executa os seguintes passos durante a sumarização:

1. Primeiramente, delimitam-se as sentenças do documento-fonte fazendo uso dos sinais de pontuação tradicionais (ponto final, de exclamação e de interrogação) e das etiquetas HTML do documento. Essas sentenças formarão o conjunto de segmentos textuais;

2. Posteriormente, delimitam-se as palavras-chave do documento-fonte usando as etiquetas HTML presentes no documento. Todas as palavras que estiverem marcadas por determinadas etiquetas HTML (explicitadas na próxima seção) são consideradas palavras-chave. Por exemplo, considerando a etiqueta de negrito <B>, e um trecho de código *<B>curso gratuito online</B>*, as palavras curso, gratuito, e online seriam consideradas palavras-chave;
3. As sentenças são então ranqueadas, isto é, ordenadas pelos seus pontos calculados pelo método de ranqueamento descrito na Seção 4.1.1.2. As sentenças com maiores pontuações no ranque comporão o sumário. Nesta etapa, para aprimorar os resultados do processo de sumarização, utiliza-se uma *stoplist*, ou seja, uma lista de palavras muito comuns (e, portanto, normalmente sem relevância para a sumarização);
4. Por fim, respeitando-se a taxa de compressão especificada, selecionam-se as sentenças que formarão o sumário. Vale ressaltar que a taxa de compressão, calculada sobre o tamanho (número de palavras) dos extratos e documentos-fonte respectivos, pode levar a ligeiras variações em relação ao modelo numérico, já que, por tratarem-se métodos extrativos, a unidade mínima considerada é sentencial.

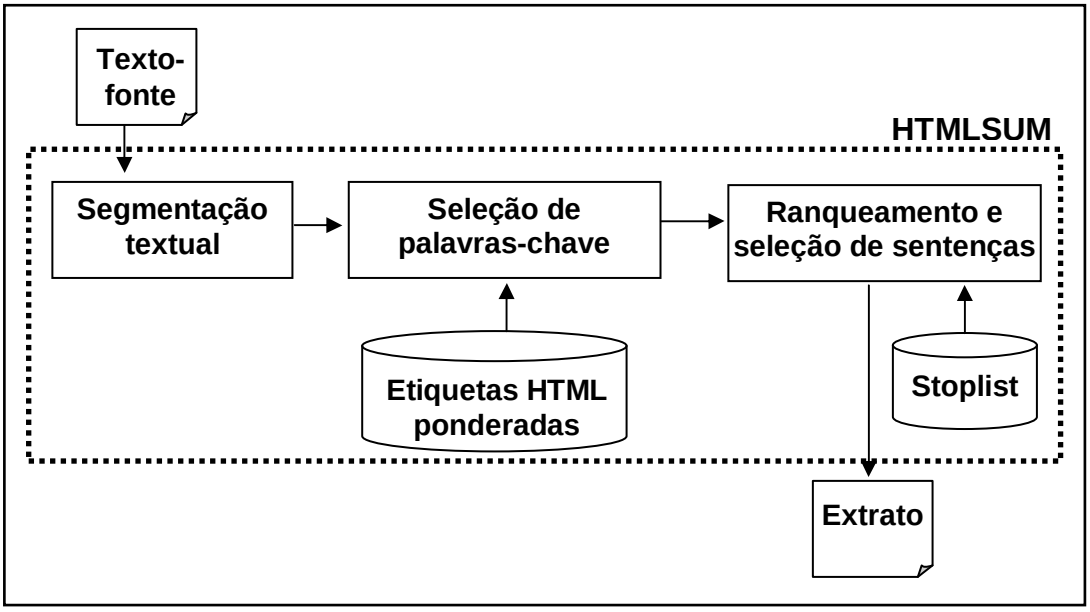


Figura 11. Arquitetura do HTMLSUMM

4.1.1.2 Metodologia de sumarização automática do HTMLSUMM

O modelo de sumarização do HTMLSUMM é baseado em três componentes principais: um conjunto de segmentos textuais, a partir do qual todas as sentenças potenciais para composição do extrato são retiradas; um conjunto de palavras-chave, que representa as informações relevantes que devem ser incluídas na saída final, e um mapeamento entre palavras-chave e segmentos textuais. Primeiro, assume-se a existência de um conjunto T de segmentos textuais $t_1,t_2,...,t_n$, cujo subconjunto formará o extrato. Cada t_i corresponde a uma sentença do documento-fonte. Segundo, supõe-se a existência de um conjunto finito C de palavras-chave $c_1,c_2,...,c_m$. As palavras-chave codificam a informação que deve ser apresentada no extrato. Para sumarizar o documento, o HTMLSUMM atribui um alto escore para os segmentos textuais que contêm as palavras-chave, já que neste modelo, elas são consideradas as informações mais relevantes do documento. Para isso, leva em consideração a estrutura HTML do documento-fonte. Para cada palavra-chave, assume-se a existência de um peso w_i, que

indica quão importante uma palavra-chave c_i é para o extrato como um todo. Este peso é dependente do tipo de etiqueta HTML que delimita a palavra-chave, já que diferentes etiquetas sugerem diferentes graus de relevância (CUTLER et al., 1997; FRESNO; RIBEIRO, 2004). Nossa hipótese é atribuir pesos maiores àquelas etiquetas mais comumente usadas pelos autores para ressaltar as informações mais salientes e isso contribui na seleção de informações e na composição das descrições desses documentos. O peso w_i associado a cada etiqueta foi determinado com base em uma análise de corpus que relatamos na próxima seção.

No HTMLSUMM, o conjunto T de segmentos textuais é obtido via processo de segmentação sentencial que delimita todas as sentenças do documento-fonte observando os sinais de pontuação tradicionais e/ou as etiquetas HTML mostradas na Tabela 4. Estas etiquetas têm a função de estruturar o texto em blocos de informações (PEDREIRA-SILVA; RINO, 2005), ou seja, são responsáveis pelo fluxo visual do texto (quebras de linhas, parágrafos, etc.) e, por isso, foram usadas no HTMLSUMM para segmentar o documento.

Tabela 4. Conjunto de etiquetas HTML para extração de segmentos textuais

<LI>	<DT>	<DD>	<P>
<PRE>	 	<H1>	<H2>
<H3>	<H4>	<H5>	<H6>
<TD>	<TH>	<TR>	
<DIV>			

Para ilustrar o processo de segmentação considere, por exemplo, o código HTML seguinte:

```
<DL>
<DT><STRONG>Lower cost</STRONG>
<DD>The new version of this product costs significantly less
than the
previous one!
<DT><STRONG>Easier to use</STRONG>
<DD>We've changed the product so that it's much easier to use!
<DT><STRONG>Safe for kids</STRONG>
<DD>You can leave your kids alone in a room with this product
and they won't get hurt (not a guarantee).</DL>
```

O conjunto T de segmentos textuais seria composto por: *T = {Lower cost, The new version of this product costs significantly less than the previous one, Easier to use, We've changed the product so that it's much easier to use, Safe for kids, You can leave your kids alone in a room with this product and they won't get hurt (not a guarantee)}*.

O conjunto C de palavras-chave é definido a partir do conjunto de etiquetas relacionadas na Tabela 5. As palavras-chave representam, portanto, as palavras que no código-fonte HTML do documento estão delimitadas por estas etiquetas - usadas pelos autores de documentos *Web* para indicar conteúdos relevantes – e são potencialmente úteis para a sumarização automática (PEDREIRA-SILVA; RINO, 2005).

Tabela 5. Conjunto de etiquetas HTML para extração de palavras-chave

<META NAME="DESCRIPTION>	<META NAME="KEYWORDS">	<TITLE>	<A>
<H1>	<H2>	<H3>	<H4>
<H5>	<H6>	<EM>	<B>
<STRONG>	<I>	<U>	

Essas etiquetas foram escolhidas pelos seguintes motivos:

- <META NAME="DESCRIPTION> e <META NAME="KEYWORDS"> - são utilizadas pelos autores, respectivamente, para fornecer uma curta descrição do documento e um conjunto de palavras-chave. Logo, as palavras contidas nessas meta-etiquetas carregam um importante conteúdo semântico.
- <TITLE>, <H1>, <H2>, <H3>, <H4>, <H5> e <H6> - indicam títulos e cabeçalhos do documento. Essas etiquetas podem ser úteis para fornecer indícios sobre os tópicos do documento.
- <EM>, <B>, <I>, <STRONG> e <U> – são usadas para indicar ênfase de fragmentos do texto. A ênfase é um recurso utilizado pelos autores para destacar informações relevantes.
- <A> - é utilizada pelos autores para delimitar textos âncoras. Eles geralmente trazem indícios sobre o conteúdo do documento para o qual apontam.

Para ilustrar o processo de seleção de palavras-chave de um documento, considere, por exemplo, o trecho do código-fonte seguinte retirado de um documento *Web*:

```
<META NAME="Keywords" content="babylon, tradutor, tradução, traduzir, translator, translations, translate, inglês, português, espanhol, alemão, francês, italiano, holandês, japonês, hebraico, online, idioma, dicionário, dictionary, babel">
```

O conjunto C de palavras-chave, neste caso, será composto pelos itens que fazem parte do conteúdo (campo *content*) da etiqueta <META NAME="Keywords">, ou seja, *C={babylon, tradutor, tradução, traduzir, translator, translations, translate, inglês, português, espanhol, alemão, francês, italiano, holandês, japonês, hebraico, online, idioma, dicionário, dictionary, babel}*. Esse processo será repetido para as demais etiquetas, até que o conjunto C esteja completo. Caso uma palavra já exista no conjunto C, ela não será incluída novamente.

Uma vez definido o conjunto T e C de segmentos textuais e palavras-chave do documento-fonte, realiza-se um processo de mapeamento entre esses conjuntos. Esse mapeamento é feito por meio de uma função f: T X C $\rightarrow$ [0,1], que diz quão bem cada palavra-chave é coberta por um dado segmento textual, isto é, indica a quantidade de informação coberta por um segmento textual. Seguindo o trabalho de Filatova e Hatzivassiloglou (2004), definimos a função *f* entre T e C limitando-a aos valores 0 ou 1, isto é, cada segmento textual ou contém ou não contém certa palavra-chave. Então, o total de informação coberta por qualquer subconjunto S de T (um extrato proposto) no HTMLSUMM é definido por:

$$I(S) = \sum_{i=1,...,m} w_i \cdot \delta_i \quad (1)$$

em que m é o número de palavras-chave, w_i é o peso de uma palavra-chave c_i e δ_i assume o valor 0 ou 1 de acordo a presença ou não da palavra-chave em determinado segmento textual:

$$\delta_i = \begin{cases} 1, & se\ \exists j \in \{1,...,m\}\ tal\ que\ f(tj, ci) = 1 \\ 0, & caso\ contrário \end{cases}$$

Em outras palavras, a pontuação de cada segmento textual corresponde ao somatório dos pesos das palavras-chave que ele cobre.

A Equação 1 mede a informação coberta por uma coleção de segmentos textuais considerando o mapeamento que verifica a correspondência total entre as palavras-chave e segmentos textuais. Obviamente, deseja-se maximizar a quantidade de informação, isto é, deseja-se gerar extratos que sejam formados por aqueles segmentos textuais que contenham o maior número possível de palavras-chave. Entretanto, isto pode acontecer somente quando fatores limitantes adicionais sobre o número dos segmentos textuais são introduzidos; de outra maneira, o conjunto completo de segmentos textuais, ou seja, o próprio documento completo, seria a solução que proveria o valor máximo para a Equação 1, isto é, $\forall S \subset T, I(S) \leq I(T)$. Isso é feito associando-se um custo p_i para cada segmento textual t_i ($i=1,2...,n$) incluído no extrato e definindo uma função P sobre o conjunto de segmentos textuais, que provê a penalidade total associada à seleção daqueles segmentos textuais para inclusão no extrato. A definição de p_i e P é dada por:

$$p_i=1$$

$$P(S) = \sum_{ti \in S} p_i$$

No HTMLSUMM é associado um custo fixo para cada unidade textual, ou seja, o custo de incluir um segmento textual no extrato é sempre igual a 1. Dessa forma, a penalidade total é igual ao tamanho do extrato, isto é, corresponde ao número de sentenças que ele possui.

Com a introdução da função de custo P(S) o modelo apresenta dois componentes gerais. Estes componentes podem então ser relacionados, nesse caso, otimizando I(S) enquanto P(S) é mantido dentro de certo limite definido pela taxa de compressão desejada. Desta forma, primeiramente é calculada a pontuação de cada segmento textual e, então, definido o ranque decrescente destes segmentos. Neste modelo, a pontuação de cada segmento textual corresponderá à soma dos pesos das palavras-chave que ele cobre. O peso de cada palavra-chave corresponde ao peso associado à etiqueta HTML que marca esta palavra-chave. Caso uma palavra-chave esteja marcada por mais de uma etiqueta, seu peso será o da etiqueta de maior peso. Através deste ranque são selecionados aqueles segmentos textuais que melhor contribuem para I(S) com relação às palavras-chave; esse procedimento é repetido até se atingir o valor de custo P(S) desejado.

4.1.1.2.1 Associação de pesos às etiquetas HTML

No modelo do HTMLSUMM, cada etiqueta tem associada a si um peso w_i que contribui para indicar quão importante a palavra-chave é no documento. Este peso associado a cada etiqueta foi determinado empiricamente com base em uma análise de corpus. Para isso, realizamos uma análise estatística tendo como foco duas dentre as

várias medidas comumente utilizadas por mecanismos de busca para computar a relevância das palavras de um documento *Web*, são elas: densidade e proeminência de palavras-chave (MARCKINI, 2001; SEDIGH; ROUDAKI, 2003).

A densidade refere-se ao número de palavras que aparecem em uma página *Web* em proporção ao número total de palavras nesta página. Ela indica quão relevante uma palavra é para um documento *Web*. A semântica por trás desta métrica assume que a importância de um termo é diretamente proporcional à sua freqüência no documento (LUHN, 1958; NOMOTO; MATSUMOTO, 1996). A proeminência mede a localização de uma palavra no código HTML da página. Quanto antes em uma página uma palavra em particular aparecer, mais proeminente ela é. Esta medida baseia-se no fato de que independentemente do conteúdo da página, aquelas que estejam mais próximas do topo são consideradas as mais relevantes pelos mecanismos de busca (MARCKINI, 2001). Este princípio se aplica principalmente a palavras delimitadas por etiquetas HTML, tais como <TITLE>, <META>, ou <H1-H6>, já que usualmente são definidas nas posições iniciais. Claramente para alguns gêneros de documentos, isso nem sempre é válido, porém, como nosso sistema considera apenas a estrutura HTML, estes aspectos foram desconsiderados. Assumimos que, de modo geral, nas páginas *Web*, as palavras mais relevantes costumam vir no topo das páginas e apresentar maior freqüência (MARCKINI, 2001).

As medidas de densidade e proeminência são definidas da seguinte maneira:

$$\text{Densidade} = \frac{FP}{TP}$$

$$\text{Proeminência} = TP - \frac{SP - 1}{NP} * \frac{100}{TP}$$

em que *FP* é o número de ocorrências da palavra em foco, *TP* corresponde ao número total de palavras no documento, *SP* é a soma de cada posição onde ocorre a palavra que está sendo analisada e *NP* é número de posições consideradas.

Nos sistemas de busca, as medidas de densidade e proeminência são combinadas e, juntas, contribuem para indicar relevância de uma palavra em um documento. O problema é que a forma de combinar essas medidas faz parte de algoritmos de ranqueamento que não estão disponíveis publicamente. Assim, decidimos combiná-las através da média harmônica, o que resulta em uma medida única que indica a relevância y_i de uma palavra. Logo, as palavras mais relevantes para a nossa análise seriam aquelas com um bom balanceamento entre *Proeminência* e *Densidade*:

$$y_i = \frac{2 * \text{Densidade} * \text{Proeminência}}{\text{Densidade} + \text{Proeminência}} \quad (2)$$

Essa medida é interessante para a SA já que ela parte da idéia de que, se uma palavra pode ser significativa em um documento-fonte por sua densidade e proeminência, ela também pode indicar a importância relativa dos segmentos textuais que a contêm.

No nosso contexto de uso de etiquetas HTML, a média harmônica serviu como um indicativo de segmentos textuais relevantes do corpus de análise. Uma palavra relevante neste contexto é aquela que é recorrente na página (sendo muito possivelmente um tópico do documento) e ao mesmo tempo aparece nas regiões consideradas mais visíveis (próximas ao topo da página) (MARCKINI, 2001).

Considerando essa média harmônica, montamos um corpus de documentos *Web* com objetivo de verificar quais as etiquetas, dentre aquelas exibidas na Tabela 5, foram mais usadas e em que proporção elas ocorreram. O corpus foi formado por 1245 segmentos textuais demarcados pelas etiquetas da Tabela 5, coletados de 50 documentos *Web* distintos. Os segmentos foram então analisados com relação à sua marcação HTML. A freqüência dessas etiquetas no corpus foi computada e depois normalizada (no intervalo de 0 a 1). É importante ressaltar que a normalização foi feita para que as freqüências pudessem ser comparáveis umas com as outras, indicando a proporção de ocorrência das etiquetas no corpus. Esses valores normalizados (Tabela 6) foram utilizados como pesos (w_i) das etiquetas no processo de sumarização do HTMLSUMM. Conforme pode ser visto na Tabela 6, optamos por agrupar algumas etiquetas, por exemplo, <B> e <STRONG>, computando a freqüência conjuntamente já que, funcional e visualmente, elas são similares (CUTLER et al., 1997).

Tabela 6. Peso das etiquetas HTML

Etiquetas	Peso (w_i)
<B>, <STRONG>	0,516393
<META NAME=”DESCRIPTION”>	0,713115
<H1>, <H2>, <H3>, <H4>, <H5>, <H6>	0,139344
<I>,<EM>,<U>	0,090164
<META NAME=”KEYWORDS”>	0,622951
<A>	0,934426
<TITLE>	1

Esses pesos são usados para ponderar as palavras-chave dos documentos *Web*. Esse modelo de pesos e o próprio modelo de sumarização do HTMLSUMM foram avaliados conforme descreveremos na próxima seção.

4.1.2 Avaliação do sistema HTMLSUMM

O objetivo do experimento foi verificar a aplicabilidade do modelo proposto para a sumarização automática de documentos *Web*, com especial enfoque na premissa

desta avaliação, isto é, a utilidade das etiquetas HTML para seleção de informações relevantes para composição de extratos. Para atingir este objetivo, os extratos gerados pelo HTMLSUMM foram comparados com descrições geradas por ferramentas comerciais, em termos de informatividade semântica (MANI, 2001), a qual indica o quanto de informação apresentada no documento-fonte foi reproduzida na descrição. No nosso contexto de documentos *Web*, a informação é expressa por palavras e cada palavra tem associada a si um valor de relevância de acordo com a sua média harmônica entre proeminência e densidade no documento.

A medida de informatividade semântica é calculada de tal forma que, quanto mais curta uma descrição e mais conteúdo informativo (palavras relevantes) ela possuir, maior sua pontuação e vice-versa. Se representarmos um documento-fonte D como uma seqüência $M(D)=M_1...M_n$ de palavras, então uma descrição informativa, ou extrato, S pode ser vista como um subconjunto M(S) de palavras relevantes de D. A informatividade semântica é medida da seguinte forma:

$$\text{informatividade semântica} = \left(1 - \frac{Tamanho(S)}{Tamanho(D)}\right)\left(\frac{SomaR(M(S))}{SomaR(M(D))}\right)$$

em que *Tamanho* é dado pelo número de palavras e *SomaR*, pela soma dos valores individuais de relevância das palavras.

Os extratos do HTMLSUMM foram comparados aos de três outros sistemas: o sumarizador Copernic Summarizer[13], o mecanismo de busca Google e um sistema baseline. O Copernic Summarizer é uma ferramenta comercial que faz a sumarização de documentos *Web* em inglês, francês, espanhol e alemão, permitindo que o usuário tenha uma idéia geral do seu conteúdo. A ferramenta usa um algoritmo de extração que

[13] http://www.copernic.com/en/products/summarizer/

envolve métodos estatísticos e características lingüísticas do texto para encontrar os conceitos principais e extrair as sentenças mais relevantes do documento. Na documentação oficial do sistema não são citados quais características lingüísticas ou métodos estatísticos são utilizados, mas ela foi escolhida por ser uma das principais ferramentas do gênero (FRESNO; RIBEIRO, 2004). O Google é o mais utilizado dos mecanismos de busca, tratando-se de uma ferramenta considerada de alta qualidade (SULLIVAN, 2004), por isso optamos por sua escolha. A criação de descrições pelo Google é completamente automatizada e leva em consideração a presença dos termos da consulta do usuário. O sistema baseline considera extratos as próprias descrições manuais feitas pelos autores das páginas, que são indicadas pela etiqueta <META NAME="Description">, quando estas existem no documento. Quando não existem, nenhuma descrição é retornada ao usuário.

Além da versão base do HTMLSUMM, que tem como estratégia de sumarização o modelo proposto anteriormente, resolvemos incluir na avaliação pequenas variações no processo de sumarização (por exemplo, desconsiderando alguma etiqueta em particular) para determinar se havia alguma configuração melhor para o sistema. Ou seja, procuramos verificar se alguma variação na estratégia de sumarização proposta inicialmente poderia prover melhores extratos. As variações da estratégia de sumarização, identificadas por V1 a V7, são as seguintes:

- V1: o sistema utiliza um peso único (igual a 1) para todas as etiquetas HTML. Desta forma, a seleção de palavras-chave não será influenciada pelo tipo da etiqueta. A inclusão dessa variação visou validar a nossa hipótese sobre o uso de pesos diferenciados para etiquetas HTML.
- V2: o sistema desconsidera como relevante o conteúdo delimitado pela etiqueta de *links* <A>. Nesta situação, a etiqueta <A> não é utilizada na

seleção de palavras-chave. Resolvemos verificar a utilidade desta etiqueta para a sumarização já que, apesar de não indicar ênfase, o texto âncora pode indicar tópicos do documento (CUTLER et al., 1997).

- V3: o sistema procura privilegiar as sentenças mais curtas, penalizando as maiores. Isto é feito dividindo a pontuação final da sentença pela quantidade de palavras que ela contém. A inclusão desta variação permitiu verificar se a informatividade dos extratos tem relação com o tamanho das sentenças incluídas.
- V4: o sistema gera descrições com um número mínimo de 20 palavras, que equivale ao tamanho médio das descrições fornecidas pelo Google. Nosso intuito foi verificar a influência da quantidade de palavras na informatividade das descrições.
- V5: o sistema exibe como resposta as descrições manuais feitas pelos autores dos documentos. Elas são recuperadas por meio do conteúdo da etiqueta <META NAME=”Description”>, sempre que esta estiver definida no documento-fonte. A razão desta variação foi verificar se a inclusão de descrições manuais contribui para a informatividade.
- V6: o sistema permite a inclusão de termos repetidos no conjunto de palavras-chave. Assim, se uma palavra aparecer em diferentes partes do documento, ela terá maior influência no cálculo da pontuação dos segmentos textuais. Nosso intuito foi verificar se considerar a freqüência das palavras-chave traria melhorias com relação à informatividade.
- V7: o sistema não utiliza uma lista de *stopwords*. A intenção foi verificar se a utilização desse recurso lingüístico traria melhorias para o modelo de sumarização proposto.

Considerando as variações descritas anteriormente, incluímos nesta avaliação 10 versões modificadas da versão base do HTMLSUMM. Desta forma, testamos no total 14 sistemas, incluindo o HTMLSUMM e suas 10 variações, o sistema baseline e os dois sistemas comerciais (Google e Copernic Summarizer). Os sistemas testados foram identificados por S1 a S14 conforme a lista seguinte:

- S1: Google;
- S2: Copernic Summarizer;
- S3: HTMLSUMM sem modificações;
- S4: Baseline;
- S5: Versão modificada do HTMLSUMM, incorporando as variações V1, V2, V6 e V7;
- S6: Versão modificada do HTMLSUMM, incorporando a variação V6;
- S7: Versão modificada do HTMLSUMM, incorporando a variação V3;
- S8: Versão modificada do HTMLSUMM, incorporando as variações V2 e V3;
- S9: Versão modificada do HTMLSUMM, incorporando as variações V2 e V6;
- S10: Versão modificada do HTMLSUMM, incorporando as variações V1, V6 e V7;
- S11: Versão modificada do HTMLSUMM, incorporando as variações V1, V2, V3, V4 e V5;
- S12: Versão modificada do HTMLSUMM, incorporando as variações V1, V2, V3 e V4;
- S13: Versão modificada do HTMLSUMM, incorporando as variações V2, V3, V4 e V5;

- S14: Versão modificada do HTMLSUMM, incorporando as variações V1, V2, V5 e V6;

Destacamos que a forma de combinar as variações foi feita intuitivamente, sem seguir nenhum critério previamente estabelecido, combinando aquelas variações que, a nosso ver, poderiam trazer melhorias para o sistema. Procuramos manter algumas configurações bem similares, diferenciando em uma ou duas variações, com intuito de detectar algum tipo de melhoria pela inclusão ou exclusão de variações. Outra preocupação foi a de verificar se havia diferenças entre utilizar uma única variação ou um número maior delas, daí o fato de considerarmos sistemas com uma única alteração ou várias com relação ao sistema base.

4.1.2.1 Construção do corpus para a avaliação

Para realizar as comparações entre os sistemas, procedemos à coleta de documentos e montagem de um corpus de teste. Para montagem do corpus, utilizamos 21 documentos *Web* recuperados a partir de 7 consultas distintas submetidas ao mecanismo de busca Google. Foram recuperados 3 documentos por consulta. Para cada documento recuperado, foi gerada uma descrição (extrato) por cada um dos 14 sistemas testados (S1-S14), totalizando 294 descrições. Como o Google possui um tamanho de descrição fixa (geralmente 2 sentenças), as descrições geradas pelos outros sistemas tentaram reproduzir aproximadamente esse tamanho, considerando também extratos com 2 sentenças, de forma a manter a comparação justa.

Para tentarmos ser o mais objetivos possível e simular da melhor maneira o contexto de uma pesquisa em um mecanismo de busca, consultas curtas foram utilizadas. As consultas foram elaboradas com dois ou três termos, porém a escolha dos domínios aos quais elas se referiram foi aleatória. Elas foram elaboradas sem nenhum

tipo de consideração sobre quais ou quantos domínios seriam utilizados, ou quantas consultas estariam relacionadas a cada domínio. A única restrição é que as consultas estivessem em inglês, devido à impossibilidade de o Copernic Summarizer gerar sumários em língua portuguesa. Apesar de o nosso corpus de avaliação ter sido montado com uma coleção em língua inglesa, acreditamos que os resultados obtidos podem ser generalizados para o português já que, teoricamente, o HTMLSUMM pode ser aplicado à sumarização de documentos *Web* em outros idiomas, visto que o método utilizado é independente de língua. Para isto basta simplesmente a substituição do repositório de *stopwords* de acordo com a língua-alvo do documento.

4.1.2.2 Síntese da avaliação do HTMLSUMM

A Tabela 7 mostra a informatividade semântica média, variando de 0 a 1, obtida por cada sistema, sendo os sombreados os métodos simples considerados nesse trabalho.

