UNIVERSIDADE FEDERAL DO AMAZONAS

PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO

DEPARTAMENTO DE APOIO A PESQUISA

PROGRAMA INSTITUCIONAL DE INICIAÇÃO CIENTÍFICA

INICIAÇÃO CIENTÍFICA EM EXTRAÇÃO E UNIFORMIZAÇÃO DE RESUMOS ESTRUTURADOS DE PÁGINAS DE ÍNDICE

Bolsista: Erickson Higor da Silva Alves, CNPq

Manaus

2013

INICIAÇÃO CIENTÍFICA EM EXTRAÇÃO E UNIFORMIZAÇÃO DE RESUMOS ESTRUTURADOS DE PÁGINAS DE ÍNDICE

UNIVERSIDADE FEDERAL DO AMAZONAS

PRO REITORIA DE PESQUISA E PÓS-GRADUAÇÃO

DEPARTAMENTO DE APOIO A PESQUISA

PROGRAMA INSTITUCIONAL DE INICIAÇÃO CIENTÍFICA

RELATÓRIO FINAL

PIB - E/0119/2012

INICIAÇÃO CIENTÍFICA EM EXTRAÇÃO E UNIFORMIZAÇÃO DE RESUMOS ESTRUTURADOS DE PÁGINAS DE ÍNDICE

Bolsista: Erickson Higor da Silva Alves, CNPq

Orientador: Prof. Dr.David Braga Fernandes de Oliveira

MANAUS

2013

Todos os direitos deste relatório são reservados à Universidade Federal do Amazonas, ao Núcleo de Estudo e Pesquisa em Ciência da Informação e aos seus autores. Parte deste relatório só poderá ser reproduzida para fins acadêmicos ou científicos.

Esta pesquisa, financiada pelo Conselho Nacional de Pesquisa – CNPq, através do Programa Institucional de Bolsas de Iniciação Científica da Universidade Federal do Amazonas, foi desenvolvida pelo Núcleo de Estudo e Pesquisa em Ciência da Informação e se caracteriza como sub projeto do projeto de pesquisa Bibliotecas Digitais

RESUMO

Nos dias de hoje percebe-se a grande variedade em layouts de sites, reflexo da mesma variedade de desenvolvedores e designers e da informatização de lojas por meio de sites. Nestes sites, as informações básicas de um produto exibido na página são conhecidas como resumos estruturados. Esses resumos estruturados são escolhidos pelo time de desenvolvimento do site, não seguindo nenhum padrão comum a não ser o próprio bom senso. Variando de site para site, é notória a diferença de informação apresentada para o usuário, de forma que ao efetuar uma busca de loja em loja virtualmente, há divergências na localização da informação, diferentemente do aspecto real.

A importância da padronização desses resumos estruturados tem ligação com a representação do mundo real, já que no âmbito de uma loja real de livros por exemplo, ao mudar de loja, as informações são as mesmas e encontradas da mesma forma, diferindo das lojas virtuais.

O projeto tem como objetivo montar uma estratégia para a padronização de resumos estruturados para sites de determinado seguimento, que contém as mesmas informações, mas não estão dispostos na mesma maneira para o cliente.

Para essa realização, um algoritmo rotulador implementado em Java define cada informação extraída de sites diferentes e uma página em PHP foi montada para a representação pós-tratamento pelo algoritmo.

ÍNDICE DE FIGURAS

Figura 1 - Página de resultados de uma máquina de busca 8
Figura 2 - Exemplo de snippet estruturado 9
Figura 3 - Exemplo de página de índice e seu resumo estruturado 9
Figura 4 - Representação do método utilizado 13
Figura 5 - Exemplo de resumo estruturado padronizado 14

SUMÁRIO

RESUMO 1

ÍNDICE DE FIGURAS 6

INTRODUÇÃO 8

REVISÃO BIBLIOGRÁFICA 11

MÉTODOS UTILIZADOS 12

Identificação de rótulos 12

RESULTADOS E DISCUSSÕES 13

Arquivos de entrada 13

Algoritmo rotulador (Labeler algorithm) 13

Padronização dos dados 14

CONCLUSÕES 16

REFERÊNCIAS BIBLIOGRÁFICAS 17

INTRODUÇÃO

Máquinas de busca para a Web são as mais populares ferramentas para encontrar informação útil sobre um determinado assunto de interesse. A justificativa para essa popularidade é a forma simples e natural com a qual as pessoas interagem com elas. O usuário submete uma consulta usando palavras-chaves e recebe como resposta uma lista de páginas relacionadas com a consulta formulada, com a expectativa de que tais documentos sejam relevantes para as necessidades de informação do usuário [4].