Tabela 7. Informatividade semântica para o corpus de avaliação

Sistema	Informatividade Semântica média
S3	0,17
S13	0,15
S14	0,15
S11	0,14
S10	0,14
S5	0,14
S12	0,13
S9	0,13
S1	0,12
S6	0,1
S2	0,09
S4	0,08
S8	0,06
S7	0,05

Conforme podemos observar na Tabela 7, os valores médios de informatividade obtidos por todos os sistemas foram muito baixos. Isso pode ser explicado pelo fato de as descrições analisadas serem muito curtas, com apenas duas sentenças, já que optamos por reproduzir descrições que caracterizassem bem aquelas que são exibidas pelos

mecanismos de busca. Considerando esse tamanho padrão de duas sentenças, conclui-se que as descrições geradas são ruins, o que pode comprometer sua utilização na apresentação de resultados de buscas. Porém, cabe ressaltar que, apesar do baixo desempenho a versão base do HTMLSUMM (S3) foi o sistema que apresentou o melhor resultado médio de informatividade semântica, o mesmo podendo ser observado para suas versões modificadas frente aos sistemas comerciais e ao baseline. Esse dado é interessante porque demonstra que existe um potencial a ser explorado no que se refere à aplicação das etiquetas HTML na SA.

A diferença para o Google pode ser explicada pelo fato de que ele gera as descrições apenas considerando a presença dos termos da consulta do usuário. Nesse caso, ele pode recuperar excertos do documento que não necessariamente sejam os mais proeminentes (localizados no topo da página). Além disso, os termos da consulta do usuário podem não coincidir com as palavras com maior densidade no documento. Já o baseline, apesar de recuperar o conteúdo da etiqueta <META NAME=”Descripton”> que traz muitas palavras proeminentes devido à sua localização nas primeiras linhas do código-fonte, não necessariamente traz na descrição palavras de alta densidade, já que elas podem não se repetir em outros trechos do documento; além disso, a presença desta etiqueta não é obrigatória nos documentos.

Mesmo fazendo uso de características lingüísticas, o desempenho médio do Copernic Summarizer ficou abaixo do Google e do HTMLSUMM. Seu desempenho inferior provavelmente seja em virtude de o seu algoritmo de sumarização não considerar a estrutura HTML do documento. Esse resultado aponta que o conjunto de etiquetas utilizado pelo HTMLSUMM e sua estratégia de sumarização conseguem gerar descrições mais informativas que os outros sistemas, do ponto de vista de trazer o

conteúdo mais denso e proeminente e isso justifica a continuidade da investigação da estratégia utilizada, apesar do baixo valor de informatividade alcançado.

A Figura 12 mostra os valores da informatividade semântica obtidos pelo HTMLSUMM (S3) e os outros sistemas para cada um dos documentos do corpus.

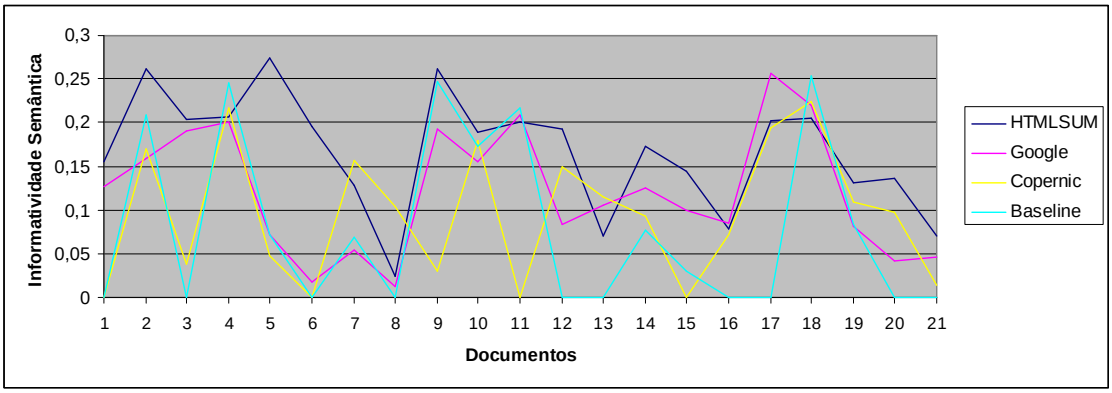


Figura 12. Informatividade dos sistemas para a coleção de documentos

Como podemos notar pelas curvas, existe uma grande variação da informatividade dependendo do documento sumarizado. Isso evidencia que o desempenho dos sistemas é dependente de algumas características dos documentos que merecem investigações mais profundas para sua detecção. Possivelmente para o HTMLSUMM esse comportamento pode estar relacionado à riqueza de etiquetas do documento, ou seja, os valores mais baixos das curvas podem indicar os documentos cujas etiquetas HTML têm pouca correspondência com aquelas definidas no modelo do HTMLSUMM. Observamos que para 13 dos 21 documentos (62% dos casos), o HTMLSUMM conseguiu obter o melhor valor de informatividade, isso aponta que a sua superioridade é constante para a maior parte da coleção. Para 5 documentos o Copernic superou o HTMLSUMM. Como ele considera características lingüísticas, esse dado pode indicar que, para certos documentos, considerar somente a estrutura HTML seja insuficiente. Logo, a metodologia utilizada pode ser aprimorada considerando informações de outra natureza, por exemplo, as lingüísticas, e não só as das etiquetas HTML.

Analisando o desempenho médio dos sistemas que corresponderam a variações do HTMLSUMM, podemos observar que a introdução dessas variações não trouxe melhorias com relação ao sistema base (S3).

Apesar da nossa hipótese sobre a utilidade de se estabelecerem pesos diferenciados para as etiquetas HTML, em alguns casos, não houve diferença significativa com relação à informatividade semântica entre os sistemas quando estes utilizavam peso homogêneo igual a 1.

De maneira geral, desconsiderar a etiqueta de *link* <*A*>, não traz prejuízos ao desempenho do sistema. Essa situação é evidenciada, por exemplo, pelo sistema S14, que teve um resultado muito próximo àquele alcançado pelo sistema base. Isso pode ser explicado pelo fato de que os *links*, geralmente, contêm informações que descrevem bem o documento para o qual eles apontam, sendo algumas vezes pouco úteis para indicar o conteúdo do documento em que eles estão definidos (MCBRYAN, 1994).

Tentar privilegiar sentenças mais longas ou mais curtas não parece influenciar a informatividade de uma descrição de um documento *Web*. Isto é evidenciado na Tabela 7, para os sistemas S7 e S8, que obtiveram um valor médio de informatividade muito similar e apresentam como única diferença justamente a variação V3, que privilegia as sentenças mais curtas em detrimento das mais longas.

Os resultados mostram que o processo de remoção de *stopwords* (palavras muito comuns ou sem conteúdo que são consideradas irrelevantes) durante a seleção e extração de informações contribui para que os sistemas gerem descrições mais informativas. Isso é evidenciado na Tabela 7, já que os sistemas com melhores desempenhos médios fazem esse processo.

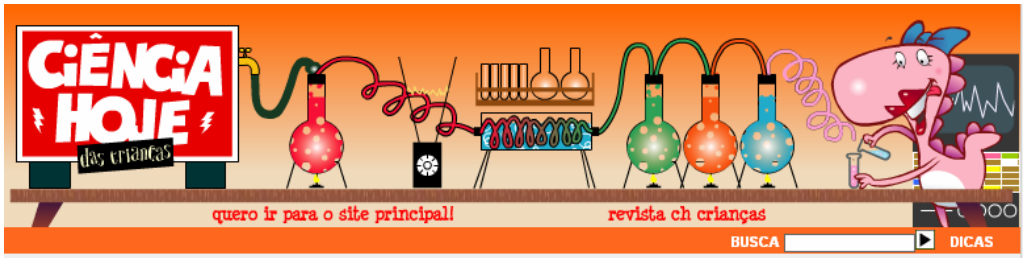
Apesar da baixa informatividade semântica média obtida pelo HTMLSUMM, esse sistema apresentou, dentre os demais, o melhor desempenho médio, superando

mesmo o desempenho do Copernic e do Google, demonstrando que existe um potencial a ser explorado. A conclusão a que chegamos é que o uso de informações estruturais pode ser um fator complementar importante para a geração de extratos de documentos *Web*, pelo seu potencial de produzir conteúdo mais denso e proeminente do que os outros sistemas. Isso justifica o estudo e o aprimoramento desse sistema, principalmente no que diz respeito à combinação e avaliação de outras variações além das aqui consideradas, uma vez que não realizamos uma combinação exaustiva de todas elas. Também pode ser aprimorado o modelo de pesos das etiquetas HTML.

Apesar do potencial evidenciado pelos resultados sobre a utilidade das etiquetas HTML, existem algumas limitações claras referentes ao modelo de sumarização utilizado pelo HTMLSUMM, sendo a principal delas a necessidade da presença no documento das mesmas etiquetas HTML utilizadas pelo sistema. Isso pode ser um problema para o processo de sumarização na forma como ele é feito no HTMLSUMM, na medida em que os documentos não utilizarem aquelas etiquetas definidas no modelo do sistema. Além disso, esse processo de sumarização lida somente com a estrutura de formatação do texto, não fazendo nenhum tipo de processamento de conteúdo que não esteja explicitamente ressaltado por etiquetas, o que pode levar a situações em que grande parte do conteúdo do documento seja ignorado para geração dos extratos. Diante dessas limitações, propusemos outro modelo de sumarização com foco no conteúdo textual de documentos *Web*. O modelo proposto baseia-se em conhecimento ontológico para realizar o processamento semântico do conteúdo textual. Esse modelo é descrito na Seção 4.2. A seguir apresentaremos alguns exemplos de extratos de documentos *Web* gerados pelo HTMLSUMM.

4.1.2.3 Geração de extratos de documentos Web utilizando o HTMLSUMM: ilustração e análise

Nesta seção apresentaremos exemplos de extratos de documentos *Web* gerados pelo HTMLSUMM e uma análise desses extratos. Os documentos-fonte utilizados (identificados por DF1 e DF2) e extratos automáticos correspondentes (identificados por E1H e E2H) são mostrados a seguir.



CH DAS CRIANÇAS ON-LINE
A Turma do Zé Neurim
Antropologia
Arqueologia e Paleontologia
Artes e Literatura
Astronomia e Exploração Espacial
Bichos e plantas
Biologia
Corpo Humano e Saúde
Ecologia e Meio Ambiente
Eventos, Festas e Exposições
Física e Química
Geografia
Grandes Cientistas
História
Matemática
Mágica para os olhos
Tecnologia e Invenções

Artes e literatura

Júlio Verne, um escritor apaixonado pela ciência

O pai da ficção científica escreveu livros que até hoje encantam leitores do mundo inteiro!

Muitos acreditam que ciência é assunto só de cientistas. Grande engano. Ciência é um tema que pode render ótimas histórias. Júlio Verne que o diga! O escritor francês □ que há exatos 100 anos faleceu e, por isso, tem sido lembrado em todo o mundo em 2005 □ é considerado um dos pais da ficção científica. Você sabe o que é isso?

O escritor francês Júlio Verne (1828-1905)

"Ficção científica é um gênero literário dedicado a criar mundos fictícios que, de alguma forma, são diferentes do mundo real em que vivem seus autores", explica Lucia de La Rocque, pesquisadora da Fundação Oswaldo Cruz. "Esses mundos inventados, em geral, são mais avançados nas áreas da ciência e da tecnologia. Isso porque a ficção científica se dá ao luxo de inventar coisas mirabolantes, que os cientistas ainda não têm como realizar! Afinal, ela é literatura e pode usar e abusar da imaginação." Foi o que Júlio Verne fez: em seus livros, criou inventos que, na época, eram impossíveis de produzir!

Nascido em 1828 na cidade portuária de Nantes, na França, Júlio Verne desde criança gostava de observar os navios, o mar e os viajantes. Aos vinte anos, foi estudar direito em Paris. Lá, começou sua carreira literária, com a publicação de algumas peças de teatro. Em 1963, um dos seus contos, *Cinco semanas em um balão*, teve sucesso ao ser publicado. A partir daí, Júlio Verne passou a se dedicar exclusivamente à escrita.

Com histórias futuristas e muito reais, os livros de Verne tornaram-se populares em todo o mundo. O mais famoso, considerado sua obra-prima, é *Vinte mil léguas submarinas*, que conta a história do capitão Nemo e seu submarino, Nautilus. Júlio Verne escreveu essa história em 1873, quando não havia tecnologia para construir um submarino! O primeiro veículo desse tipo só foi feito 25 anos após a publicação do texto.

Mas como um escritor poderia saber tanto sobre ciência a ponto de prever diversas invenções que só viriam a se concretizar no futuro? Sem a formação de um cientista e sem a experiência de um viajante, Verne pesquisava bastante antes de escrever suas histórias. Por isso, bolou ficções científicas cheias de detalhes e que pareciam reais. Além disso, conseguiu retratar muito bem a época em que viveu!

Capa de edições brasileiras de algumas obras de Júlio Verne.

Você já ouviu falar na Revolução Industrial, que ocorreu no século 19? Nesse período, sobretudo na Europa, os investimentos em tecnologia ficaram mais intensos e foram construídas as primeiras máquinas industriais. Havia um clima de progresso no ar, que está presente nas histórias fantásticas que Verne escreveu até a década de 1880. A partir daí, nos romances que o autor publicou, encontramos algo bem diferente: uma atmosfera de pessimismo e insegurança, que refletia o clima do final de século na Europa.

Júlio Verne escreveu muito durante toda a vida. *Vinte mil léguas submarinas*, *Viagem ao centro da Terra*, *A volta ao mundo em oitenta dias* e *Viagem da Terra à Lua* são considerados os livros mais importantes de sua obra. Se você ainda não leu nenhum deles, procure já nas bibliotecas ou livrarias. Ah! E se quiser saber que avanços o autor previu em alguns dos seus livros, clique aqui e fique de queixo caído!

Clara Meirelles
Ciência Hoje das Crianças
15/04/05

enviar matéria para amigo

INÍCIO | O INSTITUTO | CH ON-LINE | REVISTA CH | CH DAS CRIANÇAS | APOIO À EDUCAÇÃO | CONTATO
Instituto Ciência Hoje – Av. Venceslau Brás, 71 / casa 27 – 22.290-140 Rio de Janeiro/RJ – Fone: (21) 2109-8999

Figura 13. Documento-fonte 1 (DF1)

HTMLSUMM - Extrato Automático (90% de compressão): E1H

SH1 O pai da ficção científica escreveu livros que até hoje encantam leitores do mundo inteiro!

SH2 Vinte mil léguas submarinas, Viagem ao centro da Terra, A volta ao mundo em oitenta dias e Viagem da Terra à Lua são considerados os livros mais importantes de sua obra.

SH3 Ciência Hoje das Crianças

SH4 INÍCIO | O INSTITUTO | CH ON-LINE | REVISTA CH | CH DAS CRIANÇAS | APOIO À EDUCAÇÃO | CONTATO

Analisando o extrato E1H referente ao documento-fonte DF1 (Figura 13), cujas sentenças estão identificadas por SH1 a SH4, observamos que SH1 introduz uma referência que não pode ser resolvida no extrato, pois ao citar “O pai da ficção”, o leitor não tem como recuperar a quem esta sentença se refere. A sentença SH2 também introduz problemas ao extrato, se considerado seu contexto original, pois faz referência com “de sua obra” ao agente “Júlio Verne”, contido na sentença anterior (“Júlio Verne escreveu muito durante toda a vida”), omitida em E1H.

Outro problema claro do extrato refere-se a sua textualidade. As quatro sentenças incluídas em E1H são desconexas, prejudicando sua coerência e coesão. Além disso, se considerarmos que o tópico principal de DF1 refere-se ao escritor Júlio Verne e as suas obras de ficção científica, somente as sentenças SH1 e SH2 trazem, em seu conteúdo, alguma relação com este tópico, já que SH3 e SH4 introduzem no extrato informações completamente marginais. Assim, E1H traz 50% de sentenças com conteúdos irrelevantes, que não tem nenhuma relação com a temática principal de DF1. De modo geral, E1H deixa muito a desejar com relação à informatividade, pois omite o agente, personagem ilustre sobre o qual versa o documento, e o evento realizado.

Se observarmos os dados da Tabela 8, que traz os pesos normalizados no intervalo de 0 a 1, a inclusão das sentenças SH3 e SH4 pelo HTMLSUMM é justificável já que, apesar de serem irrelevantes do ponto de vista da manutenção do tópico

principal, tanto SH3 (com peso 35%) quanto SH4 (com peso 100%) contêm várias palavras-chave provenientes das etiquetas de título (<TITLE>) e *links* (<A>).

Tabela 8. Síntese das informações consideradas pelo HTMLSUMM para DF1

E1H	Etiquetas HTML consideradas	Palavras da sentença relacionadas à etiqueta	Peso de cada sentença [0-1]
SH1	<TITLE>	hoje	0.67
	<STRONG>	pai, ficção, científica, escreveu, livros, hoje, encantam, leitores, mundo, inteiro	
SH2	<I>	vinte, mil, léguas, submarinas, viagem, centro, terra, volta, mundo, oitenta, dias, terra, lua,	0.32
SH3	<TITLE>	ciência, hoje, crianças	0.35
SH4	<A>	início, instituto, ch, revista, apoio, educação	1
	<TITLE>	crianças	

Figura 14. Documento-fonte 2 (DF2)

HTMLSUMM - Extrato Automático (90% de compressão): E2H

SH1 Iraque condena cinco à morte e 35 à prisão por terrorismo
SH2 Na terça-feira o mesmo tribunal condenou 25 pessoas, entre elas cinco iraquianos, quatro sauditas, um sírio e um bengali a penas de morte e prisão, também por terrorismo, no caso dos iraquianos, e por entrada ilegal no Iraque, no caso dos estrangeiros.
SH3 Leia o que já foi publicado sobre pena de morte no Iraque
SH4 Leia cobertura completa sobre o Iraque sob tutela
SH5 Promotor aposta em pena mais severa para Cristian Cravinhos
SH6 Novas ações de Israel matam 23 no Líbano; Hizbollah lança foguetes
SH7 Sandy e Junior se estranham nos bastidores de show
SH8 Globo admite que "Páginas da Vida" apela e promete controle

Observando o documento-fonte DF2 (Figura 14) e o extrato correspondente gerado pelo HTMLSUMM, temos que em E2H a sentença SH2 introduz um referente que não pode ser resolvido no extrato já que ela cita “o mesmo tribunal”, não sendo possível recuperar a que tribunal se refere a sentença.

Do ponto de vista da textualidade, as sentenças incluídas em E2H são desconexas, prejudicando sua coerência e coesão.

Considerando que o tema principal do documento é “a condenação imposta pelo tribunal iraquiano”, as duas primeiras sentenças de E2H são as mais relevantes do extrato. A sentença SH1 escolhida é bem informativa neste aspecto, embora SH2 introduza problemas de quebra de referência, ela ajuda a complementar a informação trazida por SH1 já que introduz detalhes sobre o tópico. A escolha de SH1 e SH2 pelo HTMLSUMM se deve ao peso derivado do título (“Iraque condena cinco à morte e 35 à prisão por terrorismo”), já que elas compartilham várias dessas palavras-chave, conforme vemos na Tabela 9. Por outro lado, as demais sentenças de E2H são completamente irrelevantes e em nada contribuem para o extrato.

Tabela 9. Síntese das informações consideradas pelo HTMLSUMM para DF2

E2H	Etiquetas HTML consideradas	Palavras da sentença relacionadas à etiqueta	Peso de cada sentença [0-1]
SH1	<TITLE>	iraque, condena, cinco, morte, prisão, terrorismo	0.92
SH2	<TITLE>	morte, cinco, prisão, terrorismo, iraque	0.76
SH3	<TITLE>	iraque, morte	0.73
	<A>	leia, foi, publicado, pena	
SH4	<TITLE>	iraque, morte	0.73
	<A>	leia, cobertura, completa, tutela	
SH5	<A>	promotor, aposta, pena, severa, cristian, cravinhos	0.6
SH6	<A>	novas, ações, israel, matam, líbano, hizbollah, lança, foguetes	1
SH7		Sandy, junior, estranham, bastidores, show	0.69
SH8	<A>	globo, admite, páginas, vida, apela, promete, controle	0.86

Se considerado como um todo, somente 25% das sentenças de E2H são relevantes. Assim, o extrato E2H deixa muito a desejar nesse aspecto já que mesmo as sentenças marginais apresentam pesos relativamente altos, conforme mostra a Tabela 9, por exemplo, SH6 que tem peso máximo de 100%. Observe que as sentenças de SH3 a SH8 foram selecionadas por incluírem diversas palavras-chave provenientes da etiqueta de *links* <A>, porém essas palavras-chave não são boas representantes do documento-fonte já que se referem a tópicos irrelevantes. Na próxima seção descrevemos o modelo baseado em conhecimento ontológico.

4.2 Sumarização baseada em ontologia

O sumarizador extrativo baseado em ontologia, denominado GEO (**G**erador de **E**xtratos baseado em conhecimento **O**ntológico), combina características lingüísticas e estatísticas para a seleção das informações relevantes de um texto em português. Nossa hipótese é de que a informação semântica recuperada da ontologia permite que o sistema determine quais tópicos são relevantes para a seleção de sentenças e geração de um extrato, aprimorando-o e, conseqüentemente, aumentando a probabilidade de eles de fato contemplarem tópicos relevantes. O sistema faz a identificação de tópicos pela

contagem de conceitos, em vez de contagem de palavras (LIN, 1995) usando a ontologia do Yahoo. Nas próximas seções, apresentaremos as justificativas para a escolha dessa ontologia e o processo de seu enriquecimento para potencializar o seu uso e permitir a sua incorporação ao sistema.

4.2.1 A ontologia do Yahoo para o português

A razão da escolha da ontologia do Yahoo para este trabalho é múltipla: (1) o ambiente Yahoo é uma das principais ferramentas usadas por internautas para encontrar informações hierarquicamente organizadas em categorias na Internet; (2) essa ontologia é uma das maiores já compiladas por humanos: seu conteúdo é organizado por editores que visitam, analisam e incluem sites, organizando-os em 16 categorias principais e subcategorias de acordo com o assunto; (3) seu conteúdo está disponível também em língua portuguesa.

As principais categorias, ou conceitos, do Yahoo incluem: *Artes e Cultura, Esportes, Educação, Ciência, Regional, Business to Business, Fontes de Referência, Saúde, Compras e Serviços, Lazer, Informática, Internet, Notícias, Finanças, Governo e Sociedade*. Cada conceito é descrito por um conjunto de palavras-chave que o caracteriza. As palavras-chave, por sua vez, delineiam um caminho que indica a posição do conceito na hierarquia. Em outras palavras, um subconceito é descrito adicionando-se uma palavra-chave ao conjunto de palavras-chave que caracterizam seu superconceito. Considerando essa sucessão de atribuição de conceitos a cada nó da hierarquia, todos os nós da hierarquia herdarão de forma crescente os conceitos de seus sucessores. Um exemplo de caminho com cinco conceitos inter-relacionados é indicado por “>>”, como segue. O superconceito, neste caso, é *Artes e Cultura* e o subconceito mais elementar ou folha é *Bibi Ferreira*:

Artes e Cultura>>Artes Cênicas>>Artistas>>Atores e Atrizes>>Bibi Ferreira

São os conceitos como esses que são utilizados como possíveis tópicos de um documento *Web* nesta nossa proposta. Porém, a incorporação da ontologia do Yahoo não se deu de forma imediata ao nosso modelo, sendo necessário um processo de enriquecimento que descrevemos a seguir.

4.2.2 O enriquecimento da ontologia do Yahoo para o português

A descrição de conceitos na ontologia do Yahoo não segue um modelo específico. Os itens lexicais utilizados para descrever cada conceito apresentam variações com relação às suas formas. São encontrados, por exemplo, itens que descrevem os conceitos e que são formados por palavras isoladas (como *artesanato*, *dança*, *design*) ou composições de palavras (por exemplo, *Cinema e Filmes* ou *Centros Culturais*). Variações de número e gênero também podem ser percebidas, como nos conceitos *"artistas"*, *"Atores e Atrizes"* e *"dança"*. Nossa decisão de enriquecimento se deve ao fato de que a ontologia do Yahoo do inglês ser considerada pobre (CHEN, 1994; CHEN et al., 1997; FURNAS et al., 1987; TIUN et al., 2001) para dar conta do vocabulário de textos livres (isto é, textos de autoria) como os que estão em foco neste trabalho. O mesmo problema se aplica à ontologia do Yahoo em português, já que ela, basicamente, é uma tradução da versão em língua inglesa. Isso é conhecido como "problema do vocabulário". Esse problema ocorre porque o processo de identificação de tópicos de textos envolve o mapeamento das palavras da língua natural para conceitos ontológicos, assim diferentes palavras se referem muitas vezes ao mesmo conceito ou vice-versa. Claramente fazer esse mapeamento entre conceitos e palavras é extremamente difícil já que a língua natural permite uma série de variações devido aos sinônimos (palavras diferentes com o mesmo significado), à polissemia (a mesma palavra com diferentes significados), às variações léxicas (uso de radicais, conjugações

verbais, variações de gênero e número) e aos chamados quase-sinônimos (palavras correlatas, como "bomba" e "explosão"). O conceito *"Artes e Cultura"*, por exemplo, poderia ter como descritores diversas palavras: artes, cultura, "artes e cultura", música, artesanato, etc.

Por essas razões, optamos por enriquecer nossa ontologia seguindo procedimentos adotados também em outros trabalhos (CHRISTOPHI, 2004; FAATZ; STEINMETZ, 2002; MLADENIC; GROBELNIK, 1999; TIUN et al., 2001; WIVES, 2004). Definimos como enriquecimento, no contexto deste trabalho, o processo de descrever um conceito da ontologia do Yahoo por meio de palavras da língua natural em foco. Nossa metodologia envolveu a coleta manual de um vocabulário externo, utilizando como descritores de conceitos palavras que tenham algum tipo de relação semântica com os conceitos ontológicos. Estas relações podem ser de diversos tipos, por exemplo, sinonímia, hiponímia e hiperonímia. Este enriquecimento permitiu aumentar o poder de generalização da ontologia (por exemplo, dizer que *Bogotá* e *Medellin* remetem ao superconceito derivado do país comum – *Colômbia*). Por questões de simplicidade, o enriquecimento da ontologia do Yahoo para português foi realizado restringindo-se a descrição dos conceitos às palavras encontradas no *thesaurus* Diadorim e na Wikipédia. Itens lexicais coletados corresponderam às palavras de classe aberta (substantivos, verbos e adjetivos), convertidos manualmente para a forma canônica. Além da forma canônica, para substantivos e adjetivos incluímos as variações léxicas mais comuns (forma feminina e plural). Desta forma, um único conceito da ontologia passou ser descrito por itens lexicais diferentes. Por exemplo, o conceito denominado *Atleta* foi descrito por "atleta", "atletas", "desportista", "esportista" etc.

Em um primeiro passo de enriquecimento, utilizamos como descritores as palavras que identificavam os conceitos na ontologia e, posteriormente, seguindo

procedimentos similares aos utilizados por Lin (1995) e Tiun et al. (2001), consideramos a relação semântica de sinonímia, para completar o enriquecimento da ontologia original.

Em uma segunda etapa de enriquecimento, utilizamos os próprios documentos da Internet como fontes externas de conhecimento, para extrair deles os subsídios para enriquecer a ontologia. Mais particularmente, a fonte de conhecimento usada foi a Wikipédia, uma enciclopédia livre em construção por milhares de colaboradores de todo o mundo. As razões de escolha dessa enciclopédia foram as seguintes:

- A Wikipédia é considerada a enciclopédia mais popular do gênero, e possui um amplo conteúdo em língua portuguesa.
- Ela cresceu rapidamente. Em março de 2004, já contava com 70.000 usuários registrados, e mais de 6.000 editores ativos. Já haviam sido criados mais de 1 milhão de documentos em dezenas de línguas (73.621 documentos em português), o que torna o seu conteúdo rico, atualizado e abrangente.
- É consultada como uma fonte de informação séria por muitos leitores e seu material é citado por diversas fontes (LIH, 2004).
- A qualidade de seu material é comparável à qualidade da Enciclopédia Britânica, conforme apontado pela versão *on-line* da revista *Nature* (GILES, 2005).
- Finalmente, o sucesso alcançado pela Wikipédia demonstra que seu conteúdo supre as necessidades dos usuários com relação à confiabilidade e atualização do seu conteúdo (GILES, 2005) e, assim, ela pode melhorar fontes tradicionais de informação (EMIGH et al., 2005; LIH, 2004).

O enriquecimento da ontologia do Yahoo foi feito da seguinte forma: para cada conceito foram coletados manualmente, diretamente da Wikipédia, documentos cuja temática tinha relação com a idéia expressa pelo conceito. Uma vez recuperados, a

seleção das informações relevantes para o enriquecimento foi feita manualmente por um engenheiro do conhecimento. Apesar do caráter subjetivo da intervenção humana na descrição de conceitos alguns trabalhos, por exemplo, (LOH, 2001), sugerem que essa intervenção pode facilitar o processo de descrição de conceitos e melhorar os resultados finais. Após a leitura dos documentos recuperados, o engenheiro de conhecimento escolheu as palavras que melhor descreviam os conceitos já existentes na ontologia do Yahoo, relacionando-os à ontologia. Assim, procuramos garantir um conjunto mínimo de 26.300 descritores acrescentados a essa ontologia. Cabe destacar que esses descritores podiam ser compostos por mais de uma palavra, como por exemplo, “Bibi Ferreira”.

O enriquecimento da ontologia do Yahoo para o português, por meio de duas fontes de natureza distinta – a Wikipédia e o Diadorim – teve um aspecto positivo no processo de generalização de conceitos. Ambos os recursos são complementares, já que a Wikipédia é enciclopédica, enquanto o Diadorim agrega somente informações paratáticas de sinonímia e antonímia. Estas, claramente, não permitem generalizações. Por exemplo, a utilização pura e simples de um *thesaurus*, sem o tipo de léxico agregado pelo engenheiro de conhecimento, usando uma fonte como a Wikipédia, tornaria impossível generalizar e determinar, por exemplo, o conceito *restaurante* a partir de palavras como "garçom", "cliente", "comida" e "menu". Neste exemplo, as palavras "garçom", "cliente", "comida" e "menu" remetem claramente a um local que é expresso por meio do conceito *restaurante*. Esse tipo de informação poderia ser coletado de um texto na Wikipédia que tratasse, por exemplo, da origem e da definição da palavra restaurante. Já o *thesaurus* poderia acrescentar informações mais limitadas, como por exemplo, indicar que a palavra “menu” é sinônimo de “cardápio”.

Atualmente, a ontologia enriquecida por esse processo possui aproximadamente 5.500 conceitos e cerca de 26.300 descritores associados aos conceitos, utilizando a técnica de enriquecimento aqui descrita. Destacamos que o enriquecimento foi feito para 2.500 conceitos originais da ontologia do Yahoo (aproximadamente metade da coleção). Não completamos essa coleção devido ao esforço de se realizar esse enriquecimento manualmente. Embora não tenhamos executado exaustivamente esse processo, a metodologia adotada pode ser replicada para tratar o conjunto de conceitos restantes. Esse repositório foi agregado ao GEO, cuja descrição apresentamos a seguir.

4.2.3 O sistema de geração de extratos GEO

Apresentamos aqui a arquitetura do GEO, que é baseado na ontologia do Yahoo enriquecida, assim como sua metodologia de reconhecimento de informações relevantes para composição de extratos.

4.2.3.1 A arquitetura do GEO

A arquitetura do GEO é exibida na Figura 15.

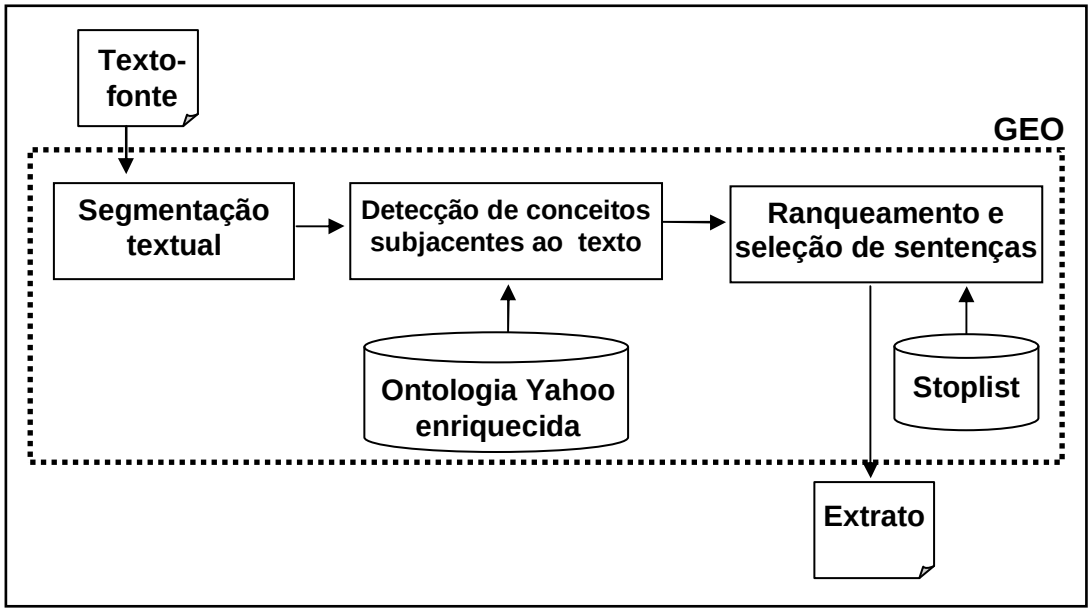


Figura 15. Arquitetura do GEO

De acordo com esta arquitetura, um texto-fonte sem nenhuma formatação HTML é dado como entrada para o sistema e os seguintes passos são realizados para geração do extrato:

1. Inicialmente o texto-fonte é pré-processado, sendo segmentado em sentenças por meio de regras simples baseadas na ocorrência dos sinais de pontuação tradicionais (por exemplo, ponto final, ponto de exclamação e ponto de interrogação);
2. Conceitos subjacentes ao texto são detectados com base na ontologia do Yahoo conforme será descrito na próxima seção.
3. As *stopwords* das sentenças são removidas e, então, essas sentenças são ranqueadas e selecionadas de acordo sua proximidade com os conceitos detectados.
4. Finalmente, as sentenças selecionadas são utilizadas para formar o extrato.

4.2.3.2 Metodologia de sumarização automática do GEO

O primeiro passo do método de sumarização proposto é determinar os tópicos mais importantes do documento a ser sumarizado e estimar sua relevância. Isto é feito verificando se as palavras presentes no texto correspondem àquelas que descrevem os conceitos ontológicos. Esse procedimento de verificação de correspondência é o que chamamos de mapeamento. Sempre que essa correspondência ocorrer, assumimos que aquele conceito é subjacente ao texto e representa, portanto, um de seus tópicos.

Como fator de discriminação da importância dos conceitos, inicialmente é calculado o peso de todos eles. O cálculo do peso de um conceito é feito tendo por base a freqüência das palavras no documento que são mapeadas no conceito, ou seja, que correspondem aos descritores dos conceitos. A freqüência é identificada pela contagem absoluta das ocorrências da palavra no documento. O cálculo do peso de um conceito,

com base na freqüência das palavras, parte do princípio de que a repetição de palavras em um texto é feita pelos autores com o intuito de enfatizar algum assunto e pode ser um indicador de significância das palavras (SALTON; MACGILL, 1983). No GEO, sempre que uma palavra do texto corresponder a um descritor de um conceito, o peso deste conceito é incrementado em 1 unidade; esse processo é cumulativo, ou seja, toda vez que a mesma palavra aparece no texto adiciona-se 1 unidade.

Considerando a estrutura hierárquica da ontologia e os relacionamentos entre os conceitos, a detecção de um conceito subjacente ao documento implica indiretamente a presença de seu conceito-pai também no documento (TIUN et al., 2001). Por exemplo, se considerarmos uma relação ontológica entre os conceitos *Futebol* e *Esporte*, em que *Esporte* é pai de *Futebol*, a presença do conceito *Futebol* em um texto indica que, em um nível mais genérico, o conceito *Esporte* também está presente no texto. Com base no exposto, para tentar modelar esse processo de generalização em que um conceito é detectado a partir de seus conceitos-filhos, sempre que o peso de um conceito é incrementado devido ao mapeamento de uma palavra, o peso de seu pai também é incrementado. Cabe destacar que, no contexto considerado nesta investigação, esse processo de generalização é considerado somente de conceito-filho para seu conceito-pai (contexto imediato). A decisão de considerar somente o contexto imediato foi tomada em virtude da própria estrutura da ontologia do Yahoo. Conforme acusa Fensel et al. (2002), a ontologia do Yahoo provê uma noção básica de generalização e especialização com a maior parte da relações do tipo "é-um". Por exemplo, considerando uma relação entre conceitos indicada por >>, onde o conceito da esquerda é o pai do conceito da direita, em *Esporte>>Artes Marciais>>Capoeira*, temos que *Capoeira* é uma *Arte Marcial*. No esquema geral de organização da ontologia do Yahoo, tipicamente um conceito-filho de um conceito mais específico também mantém

a relação do tipo "é-um" com os conceitos mais genéricos (para o exemplo anterior, podemos considerar que *Capoeira* é também um *Esporte*). Entretanto, nem todas as relações desta ontologia seguem estritamente esse tipo de relação "é-um", fazendo com que o processo de generalização que propomos não seja aplicável sem considerarmos um limite. Por exemplo, considerando um conjunto de conceitos relacionados na ontologia, indicados por *Vestuário>>Vestuário Feminino>>Acessórios Femininos>>Maquiagem*; podemos considerar que *Maquiagem* é um tipo de *Acessório Feminino*, mas já não poderíamos dizer que é um tipo de *Vestuário*. Adicionalmente a esse problema, devido ao grande número de conceitos, a propagação para todos os níveis da ontologia tornaria o processo computacionalmente ineficiente, sobretudo se considerarmos o ambiente *Web* que é o nosso foco final.

Considerando que o mapeamento é um processo recorrente sobre a estrutura hierárquica da ontologia, por esse procedimento de propagação de pesos os conceitos mais próximos da raiz obteriam os maiores pesos e claramente os conceitos terminais seriam prejudicados. Com o intuito de evitar essa situação, procedemos à propagação de pesos entre conceitos pais e filhos de modo similar à usada por Tiun et al. (2001). Assim, sempre que um peso é propagado do conceito-filho para o conceito-pai seu valor é reduzido. No GEO, a propagação de pontos para o conceito-pai corresponde a 25% do peso obtido pelo conceito-filho. Esse valor de redução foi empiricamente escolhido.

Para ilustrar esse processo de determinação de pesos de conceitos considere o texto seguinte, composto por 7 sentenças (Figura 16):

Atletas contra artistas! [1]

Esta disputa aconteceu na última quarta-feira, dia 21, no Mineirão, em Belo Horizonte, através de um jogo de futebol beneficente. [2]

Além da participação de Renato Aragão e Romário, a partida contou com vários atores e com o cantor Daniel. [3]

Além deles, o piloto de Fórmula 1 Michael Schumacher também participou. [4]

O vocalista do grupo Skank, Samuel Rosa, marcou um gol. [5]

Ele mostrou que entende muito além de músicas, pois demonstrou sua habilidade com a bola em campo. [6]

Há quem diga que o resultado da partida não importa e sim o ato, mas a equipe vermelha de Didi perdeu da equipe azul por 3 a 2. [7]

Figura 16. Texto exemplo

A Figura 17 mostra as palavras do texto que foram mapeadas nos respectivos conceitos da ontologia. Na figura, os círculos representam os conceitos da ontologia e as palavras fora dos círculos são as palavras do texto mapeadas; entre parênteses se encontram as freqüências dessas palavras no texto. Como o objetivo é apenas ilustrar o processo de mapeamento, somente consideramos um recorte simples da ontologia com 20 conceitos. Cabe destacar que, apesar de considerarmos as categorias gramaticais substantivos, adjetivos e verbos durante o enriquecimento da ontologia, não necessariamente todas as palavras do texto pertencentes a uma dessas categorias serão mapeadas em algum conceito ontológico. Isso ocorre porque o mapeamento entre palavras e conceitos só é feito se a palavra fizer parte dos descritores dos conceitos associados no processo de enriquecimento descrito na seção anterior. Daí o fato de algumas palavras do texto exemplo, como “disputa”, “aconteceu”, “beneficente” não serem mapeadas em nenhum conceito ontológico.

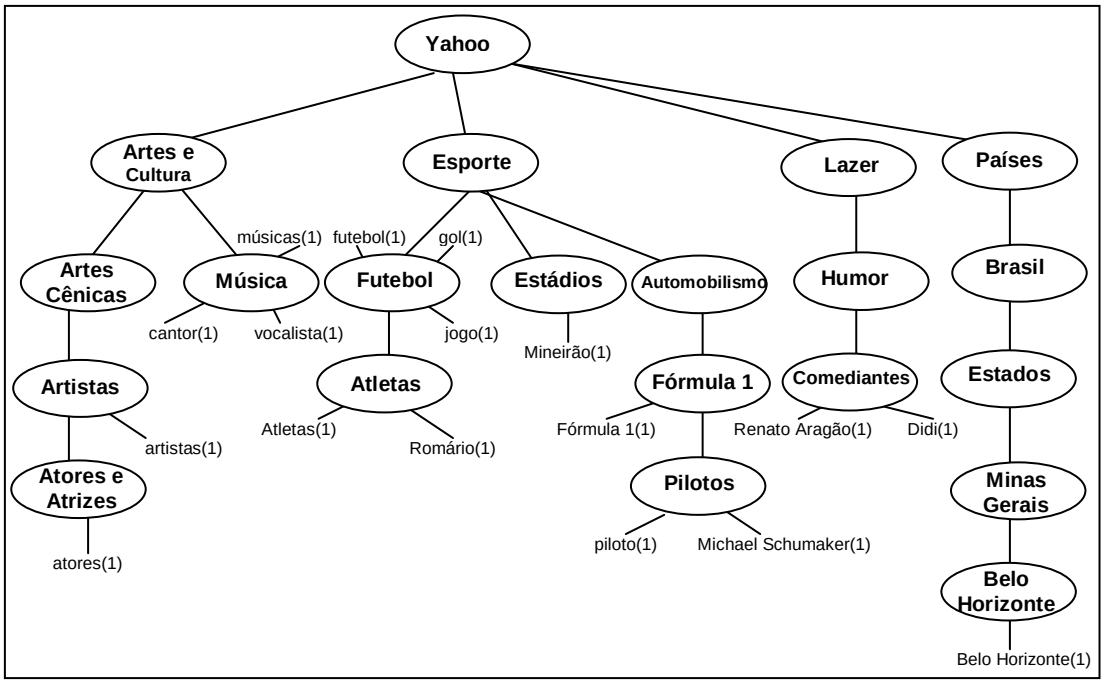


Figura 17. Palavras mapeadas em conceitos ontológicos

Considerando o processo de mapeamento descrito, na Tabela 10 são mostrados os 16 conceitos identificados no texto e seus respectivos pesos.

Tabela 10. Pesos dos conceitos ontológicos

Conceito	Peso	Origem do peso
Futebol	3.5	3 (proveniente das palavras gol, jogo e futebol) + 0.5 (proveniente da propagação do peso do conceito *Atletas*)
Música	3	3 (proveniente das palavras cantor, vocalista e músicas)
Atletas	2	2 (proveniente das palavras atletas e Romário)
Pilotos	2	2 (proveniente dos termos Michael Schumaker e piloto)
Comediantes	2	2 (proveniente dos termos Renato Aragão e Didi)
Fórmula 1	1.5	1 (proveniente do termo Formula 1) + 0.5 (proveniente da propagação do peso do conceito *Pilotos*)
Artistas	1.25	0.25 (proveniente da propagação do peso do conceito *Atores e Atrizes*) + 1 (proveniente da palavra artistas)
Atores e Atrizes	1	1 (proveniente da palavra atores)
Estádios	1	1 (proveniente da palavra Mineirão)
Esporte	1	0.75 (proveniente da propagação do peso do conceito *Futebol*) + 0.25 (proveniente da propagação do peso do conceito *Estádios*)
Belo Horizonte	1	1 (proveniente do termo Belo Horizonte)
Artes e Cultura	0.75	0.75 (proveniente da propagação do peso do conceito *Música*)
Humor	0.5	0.5 (proveniente da propagação do peso do conceito *Comediantes*)
Artes Cênicas	0.25	0.25 (proveniente da propagação do peso do conceito *Artistas*)
Automobilismo	0.25	0.25 (proveniente da propagação do peso do conceito *Fórmula 1*)
Minas Gerais	0.25	0.25 (proveniente da propagação do peso do conceito *Belo Horizonte*)

Cabe observar que o método de mapeamento aqui proposto não trata de fenômenos lingüísticos que podem ocorrer durante este processo, como por exemplo, o fenômeno da ambigüidade lexical introduzida pela polissemia, já que uma mesma palavra pode aparecer como descritor de mais de um conceito, não sendo possível distinguir em que conceito ela deve ser mapeada. Segundo Wives (2004), a melhor alternativa quando não se dispõe de um mecanismo de desambiguação, é utilizar todos conceitos possíveis. Assim, para tentar minimizar esse problema é adotada a estratégia seguinte: uma palavra ambígua é mapeada em todos os conceitos que são possíveis de

serem descritos por tal palavra. Como exemplo, considere as palavras “lutador” e “chave-de-braço” que podem descrever simultaneamente os conceitos *Judô* e *Jiu-Jitsu*. Se estas palavras fossem encontradas descontextualizadas em um documento, não seria possível identificar a qual dos conceitos elas se refeririam. Por essa razão, o mapeamento das palavras em um único conceito não seria possível. Neste caso, ambos os conceitos são considerados em nosso método.

Após identificar os conceitos e seus pesos, nosso sistema seleciona as sentenças que comporão o extrato. Primeiramente, todas as sentenças do documento são ponderadas com base na sua proximidade com os conceitos selecionados. O ranque é feito da seguinte maneira (WU; LIU, 2003):

1 – Pontuam-se as sentenças baseando-se nos conceitos identificados no texto. Para isso é verificado se as sentenças contêm palavras que tenham sido mapeadas nos conceitos ontológicos identificados e, então, é associada a cada sentença uma pontuação referente à quantidade de palavras mapeadas em relação a cada um dos conceitos. Este número de palavras mapeadas é multiplicado pelo peso do conceito. Por exemplo, suponha que uma sentença tenha 4 palavras que tenham sido mapeadas em um conceito cujo peso seja 5, sua pontuação com relação àquele conceito será de 20 pontos (4*5). Assumindo-se que existam *n* conceitos identificados, cada sentença terá *n* escores. O somatório desses escores será a pontuação total da sentença.

2- Computa-se a soma da freqüência de cada uma das palavras (excluindo-se as *stopwords*) da sentença no texto inteiro (BLACK; JOHNSON, 1988). Esta soma, indicada na Equação 3 por f_i, será adicionada aos pontos obtidos no passo 1 pela sentença.

3- Ranqueiam-se as sentenças, e selecionam-se aquelas que comporão o extrato, de acordo com a taxa de compressão desejada.

De forma resumida, é utilizada a seguinte fórmula para pontuar as sentenças:

$$S_i = \left(\sum_{j=1}^{n} w_j * t_{ij} \right) + f_i \quad (3)$$

em que S_i é o escore da i-ésima sentença, w_j é o peso do j-ésimo conceito, t_{ij} é o número de palavras da sentença mapeadas no j-ésimo conceito subjacente ao texto, n é o número de conceitos subjacentes ao texto, e f_i é o somatório das freqüências de todas as palavras da sentença em relação ao texto.

Segundo Wu e Liu (2003) considerar f_i, isto é, a freqüência dos termos em associação à detecção de conceitos com base na ontologia, consiste em uma estratégia positiva na seleção das sentenças, pois melhora a qualidade dos extratos gerados, complementando a estratégia que usa a ontologia, uma vez que a contagem de freqüência é capaz de identificar as palavras mais citadas no texto, mesmo que elas não tenham sido mapeadas em conceitos ontológicos. Já a ontologia pode indicar mais precisamente a relação conceitual subjacente ao texto, que não tem a ver com a freqüência, mas pode ser até mais importante para a extração de informação.

Para ilustrar esse processo de ranqueamento, considere a Equação 3 anterior e o texto da Figura 16, cujas sentenças são reproduzidas a seguir, com a freqüência de cada palavra significativa no texto. Para o cálculo a seguir, consideramos os pesos dos conceitos indicados na Tabela 10 que foram calculados no exemplo anterior.

S1: *Atletas(1) contra artistas(1)!*

S2: *Esta disputa(1) aconteceu(1) na última(1) quarta-feira(1), dia(1) 21(1), no Mineirão(1), em Belo(1) Horizonte(1), através de um jogo(1) de futebol(1) beneficente(1).*

S3: *Além da participação(1) de Renato(1) Aragão(1) e Romário(1), a partida(2) contou(1) com vários(1) atores(1) e com o cantor(1) Daniel(1).*

S4: *Além deles, o piloto(1) de Fórmula(1) 1(1) Michael(1) Schumacher(1) também participou(1).*

S5: *O vocalista(1) do grupo(1) Skank(1), Samuel(1) Rosa(1), marcou(1) um gol(1).*

S6: *Ele mostrou(1) que entende(1) muito além de músicas(1), pois demonstrou(1) sua habilidade(1) com a bola(1) em campo(1).*

S7: *Há(1) quem diga(1) que o resultado(1) da partida(2) não importa(1) e sim o ato(1), mas a equipe(2) vermelha(1) de Didi(1) perdeu(1) da equipe(2) azul(1) por 3(1) a 2(1).*

A pontuação de S1, por exemplo, é dada por:

Pontuação(S1): (1*0) + (1.25*1) + (0.25*0) + (3*0) + (0.75*0) + (2*1) + (3.5*0) + (1*0) + (1*0) + (2*0) + (1.5*0) + (0.25*0) + (2*0) + (0.5*0) +(1*0) + (0.25*0) + (1+1) = 5.25

Considerando o cálculo anterior para S1, onde temos (3.5*0) significa que o conceito *"Futebol"* cujo peso é 3.5 não foi mapeado por nenhuma palavra de S1 (daí o número 0), já (1.25*1) indica que o conceito *"Artistas"* foi mapeado por uma única palavra dessa sentença (neste exemplo, a palavra artistas). O último fator (1+1) corresponde ao termo $f_{i,}$ da Equação 3, representado o somatório da freqüência das palavras de S1. A pontuação das outras sentenças é vista na Tabela 11.

Tabela 11. Pontuação das sentenças

Sentença	Pontuação
S1	5.25
S2	21
S3	19
S4	11.5
S5	13.5
S6	10
S7	17

Considerando as pontuações das sentenças frente aos 16 conceitos identificados, a geração de um extrato priorizaria as sentenças na seguinte ordem: S2, S3, S7, S5, S4, S6, S1. A próxima seção relata a avaliação deste modelo de sumarização proposto.

4.2.3.3 Avaliação do GEO

Para avaliar o potencial do GEO optamos por uma avaliação diferente daquela utilizada para avaliar o HTMLSUMM. A razão de não replicarmos o cálculo da informatividade semântica é que o cálculo da relevância das palavras é baseado na sua proeminência e densidade nos documentos-fonte já anotados com etiquetas HTML. Como os textos de entrada do GEO não são formatados, não é possível calcular a informatividade semântica do mesmo modo realizado anteriormente. Desta forma, adotamos as medidas padrão de precisão e cobertura, definidas, respectivamente, por: a) razão entre o número total de sentenças relevantes incluídas no extrato e seu número total de sentenças; b) razão entre o número total de sentenças relevantes incluídas no extrato e o número total de sentenças do extrato ideal. As sentenças relevantes são aquelas coincidentes com sentenças dos extratos ideais, sendo estes gerados por um sistema particular, como descrito abaixo. Essas medidas, juntamente com sua *f-measure*, foram comparadas àquelas obtidas por outros sumarizadores disponíveis para o português, em um experimento que reproduz sua comparação prévia (RINO et al., 2004).

Mantendo o mesmo cenário prévio, portanto, o corpus de teste escolhido permanece sendo o TeMário (PARDO; RINO, 2003) e os extratos são gerados pelo GEO com a mesma taxa de compressão, tendo um tamanho aproximado de 30% do texto-fonte.

O TeMário é formado por textos jornalísticos e foi especialmente construído para tarefas de avaliação. É composto por três corpora distintos, cada um de 100 textos:

- O corpus de textos-fonte, propriamente dito, que compreende 60 textos da Folha de São Paulo on-line e estão distribuídos igualmente nas seções Especial, Mundo e Opinião e 40 textos do Jornal do Brasil, também on-line, uniformemente distribuídos nas seções Internacional e Política.
- O corpus de referência, de sumários manuais, produzidos por um especialista em língua portuguesa.
- O corpus de extratos ideais, gerados automaticamente pelo GEI (PARDO; RINO, 2004), um gerador de extratos ideais baseado no modelo do espaço vetorial e da medida de similaridade de Salton (1988). Segundo este modelo, o extrato ideal é formado por sentenças do texto-fonte que são mais similares às sentenças do sumário manual correspondente.

Nesse experimento foram consideradas duas versões do GEO, que remetem, na verdade, a duas versões da ontologia, dando origem aos sistemas aqui denominados OntoV1 e OntoV2. Suas características são apresentadas na Tabela 12.

Tabela 12. Características das Ontologias

Versão da Ontologia	Número de conceitos	Número de descritores associados após enriquecimento
OntoV1	5.556	9.415
OntoV2	5.556	26.372

O OntoV1 incorpora a ontologia do Yahoo enriquecida somente com um vocabulário proveniente das relações de sinonímia indicadas no *thesaurus* Diadorim e dos próprios identificadores dos conceitos. O OntoV2 adiciona à OntoV1 também vocabulário proveniente da Wikipédia. Nosso intuito com as duas versões ontológicas

foi verificar se haveria diferença de desempenho dos sumarizadores automáticos, em relação a diferentes formas de enriquecimento.