Para auxiliar o usuário a decidir quais documentos satisfazem suas necessidades de informação, as máquinas de busca apresentam um resumo de cada página presente em sua lista de resposta (vide Figura 1). A geração desses resumos está entre os principais temas de pesquisa da área de recuperação de informação, uma vez que eles podem influenciar a percepção de relevância de um documento. Se um resumo de baixa-qualidade é gerado para um documento muito relevante, o usuário pode entender o documento como não relevante e não fazer acesso ao seu conteúdo, e vice-versa [5].

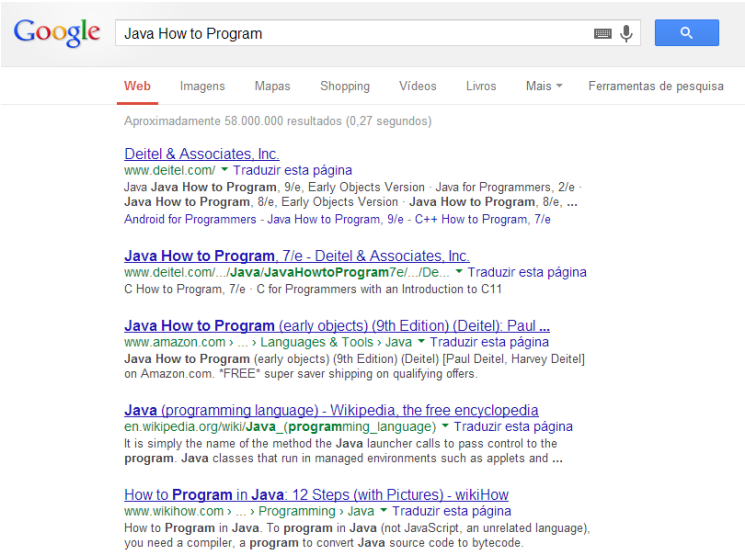

Figura 1 - Página de resultados de uma máquina de busca

Esses resumos são mais conhecidos como snippets, e geralmente são formados pela concatenação de sentenças textuais extraídas das páginas Web. Tradicionalmente, os snnipets são extratos não estruturados de textos retirados de cada página de resposta, mas estudos recentes mostraram que páginas de resposta com snnipets estruturados são capazes de dar aos usuários um

melhor entendimento sobre as páginas para os quais apontam [2]. Snnipets estrutados são aqueles que dividem as informações apresentadas em um ou mais blocos (vide Figura 2).

Figura 2 - Exemplo de snippet estruturado

Neste projeto, apresentamos um método de geração de snnipets estruturados, agregando fragmentos textuais e metadados das páginas Web. Este método faz uso das páginas de índice, que são páginas Web contendo resumos estruturados de outras páginas do mesmo Web site. Por exemplo, a Figura 3 mostra uma página de índice contendo uma listagem de livros, sendo que cada item desta listagem possui um resumo do conteúdo da página ao qual o item aponta.