A Tabela 13 mostra o desempenho do GEO com essas versões, comparado diretamente com os demais sumarizadores do experimento prévio já mencionado, dos quais *From-Top* e *Random order* são sistemas baseline e os demais são todos distintos entre si, envolvendo técnicas de aprendizado de máquina (NeuralSumm, ClassSumm, SuPor) ou métodos estatísticos (GistSumm, TF-ISF-Summ, SuPor).

O sistema *Random Order* não segue nenhum critério para escolha de sentenças, selecionando-as aleatoriamente. O *From-Top* gera os extratos selecionando as primeiras sentenças do texto.

O método de sumarização do SuPor (MÓDOLO, 2003) utiliza um classificador Bayesiano para calcular a probabilidade de uma sentença ser incluída no extrato com base em suas características (*features*). As características usadas são: tamanho da sentença; presença de sintagmas sinalizadores; localização da sentença no parágrafo; freqüência das palavras; cadeias lexicais e presença de nomes próprios. Depois de uma fase de treinamento, o SuPor trabalha do seguinte modo: primeiro, o conjunto de *features* de cada sentença do texto é extraído; depois as sentenças são classificadas e selecionadas para o extrato de acordo com as probabilidades indicadas pelo classificador Bayesiano. Também seguindo esse método de sumarização, o sistema SuPor-v2 utiliza as *features* de um modo mais refinado, garantindo uma combinação que torna o seu desempenho superior ao SuPor.

A exemplo do SuPor, o ClassSumm também utiliza um classificador Bayesiano considerando várias características estatísticas e lingüísticas para a geração de extratos. No total ele considera 16 características que vão desde posição da sentença no texto até ocorrência de pronomes e anáforas. Essas características são associadas às sentenças e

estas, por sua vez, são classificadas em relevantes ou não-relevantes para inclusão no extrato de acordo com as regras aprendidas pelo sistema. O NeuralSumm que também é baseado em aprendizagem não supervisionada, usa um conjunto de características similares às do SuPor para treinar uma rede neural do tipo SOM que é utilizada durante a sumarização. A rede neural é responsável por dividir o texto em dois grupos de sentenças: as relevantes e as não-relevantes. As sentenças relevantes são justapostas para formar o extrato. As características utilizadas por esse sistema envolvem tamanho e posição das sentenças além da presença de palavras-chave e palavras significativas.

O TF-ISF-Summ usa como métrica de relevância de sentenças a TF-ISF (*Term-Frequency Inverse-Sentence-Frequency*). Para cada um dos termos das sentenças é calculado a freqüência do termo no documento (TF) e o número de sentenças em que ele aparece (ISF), finalmente a TF-ISF de cada sentença é calculada como a média aritmética de todos os valores individuais das TF-ISF dos termos. O GistSumm seleciona as sentenças de acordo com sua proximidade com a sentença que expressa a idéia principal do texto: a sentença *gist*. Pata determinar a sentença *gist* o sistema pontua cada sentença somando a freqüência de suas palavras, a de maior escore será considerada a *gist*. Posteriormente as outras sentenças são selecionadas verificando a co-ocorrência de palavras com relação a *gist*, privilegiando a coesão lexical.

Cabe destacar que não reprocessamos os outros sistemas, simplesmente acrescentamos à tabela prévia os novos resultados do GEO (destacados em sombreado). Adicionalmente, também incluímos os resultados de Leite e Rino (2006), que reproduziram da mesma forma esse experimento para o sistema SuPor-v2, uma versão melhorada do sistema SuPor.

Embora o GEO também use métodos estatísticos (*term frequency*) para identificar sentenças relevantes, seu diferencial com relação a todos os sistemas citados

é o processamento ontológico, o que deve explicar o seu melhor desempenho. Podemos observar que tanto OntoV1 quanto OntoV2 superam os resultados obtidos pelo SuPor, o sistema mais bem classificado na avaliação do experimento anterior.

Tabela 13. Comparação com a avaliação de Rino et al. (2004)

Sistemas	P	R	F
OntoV2	48.4	45.2	46.7
OntoV1	48.1	45.0	46.5
SuPor-v2	47.4	43.9	45.6
SuPor	44.9	40.8	42.8
ClassSumm	45.6	39.7	42.4
From-Top	42.9	32.6	37.0
TF-ISF-Summ	39.6	34.3	36.8
GistSumm	49.9	25.6	33.8
NeuralSumm	36.0	29.5	32.4
Random order	34.0	28.5	31.0

Podemos reconhecer algumas semelhanças entre as estratégias utilizadas pelo SuPor e aquelas empregadas em nossa proposta, se observarmos as seguintes combinações de características, respectivamente: freqüência de palavras e sintagmas sinalizadores; freqüência de palavras e ontologia. Em ambos os casos, essas combinações ajudam a determinar os conceitos importantes subjacentes ao texto que vão indicar a relevância de sentenças. Outra característica similar é a utilização de um *thesaurus*, visando explorar a relação de sinonímia entre as palavras. No SuPor ele foi empregado para a computação das cadeias lexicais; no nosso caso, para o enriquecimento da ontologia

Embora o OntoV2 pudesse levar a maior impacto no desempenho do GEO, já que as relações hierárquicas foram consideradas, esses resultados mostram que não há melhora significativa do GEO em relação ao uso exclusivo de relações de *thesaurus.* No entanto, ambas as versões que incorporam ontologia têm melhor desempenho que o dos demais sistemas, evidenciando que o conhecimento ontológico influencia positivamente a inclusão de informações relevantes nos extratos. Ou seja, o GEO capta de modo

significativo a conectividade semântica entre os componentes textuais, reconhecendo de forma mais refinada a relevância de sentenças a serem incluídas em seus extratos.

Comparado ao SuPor e ao Supor-v2, o OntoV2 apresenta uma melhora de aproximadamente 4% e 1%, respectivamente, com relação à *F-measure*. Essa diferença sugere que esse resultado tenha sido em decorrência da ontologia, embora tenhamos também a influência do método estatístico utilizado pelo GEO. Os resultados indicam que a ontologia é capaz de reconhecer conceitos subjacentes ao texto, e, conseqüentemente, determinar as sentenças mais importantes para composição do extrato de modo ligeiramente superior ao feito pelo SuPor e pelo SuPor-v2, que não tratam a informação no nível semântico.

Se compararmos o desempenho do OntoV2 aos dois sistemas baselines, a melhora de *F-measure* é de cerca de 13%, variando de 10% (*From-Top*) a 16% (*Random Order*). A diferença significativa para a *Random Order* era esperada, já que ele apenas seleciona as sentenças de modo aleatório. O *From-Top*, apesar de sua metodologia de seleção de sentenças bem simples, está apenas 10% abaixo do desempenho de OntoV2 e OntoV1. Isso pode ser explicado pelo fato de o experimento ter sido realizado com artigos jornalísticos, que geralmente apresentam as primeiras sentenças como as mais relevantes.

Para os sistemas que usam aprendizado de máquina, ClassSumm e NeuralSumm, a melhora da *F-measure* é de aproximadamente 10%, variando respectivamente 5% e 15%. O ClassSumm também não faz nenhum tipo de consideração sobre a conectividade semântica entre elementos textuais, o que deve explicar seu desempenho inferior. O desempenho inferior do NeuralSumm pode ser explicado, em parte, pelo uso de conhecimento ontológico, mas também devemos considerar que por se tratar de um

sistema conexionista, o pequeno número de textos do TeMário pode ter influenciado negativamente seu treinamento.

Os sistemas que usam métodos estatísticos (TF-ISF-Summ, GistSumm) têm um desempenho médio inferior ao OntoV2 de cerca de 11%, respectivamente esse desempenho é 10% e 13% inferior ao nosso. Os sistemas TF-ISF-Summ e GistSumm são similares, tendo sua metodologia centrada fortemente na freqüência de palavras. Neste aspecto, eles se assemelham ao GEO, já que ele também faz uso de freqüência, porém o diferencial está sobretudo na ontologia, o que deve explicar os resultados superiores obtidos por OntoV1 e OntoV2.

Para verificarmos o desempenho do GEO considerando taxas de compressão maiores, repetimos o experimento para todos os 100 textos do TeMário. Somente calculamos a precisão, considerando uma taxa de compressão de 80% e também a seleção de uma única sentença. Conforme podemos ver na Tabela 14, o desempenho do GEO, utilizando OntoV1, mostrou-se relativamente satisfatório.

Tabela 14. Precisão do OntoV1 para extratos curtos

Precisão (%)	
80% de compressão	51.1
Seleção de 1 sentença	64.0

Os dados da Tabela 14 sugerem que o conhecimento ontológico pode influir positivamente na determinação de informações relevantes para a geração de extratos, mesmo que estes sejam significantemente comprimidos. Isso é particularmente interessante para nossa investigação, sugerindo que ele tem potencial para ser utilizado no contexto de exibição de mecanismos de busca, onde extratos curtos são utilizados como padrão para a apresentação de resultados.

Apesar de ter superado os outros sistemas, a desvantagem do GEO é que ele exige conhecimento profundo enquanto os outros sistemas, quando exigem treinamento, este é feito de modo não supervisionado. Diante desse fato, sistemas como, por exemplo, o SuPor-v2 que ficou a apenas 1% do GEO, podem vir a superá-lo já que possuem maior potencial de treinamento.

A partir dos resultados comparativos aqui apresentados, resolvemos incorporar a ontologia do Yahoo enriquecida ao HTMLSUMM, resultando no ExtraWeb. O ExtraWeb representa aquela que é a maior contribuição desta investigação: um sistema de sumarização extrativo de documentos *Web*. Nosso objetivo, ao propormos a incorporação das metodologias utilizada pelo GEO e pelo HTMLSUMM em um único sistema, foi o de gerar extratos de documentos *Web* que pudessem preservar o conteúdo mais relevante considerando tanto a estrutura do documento (indicada pelas etiquetas HTML) quanto o seu conteúdo. Ao lidar com essas duas dimensões, acreditamos que o ExtraWeb possa gerar extratos que sejam bons representantes de documentos, já que utiliza as estratégias do GEO e do HTMLSUMM paralelamente para a seleção de sentenças. Particularmente para documentos *Web* esta nos pareceu ser uma estratégia com potencial a ser investigado já que, sobretudo na Internet, a forma de um documento é tão importante quanto o seu conteúdo. O ExtraWeb é descrito no Capítulo 5. A seguir apresentaremos alguns exemplos de extratos de documentos *Web* gerados pelo GEO.

4.2.3.4 Geração de extratos de documentos Web utilizando o GEO: ilustração e análise

Nesta seção apresentaremos exemplos de extratos de documentos *Web* gerados pelo GEO e uma análise desses extratos. Cabe destacar que os documentos-fonte utilizados (identificados por DF1 e DF2) são os mesmos das Figuras 13 e 14 (da Seção

4.1.2.3) utilizados na geração de extratos pelo HTMLSUMM e por esse motivo eles não foram reproduzidos nessa seção. Os extratos automáticos (identificados por E1G e E2G) gerados pelo GEO, que correspondem aos documentos DF1 e DF2 respectivamente, são mostrados a seguir.

GEO - Extrato Automático (90% de compressão): E1G

SG1 O pai da ficção científica escreveu livros que até hoje encantam leitores do mundo inteiro!

SG2 "Ficção científica é um gênero literário dedicado a criar mundos fictícios que, de alguma forma, são diferentes do mundo real em que vivem seus autores", explica Lucia de La Rocque, pesquisadora da Fundação Oswaldo Cruz.

SG3 Foi o que Júlio Verne fez: em seus livros, criou inventos que, na época, eram impossíveis de produzir!

Analisando o extrato E1G gerado pelo GEO referente ao documento-fonte DF1 (Figura 13), cujas sentenças estão identificadas por SG1 a SG3, observamos que em E1G a sentença SG3 introduz uma incoerência, se considerado seu contexto original, pois faz referência (com "Foi o que Júlio Verne fez") ao "dar-se o luxo de inventar coisas mirabolantes", contido na sentença anterior, omitida em E1G. No entanto, a primeira sentença desse parágrafo, por ser mais genérica e introduzir claramente a criação de mundos fictícios, permite o estabelecimento da relação entre SG3 e SG2, muito embora ligeiramente modificada.

A sentença SG1 poderia introduzir outro problema, se não houvesse a referência catafórica a Júlio Verne em SG3, já que um leitor não saberia de quem o texto fala.

De modo geral, o extrato S1G é coerente e coeso. As sentenças incluídas nesse extrato são bastante informativas em relação ao agente e evento realizado, remetendo bastante satisfatoriamente ao tópico principal do documento *Web*. Observando os dados da Tabela 15, as três sentenças incluídas no extrato estão associadas a conceitos que são realmente relevantes para DF1 como, por exemplo, "Livros", "Ficção Científica", "Gênero Literário" e "Júlio Verne".

Tabela 15. Síntese das informações consideradas pelo GEO para DF1

E1G	Conceitos considerados[14]	Palavras da sentença mapeadas no conceito	Peso de cada sentença
SG1	Compras e Serviços>>**Livros**	livros	0.83
	Artes e Cultura>>Literatura>>Gêneros literários>>**Ficção Científica**	ficção científica	
	Sociedade>>Família>>**Pais e Filhos**	pai	
SG2	Artes e Cultura>>Literatura>>**Autores**	autores	0.84
	Construção Civil>>Trabalho de Campo da construção civil>>**Fundações**	fundação	
	Artes e Cultura>>Literatura>>**Gêneros literários**	gênero literário	
SG3	Compras e Serviços>>**Livros**	livros	1
	Artes e Cultura>>Literatura>>Autores>>**Júlio Verne (1828-1905)**	júlio verne	

GEO - Extrato Automático (90% de compressão): E2G

SG1 Entre os estrangeiros estão um cidadão sírio e seis sauditas, que foram condenados a penas de prisão de entre dez e quinze anos por ter entrado de modo ilegal em território iraquiano.

SG2 Na terça-feira o mesmo tribunal condenou 25 pessoas, entre elas cinco iraquianos, quatro sauditas, um sírio e um bengali a penas de morte e prisão, também por terrorismo, no caso dos iraquianos, e por entrada ilegal no Iraque, no caso dos estrangeiros.

SG3 As autoridades iraquianas sustentam que a maioria desses combatentes são membros do braço iraquiano da organização terrorista Al Qaeda, cujo líder, Abu Musab al Zarqawi, morreu em um ataque aéreo americano em junho ao norte de Bagdá.

Observando o documento DF2 (Figura 14) e o extrato E2G, temos que a sentença SG2 introduz uma referência não resolvida no extrato (“o mesmo tribunal”) que está relacionada com “tribunal iraquiano”, citada na primeira sentença de DF2, mas que foi omitida de E2G. A sentença SG1 faz uma citação a “entre os estrangeiros” que se refere a “oito estrangeiros”, também mencionado na primeira sentença de DF2.

Do mesmo modo que as outras duas sentenças, SG3 também introduz uma referência não resolvida, pois cita a “maioria desses combatentes” que se refere, em DF2, à informação “centenas de combatentes estrangeiros” contida em uma sentença

[14] Os conceitos considerados são aqueles em negrito. O caminho completo da ontologia foi incluído para facilitar a contextualização

omitida do extrato. A ausência desses antecedentes prejudica a compreensão do extrato, tornando-o pouco coeso.

Embora os problemas citados prejudiquem a textualidade do extrato E2G, do ponto de vista da informatividade, ele consegue filtrar do documento-fonte sentenças que, ainda que desconexas, se relacionam ao tópico principal de DF2 que é "a condenação imposta pelo Iraque por terrorismo". Observando os dados da Tabela 16, a maioria dos conceitos relacionados às sentenças são relevantes, pois têm relação direta com o tópico principal de DF2. As duas sentenças de maior peso, SG2 (97%) e SG3 (100%), por exemplo, estão associadas aos conceitos "Iraque" e "Terrorismo", que são centrais no documento-fonte.

Tabela 16. Síntese das informações consideradas pelo GEO para DF2

E2G	Conceitos considerados	Palavras da sentença mapeadas no conceito	Peso de cada sentença
SG1	Ciência>>Agropecuária>>**Culturas e Solos**	território	0.92
	Regional>>Países>>**Síria**	sírio	
	Regional>>Países>>**Arábia Saudita**	sauditas	
	Regional>>Países>>**Iraque**	iraquiano	
SG2	Regional>>Países>>**Iraque**	iraquianos, Iraque	0.97
	Regional>>Países>>**Arábia Saudita**	sauditas	
	Sociedade>>**Morte**	morte	
	Sociedade>>Crime>>Tipos de Crime>>**Terrorismo**	terrorismo	
	Regional>>Países>>**Síria**	sírio	
	Entretenimento>>**Ingressos**	entrada	
SG3	Regional>>Países>>**Iraque**	Iraquianas, iraquiano, Bagdá	1
	Sociedade>>Crime>>Tipos de Crime>>**Terrorismo**	terrorismo, Al Qaeda, terrorista	
	Business to Business>>**Firmas e Escritórios**	organizações	
	Regional>>**Agências e Empresas**	organizações	
	Organizações e Associações	organizações	
	Regional>>Países>>**Estados Unidos**	americano	

4.3 Geração de extratos de documentos Web: comparação de desempenhos do HTMLSUMM e GEO

O que se nota pelos exemplos é que a geração dos extratos pelo HTMLSUMM é influenciada pelo *layout* das páginas *Web*. Páginas com menus, imagens e propagandas, que trazem geralmente informações marginais sem nenhum tipo de relação direta com o

conteúdo relevante da página, têm o processo de seleção de sentenças claramente prejudicado. O documento-fonte DF2 (Figura 14) é um exemplo desse tipo, no qual podemos distinguir visualmente uma região central, onde está localizado o conteúdo relevante do documento, circundado por menus, figuras e propagandas que, apesar de representarem a maior parte da informação do documento, são, de fato, informações irrelevantes. Os menus trazem tipicamente informações delimitadas pela etiqueta de *links* <A>, que é usada pelo HTMLSUMM para seleção de palavras-chave. Nessas situações, em vez de ajudar na seleção de conteúdos relevantes, os *links* acabam introduzindo informações marginais nos extratos, vide, por exemplo, sentenças SH3 a SH8 no extrato E2H (Seção 4.1.2.3). Portanto, em páginas com este *layout* a estratégia que utiliza etiquetas HTML, sobretudo influenciada pela etiqueta de *links* <A>, parece relativamente menos útil para gerar os extratos já que ela pode introduzir informações marginais. Os títulos das páginas, geralmente muito úteis para contextualizar (SH1 em E2H (Seção 4.1.2.3), por exemplo), pode também ter uma influência negativa no modelo do HTMLSUMM quando eles trouxerem, por exemplo, informações marginais como o nome do autor do site (vide SH3 em E1H (Seção 4.1.2.3)). Nessas situações, o modelo não é bom o suficiente para excluir as informações marginais, já que ele apenas considera a etiqueta para extração de palavras-chave sem fazer nenhuma consideração sobre o *layout* da página, ou seja, sobre as regiões que são potencialmente fornecedoras de informações relevantes.

Uma alternativa para aprimorar o modelo seria classificar as regiões do documento, detectando o maior corpo da página e somente extrair sentenças dessa região, também pode-se considerar excluir palavras isoladas e as sentenças muito curtas que indicam, em muitos casos, os menus. Outro problema claro da estratégia é que a seleção das palavras-chave está restrita apenas a algumas poucas regiões da página, e

não necessariamente àquelas regiões mais relevantes. Considerando as limitações do modelo, essa estratégia de sumarização possivelmente poderá desempenhar melhor para documentos compostos por um corpo simples, sem muitas informações marginais. Nesse caso, a seleção de palavras-chave para composição dos extratos pelo HTMLSUMM tende a ser mais refinada, fazendo com que ele selecione as sentenças mais relevantes.

A estratégia do GEO que é baseada em conhecimento ontológico mostra que a ontologia consegue captar alguns tópicos relevantes para os documentos, conseguindo selecionar, na maior parte dos casos, o conteúdo mais importante da página, filtrando grande parte das informações marginais (por exemplo, os extratos E1G e E4G (Seção 4.2.3.4)). O que se nota é que desempenho do sistema, diferentemente do HTMLSUMM, parece não sofrer muita influência do *layout* da página, mesmo o mapeamento sendo realizado também com palavras de regiões marginais do documento. De todo modo, o sucesso dessa estratégia depende de que a ontologia consiga detectar corretamente os tópicos mais relevantes, o que só ocorre quando há explicitamente o mapeamento entre as palavras do documento e os conceitos ontológicos. Apesar das limitações da ontologia, o que se observa, pelos exemplos, é que os extratos gerados pelo GEO conseguem ser relativamente informativos, principalmente se comparados àqueles gerados pelo HTMLSUMM. No próximo capítulo é apresentado o sistema ExtraWeb.

5 ExtraWeb: um sumarizador de documentos Web baseado em etiquetas HTML e ontologia

O ExtraWeb, sigla para **Extra**tos de documentos **Web**, é um sistema de sumarização automática de documentos que faz uso de informações provenientes da ontologia do Yahoo enriquecida e das etiquetas HTML do próprio documento para a geração de extratos. Ele aplica, independentemente, os métodos de sumarização do GEO e do HTMLSUMM, descritos anteriormente, para gerar os extratos dos documentos.

5.1 Arquitetura do ExtraWeb

A arquitetura do ExtraWeb (Figura 18) agrega as arquiteturas do GEO e do HTMLSUMM. Agora, um texto-fonte é segmentado como no HTMLSUMM, mas suas sentenças são ranqueadas independentemente usando os métodos do GEO e do HTMLSUMM. A pontuação final das sentenças pelo ExtraWeb é, então, calculada como segue:

1. As pontuações obtidas, independentemente, são normalizadas para que os dois métodos tenham a mesma influência sobre a relevância das sentenças.
2. As pontuações de cada sentença são, então, somadas.
3. Uma vez classificadas, elas dão origem ao extrato final.

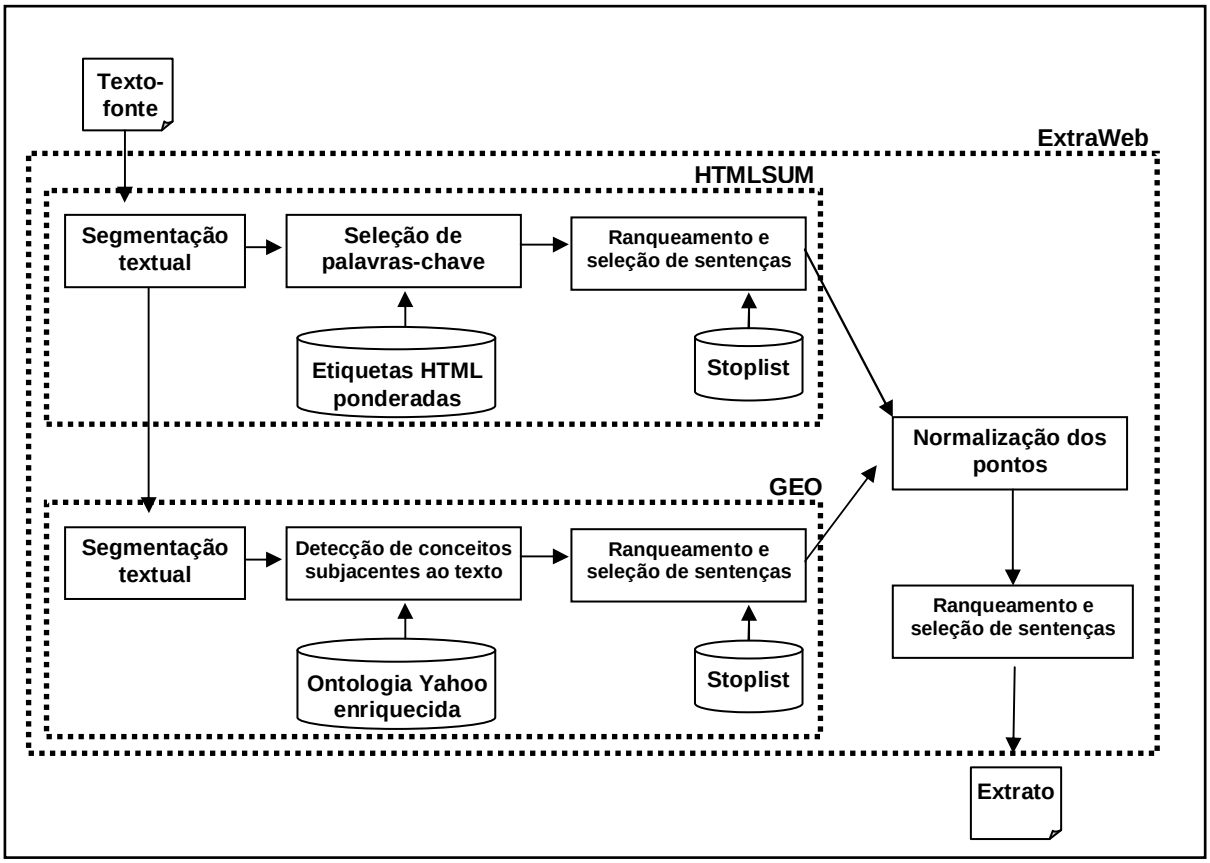


Figura 18. Arquitetura do ExtraWeb

A escolha por utilizar uma função de soma simples para determinar a pontuação final das sentenças foi feita devido a este tipo de função ser utilizado, em trabalhos clássicos da área de SA, para combinar diferentes métodos de seleção de sentenças, vide, por exemplo, (EDMUNDSON, 1969).

Para avaliar o potencial do sumarizador ExtraWeb, propusemos um experimento que mediu sua utilidade para tarefas de recuperação de documentos, em um contexto de busca na Internet, envolvendo a participação de juízes humanos. Diferentemente do nosso experimento anterior, em que a precisão e a cobertura foram o foco central, medindo a capacidade de o sistema recuperar informações relevantes, propusemos uma sistemática de avaliação que pudesse fornecer alguma indicação sobre as preferências do usuário e sua interação com a forma como os documentos são apresentados por meio

de suas descrições. Coletar essas informações é relevante, visto que a interação dos usuários com os sumários produzidos por um sistema de SA é importante, não podendo ser capturada apenas com as avaliações automáticas, sem nenhum tipo de *feedback* dos usuários. O experimento proposto é descrito na próxima seção.

5.2 Avaliação do ExtraWeb: a relevância de documentos indicada por seus extratos

Já que a utilidade de um sumário pode ser medida em uma tarefa extrínseca de julgamento de relevância (DORR et al., 2005), optamos por uma avaliação envolvendo juízes humanos e extratos gerados pelo ExtraWeb. Extratos de documentos resultantes de acesso direto a *websites* foram apresentados a um comitê de juízes, usuários da *Web*, para avaliação de sua utilidade como descrições dos documentos completos. Assim, procuramos verificar se os extratos apresentados foram úteis para indicar a relevância dos documentos. A opção por esta avaliação extrínseca, diferente das avaliações do GEO e do HTMLSUMM, foi feita em decorrência dos resultados inexpressivos obtidos com a informatividade semântica e da impossibilidade de usar o TeMário, como no GEO, já que ele não contempla documentos formatados em HTML.

Testes de significância estatística, para as respostas dos juízes, foram feitos com base no valor de p do teste de Mann-Whitney. Somente quando $p<5\%$, a comparação entre dados é estatisticamente significante. Adotamos um cenário de avaliação, usado anteriormente em outros trabalhos (AMITAY, 2001; WHITE et al., 2002), que simula uma sessão de usuário em um contexto de recuperação de informação, conforme descrito a seguir. A maior parte da configuração do experimento foi herdada do trabalho de Amitay (2001).

5.2.1 Objetivos do experimento

O objetivo principal do experimento foi avaliar as descrições produzidas pelo ExtraWeb em relação àquelas produzidas pelo Google e por um sistema baseline, no contexto de um mecanismo de busca. Esta é uma comparação relativa, já que não existem diretrizes que definam com exatidão o que seja uma boa descrição, e este conceito de bom está intimamente ligado à percepção do usuário, sendo este avaliado de forma completamente subjetiva.

Procuramos avaliar, portanto, não a eficácia de recuperação dos sistemas, mas sim se as formas de apresentação de resultados são adequadas para atender as necessidades de informação dos usuários, isto é, se elas trazem informações dos documentos-fonte que são úteis ou relevantes para o usuário.

Para isso, os extratos do ExtraWeb foram comparados aos de dois outros sistemas: o Google e um baseline. As descrições do Google são formadas com base na consulta dos usuários, onde um termo de consulta e seu contexto (palavras próximas) é extraído do documento-fonte recuperado pelo sistema. As descrições utilizadas no experimento foram coletadas diretamente do Google, representando, portanto, descrições reais.

O modo como as descrições foram exibidas pelo sistema baseline correspondeu a um estilo utilizado em conferências de avaliações de sumarizadores como, por exemplo, a SUMMAC (MANI et al., 2002). Nesse estilo, as descrições foram formadas exibindo o conteúdo inicial do documento-fonte (primeiras linhas), limitando-as a um determinado número de palavras. Neste experimento, o tamanho dos extratos baseline, assim como das demais descrições, foi limitado a, aproximadamente, 22 palavras, lembrando da restrição à unidade mínima sentencial manipulada já mencionada (Seção 4.1.1.1).

5.2.2 Hipóteses do experimento

Ao planejarmos esse experimento, os seguintes pontos foram observados:

a) O experimento deveria examinar um problema real de um cenário de busca na Internet:

1 – Mecanismos de busca geralmente apresentam poucos resultados por página (INKTOMI, 2003; SPINK et al., 2000);

2 – Existe mais que um documento relevante;

3 – O usuário precisa analisar as descrições dos documentos para verificar a sua relevância;

4 – O usuário deve decidir que documento seguir primeiro;

b) O experimento deveria tentar responder às seguintes questões:

1 – Que tipo de descrições é o preferido?

2 – O tipo de descrição influencia a interação do usuário com os resultados?

3 – Qual a utilidade das descrições providas pelo ExtraWeb?

4 – Como esta utilidade está com relação àquela demonstrada em relação aos outros sistemas?

Seguindo as considerações anteriores, o experimento foi conduzido de modo a permitir que os usuários acessassem os documentos através dos extratos. Por questões de simplicidade, nesta investigação, não distinguimos simples trechos de documentos de extratos baseados em sentenças reais, embora os primeiros sejam totalmente não textuais (isto é, somente um fragmento de informação) e os últimos geralmente coesos e coerentes.

A nossa hipótese principal neste experimento foi que a simulação de um cenário de busca, tendo como foco apenas os extratos para recuperação de documentos, seria capaz de detectar a satisfação dos juízes com a utilidade dos extratos, e que esta utilidade, indicada pelos juízes respondendo a um conjunto de questões, refletiria diretamente a qualidade dos extratos.

Para cada documento e consulta, três extratos diferentes foram produzidos, sendo gerados por três sumarizadores distintos: ExtraWeb, Google e baseline. O Google foi escolhido porque ele é amplamente usado e é considerado, por muitos, o melhor sistema de busca (GRIESBAUM, 2004). Ele geralmente produz somente extratos baseados na consulta do usuário que não têm nenhum tipo de compromisso com a coerência ou com o conteúdo principal do documento. O sumarizador Baseline somente coleta as primeiras linhas do documento. Ambos os sistemas mostram extratos de 1 a 2 linhas. Os documentos-fonte usados na tarefa de avaliação foram obtidos, previamente, usando o Google. Somente os *top* 3 documentos foram considerados na avaliação sob a hipótese de que eles são os mais relevantes. Após exibir os extratos para os juízes, sua tarefa foi compará-los e responder a um questionário, como será explicado posteriormente.

Claramente esta é uma avaliação temporal e subjetiva, mas nossa hipótese é que ela sirva como parâmetro para medir o contentamento do participante em relação aos documentos e seus extratos, associados a uma tarefa de busca e, principalmente, para indicar a utilidade dos extratos gerados pelo sistema ExtraWeb.

5.2.3 Metodologia

Sessenta pessoas foram convidadas a contribuir com a avaliação, atuando como juízes. A escolha dos juízes foi livre: o experimento foi disponibilizado na Internet, podendo ser acessado através de qualquer navegador. O experimento foi realizado do

seguinte modo: ao juiz foi apresentada uma tarefa inicial de busca de informação sobre um determinado tópico (aleatoriamente atribuído). Então, ele foi informado sobre um termo de busca X usado para recuperar documentos relativos à sua tarefa, que foi submetido a um mecanismo de busca (a consulta fazia parte da tarefa do experimento e não podia ser modificada ou controlada pelo juiz). Posteriormente, uma lista com três resultados de pesquisa (*links* de sites) com uma curta descrição associada foi apresentada ao juiz. Foi pedido que ele escolhesse apenas um resultado que ele acreditasse ser o mais relevante para sua tarefa. Após visualizar o conteúdo do documento escolhido, cada juiz deveria responder a cinco questões, graduando suas respostas numa escala de 1 a 7 pontos. O fluxo do experimento é mostrado na Figura 19. A primeira tela do experimento, com as instruções dadas aos participantes, é mostrada na Figura 20. As seções seguintes detalham alguns aspectos relacionados ao experimento.

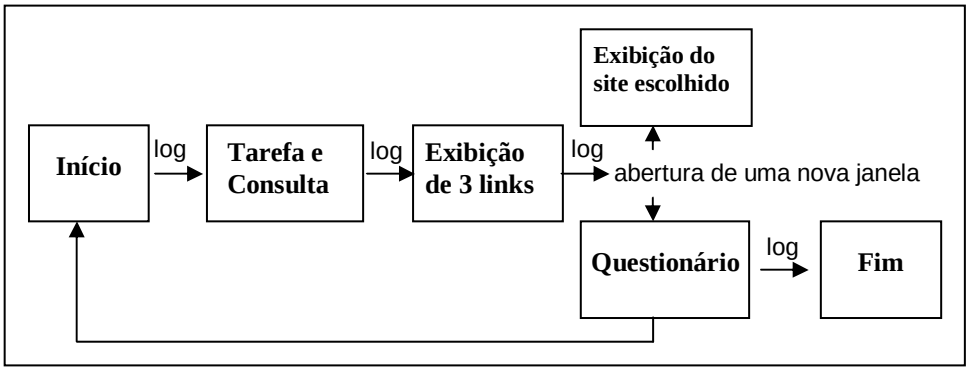


Figura 19. Fluxo do experimento

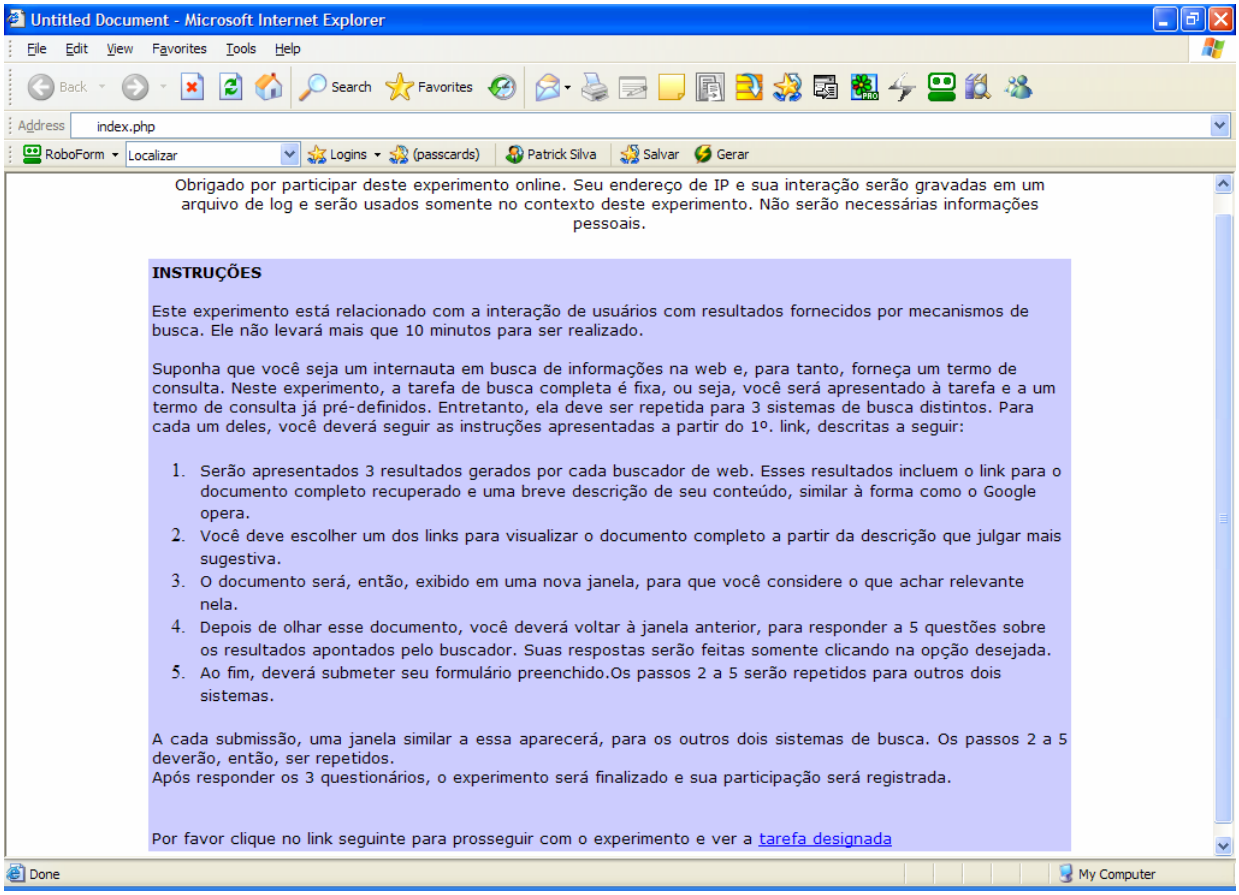


Figura 20. Tela de instruções

5.2.3.1 Escolha das consultas

Em nosso experimento, as consultas foram incorporadas às tarefas de avaliação previamente à interação dos juízes e não podiam ser modificadas. Isso foi feito para garantir que eles interagissem com os mesmos conjuntos de resultados relativos às mesmas consultas. Desta forma, poderíamos comparar três diferentes tipos de descrições. Essa associação permitiu-nos focar a avaliação das diferentes descrições geradas pelos sistemas sem nos preocuparmos com o modo como os juízes deveriam submeter consultas aos três sistemas para alcançar o mesmo conjunto de resultados (o que seria impossível de ser feito em tempo real).

Consultas curtas (1-3 palavras) foram fixadas, estando relacionadas a temas como nome de pessoas famosas, atividades de lazer popular, temáticas de pesquisa,

saúde, etc. Esses são os mesmos temas de Amitay (2001). As consultas, nesse caso, representaram consultas típicas da *Web* (SPINK et al., 2000; WOLFRAM et al., 2001). Para cada consulta foi associada uma tarefa que explicitava seu contexto de submissão ao mecanismo de busca, indicando seu propósito. As cinco tarefas e consultas correspondentes foram as seguintes:

Tarefa 1. Você está ajudando um estudante do segundo grau em uma pesquisa sobre Albert Einstein. Com o objetivo de encontrar um documento dedicado ao assunto você submete o termo "Albert Einstein" a um mecanismo de busca.

Tarefa 2. Um amigo seu escutou uma notícia sobre o Projeto Genoma e perguntou se você poderia ajudá-lo a encontrar mais informações sobre esse projeto. Com o objetivo de encontrar um documento dedicado ao assunto, você submete o termo "Projeto Genoma" a um mecanismo de busca.

Tarefa 3. Você deseja aprender sobre a arte de dobraduras de papel (origami). Com o objetivo de encontrar um documento dedicado ao assunto, você submete o termo "origami" a um mecanismo de busca.

Tarefa 4. Você deseja encontrar boas imagens e boas fontes de informação sobre vida selvagem para disponibilizar links em sua página pessoal. Com o objetivo de encontrar um documento dedicado ao assunto, você submete o termo "vida selvagem" a um mecanismo de busca.

Tarefa 5. Você deseja explicar o conceito de fusos horários para um amigo, procurando um jeito simples de demonstrá-lo. Com o objetivo de encontrar um documento dedicado ao assunto você submete o termo "fusos horários" a um mecanismo de busca.

Cada juiz recebeu uma tarefa que foi aleatoriamente atribuída por um *script*. O *script* também gravou qual tarefa foi atribuída ao participante e os horários de interação

com o experimento. Essas informações foram incluídas em um arquivo de *log* juntamente com o número do IP do juiz, a tela associada e suas respostas. A Figura 21 é um exemplo de uma tela de tarefa exibida aos participantes.

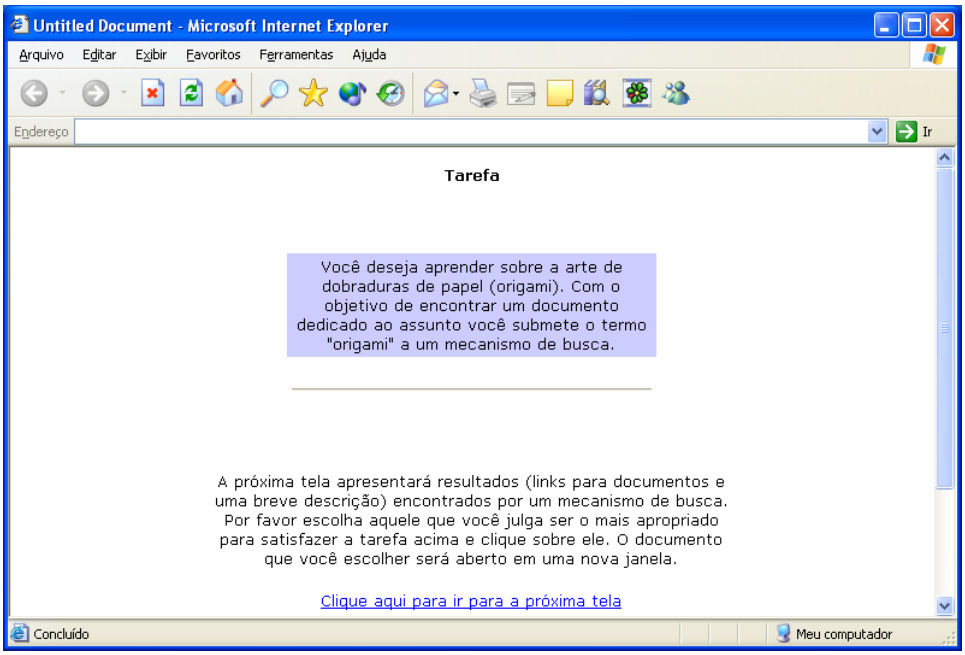


Figura 21. Exemplo de uma tarefa do experimento

5.2.3.2 Geração das descrições utilizadas no experimento

No experimento foram exibidos três resultados por página (um número dentro do limite padrão de 10 exibidos pelos mecanismos de busca) para os juízes. Eles tiveram acesso a *links* de documentos considerados relevantes à consulta pelo Google. Somente foram considerados documentos escritos em língua portuguesa. No total, foram recuperados cinco conjuntos de resultados (cada um correspondendo a uma tarefa), com três URLs simples e uma curta descrição associada.

Depois que cada participante leu a tarefa e a consulta associada, ele visualizou três páginas de resultados distintas, cada uma delas contendo três descrições de documentos. As descrições apresentadas foram geradas para os mesmos documentos-

fonte por cada um dos três sistemas testados, ou seja, para quaisquer documentos a, b e c, a primeira página de resultados trazia descrições geradas pelo Google; a segunda trazia descrições geradas pelo ExtraWeb e a terceira trazia descrições geradas pelo sistema baseline para estes mesmos documentos. A Figura 22, a seguir, mostra uma página de resultados com descrições geradas pelo ExtraWeb.

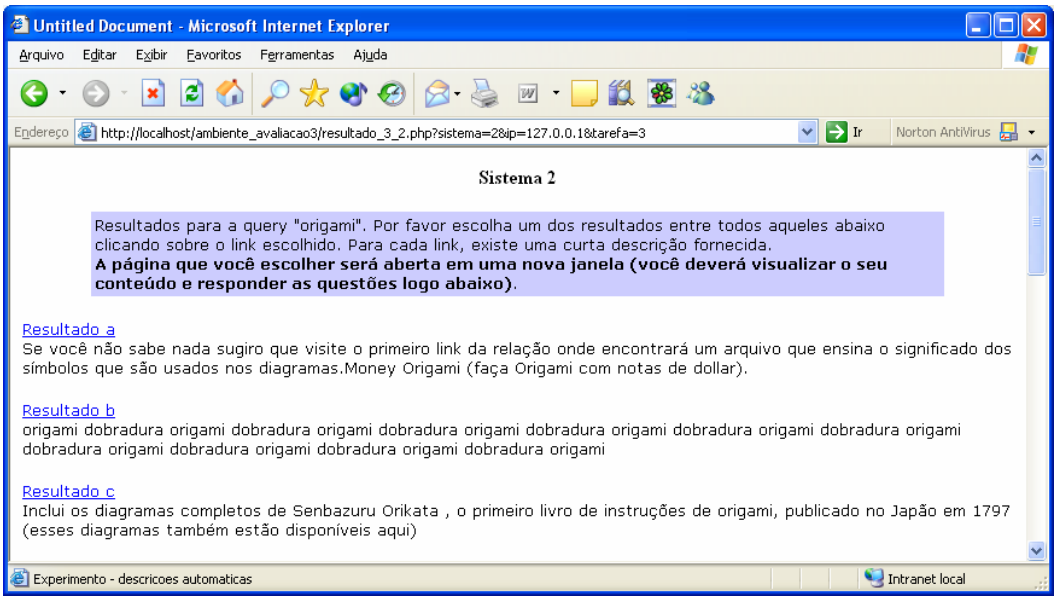


Figura 22. Exemplo de uma página de resultados gerada pelo ExtraWeb

Muitas pessoas tentam compensar a falta de informações providas pelas descrições usando dados encontrados nos títulos e nomes que aparecem na URL (tais como, wikipedia.com, aprendaespanhol.com, downloads.com, etc.). Por este motivo, para evitar que esta informação interferisse na meta do experimento, optamos por não exibir títulos e URLs. Em vez disso, rotulamos os resultados de a, b ou c como mostra a Figura 22.

5.2.3.3 O questionário de avaliação

O questionário utilizado foi formulado considerando dois aspectos: objetividade e brevidade. Tivemos esse cuidado porque as pessoas preferem responder rapidamente às questões e finalizar rapidamente o experimento (LEWIS, 1995; PERLMAN, 2001).

Esse é o motivo principal do número limitado de questões (somente cinco). Para a formulação das questões consideramos a seguinte hipótese:

- Se as descrições forem boas representantes dos documentos-fonte, ou seja, se elas forem realmente úteis para ajudar na escolha de um documento que atenda à necessidade de informação, isso será refletido pela satisfação dos juízes com o documento-recuperado. Por sua vez, o grau de satisfação indicado pelos juizes reflete a utilidade das descrições.

Esta hipótese nos levou a definir as questões seguintes que foram feitas aos juízes:

1) Você está satisfeito com esse resultado?
1- Não, completamente insatisfeito
7 - Sim, completamente satisfeito

2) A descrição escolhida por você correspondeu ao conteúdo da página apontada pelo link?
1- Não
7 - Sim, completamente

3) Neste experimento eu li todas as descrições de site fornecidas com os resultados para tomar a decisão.
1- Discordo totalmente
7 - Concordo totalmente

4) Com que freqüência você costuma ler as descrições fornecidas pelos mecanismos de busca?
1- Nunca
7 - Sempre

5) Que nota você atribui à utilidade das descrições fornecidas, para ajudar na sua escolha?
1- Péssima
7 - Ótima

As questões foram focadas em diversas hipóteses. As duas primeiras (Figura 23) visaram detectar a satisfação dos juízes com a utilidade das descrições escolhidas e sua correspondência com o respectivo documento *Web*. Isso coincide com o objetivo principal de qualquer mecanismo de busca. A terceira e a quarta questões, ao contrário, foram focadas nos hábitos dos participantes, principalmente sobre sua regularidade

durante o experimento. A Questão 3 (Figura 23) referiu-se particularmente ao experimento, enquanto a Questão 4 (Figura 24) foi mais ampla. As respostas a essas questões nos permitiram detectar se os juízes escolheram as descrições baseando-se na comparação entre elas. Tais questões foram, assim, independentes de conteúdo. Assumindo que a resposta para a Questão 3 fosse a opção 7, a última questão visou a uma avaliação geral das descrições automáticas exibidas na tela para cada sumarizador e tarefa. Com exceção das questões 3 e 4, as outras buscaram medir somente a utilidade das descrições com relação à decisão feita pelos usuários *Web*.

A escala de pontos de 1 a 7 usada nas respostas às questões foi escolhida para manter a mesma configuração do experimento de Amitay (2001) e por ser recomendada para medir, em um sistema interativo de recuperação de informação, a relevância dos documentos para uma consulta (TANG, 1999).

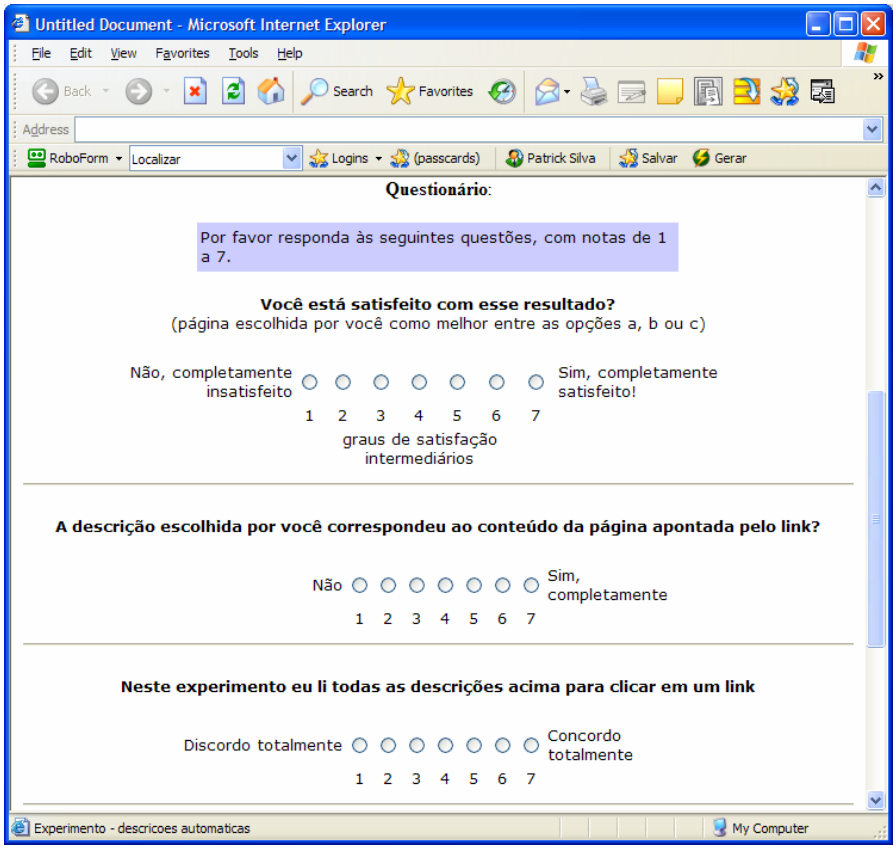


Figura 23. Questionário – perguntas 1 a 3

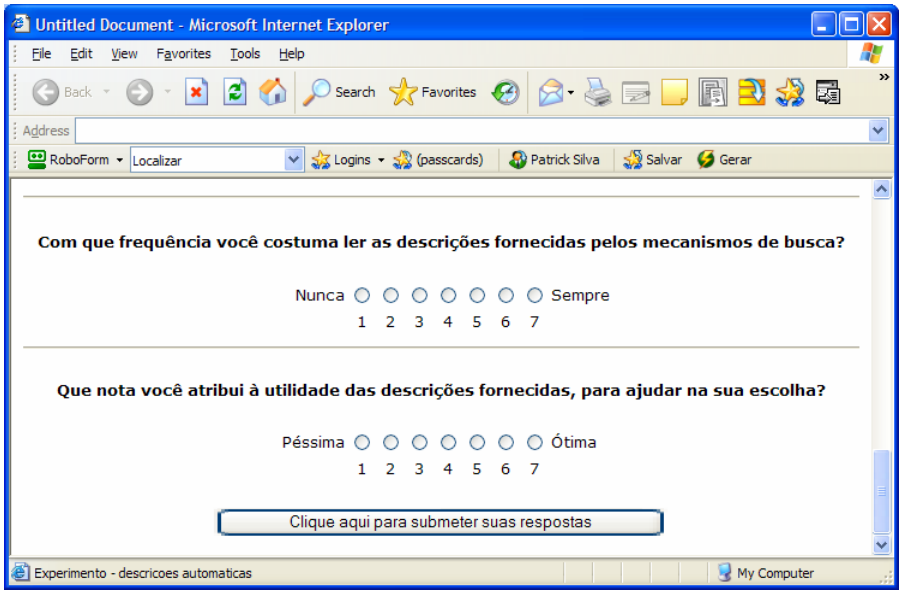


Figura 24. Questionário – perguntas 4 e 5

5.2.3.4 Seleção de respostas válidas

Para compilar as respostas dos juízes, primeiramente recuperamos os arquivos de *log* e verificamos se cada um deles consistentemente respondeu ao questionário. Também verificamos se foram seguidas corretamente as instruções, isto é, se eles responderam a todas as questões em uma interação simples e interagiram com cada sistema somente uma vez. Em casos em que mais que uma interação originou-se de um mesmo número IP, somente a primeira foi considerada. Interações consideradas indesejáveis foram aquelas que não provieram todas as respostas para as cinco questões. Optamos por não desconsiderar os dados dos participantes que não responderam ao questionário para os 3 sistemas (experimento incompleto), já que isto reduziria nossa amostragem. Após a filtragem, restaram 169 respostas válidas para analisar o desempenho do ExtraWeb.