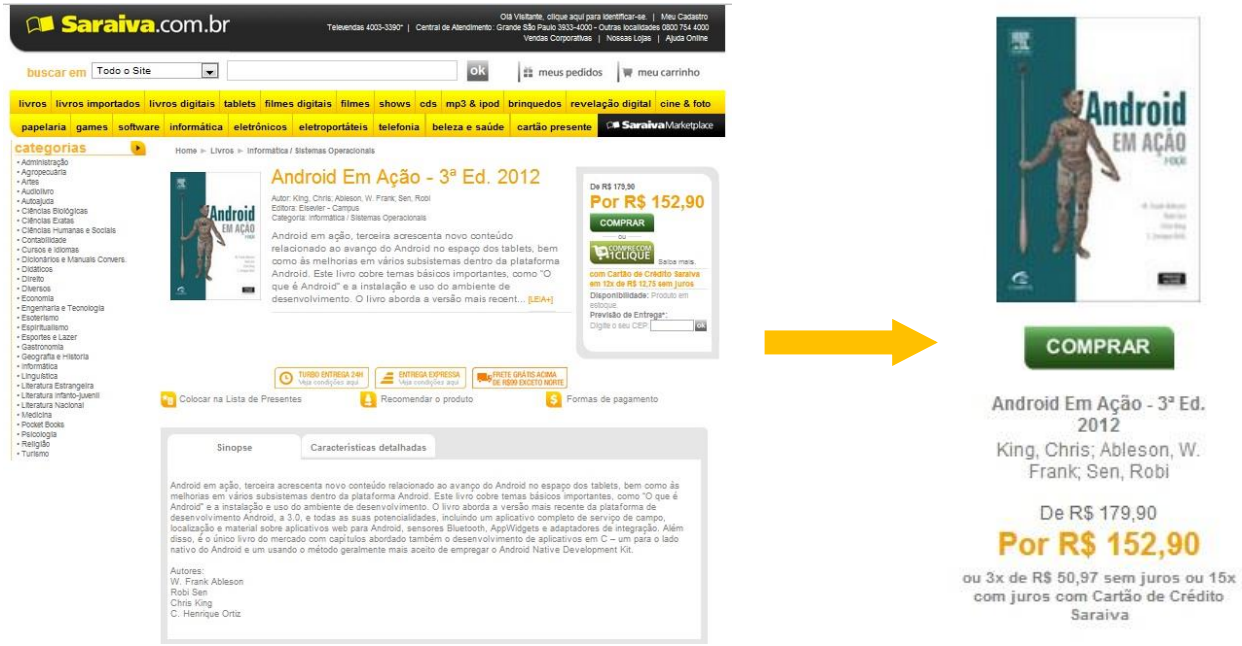

Figura 3 - Exemplo de página de índice e seu resumo estruturado

Um problema com a extração direta de resumos estruturados a partir das páginas de índice é que o formato dos resumos pode variar muito de site para site. Por exemplo, cada resumo da Figura 3 apresenta uma imagem (capa de um livro) acima das demais informações do resumo, mas outras livrarias virtuais poderiam optar dispor as imagens a esquerda ou a direita do

resumo. Isso é um problema, porque apresentar snnipets com diferentes estruturas em uma mesma página de respostas pode confundir os usuários das máquinas de busca, dificultando a identificação dos documentos relevantes para suas consultas.

Neste trabalho, apresentamos um método para padronizar a disposição das informações dos resumos extraídos a partir de páginas de índice. Para tanto, criamos formatos padrões de resumos para determinados domínios de aplicação (produtos, filmes, notícias, etc), de forma que os resumos de um mesmo domínio sejam sempre reestruturados em um mesmo formato. Desta forma, uma vez que um resumo foi extraído de uma página de índice, suas informações são rearranjadas conforme o domínio da aplicação do resumo. Para fazer isso, identificamos rótulos capazes de identificar o tipo de informação presente em cada bloco do resumo estruturado, e então preenchemos o modelo de domínio correspondente com base nos rótulos encontrados. Para tanto, adotamos a estratégia de rotulamento de informações proposta em [3]. Adotando esta estratégia, obtivemos uma precisão de cerca de 90% nos rótulos encontrados.

REVISÃO BIBLIOGRÁFICA

Bando et al [1] mostrou a diferença na construção de snippets por humanos e gerados por sistemas. Com o seu estudo, ficou evidente a divergência entre os resumos gerados. Os humanos tendem a repetir palavras de acordo com o seu uso pessoal, já os sistemas tendem a trazer informações mais mecânicas. Também pode-se notar que os humanos geram resumos que possam responder ao problema de forma sucinta e às vezes mais precisa que os resumos gerados automaticamente.

Haas et al [2] desenvolve sobre o aprimoramento de pesquisas na web, de forma a trazer resultados mais precisos e mais diretos, sem que tenha a necessidade de navegação através de páginas e páginas de resultado. Alterações na forma de pesquisar, acréscimo de palavras-chave, pesquisa com objetos diferentes de texto, são formas de aumentar a potencialidade da pesquisa feita por um usuário. A questão envolvida é o aumento do poder de busca de uma ferramenta através de pequenas alterações na forma de pesquisar, fato que é desconhecido pelos usuários mas que pode aprimorar os resultados.