5.2.4 Resultado do julgamento de relevância

A Tabela 17 mostra a distribuição das respostas dadas pelos juízes em relação aos diferentes estilos de descrição e às diferentes tarefas.

Tabela 17. Distribuição das respostas dos participantes

	No. Respostas	Tarefa 1	Tarefa 2	Tarefa 3	Tarefa 4	Tarefa 5
Estilo Google	60	9	19	11	13	8
Estilo ExtraWeb	56	9	16	10	13	8
Estilo Baseline	53	9	16	10	13	5
Total	169	27	51	31	39	21

Como podemos ver, as respostas foram relativamente bem distribuídas entre os três sistemas. Porém, os juízes completaram mais vezes as tarefas 2 e 4.

As respostas dos juízes referentes à Questão 5 (Figura 25) não levam a uma diferença significativa entre o desempenho do Google (média de 5.4) e ExtraWeb (média de 5.2). A diferença entre esses resultados não é estatisticamente significante, o valor p (Teste de Mann-Whitney) é igual a aproximadamente 35%, bem acima do limite usual de 5%. A proximidade entre os dois sistemas também é evidenciada comparando as notas atribuídas pelos juízes para o ExtraWeb e o Google, 63% deles apontaram que as descrições do ExtraWeb eram equivalentes (mesma nota) ou superiores à do Google, para esta mesma questão. O ExtraWeb superou o baseline (média de 4.2), conseguindo um resultado estatisticamente significante – o valor p (Teste de Mann-Whitney) foi igual a aproximadamente 1%.

Considerando que os juízes indicaram que o desempenho do ExtraWeb é similar ao do Google, o fato de propormos um novo método de gerar descrições poderia ser questionável. Entretanto, para uma das tarefas (Tarefa 4), o ExtraWeb superou o Google e o baseline, como visto na Figura 26. Isto indica que a utilidade do sistema pode ser dependente de consulta e, assim, merece investigação mais profunda.

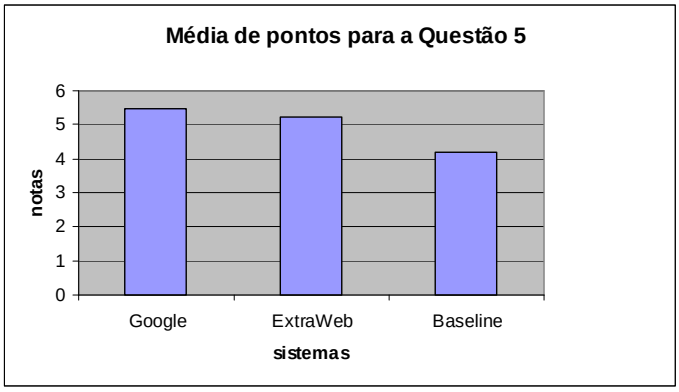


Figura 25. Pontuação média para a Questão 5

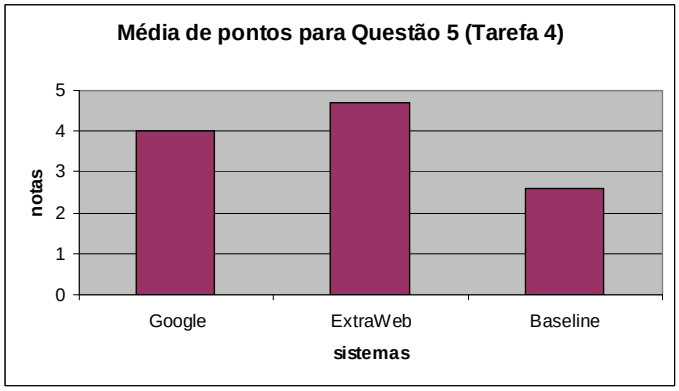


Figura 26. Pontuação média para Questão 5 (Tarefa 4)

A maioria dos juízes leu todos os extratos (veja as médias para a Questão 3 na Tabela 18), apesar do fato de que somente em aproximadamente 83% das interações os juízes costumam lê-los (Questão 4). Isso é importante porque como a maioria leu todos os extratos, as respostas dadas são mais próximas da realidade.

Tabela 18. Pontuação dos sistemas para o conjunto de questões[15]

	Google	ExtraWeb	Baseline
Questão 1	5,3	4,8	5,0
Questão 2	5,6	5,2	4,7
Questão 3	6,7	6,6	6,4
Questão 4	5,8	5,8	5,7
Questão 5	5,4	5,2	4,2

A Tabela 18 também mostra que os julgamentos foram bastante consistentes. As questões 1 e 5 apresentaram as mesmas médias para o Google e o ExtraWeb.

[15] Valores absolutos estão na escala de 7 pontos. Porcentagens foram calculadas normalizando estes valores para a escala de percentual.

Considerando a Questão 1 em particular, houve uma baixa de satisfação dos juízes com os resultados apresentados pelo ExtraWeb (69% somente). Isto pode ser explicado pelo fato de que os usuários não foram capazes de submeter suas próprias consultas, em um cenário real, de qualquer forma, a satisfação poderia aumentar. O próprio modelo do ExtraWeb pode ter sido negativamente afetado pela estrutura e linguagem dos documentos-fonte. Muitos autores de páginas não usam o HTML consistentemente ou amplamente, dificultando a identificação de fragmentos relevantes. A linguagem dos documentos *Web* pode ser pobre (por exemplo, erros ortográficos ou gírias) prejudicando a correspondência com o modelo ontológico do nosso sistema. Claramente isso evidencia que esses são pontos fracos do sistema que necessitam ser revistos para que o ExtraWeb possa ser aprimorado em investigações futuras. Apesar deste desempenho ruim, obtendo praticamente o mesmo resultado do baseline (71%), a diferença com relação à satisfação obtida pelo Google (76%) foi de apenas 7%. Diante desta proximidade e das vantagens já apresentadas sobre o uso de etiquetas HTML e conhecimento ontológico, acreditamos que o sistema tem potencial e que seus resultados podem ser melhorados, justificando a continuidade da investigação.

Médias para a Questão 2 mostram que em 80% dos casos os juízes consideraram que os extratos do Google refletiram o conteúdo dos documentos, para o ExtraWeb esse valor é de 74%, já para o baseline é de 67%. Esta diferença entre Google e ExtraWeb não é estatisticamente significante (p=22%), evidenciando o potencial do sistema na seleção de informações relevantes.

O desempenho ligeiramente superior do Google, quando comparado ao ExtraWeb, pode ser simplesmente pelo fato de que o ExtraWeb trabalhou com documentos previamente coletados pelo próprio Google. Neste caso, o Google atua como um mecanismo de busca e não como um sumarizador. Como conseqüência, o

ExtraWeb pode acumular as fragilidades de ambos os sistemas. Outra razão pode ser o mapeamento inadequado de itens lexicais em conceitos da ontologia do Yahoo. Conforme destacamos no capítulo anterior, essa ontologia tem uma quantidade limitada de conceitos e seu enriquecimento foi feito de modo bastante restrito, limitando o processamento do ExtraWeb. Nenhum desses dois problemas foi profundamente explorado.

Também vale a pena considerar o modo como a informação fornecida pelos extratos influenciou a análise dos usuários. A apresentação dos resultados para os usuários, fora de um contexto de busca real, pode forçá-los a lerem os extratos mais cuidadosamente do que o usual. Ao mesmo tempo em que isto contribui para manter os usuários focados nos extratos, conforme objetivava o experimento, pode evitar que eles façam as suas escolhas de modo mais próximo da realidade.

Embora o experimento não tentasse controlar a homogeneidade da população de juízes ou sua subjetividade ao avaliar os resultados, a análise dos nossos dados mostra que o julgamento geral foi bastante consistente. Entretanto, a mesma avaliação extrínseca pode fornecer resultados diferentes quando uma quantidade maior de ambos, juízes e documentos recuperados, é levada em conta. Como conseqüência, seu julgamento pode ser mais acurado. Isso provavelmente seria evidenciado com o aumento de juízes e documentos do experimento. Apesar dessas limitações, foi possível avaliar o potencial do ExtraWeb.

Uma das principais conferências de avaliação de sumarizadores, a DUC, em seus resultados mais recentes (DUC 2005), aponta uma utilidade média, para a utilização de sumários em tarefas de recuperação de documentos, de 48% para os sumários automáticos e de 93% para sumários manuais (HACHEY et al., 2005). Confrontando esses resultados com os nossos, a média de utilidade alcançada por nosso sistema foi de

72% (considerando os resultados conjuntamente para questões 1, 2 e 5), o que dá indícios de que o desempenho do nosso sistema de sumarização tenha potencial, apesar de seu resultado mediano. Certamente esta é uma comparação muito simplista, devido às diferenças profundas entre as nossas tarefas e aquelas realizadas pela DUC. Entretanto, todo o experimento é bastante similar ao da DUC, por exemplo, é orientado ao usuário, verifica a utilidade de sumários através de simulações de necessidade de informação expressa por um tópico (consulta), utiliza juízes humanos que atribuem notas aos sumários. Para tornar esses nossos resultados estatisticamente significantes, devemos investir em robustez (por exemplo, aumentando o número de juízes e de respostas dos sistemas de busca).

Apesar de ter obtido apenas resultados intermediários, ficando em diferentes situações posicionado abaixo do baseline como, por exemplo, na questão da satisfação dos juízes, o fato de ter um desempenho próximo ao Google em termos de utilidade, demonstra que o ExtraWeb pode ser tão útil quanto o Google para os usuários tomarem decisões de recuperação de documentos. Isso evidência que o ExtraWeb, apesar das limitações já apontadas, tem potencial, justificando a proposta deste sistema. Com o próprio advento da *Web* Semântica, que visa dar significado semântico ao conteúdo das páginas *Web*, a tendência é que o processamento semântico de documentos se torne cada vez mais necessário, fazendo com que recursos como, por exemplo, as ontologias sejam cada vez mais importantes (FREITAS, 2003). Esse fato gera boas perspectivas para melhorias e futuras investigações do ExtraWeb. A seguir apresentaremos alguns exemplos de extratos de documentos *Web* gerados pelo ExtraWeb.

5.2.5 Geração de extratos de documentos Web utilizando o ExtraWeb: ilustração e análise

Nesta seção apresentaremos exemplos de extratos de documentos *Web* gerados pelo ExtraWeb e uma análise desses extratos. Cabe destacar que os documentos-fonte utilizados (identificados por DF1 e DF2) são os mesmos das Figuras 13 e 14 (da Seção 4.1.2.3) utilizados na geração de extratos pelo HTMLSUMM e pelo GEO (Seção 4.2.3.4) e por esse motivo eles não foram reproduzidos nessa seção. Os extratos automáticos (identificados por E1E e E2E) gerados pelo ExtraWeb, que correspondem aos documentos DF1 e DF2 são mostrados na seqüência.

ExtraWeb - Extrato Automático (90% de compressão): E1E

SE1 O pai da ficção científica escreveu livros que até hoje encantam leitores do mundo inteiro!

SE2 Júlio Verne escreveu essa história em 1873, quando não havia tecnologia para construir um submarino.

SE3 Foi o que Júlio Verne fez: em seus livros, criou inventos que, na época, eram impossíveis de produzir!

SE4 INÍCIO | O INSTITUTO | CH ON-LINE | REVISTA CH | CH DAS CRIANÇAS | APOIO À EDUCAÇÃO | CONTATO

Analisando o extrato E1E referente ao documento-fonte DF1 (Figura 13), cujas sentenças estão identificadas por SE1 a SE4, observamos que sentença SE3 introduz uma referência não resolvida no extrato, já que cita “Foi o que Júlio Verne fez”) que em DF1 refere-se a “dar-se o luxo de inventar coisas mirabolantes”, informação que foi omitida em E1E. A sentença SE1 poderia introduzir esse mesmo problema, se não houvesse a referência catafórica a Júlio Verne em SE2 e SE3.

Do mesmo modo que SE3, a sentença SE2 prejudica a coesão do texto, pois ela, ao citar “essa história”, exige o referente “Vinte mil léguas submarinas” que não pode ser encontrado no extrato.

Claramente o extrato apresenta problemas com relação à sua textualidade. Todas as sentenças incluídas em E1E são desconexas, e duas delas apresentam problemas de quebra de referência, prejudicando sua coerência e coesão. Se considerarmos que o tópico principal de DF1 é escritor Júlio Verne e as suas obras de ficção científica, SE1, SE2 e SE3 trazem em seu conteúdo alguma relação com este tópico, nesse caso, E1E consegue, de alguma forma, focar a temática principal. O que se observa pela Tabela 19, que traz os pesos atribuídos pelos ExtraWeb a cada sentença (normalizados em uma escala de 0 a 2), é que para as três primeiras sentenças, tanto as palavras-chave quanto os conceitos considerados têm relação com o tópico principal.

Tabela 19. Síntese das informações consideradas pelo ExtraWeb para DF1

E1E	Etiquetas HTML consideradas	Palavras da sentença relacionadas à etiqueta	Conceitos considerados	Palavras da sentença mapeadas no conceito	Peso de cada sentença [0-2]
SE1	<TITLE>	hoje	Compras e Serviços>>**Livros**	livros	1.5
	<STRONG>	pai, ficção, científica, escreveu, livros, hoje, encantam, leitores, mundo, inteiro	Artes e Cultura>>Literatura>>Gêneros literários>>**Ficção Científica**	ficção científica	
			Sociedade>>Família>>**Pais e Filhos**	pai	
SE2	<STRONG>	escreveu	Ciência>>Ciências Humanas>>**História**	história	0.99
	<A>	história	Artes e Cultura>>Literatura>>Autores>>**Júlio Verne (1828-1905)**	júlio verne	
SE3	<STRONG>	livros	Compras e Serviços>>**Livros**	livros	1.06
			Artes e Cultura>>Literatura>>Autores>>**Júlio Verne (1828-1905)**	júlio verne	
SE4	<A>	início, instituto, ch, revista, apoio, educação	**Educação**	educação	1.58
	<TITLE>	crianças	Notícias e Mídia>>**Revistas**	revista	
			Sociedade>>Grupos e Culturas>>**Crianças**	crianças	
			Ciência>>Ciências Humanas>>**Educação e Formação**	educação	
			Educação>>**Institutos, Faculdades**	instituto	
			Ciência>>**Institutos**	instituto	

A inclusão da sentença SE4 em E1E, apesar de seu alto peso (79%) na escala normalizada de 0 a 2, claramente introduz informações irrelevantes já que seu conteúdo não tem nenhuma relação com a temática principal do documento. Observando os dados

da Tabela 20, a inclusão de SE4 pelo ExtraWeb deveu-se principalmente ao HTMLSUMM já que ela foi considerada a sentença mais relevante por esse sistema, obtendo o peso máximo (100%). A alta pontuação atribuída pelo HTMLSUMM é justificável já que, apesar de irrelevante para o extrato do ponto de vista da manutenção da informatividade, a sentença SH4 contém várias palavras-chave provenientes das etiquetas de título (<TITLE>) e *links* (<A>), conforme mostra a Tabela 19. Não se pode deixar de observar que esta mesma sentença também obteve um peso acima da média atribuído pelo GEO (58%), conforme indicado na Tabela 20. O problema, nesse caso, é que todos os conceitos associados a SE4 são marginais no documento e em nada contribuem para seleção do conteúdo mais informativo. No entanto, o GEO não a incluiu no extrato correspondente (E1G, mostrado na Seção 4.2.3.4). Isso indica que, de fato, a inclusão de SE4 pelo ExtraWeb se deve exclusivamente à influência do modelo de pesos do HTMLSUMM.

Tabela 20. Sentenças incluídas no ExtraWeb para DF1

Sentença do ExtraWeb	Correspondência com as sentenças dos outros sistemas	Peso no HTMLSUMM	Peso no GEO	Peso no EXTRAWEB
SE1	SH1, SG1	0.67	0.83	1.5
SE2	-	0.27	0.72	0.99
SE3	SG3	0.06	1	1.06
SE4	SH4	1	0.58	1.58

Embora o extrato E1E apresente problemas de textualidade devido a presença de sentenças desconexas, ele inclui sentenças bastante informativas em relação ao agente e evento realizado.

ExtraWeb - Extrato Automático (90% de compressão): E2E

SE1 Iraque condena cinco à morte e 35 à prisão por terrorismo.

SE2 Um tribunal iraquiano condenou nesta quarta-feira cinco pessoas à morte e outras 35 à prisão, entre elas oito estrangeiros, por atos terroristas, assassinatos, seqüestros e por entrar ilegalmente no país, informaram fontes judiciais locais.

SE3 As fontes judiciais informaram que, entre os 32 iraquianos, estão cinco que foram condenados à forca, 16 à prisão perpétua e 11 a penas de detenção que oscilam entre um e seis anos.

SE4 Na terça-feira o mesmo tribunal condenou 25 pessoas, entre elas cinco iraquianos, quatro sauditas, um sírio e um bengali a penas de morte e prisão, também por terrorismo, no caso dos iraquianos, e por entrada ilegal no Iraque, no caso dos estrangeiros.

Analisando o extrato E2E referente ao documento-fonte DF2 (Figura 14) observamos que ele não apresenta problemas com relação à textualidade. De fato, o texto apresentado está bastante coeso e coerente. A sentença SE3 que faz uma referência a “o mesmo tribunal” não tem sua interpretação prejudicada já que ela encontra seu referente em SE2.

Considerando que o tema principal do documento é “a condenação imposta pelo tribunal iraquiano”, todas as sentenças de E2E são bem informativas neste aspecto, já que permitem que o leitor tenha uma informação bastante clara sobre o tópico principal do documento.

Observando os dados da Tabela 21, vemos que tanto as palavras-chave quanto os conceitos associados às sentenças correspondem, em sua maioria, ao tópico principal do texto, o que as torna bastante relevantes.

Tabela 21. Síntese das informações consideradas pelo ExtraWeb para DF2

E2E	Etiquetas consideradas	Palavras da sentença relacionadas à etiqueta	Conceitos considerados	Palavras da sentença mapeadas no conceito	Peso de cada sentença [0-2]
SE1	<TITLE>	iraque, condena, cinco, morte, prisão, terrorismo	Regional>>Países>>**Iraque**	Iraque	1.37
			Sociedade>>Crime>>Tipos de Crime>>**Terrorismo**	terrorismo	
			Sociedade>>**Morte**	morte	
SE2	<TITLE>	cinco, morte, prisão	Regional>>Países>>**Iraque**	iraquiano	1.31
			Sociedade>>**Morte**	morte	
			Sociedade>>Crime>>Tipos de Crime>>**Terrorismo**	terroristas	
			Sociedade>>Crime>>**Tipos de Crime**	seqüestros	
			Regional>>**Países**	país	
SE3	<TITLE>	prisão	Governo>>Brasil>>**Poder Judiciário**	judiciais	1.2
			Regional>>Países>>**Iraque**	iraquianos	
SE4	<TITLE>	morte, cinco, prisão, terrorismo, iraque	Regional>>Países>>**Iraque**	iraquianos, Iraque	1.73
			Regional>>Países>>**Arábia Saudita**	sauditas	
			Sociedade>>**Morte**	morte	
			Sociedade>>Crime>>Tipos de Crime>>**Terrorismo**	terrorismo	
			Regional>>Países>>**Síria**	sírio	
			Entretenimento>>**Ingressos**	entrada	

A escolha de SE1 pelo ExtraWeb se deve ao peso derivado do título (“Iraque condena cinco à morte e 35 à prisão por terrorismo”), já que ela compartilha várias dessas palavras-chave, conforme vemos na Tabela 21. A inclusão do título, quando informativo como em SE1, ajuda na contextualização e no entendimento do conteúdo do documento. No entanto, sua omissão não parece prejudicial ao tópico principal, pois a segunda sentença (SE2) pode ser considerada sua paráfrase e, portanto, cobre a primeira sentença. Notadamente, conforme indica a Tabela 22, essa sentença foi incluída no ExtraWeb devido ao peso atribuído pelo HTMLSUMM (92%) já que seu peso no GEO é relativamente baixo (45%).

Tabela 22. Sentenças incluídas no ExtraWeb para DF2

Sentença do ExtraWeb	Correspondência com as sentenças dos outros sistemas	Peso no HTMLSUMM	Peso no GEO	Peso no EXTRAWEB
SE1	SH1	0.92	0.45	1.37
SE2	-	0.46	0.85	1.31
SE3	-	0.3	0.9	1.2
SE4	SH2, SG2	0.76	0.97	1.73

5.2.6 Geração de extratos de documentos Web: comparação de desempenhos do ExtraWeb, HTMLSUMM e GEO

Após analisarmos os extratos gerados pelo HTMLSUMM, GEO e ExtraWeb para os mesmos documentos-fonte (relatados nas seções 4.1.2.3, 4.2.3.4 e 5.2.5), o que se nota é que, de modo geral, todos eles apresentam, em maior ou menor grau, problemas de textualidade. Esses problemas estão relacionados à referências não resolvidas (antecedentes irrecuperáveis pelos extratos) e à justaposição de sentenças desconexas que prejudicam de algum modo a compreensão do extrato.

Os extratos apresentam também alguns problemas de informatividade, incluindo, algumas vezes, sentenças marginais que não remetem à temática principal do documento-fonte. Nesse aspecto, considerando-se os extratos independentemente dos documentos-fonte, os mais informativos e também os que apresentam melhor textualidade são aqueles gerados pelo GEO e pelo ExtraWeb, muito embora suas associações inter-sentenciais possam se apresentar ligeiramente modificadas, em relação ao documento-fonte.

Particularmente o GEO parece filtrar melhor as sentenças relevantes do documento já que o HTMLSUMM inclui, proporcionalmente, mais informações marginais. Esse desempenho ruim da estratégia que faz uso de etiquetas HTML acaba influenciando negativamente o desempenho do ExtraWeb já que ele faz a seleção das sentenças considerando os dois modelos.

A análise dos extratos gerados pelos sistemas evidencia que o modelo do ExtraWeb, que determina o peso de uma sentença por meio de uma soma simples dos pesos provenientes do HTMLSUMM e do GEO, em algumas situações é inadequado. Isso pode constituir um problema para o ExtraWeb já que se uma sentença recebe uma pontuação muito alta em uma estratégia e muito baixa em outra, ela muito possivelmente será incluída no extrato. Isso pode acarretar na inclusão de sentenças que

tenham sido mal selecionadas pelo HTMLSUMM ou pelo GEO, nos extratos do ExtraWeb. Uma alternativa para esse problema seria tirar a média dos pesos das sentenças em vez de simplesmente somar, dessa forma, seriam privilegiadas para inclusão no extrato aquelas que tivessem um peso médio maior determinado por ambas as estratégias.

O que observamos é que, de modo geral, aquelas sentenças com alto peso atribuído pelo HTMLSUMM e com um peso mais baixo atribuído pelo GEO não são tão informativas e acabam introduzindo os problemas citados nos extratos do ExtraWeb. Ao contrário, sentenças com alto peso no GEO e com baixo peso no HTML são relativamente informativas. Isso evidencia que ontologia contribui para selecionar o conteúdo mais relevante de forma superior ao método que só utiliza etiquetas HTML. Nesse sentido, o processamento ontológico parece filtrar melhor as sentenças que, de fato, são relevantes e contribuem para a informatividade dos extratos.

6 Considerações finais

Esta dissertação apresenta três abordagens para gerar extratos automáticos de documentos, as quais incluem a identificação de características relacionadas à estruturação do documento (etiquetas HTML), ao conhecimento ontológico e à combinação de ambos. De modo geral, as abordagens exploram as informações relevantes considerando características de naturezas distintas: no caso do uso de etiquetas HTML, as informações relevantes são expressas por segmentos textuais delimitados na estrutura do documento por um conjunto de etiquetas; no caso do uso de ontologia, elas são determinadas pelo mapeamento entre as palavras do texto e conceitos ontológicos, analisando-se a conectividade semântica entre os elementos textuais.

Os sumários extrativos gerados de acordo com as três abordagens foram avaliados objetivamente, considerando-se várias medidas estatísticas de desempenho, dependendo da abordagem, e subjetivamente, por usuários humanos. Para a avaliação objetiva da sumarização baseada em etiquetas HTML, a informatividade semântica foi usada na comparação do sistema HTMLSUMM com ferramentas comerciais afins, isto é, ferramentas que também têm como foco a determinação de informações relevantes, como o *Copernic Summarizer*, o Google e um sistema baseline que considera somente as descrições manuais de autores de páginas *Web*, identificadas por uma meta-etiqueta HTML particular. Variações das estratégias de reconhecimento das informações textuais relevantes do HTMLSUMM foram também exploradas. Este sistema apresentou, dentre os demais, o melhor desempenho médio em relação à informatividade semântica, superando mesmo o desempenho do Copernic e do Google. A conclusão a que chegamos é que o uso de informações estruturais pode ser um fator complementar importante para a geração de extratos de documentos *Web*, quando devidamente

etiquetados em HTML, pelo seu potencial de produzir conteúdo mais denso e proeminente do que os outros sistemas.

Para a avaliação objetiva baseada no uso de ontologia, foram usadas as medidas de precisão e cobertura, calculadas com base em um corpus de referência. Os índices obtidos com o sistema correspondente, o GEO, foram comparados com os de sete outros sumarizadores extrativos para o português. Também foram consideradas duas versões do GEO, sendo as variações correspondentes à forma como a ontologia do Yahoo foi enriquecida: com o vocabulário do Diadorim e, portanto, simples relações de sinonímia (OntoV1), ou com o vocabulário da Wikipédia adicionado à ontologia anterior (OntoV2). Embora a introdução de relações hierárquicas em OntoV2 não tenha apresentado melhora significativa em relação ao enriquecimento da ontologia por sinonímia, ambas as versões apresentaram melhor desempenho do que os outros sete sistemas. Isto demonstra que o conhecimento ontológico pode influir positivamente na determinação de informações relevantes para a geração de extratos, mesmo que estes sejam significantemente comprimidos.

A avaliação subjetiva foi realizada sobre o ExtraWeb, sistema principal proposto neste mestrado, que agrega o potencial indicado pelas avaliações das duas abordagens anteriores: o uso concomitante de etiquetas HTML e ontologia. Foi elaborada uma tarefa extrínseca de julgamento de relevância mediante o acesso direto de internautas a *websites* e o preenchimento de um questionário após a exibição dos resultados, para avaliação da utilidade dos extratos gerados pelo ExtraWeb. A utilidade, nesse caso, refere-se ao uso de um extrato como meio para se decidir se um documento é, de fato, relevante. Embora o ExtraWeb tenha alcançado um grau de utilidade mediano, a proximidade de seus índices com os do Google indica seu potencial.

Os problemas estudados nesta dissertação indicam alguns desafios para a Sumarização Automática de documentos *Web*, a saber:

- a análise semântica de conteúdo de documentos *Web* para capturar relacionamentos entre conceitos ou objetos do domínio com o intuito de que sejam produzidos sumários de melhor qualidade: os experimentos descritos nesta dissertação indicam que as soluções propostas são interessantes para contornar este desafio;

- a própria sumarização de documentos em ambiente *Web* envolvendo documentos em outras línguas que não o português: considerando-se a sumarização baseada em documentos *Web*, ou seja, o HTMLSUMM, por exemplo, seria possível utilizá-lo diretamente, já que ele é dependente da estruturação em HTML e não da língua natural, propriamente dita, bastando simplesmente a substituição do repositório de *stopwords* de acordo com a língua-alvo do documento. A própria abordagem do GEO (somente com conhecimento ontológico) poderia ser replicada, considerando-se a existência da ontologia Yahoo em outras línguas.

Conforme já destacamos e detalharemos na Seção 6.2, as abordagens propostas apresentam algumas limitações que levam a um desempenho mediano do sistema ExtraWeb. Porém, um dos pontos favoráveis, que justifica a continuidade desta investigação e que valoriza a nossa proposta, reside no fato de termos, como um dos focos centrais, o processamento semântico do conteúdo dos documentos por meio de uma ontologia. Apesar de ainda não ser uma realidade, o projeto da *Web* Semântica prevê uma mudança nos rumos do desenvolvimento dos sistemas atuais nessa mesma linha: ela provê um conjunto de iniciativas, tecnológicas em sua maioria, destinadas a criar uma futura *World Wide Web* na qual os sistemas possam processar a informação no nível semântico, isto é, representá-la, encontrá-la e gerenciá-la de modo "inteligente". A perspectiva é que os sistemas futuramente desenvolvidos estejam aptos

a executar tarefas que requeiram a interpretação da informação, analisando-a e processando-a semanticamente (FENSEL et al., 2002).

Diante desses fatos, a nossa proposta, ainda que limitada, de integrar conhecimento semântico ao processamento de documentos, resultando no desenvolvimento de um modelo de acoplamento de uma ontologia a um sistema de sumarização de documentos em HTML e, assim, na construção do ExtraWeb, vai ao encontro dessas metas.

6.1 Contribuições

Se considerarmos o foco particular do processamento da língua portuguesa, são muitas as contribuições deste trabalho, não só para a Sumarização Automática, como também para as próprias tarefas de acesso a *websites*, sobretudo porque há muito poucos recursos para o processamento do português e não existem atualmente sistemas de sumarização automática específicos para documentos *Web* nessa língua.

Os resultados alcançados pelo método que faz uso de ontologia (implementado no GEO e no ExtraWeb), que superou os de sete outros sumarizadores extrativos para o português, constituem uma contribuição para o avanço da área. Entretanto uma das principais contribuições deste trabalho de mestrado é a proposta da metodologia mista implementada no ExtraWeb. Ela permite demonstrar ser viável gerarem-se sumários de documentos *Web*, analisando seu conteúdo por meio de uma ontologia e de sua formatação HTML, para identificação de informações relevantes, isso é evidenciado pela potencialidade do ExtraWeb na geração de extratos. Essa potencialidade é aparente nas avaliações apresentadas, muito embora os resultados indiquem um largo espaço para melhora, a qual é possível de obtenção, já que as possíveis causas da performance mediana são identificáveis e solucionáveis.

As principais contribuições deste trabalho para a pesquisa em Sumarização Automática, vinculada à Recuperação da Informação, são as seguintes:

- Existência de uma ontologia enriquecida para o português brasileiro: embora ainda bastante restrito esse recurso é uma das mais importantes contribuições deste projeto, devido à inexistência de qualquer recurso similar para o português, podendo ser usado inclusive com outros fins que não a Sumarização Automática;
- A metodologia de avaliação e a própria plataforma *Web* criada como parte do experimento com o ExtraWeb podem ser utilizadas em outros contextos de avaliação de descrições de sistemas de busca.
- Disponibilização de três protótipos de sumarizadores automáticos: o GEO (**G**erador de **E**xtratos baseados em conhecimento **O**ntológico), o HTMLSUMM (**HTML Summ**arizer) e o ExtraWeb (**Extra**tos de documentos **Web**)[16];
- Disponibilização de corpora gerados com os protótipos dos sistemas.

Os protótipos e a base ontológica produzidos por este trabalho de mestrado encontram-se disponíveis para uso pela comunidade de pesquisa e poderão ser aprimorados, para a continuidade das pesquisas relacionadas a este projeto, especificamente para a sumarização automática de *websites* em português.

6.2 Limitações

Os resultados medianos dos sistemas de SA baseados em conhecimento ontológico e em etiquetas HTML propostos neste trabalho evidenciam algumas fragilidades e limitações das metodologias propostas nesta investigação. Por exemplo, o

[16] No apêndice desta dissertação apresentamos suas respectivas interfaces.

modelo baseado em HTML exige que o documento a ser sumarizado utilize exatamente as mesmas etiquetas definidas pelo sistema para coletar as informações relevantes. Isso é um problema para a sumarização do modo como ela é realizada no HTMLSUMM, à medida que os documentos podem não conter as mesmas etiquetas. O próprio conjunto de etiquetas utilizados na proposta, apesar de englobar um número maior que o de outros trabalhos correlatos, é ainda bastante reduzido. O modelo de pesagem de etiquetas HTML também pode ser uma limitação do sistema, já que os pesos são definidos estatisticamente com base em um corpus de tamanho limitado.

Uma limitação do GEO, como aplicação baseada em ontologia para processar documentos da *Web*, refere-se à definição da própria ontologia. Uma vez que a *Web* é dinâmica, a linguagem utilizada neste ambiente tende a ser muito variada. Isto dificulta o processo de mapeamento entre conteúdos textuais e conceitos ontológicos. O processo de enriquecimento da ontologia do Yahoo também foi bastante limitado, seja do ponto de vista do número de conceitos descritos quanto do ponto de vista das fontes utilizadas para o enriquecimento, quais sejam, o Diadorim e a Wikipédia. Os documentos da Wikipédia, particularmente, não foram selecionados por algum critério de qualidade individual, mas somente pela reconhecida qualidade geral da enciclopédia. No caso da *Web*, tal problema é ainda mais preocupante já que os documentos mudam rapidamente (CHEN, 1993) e, assim, mesmo que as fontes geradoras dos textos sejam confiáveis, a informação pode não ser. Isto pode, inclusive, ter contribuído para os índices medianos obtidos. As imperfeições, neste caso, remetem aos seguintes tipos prováveis de informação, que poderiam ser melhorados (PARSONS, 1996):

a) informação incompleta e, portanto, necessitando maior detalhamento;

b) informação incerta, sem possibilidade de comprovação, devido ao dinamismo da *Web;*

c) informação vaga, devido a imprecisões do vocabulário;

d) informação inconsistente, devido à existência de valores contraditórios.

Ainda com relação à base ontológica, seu enriquecimento foi influenciado diretamente pela intervenção humana, pois não sofreu nenhum tipo de refinamento e foi feito de modo completamente subjetivo e *ad hoc* por um único engenheiro de conhecimento. Os próprios itens lexicais utilizados para descrever os conceitos ontológicos não foram especificados exaustivamente (devido ao elevado custo manual) em todas as suas formas variantes, minimamente, em relação a flexões de gênero, número ou grau. Nesses casos, variações lexicais não são contempladas, resultando em um mapeamento ontológico frágil. Uma alternativa para isso seria prover o módulo de mapeamento com mecanismos de transformações lexicais, por exemplo, usando métodos de clusterização.

Tampouco foi considerada, na construção da ontologia, a questão da ambigüidade, sobretudo devida à polissemia. O mapeamento adotado neste caso, isto é, o de considerarem-se todos os conceitos relacionados a uma mesma palavra, claramente introduz distorções no cômputo dos pesos dos componentes textuais, que podem ser tão mais severas quanto o uso livre da língua, que permite uma grande variedade de uso polissêmico. Como conseqüência, pode haver também distorções na produção dos extratos, que poderiam ser minimizadas com um modelo mais sofisticado de mapeamento ontológico. Assim, as informações contidas na ontologia não podem ser consideradas como um conjunto completo, apesar de amplo. Os resultados precisam ser interpretados dentro destes limites.

As diferentes formas de avaliação utilizadas para medir o desempenho das estratégias de sumarização também constituem uma limitação já que, por terem natureza completamente distinta, torna mais difícil avaliarmos em que medida ou em que

situações, uma ou outra estratégia traz ganhos para o processo de sumarização. Com relação à avaliação do ExtraWeb, toda tarefa subjetiva, manual, envolve também limitações consideráveis, que comprometem a confiabilidade e escalabilidade de um sistema automático. Neste trabalho, o número de juízes e de documentos providos para a avaliação do ExtraWeb pode ser considerado pouco expressivo, consistindo um fator limitante para uma análise mais ampla dos resultados obtidos.

Por fim, um outro ponto a aprimorar é a própria eficiência do sistema: por termos privilegiado a modelagem do conhecimento profundo, não foi foco deste trabalho considerar as questões de indexação e busca típicas de um sistema de Recuperação de Informação. No entanto, ao pretender usar extratos como descrições de documentos *Web*, essa desvinculação torna-se imprópria.

6.3 Trabalhos Futuros

Com relação às estratégias de sumarização e recursos implementados pelo ExtraWeb, estes podem ser estendidos e aprimorados, dadas suas limitações e possibilidades não exploradas. Destacamos as seguintes linhas de investigação, para a continuidade deste trabalho:

a) Refinar a ontologia para facilitar a definição e descrição de conceitos, e, conseqüentemente, permitir melhorias no desempenho do sistema.

b) Automatizar o processo de definição de conceitos, utilizando técnicas capazes de sugerir palavras que identifiquem os mesmos, por exemplo:

   i. Análise de freqüência em coleções de documentos relacionados a certo tópico ou conceito ontológico. Os próprios documentos categorizados no Yahoo podem servir para formar uma coleção com essa finalidade.

ii. *Stemming.* O acoplamento de um *Stemming* poderia ajudar a reduzir o problema das variações lexicais não contempladas na ontologia, reduzindo a fragilidade do mapeamento ontológico. A sua inclusão pode ser dar de forma imediata uma vez que esse recurso pode ser encontrado no NILC (Núcleo Interinstitucional de Lingüística Computacional).

iii. Utilização de outras fontes externas de conhecimento *Web* além da Wikipédia. Uma sugestão seria a incorporação de informações de uma base de conhecimento como a WordNet.br (DIAS-DA-SILVA et al., 2002) que se encontra em desenvolvimento.

c) Levar em conta graus de relevância para as palavras que descrevem os conceitos. Uma possibilidade é seguir o trabalho de Wives (2004), que sugere que a definição da relevância de um termo descritor de um conceito pode seguir uma estratégia que faz distinção entre termos indicadores e termos caracterizadores. Os primeiros apontam para um conceito específico, enquanto que os últimos restringem conceitos em um conjunto. Assim, termos indicadores devem receber um peso maior, pois possuem maior força para indicar a presença do conceito (nomes próprios, por exemplo). Enquanto que os termos caracterizadores devem receber pesos relativos menores, pois, apesar de ajudarem a identificar um conceito, não dão certeza de tal.

d) Realizar o enriquecimento de todos os conceitos originais da ontologia do Yahoo já que neste projeto de mestrado o enriquecimento foi feito para 2.500 conceitos (aproximadamente metade da coleção). Apesar do alto custo de se realizar o enriquecimento manualmente, os resultados apontaram melhorias após

o enriquecimento, dessa forma, certamente será válido dar continuidade a este processo.

e) Aprimorar o processo de mapeamento para lidar com a questão da polissemia. Para isso, é necessário estudar estratégias que ofereçam meios de desambiguar essas as palavras antes de elas serem mapeadas na ontologia. Também pode ser verificada a existência de um desambiguador automático que possa ser incorporado ao sistema. Isso contribuirá para a identificação de tópicos relevantes mais precisamente.

f) Investigar e aprimorar alguns aspectos relacionados à estrutura da ontologia já que eles podem trazer algum tipo de limitação para a sumarização realizada pelo ExtraWeb. Dentre esses aspectos citamos:

i. Balanceamento: diversidade dos domínios;
ii. Abrangência: generalidade e qualidade dos conceitos ontológicos;
iii. Profundidade: qualidade e quantidade dos descritores e qualidade da hierarquia.

g) Investigar outras formas de se determinar a relevância individual das etiquetas HTML bem como investigar a possível extensão do conjunto de etiquetas utilizado a fim de refinar a seleção de informações relevantes.

h) Investigar outras maneiras de associar as estratégias do HTMLSUMM e do GEO no ExtraWeb. Pode-se, por exemplo, combiná-las considerando que uma seja mais ou menos relevante que a outra.

i) Verificar a relação entre as estratégias de sumarização do ExtraWeb e os tipos de documentos. Uma análise de casos de testes pode ajudar a identificar problemas pontuais.

j) Acoplar um categorizador de textos ao ExtraWeb de forma a permitir que o sistema selecione a estratégica mais adequada de acordo com documento.

k) Avaliar o processo de aplicação dos extratos gerados pelo ExtraWeb em um corpus maior de documentos assim como considerar um número maior de juízes, isso certamente permitirá um entendimento mais amplo dos resultados e das limitações do sistema. Para isso, o próprio ambiente de avaliação já disponível pode ser reaproveitado bastando apenas acrescentar, por exemplo, novas descrições.

l) Melhorar a ferramenta de sumarização tendo como foco a eficiência, isto é, seu tempo de resposta. Isso pode ser feito, por exemplo, re-implementando alguns processos do sistema usando outras linguagens de programação. O fato de já existir um protótipo disponível, com os algoritmos já implementados, facilita este processo, permitindo que ele possa ser feito de forma imediata.

m) Avaliar e aprimorar o método de segmentação de documentos *Web* já que este pode apresentar fragilidades que podem interferir no processo de sumarização.

n) Estender o ExtraWeb, acoplando-o a um mecanismo de busca, de forma que ele possa gerar os extratos dos documentos recuperados pelo mecanismo.

o) Aplicar o sistema a uma coleção de documentos *Web* multilingual, isto é, com textos em línguas diferentes. A hipótese é de que a abordagem pode ser usada satisfatoriamente nestas situações.

p) Verificar a utilidade da base ontológica desenvolvida também em aplicações distintas da sumarização automática como, por exemplo: auxiliar na catalogação automática de documentos em hierarquias de diretórios *Web*, organizar documentos em clusters, complementar documentos inserindo meta

dados, associar tópicos a consultas de usuários e assim filtrar documentos que tragam tópicos irrelevantes, informar os tópicos de um documento em um idioma diferente daquele usado no documento fonte, gerar descrições com foco, etc.

É importante ressaltar que a maior parte das sugestões de trabalhos futuros citados nessa seção visam melhorar os resultados do ExtraWeb, o qual, no estágio atual, é um protótipo de sumarizador de documentos *Web*. Definido o objetivo de projetar apenas um protótipo simples, embora completamente funcional, o importante foi verificar a validade da hipótese levantada, ou seja, a aplicabilidade do modelo sugerido nessa investigação para a sumarização de documentos *Web*. Em contrapartida, esse sistema, mesmo que limitado, pode servir de referência para outros trabalhos para o português, contribuindo, com isso, para o andamento das pesquisas em Lingüística Computacional.

Referências bibliográficas

AGIRRE, E. et al. Enriching very large ontologies using the WWW. In: WORKSHOP ON ONTOLOGY LEARNING, 2000, Berlin. **Proceedings**… Aachen: CEUR-WS, 2000. Disponível em: <http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-31/>. Acesso em: 20 set. 2005.

AGIRRE, E. et al. Enriching WordNet concepts with topic signatures. In: SIGLEX WORKSHOP ON WORDNET AND OTHER LEXICAL RESOURCES: APPLICATIONS, EXTENSIONS AND CUSTOMIZATIONS, 2001, Pittsburg. **Proceedings**…Cambridge: MIT Press, 2001. p. 1-7.

AMITAY, E. **What lays in the layout:** using anchor-paragraph arrangements to extract descriptions of Web documents. 2001. 147 f. Tese (Doutorado) - Division of Information and Communication Sciences, Macquarie University, Sydney, 2001.

BERGER, A. L.; MITTAL, V. O. OCELOT: a system for summarizing Web pages. In: ANNUAL INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 23., 2000, New York. **Proceedings**… New York: ACM Press, 2000. p. 144-151.

BLACK, W.; JOHNSON, F. A practical evaluation of two rule-based automatic abstraction techniques. **Expert Systems for Information Management**, v. 1, n. 3, p. 159-177, 1988.

BRIN, S.; PAGE, L. The anatomy of a large-scale hypertextual Web search engine. In: INTERNATIONAL CONFERENCE ON WORLD WIDE WEB, 7., 1998, Amsterdam. **Proceedings**… Amsterdam: Elsevier Science, 1998. p. 107-117.

CHANDRASEKARAN, B.; JOSEPHSON, J. R.; BENJAMINS, V. R. What are ontologies, and why do we need them? **IEEE Intelligent Systems**, v. 14, n. 1, p. 20-26, 1999.

CHEN, H. Collaborative systems: solving the vocabulary problem. **IEEE Computer**, v. 27, n. 5, p. 58-66, 1994. Special Issue on Computer Supported Cooperative Work.

CHEN, H. et al. A concept space approach to addressing the vocabulary problem in scientific information retrieval: an experiment on the worm community system. **American Society for Information Science**, v. 48, n. 1, p. 17-31, 1997.

CHEN, Z. Let documents talk to each other: a computer model for connection of short documents. **Journal of Documentation**, v. 49, n. 1, p. 44-54, 1993.

CHRISTOPHI, C. **Mining keywords from large topic taxonomies**. 2004. Dissertação (Mestrado) - Department of Computer Science, University of Cyprus, Nicosia, 2004.

CINTRA, L.; CUNHA, C. **Nova gramática do português contemporâneo**. Rio de Janeiro: Nova Fronteira, 2001. 726 p.

CLEVERDON, C.; MILLS, J.; KEEN, M. **Factors determining the performance of indexing systems**. Cranfield: ASLIB, 1966. 376 p.

CONKLIN, J. Hypertext: an introduction and survey. **IEEE Computer,** v. 20, n. 9, p. 17-41, 1987.

CRAVEN, T.C. HTML tags as extraction cues for Web page description construction. **Informing Science Journal**, v. 6, p. 1-12, 2003.

CUTLER, M.; SHIH, Y.; MENG, W. Using the structure of HTML documents to improve retrieval. In: USENIX SYMPOSIUM ON INTERNET TECHNOLOGIES AND SYSTEMS, 1997, Monterey. **Proceedings**… Lake Forest: USENIX, 1997. p. 241-251.

DALIANIS, H. **SweSum**: a text summarizer for swedish. Stockholm: Royal Institute of Technology (KTH), 2000. Relatório TRITA-NA-P0015.

DIAS-DA-SILVA, B. C.; OLIVEIRA, M. F.; MORAES, H. R. Groundwork for the development of the Brazilian Portuguese Wordnet. In: INTERNATIONAL CONFERENCE ON ADVANCES IN NATURAL LANGUAGE PROCESSING, 3., 2002, Faro. **Proceedings**… Londres: Springer-Verlag, 2002. p. 189-196.

DORR, B. et al. A methodology for extrinsic evaluation of text summarization: does ROUGE correlate? In: THE ACM WORKSHOP ON INTRINSIC AND EXTRINSIC EVALUATION MEASURES FOR MT AND/OR SUMMARIZATION, 2005, Michigan. **Proceedings**… New York: ACM Press, 2005. p. 1-8.

EDMUNDSON, H.P. New methods in automatic extracting. **Journal of the ACM**, v. 16, n. 2, p. 264-285, 1969.

EMIGH, W.; HERRING, S.C. Collaborative authoring on the Web: a genre analysis of online encyclopedias. In: ANNUAL HAWAII INTERNATIONAL CONFERENCE ON SYSTEM SCIENCES, 38., 2005, Big Island. **Proceedings**… Los Alamitos: IEEE Computer Society, 2005. p. 9-10.

ESTIVAL, D.; NOWAK, C.; ZSCHORN, A. Towards ontology-based natural language processing. In: ACL WORKSHOP ON NLP AND XML, 4., 2004, Barcelona. **Proceedings**… Cambridge: MIT Press, 2004. p. 59-66.

FAATZ, A.; STEINMETZ, R. Ontology enrichment with texts from the WWW. In: SEMANTIC WEB MINING: WORKSHOP AT ECML/PKDD, 2., 2002, Helsinki. **Proceedings**… Sankt Augustin: Knowledge Discovery Network of Excellence, 2002. p. 20-34.

FENSEL, D.; WAHLSTER, W.; LIEBERMAN, H. **Spinning the Semantic Web:** bringing the World Wide Web to its full potential. Cambridge: MIT Press, 2002. 392 p.

FILATOVA, E.; HATZIVASSILOGLOU, V. A formal model for information selection in multi-sentence text extraction. In: COLING: INTERNATIONAL CONFERENCE

ON COMPUTATIONAL LINGUISTICS, 20., 2004, Geneva. **Proceedings**… Cambridge: MIT Press, 2004. p. 397-403.

FREITAS, F. L. G. D. Ontologias e a Web semântica. In: CONGRESSO DA SOCIEDADE BRASILEIRA DE COMPUTAÇÃO, 23., 2003, Campinas. **Anais**… Campinas: SBC, 2003. p. 1-52.

FRESNO, V.; RIBEIRO, A. An analytical approach to concept extraction in HTML environments. **Journal of Intelligent Information Systems**, v. 22, n. 3, p. 215-235, 2004.

FURNAS, G. W. et al. The vocabulary problem in human-system communication: an analysis and a solution. **Communications of the ACM**, v. 30, n. 11, p. 964-971, 1987.

GENESERETH, R. M.; NILSSON, L. **Logical foundations of AI**. Los Altos: Morgan Kaufman, 1987.

GILES, J. Internet encyclopaedias go head to head. 2005. **Nature International Weekly Journal of Science**. Disponível em: <http://www.nature.com/nature/journal/v438/n7070/full/438900a.html>. Acesso em: 15 dez. 2005.

GREGHI, J. G.; MARTINS, R.; NUNES, M. G. V. Diadorim: a lexical database for brazilian portuguese. In: LREC: INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION, 3., 2002, Las Palmas. **Proceedings**… Paris: ELRA, 2002. v. 4, p. 1346-1350.

GRIESBAUM, J. Evaluation of three German search engines: Altavista.de, Google.de and Lycos.de. **Information Research: an international electronic journal**, v. 9, n. 4, 2004. Disponível em: <http://informationr.net/ir/9-4/paper189.html>

GRUBER, T. R. A translation approach to portable ontologies. **Knowledge Acquisition**, v. 5, n. 2, p. 199-220, 1993.

GUARINO, N.; CARRARA, M.; GIARETTA, P. Formalizing ontological commitment. In: NATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE, 12., 1994, Seattle. **Proceedings**… San Fransisco: Morgan Kaufmann, 1994. p. 560-567.

HACHEY, B.; MURRAY, G.; REITTER, D. Embra System at DUC 2005: query-oriented multi-document summarization with a very large latent semantic. In: DOCUMENT UNDERSTANDING CONFERENCE, 5., 2005, Vancouver. **Proceedings**… Vancouver: DUC, 2005. Disponível em: <http://duc.nist.gov/pubs/2005papers/uedinburgh.hachey.pdf>. Acesso em: 09 mar. 2006.

HASSEL, M. **Development of a swedish corpus for evaluating summarizers and other IR-tools**. Stockholm: Royal Institute of Technology (KTH), 2003. Relatório TRITA-NA-P0112.

INKTOMI Corp. **Web search relevance test**. 2003. Veritest. Disponível em: <http://www.veritest.com/clients/reports/inktomi/inktomi_Web_search_test.pdf>. Acesso em: 5 maio 2005.

JANSEN, B. The effect of query complexity on Web searching results. **Information Research:** an international electronic journal, v. 6, n. 1, 2001. Disponível em: < http://informationr.net/ir/6-1/paper87.html>. Acesso em: mar. 2005.

LEITE, D.S.; RINO, L.H.M. Selecting a Feature Set to Summarize Texts in Brazilian Portuguese. In: INTERNATIONAL JOINT CONFERENCE IBERAMIA/SBIA, 2006, Ribeirão Preto. **Proceedings**… Heildelberg : Springer-Verlag.

LEWIS, J.R. Computer usability satisfaction questionnaires: psychometric evaluation and instructions for use. **International Journal of Human-Computer Interaction**, v. 7, n. 1, p. 57-78, 1995.

LIH, A. Wikipedia as participatory journalism: reliable sources? Metrics for evaluating collaborative media as a news resource. In: INTERNATIONAL SYMPOSIUM ON ONLINE JOURNALISM, 5., 2004, Austin. **Proceedings**… Disponível em: <http://staff.washington.edu/clifford/teaching/readingfiles/utaustin-2004-wikipedia-rc2.pdf>. Acesso em: dez. 2005.

LIN, C. Knowledge-based automatic topic identification. In: ANNUAL MEETING ON ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 33., 1995, Morristown. **Proceedings**… Cambridge: MIT Press, 2004. p. 308-310.

LOH, S. **Uma abordagem baseada em conceitos para descoberta de conhecimento em textos**. 2001. 110 f. Tese (Doutorado) - Instituto de Informática, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2001.

LUHN, H.P. The automatic creation of literature abstracts. **IBM Journal of Research Development**, v. 2, n. 2, p.159-165, 1958.

LYMAN, P.; VARIAN, H.R. **How much information.** 2003. Disponível em: <http://www2.sims.berkeley.edu/research/projects/how-much-info-2003>. Acesso em: 19 maio 2005.

MANI, I. **Automatic summarization**. Amsterdam: John Benjamins, 2001. 285 p.

MANI, I. et al. SUMMAC: a text summarization evaluation. **Natural language engineering**, v. 8, n. 1, p. 43-68, 2002.

MANI, I.; MAYBURY, M. **Advances in automatic text summarization**. Cambridge: MIT Press, 1999. 442 p.

MARCKINI, F. **Search engine positioning**. Plano: Wordware Publishing, 2001. 576 p.

MCBRYAN, O. GENVL and WWWW: tools for taming the Web. In: INTERNATIONAL WORLD WIDE WEB CONFERENCE, 1994, Geneva. **Proceedings**… Amsterdam: Elsevier Science, 1994. p. 308-322.

MILLER, G.A. WordNet: a lexical database for english. **Communications of the ACM**, v. 38, n. 11, p. 39-41, 1995.

MLADENIC, D.; GROBELNIK, M. Assigning keywords to documents using machine learning. In: INTERNATIONAL CONFERENCE ON INFORMATION AND INTELLIGENT SYSTEMS, 10., 1999, Varazdin. **Proceedings**… Varazdin: Faculty of Organization and Informatics, University of Zagreb 1999. p. 123-131.

MÓDOLO, M. **SuPor: um Ambiente para a Exploração de Métodos Extrativos para a Sumarização Automática de Textos em Português**. 2003. Dissertação (Mestrado) - Departamento de Computação, Universidade Federal de São Carlos, São Carlos, 2003.

NOMOTO, T.; MATSUMOTO, Y. Exploiting text structure for topic identification. In: ACL-SIGDAT: WORKSHOP ON VERY LARGE CORPORA, 4., 1996, Copenhagen. **Proceedings**… Cambridge: MIT Press, 1996. p. 101-112.