Altigran et al [3] disserta e desenvolve sobre rotulação de dados extraídos da web. O problema é a não classificação de dados retornados pela web através de extração automática de dados. Para solucionar tal problema, o autor mostra que máquinas de busca são ferramentas auxiliares que trazem uma forma de solucionar o problema. Ao realizarem-se buscas utilizando o dado a ser rotulado com o acréscimo de palavras-chave, a máquina de busca retorna possíveis candidatos a rotular a informação. A partir dos candidatos, um ranking é feito com as possibilidades e o resultado são os candidatos mais propícios a definirem cada bloco de informação.

MÉTODOS UTILIZADOS

Identificação de rótulos

Para demonstrar como funciona a identificação dos rótulos em nosso método, vamos apresentar um exemplo prático explanatório para que seja possível o entendimento desse procedimento. A princípio, tem-se o seguinte conjunto de informação: “Java How to Program”, “Deitel” e “Pearson”. Neste exemplo, adotasse que esse conjunto de informações foi extraído do resumo estruturado de alguma página de índice pertencente a um site qualquer. Com essas informações inseridas em um arquivo CSV, pode-se começar o processo de identificação de rótulos.

O algoritmo iniciará seu funcionamento com a primeira coluna do bloco de informação, no caso “Java How to Program”. Essa string é utilizada como parâmetro para que uma classe programada possa realizar buscas sucessivas em uma máquina de busca pré-definida, para o nosso método, utilizamos o Google. As buscas efetuadas são do tipo “*<string> is*”, “*such as <string>*” e outras pesquisas pré-definidas para tentar definir com uma palavra-chave o que é, de fato, a string. A partir de várias buscas, todos os resultados retornados, dez por consulta, são tratados de forma a retirar palavras não significativas, como artigos, pronomes, preposições e assim por diante. Depois desse tratamento, a probabilidade de que o restante das palavras define de alguma forma a string é muito alta.

Então, essas palavras são organizadas em ranking para verificar quais as palavras aparecem mais frequentemente nos resultados. Após esse processo, é gerado um arquivo com os dez rótulos melhor ranqueados para cada coluna dos blocos de informações, ou seja, o melhor rótulo para a informação a qual se deseja rotular é o que primeiro aparece no arquivo de saída.

Voltando ao nosso exemplo, a primeira string “Java How to Program” é submetida a sucessivas buscas e o restante do processo apresentado acima. Para o arquivo de saída dessa string, obteve-se como principal resultado “*title*”. Para “Deitel” e “Pearson”, respectivamente, os resultados obtidos foram “*author*” e “*publisher*”. Com esses rótulos, podemos estruturar o resumo de acordo com os padrões abordados em nosso trabalho.

RESULTADOS E DISCUSSÕES

O projeto está divido em três partes principais, descritas pela figura abaixo. Por meio dela, é possível observar o processo completo realizado pelo projeto para a execução do mesmo.

Figura 4 - Representação do método utilizado

Arquivos de entrada

Os arquivos de entrada são dados referentes a produtos extraídos previamente de sites de compras e organizados no formato CSV, onde cada linha representa um produto diferente e cada coluna representa uma informação referente ao produto. Cada coluna deve, necessariamente, conter a mesma informação para todas as linhas do arquivo.

Algoritmo rotulador (Labeler algorithm)

O algoritmo rotulador serve para a rotulação dos dados presentes nos arquivos de entrada. Para cada conjunto de informações da mesma coluna nesses arquivos, o algoritmo retornará um outro arquivo, contendo as dez classes ordenadas que representam a possível classificação das informações, onde a primeira classe é a que contém maior probabilidade de estar correta.

O funcionamento desse algoritmo é simples. Dado uma cadeira de caracteres que representam uma informação qualquer, o algoritmo fará pesquisas na máquina de busca selecionada de forma a tentar definir o que essa informação representa. Essas pesquisas são "<cadeia> is ", "such as <cadeira>" entre outras do mesmo gênero que instigam a pesquisa a definir o que a cadeia pesquisada representa. Os resultados da pesquisa são denominados

snippets, e é a partir deles que o algoritmo continuará o trabalho. Tendo os snippets, o algoritmo trata cada conjunto deles de forma a eliminar as palavras desnecessárias, como artigos, preposições, conjunções, e mantém apenas substantivos. Esses substantivos definem a cadeia de caracteres e são ordenados por um sistema de pontuação de acordo com sua ocorrência. Ao final das execuções, tem-se o arquivo com as possíveis definições para a cadeia pesquisada.