PARDO, T. A. S.; RINO, L. H. M. **Descrição do GEI**: gerador de extratos ideais para o português do Brasil. São Carlos-SP: USP/ICM, 2004. Série de Relatórios do NILC; NILC-TR-04-07.

PARDO, T. A. S.; RINO, L. H. M. **TeMário**: um corpus para sumarização automática de textos. São Carlos-SP: USP/ICM, 2003. Série de Relatórios do NILC; NILC-TR-03-09.

PARIS, C.; AMITAY, E. Automatically summarising Web sites - is there a way around it? In: INTERNATIONAL CONFERENCE ON INFORMATION AND KNOWLEDGE MANAGEMENT, 9., 2000, McLean. **Proceedings**… New York: ACM Press, 2000. p. 173-179.

PARSONS, S. Current approaches to handling imperfect information in data and knowledge bases. **IEEE Transactions on Knowledge and Data Engineering**, v. 8, n. 3, p. 353-372, 1996.

PEDREIRA-SILVA, P.; RINO, L. H. M. **Código HTML**: contribuições para a determinação de informações relevantes de documentos. São Carlos-SP: USP/ICMC, 2005. Série de Relatórios do NILC; NILC-TR-05-06.

PERLMAN, G. **Web-based user interface evaluation with questionnaires.** 2001. Disponível em: <http://www.acm.org/~perlman/question.html>. Acesso em: 1 fev. 2006.

RAGGETT, D.; HORS, A. L.; JACOBS, I. **HTML 4.01 specification. W3C Recommendation REC-html401-19991224**. World Wide Web Consortium (W3C). 1999. Disponível em: <http://www.w3.org/TR/1999/REC-html401-19991224/>. Acesso em: 5 maio 2005.

RINO, L. H. M. et al. A comparison of automatic summarization systems for brazilian portuguese texts. In: SBIA: BRAZILIAN SYMPOSIUM ON ARTIFICIAL INTELLIGENCE, 17., 2004, São Luis. **Proceedings**… Heildelberg : Springer-Verlag, 2004. p. 235-244.

SALTON, G. **Automatic text processing:** the transformation, analysis, and retrieval of information by computer. Boston: Addison-Wesley Longman, 1988. 530 p.

SALTON, G.; MCGILL, M. J. **Introduction to modern information retrieval**. New York: McGraw-Hill, 1986. 448 p.

SAMPSON, G. Probabilistic models of analysis. In: LEECH, G.; GARSUDE, R.; SAMPSON, G., (Ed.). **The computational analysis of English**. London: Longman Harrow, 1987. p. 16-29.

SEDIGH, A. K.; ROUDAKI, M. Identification of the dynamics of the Google's ranking algorithm. In: IFAC SYMPOSIUM ON SYSTEM IDENTIFICATION, 13., 2003, Rotterdam. **Proceedings**… Oxford: Elsevier Science, 2003.

SHERMAN, C. **The search engine "perfect page" test.** 2002. Disponível em: <http://searchenginewatch.com/searchday/article.php/2161121>. Acesso em: 6 maio 2005.

SPINK, A.; JANSEN, B. J.; SARACEVIC, T. Real life, real users, and real needs: a study and analysis of user queries on the Web. **Information Processing and Management**, v. 36, n. 2, p. 207-227, 2000.

SULLIVAN, D. **Major search engines and directories**. 2004. Disponível em: <http://searchenginewatch.com/links/article.php/2156221>. Acesso em: 5 maio 2005.

TANG, R.; SHAW, W.; VEVEA, J. Towards the identification of the optimal numbers of relevance categories. **Journal of the American Society for Information Science**, v. 50, n. 3, p. 254-264, 1999.

TIUN, S.; ABDULLAH, R.; KONG, T. E. Automatic topic identification using ontology hierarchy. In: CICLING: CONFERENCE ON INTELLIGENT TEXT PROCESSING AND COMPUTATIONAL LINGUISTICS, 2., 2001, Mexico City. **Proceedings**… Heildelberg : Springer-Verlag, 2001. p. 444-453.

WHITE, R. W.; RUTHVEN, I.; JOSE, J. M. Finding relevant documents using top ranking sentences: an evaluation of two alternative schemes. In: ANNUAL INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 25., 2002, New York. **Proceedings**… New York: ACM Press, 2002. p. 57-64.

WITTEN, I.; MOFFAT, A.; BELL, T. **Managing gigabytes:** compressing and indexing documents and images. New York: Van Nostrand Reinhold, 1994.

WIVES, L.K. **Utilizando conceitos como descritores de textos para o processo de identificação de conglomerados (clustering) de documentos**. 2004. 136 f. Tese

(Doutorado) - Instituto de Informática, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2004.

WOLFRAM, D. et al. Vox populi: the public searching of the Web. **Journal of the American Society of Information Science and Technology,** v. 52, n. 12, p. 1073-1074, 2001.

WOODRUFF, A. et al. An investigation of documents from the World Wide Web. **Computer Networks and ISDN Systems**, v. 28, n. 7-11, p. 963-980, 1996.

WU, C. W.; LIU, C. L. Ontology-based text summarization for business news articles. In: ISCA: INTERNATIONAL CONFERENCE ON COMPUTERS AND THEIR APPLICATIONS, 18., 2003, Honolulu. **Proceedings**… Cary: ISCA, 2003. p. 389-392.

ZHANG, Y.; ZINCIR-HEYWOOD, N.; MILIOS, E. World Wide Web site summarization. **Web Intelligence and Agent Systems**, v. 2, n. 1, p. 39-53, 2004.

Apêndice – Interface Web dos sistemas de sumarização automática implementados

A construção de uma ferramenta de sumarização automática completa demanda tempo, principalmente devido aos recursos necessários à sua implementação. Assim, pela complexidade da tarefa, foram desenvolvidos apenas protótipos das ferramentas de sumarização descritas nesta dissertação. O HTMLSUMM, o GEO e o ExtraWeb foram desenvolvidos em linguagens próprias para o ambiente *Web* (PHP, HTML, DHTML e JavaScript), assim, uma vez implantado em um servidor *Web*, eles poderão ser acessados *on-line* a partir de qualquer navegador. Todas as três ferramentas encontram-se integradas em um único ambiente de sumarização (Figura A.1). Para utilizar um dos sistemas, o usuário deve escolher, entre as três opções disponíveis, o sumarizador que ele deseja executar.

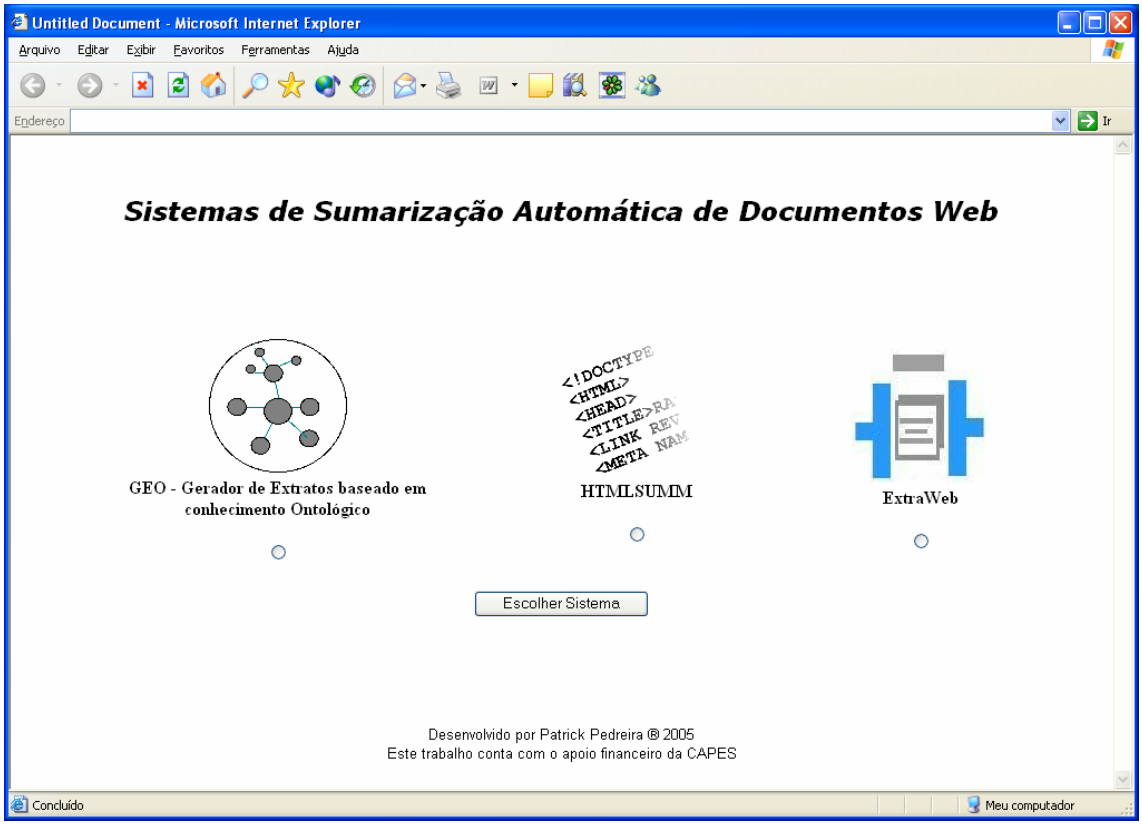


Figura A.1. Tela inicial do Sistema Integrado de Sumarização Web

Ao ser escolhido um dos sumarizadores, é exibida uma tela com três partes principais: uma janela-filha onde será exibido o documento-fonte, uma barra de endereços onde deverá ser fornecida a URL do documento e um menu onde o usuário poderá solicitar que seja gerado automaticamente o sumário, conforme mostra a Figura A.2.

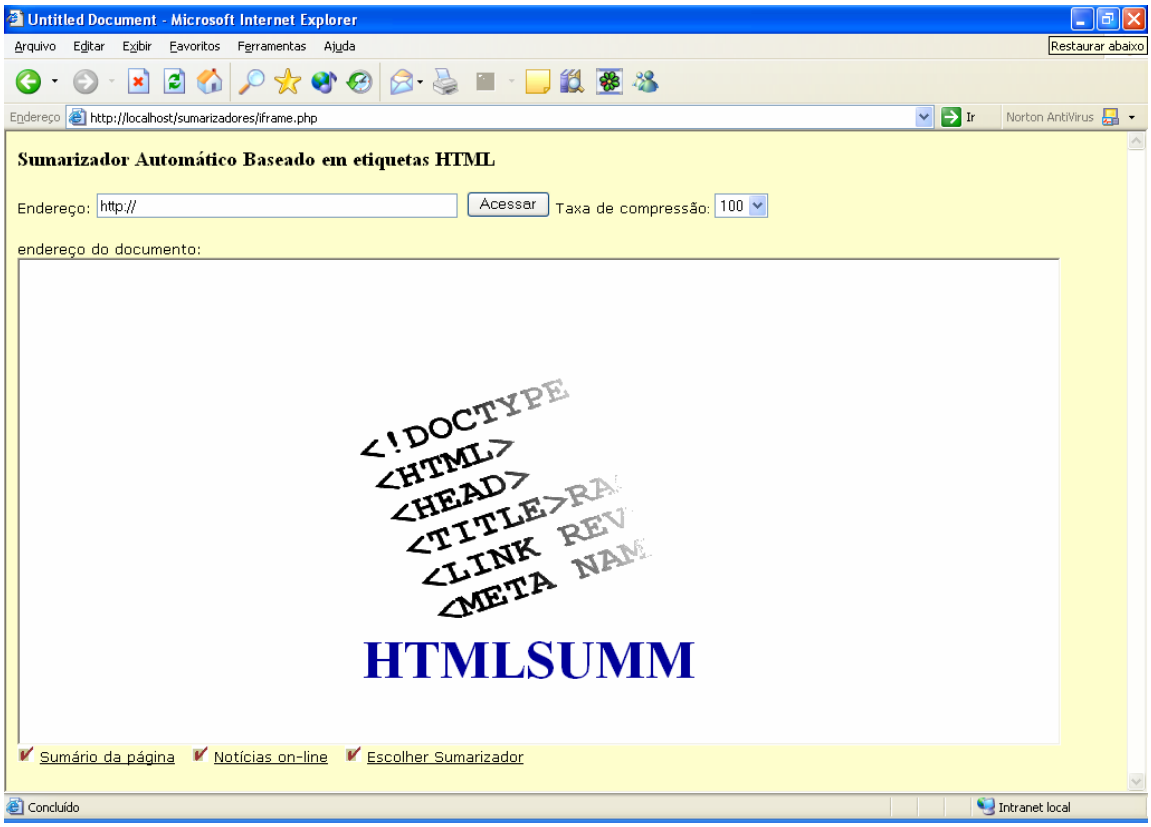


Figura A.2. Tela inicial do HTMLSUMM

Para sumarizar um documento, o usuário deve, primeiramente, fornecer a URL, escolher taxa de compressão desejada e, então, pressionar o botão *"Acessar"* como mostra a Figura A.3. Ao fazer isso, será exibido na janela principal o documento a ser sumarizado. Esse documento deve ser um arquivo HTML. A Figura A.4 mostra o ambiente de sumarização com um documento selecionado para sumarização automática. Pode-se notar que acima da janela principal, onde é carregado o documento-fonte, o sistema indica o *link* do documento em foco.

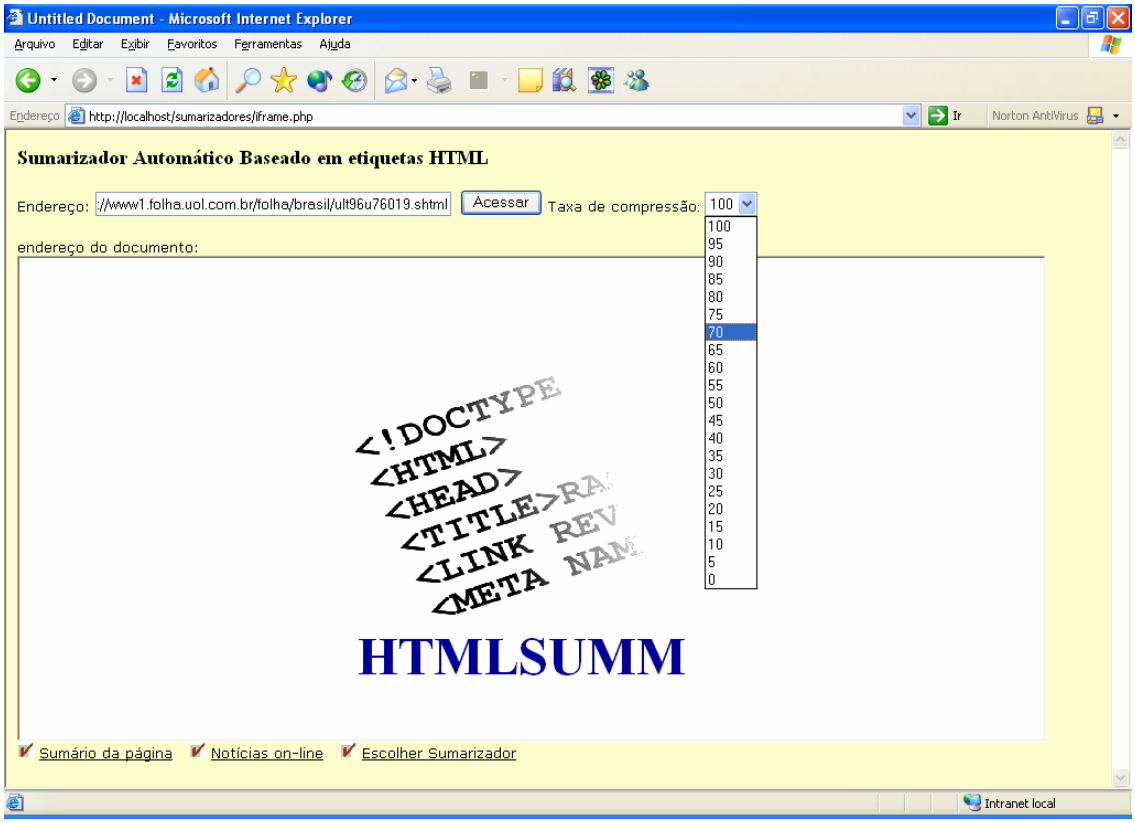


Figura A.3. Indicação do documento-fonte e seleção da taxa de compressão

Figura A.4. Documento-fonte a sumarizar

Após carregar o documento-fonte, o usuário poderá solicitar que seja gerado o sumário automático usando o menu inferior. A Figura A.5 mostra essa etapa de escolha do usuário.

Figura A.5. Sumarizando o documento

Para sumarizar o documento-fonte, basta clicar no menu *"Sumário da Página"*, conforme mostrado na Figura A.5. O sumário gerado aparecerá, então, em uma nova janela, como é mostrado na Figura A.6.

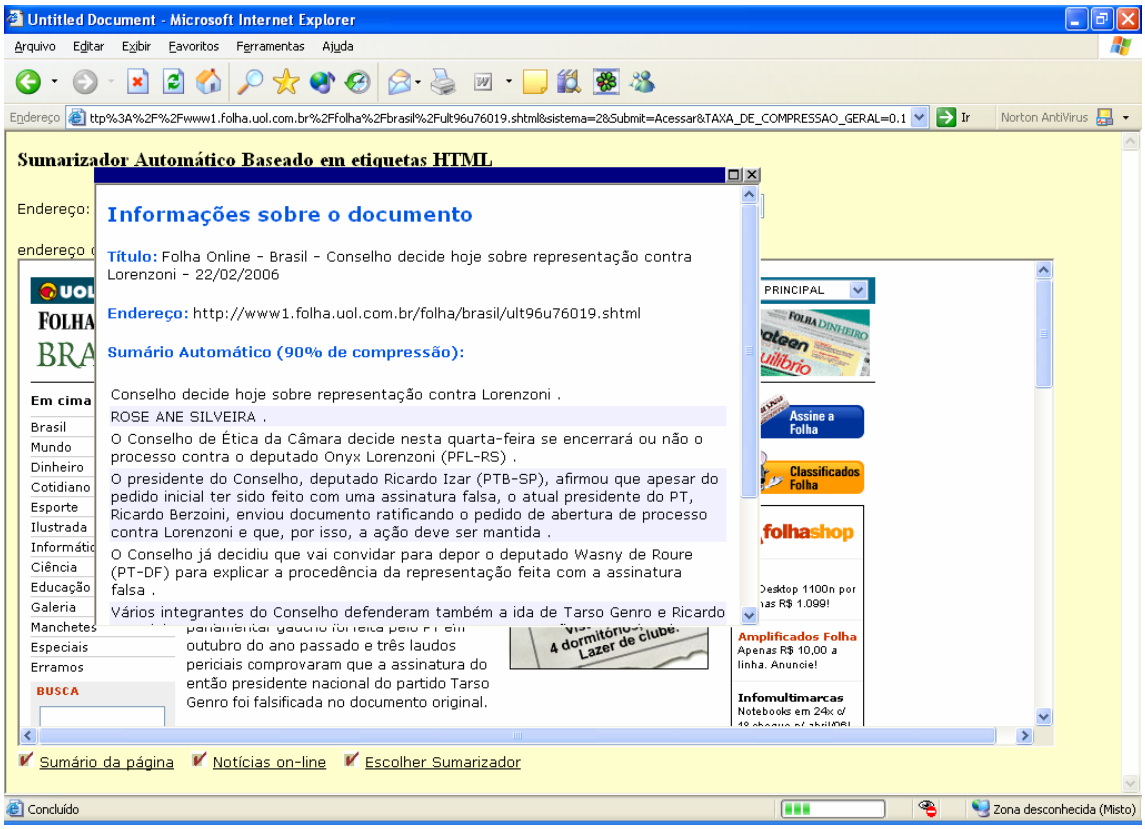


Figura A.6. Exibição do sumário gerado

Para gerar sumários de outros documentos, basta repetir os passos anteriores. Além das funções de um sistema de sumarização automática, este ambiente *Web* de sumarização funciona como um agregador de notícias no formato RSS (sigla para *Really Simple Syndication* ou *Rich Site Summary*). O formato, criado em 1997, usa a tecnologia XML (*Extensible Markup Language*) para a publicação automática de conteúdo e *links* de um site. Também permite o uso de programas conhecidos como leitores RSS ou agregadores, que reúnem o conteúdo de diversos sites em uma só interface. Os arquivos RSS trazem, geralmente, o título das notícias, além de um *link* que remete para o site do documento-fonte, permitindo a leitura do texto completo. Essa função pode ser acessada pela interface através da opção "*Notícias on-line*".

Para exemplificar o uso da tecnologia RSS integrada a um sumarizador de documentos *Web*, são disponibilizados na interface alguns arquivos RSS para as sites de notícias em português, espanhol e inglês, conforme podemos ver na Figura A.7.

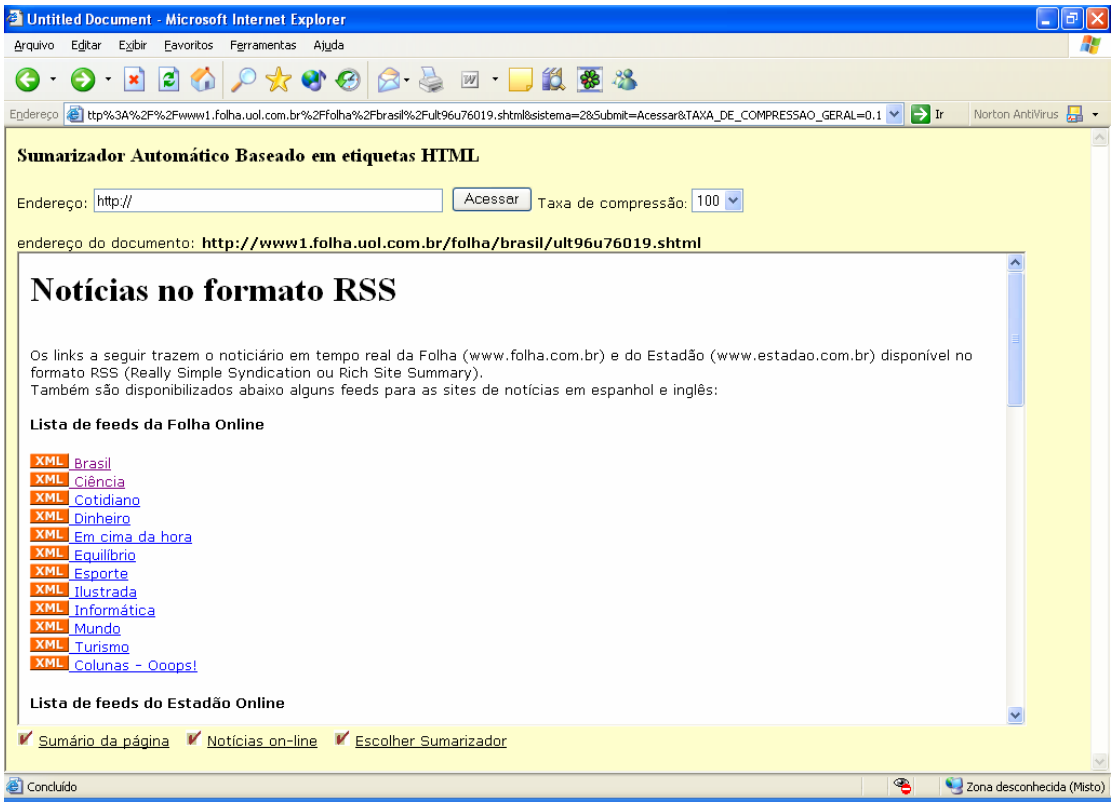


Figura A.7. Lista de arquivos RSS

Clicando sobre um dos arquivos RSS disponíveis, serão exibidas todas as notícias relacionadas ao arquivo conforme é visto na Figura A.8.

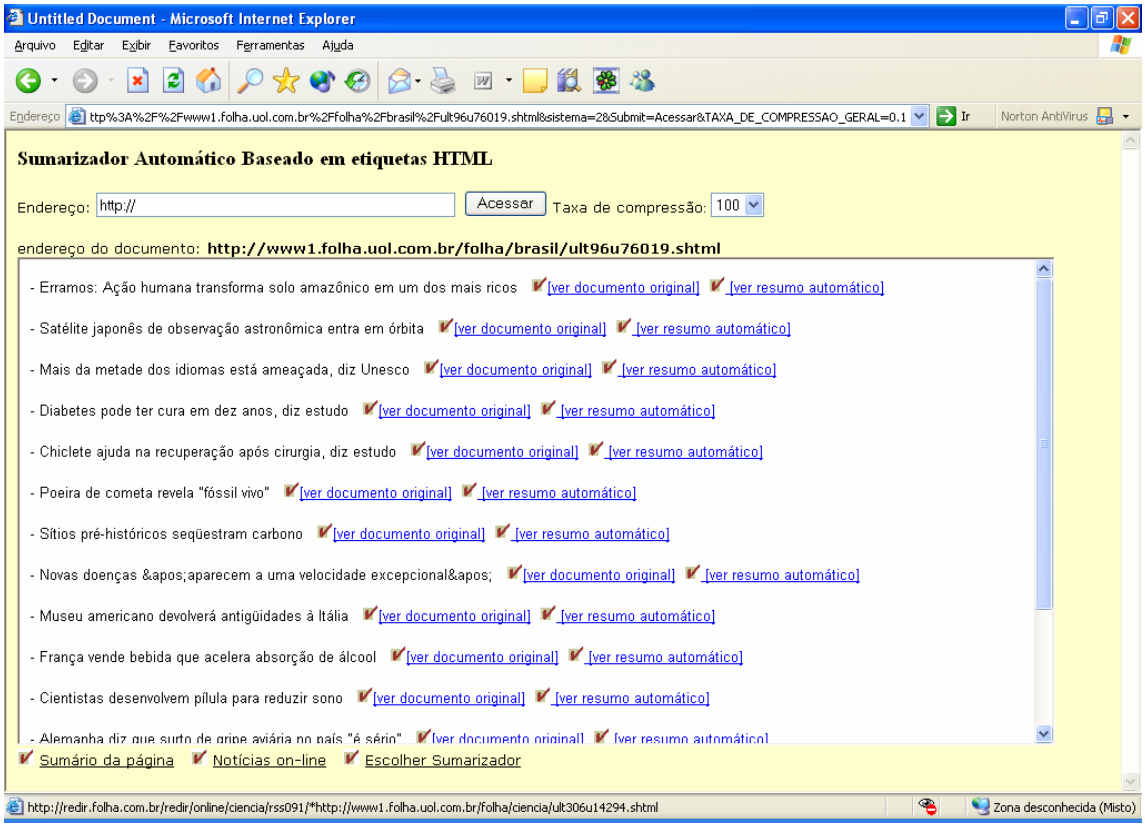


Figura A.8. Notícias vinculadas aos arquivos RSS

A partir da tela da Figura A.8 anterior, o usuário poderá solicitar a exibição do documento original ou solicitar que seja gerado o seu sumário. A opção "*Escolher Sumarizador*", disponível no menu inferior, permite que o usuário alterne entre os três sumarizadores disponíveis.