Padronização dos dados

A partir do fim da execução do algoritmo, tem-se os arquivos com as informações do arquivo de entrada classificadas em dez classes distintas. De acordo com os resultados do algoritmo, por padrão, a primeira classe representa a definição de maior probabilidade, logo, ela é a escolhida para representar aquele conjunto de atributos de cada produto. Em seguida, é montada uma página em PHP para a visualização dos dados. Contendo os dados classificados, adicionam-se os mesmos numa tabela em uma base de dados de teste e os produtos podem ser visualizados nessa página.

Esse passo representa o objetivo final do projeto, que é a padronização dos resumos estruturados para determinado conjunto de sites. Aqui se abrem as possibilidades para estudos futuros, pois o resumo estruturado padrão está concluído nesta etapa.

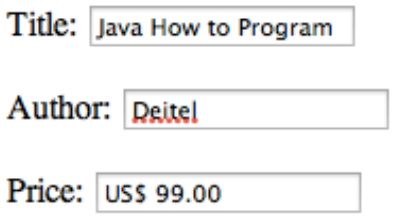

Figura 5 - Exemplo de resumo estruturado padronizado

Abaixo, tem-se uma tabela com o resumo dos nossos resultados obtidos durante o projeto:

Número de arquivos de teste	Número médio de atributos	Número médio de registros	Porcentagem de acerto
10	5	72	84%
20	6	68	91%
30	5	52	89%
40	5	54	88%
Média	5,25	61,5	88%

Tabela 1 - Resultados obtidos

Para o nosso método, demonstraremos agora por meio de um exemplo o significado de acerto e de erro. Para a string “Java How to Program” um acerto é caracterizado pelo retorno de “title” como rótulo, pois a string representa o título de um livro. Já um erro, para a string “U$ 32” é caracterizado pelo rótulo “buy” ao invés de “price”, já que essa string representa o preço de um livro.

CONCLUSÕES

O projeto surgiu para a estruturação e uniformização de diferentes resumos coletados de sites de compras, pois não havia uma padronização para os diversos sites pesquisados. A questão está relacionada com a divergência para o âmbito real, onde independente da loja que o cliente vá, a informação pertence a cada produto é a mesma. Dessa forma, surge uma deficiência em áreas que necessitam de padrões, como coletas automáticas, já que não seria possível discernir de forma direta qual a representação de cada informação.

No decorrer do projeto todos o objetivos foram alcançados de forma satisfatória. A extração de apenas dez snippets pareceu eficiente, porém existem algumas informações, por exemplo preço, que são confundidas com outras similares (para a coluna preço, o melhor resultado foi compra).

REFERÊNCIAS BIBLIOGRÁFICAS

[1] Bando, L. L.; Scholer, F. & Turpin, A. (2010). Constructing query-biased summaries: a comparison of human and system generated snippets. In Proceedings of the third symposium on Information interaction in context, IIiX '10, pp. 195--204, New York, NY, USA. ACM.

[2] Haas, K.; Mika, P.; Tarjan, P. & Blanco, R. (2011). Enhanced results for web search. In Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval, SIGIR '11, pp. 725--734, New York, NY, USA. ACM.

[3] Altigran S. Da Silva, Denilson Barbosa, João M. B. Cavalcanti, and Marco A. S. Sevalho. Labeling data extracted from the web. Em Proceedings of the 2007 OTM Confederated international conference on On the move to meaningful internet systems. OTM'07, pp. 1099-1116. Springer-Verlag, Berlin, Heidelberg,.

[4] Varlamis, I. & Stamou, S. (2009). Semantically driven snippet selection for supporting focused web searchs. Data Knowl. Eng., 68:261-277.

[5] Metzler, D. & Kanungo, T. (2008). Machine learned sentence selection strategies for query-biased summarization. Sigir Learning to Rank Workshop.