GERAÇÃO DE *SNIPPETS* ESTRUTURADOS PARA MÁQUINAS DE BUSCA DA WEB

KLEVERSON SANTANA DA PAIXÃO

GERAÇÃO DE *SNIPPETS* ESTRUTURADOS PARA MÁQUINAS DE BUSCA DA WEB

Proposta de dissertação apresentada ao Programa de Pós-Graduação em Informática do Instituto de Ciências Exatas da Universidade Federal do Amazonas como requisito parcial para a obtenção do grau de Mestre em Informática.

Orientador: David Braga Fernandes de Oliveira

Manaus

Janeiro de 2013

Resumo

Snippets tem papel fundamental na tarefa de auxiliar os usuários a julgar quais documentos *Web* são relevantes a consulta submetida à máquina de busca. Gerar *snippets* de qualidade é importante, eles influenciam na percepção do usuário da relevância do documento. Neste contexto, *snippets* considerados estruturados, por conterem além de fragmentos de textos extraídos do documento *Web*, mas também *metadados* associado ao conteúdo do documento, são mais eficazes. Contudo a geração automática desses resumos é difícil, dependem do domínio dos dados e da estrutura do documento /textitWeb. Apresentamos um método para geração automática de *snippets* estruturados pelo uso do conceito de blocos de índice.

Abstract

Snippets play a fundamental role in the task to assist the users to judge which Web documents are relevant to the query submitted to the search machine. The generation of quality snippets is important, they influence the perceiving of the user of the documents relevance. In this context, snippets called structured, by contain besides text fragments extracted from Web document, but metadatas associated to the content of document are more effective. However the automatic generation of those type of abstracts are difficult, it depends of data domain. We purpose an automatic generation of structured snippets with the concept of index block.

Lista de Figuras

1.1 Exemplo de um resultado de busca. 1

3.1 Exemplo de uma resposta para uma consulta submetida a um site. A região regular de registro é o bloco de índice. 10

Lista de Tabelas

Sumário

Resumo **v**

Abstract **vii**

Lista de Figuras **ix**

Lista de Tabelas **xi**

1 Introdução **1**

1.1 Organização do Trabalho 3

2 Trabalhos Relacionados **5**

3 Geração de *Snippets* Estruturados a partir de Blocos de Índice **9**

3.1 Geração de *Snippets* 9

3.2 Blocos de Índice 10

3.3 Arcabouço para Geração de Snippets Estruturados 11

3.3.1 Identificação dos Blocos de Índice e Extração dos Dados Estruturados 11

3.3.2 Rotulação dos Atributos 12

3.3.3 Composição do *Snippet* 12

3.4 Metodologia de Avaliação 13

Referências Bibliográficas **15**

Capítulo 1

Introdução

Máquinas de busca para Web são as mais populares ferramentas para encontrar informação útil sobre algum assunto de interesse. A justificativa para essa popularidade é a forma simples e natural que as pessoas interagem com elas. O usuário submete uma consulta usando palavras-chaves e recebe como resposta uma lista de URLs que apontam para páginas similares à consulta formulada, com a expectativa de tais documentos sejam relevantes para as necessidades de informação do usuário [Varlamis & Stamou, 2009].

Para auxiliar o usuário a decidir quais documentos satisfazem suas necessidades de busca é comum apresentar um resumo para cada documento presente na lista de resposta. O objetivo do resumo é ajudar os usuários uma rápida avaliação se o documento Web retornado é relevante a consulta realizada ou não [Haas et al., 2011] (veja figura 1.1). É necessário então apresentar resumos de qualidade, uma vez que eles influenciam 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 [Metzler & Kanungo, 2008].

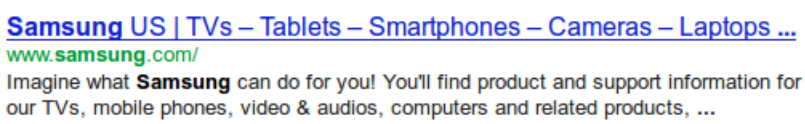

Figura 1.1. Exemplo de um resultado de busca.

Esses resumos são mais conhecidos como *snippets*, e são geralmente formados pela concatenação de sentenças extraídas do documento. Extensivas pesquisas têm focado em métodos de extração de sentenças dos documentos Web para geração de

resumos legíveis de modo eficiente. Um estudo realizado por Bando et al. [2010] sugere o desenvolvimento de novos métodos para geração de *snippets*, dado que os abordagens atuais não são satisfatórias.

Uma justificativa para a baixa qualidade dos métodos pode estar no tempo de disponível para o processamento dos resumos, que impede o uso de técnicas de processamento de linguagem natural, e nos dados providos. No geral, os documentos são considerados como não estruturados, mesmo que possuam dados ou *metadados*, estes são ignorados, e processados como *bag-of-word*, relações entre palavras e sentenças são ignoradas.

Haas et al. [2011] foi o primeiro trabalho a abordar a geração de *snippets* para documentos "estruturados"ou que possuam alguma informação estruturada. Seus resultados apontam que *snippets* formados também com dados estruturados e *metadados* têm a preferência de acesso dos usuários, e dão aos usuários um melhor entendimento sobre as páginas para os quais apontam. Mais recentemente, Zhang et al. [2012] trabalhou com o problema da seleção de dados que devem compor o resumo.

Contudo a geração estruturada de *snippets* é uma tarefa difícil, e depende do domínio dos dados do documento Web a ser resumido. Por exemplo, *sites* de livros, músicas e filmes têm diferentes entidades e estruturas associadas, dificultando o uso de métodos absolutamente automáticos.

Neste trabalho abordamos o problema da geração automática de *snippets* estruturados, desde a extração dos dados a composição do resumo. O arcabouço proposto usa o sistema de busca de informações do *site* que contém o documento Web e utiliza esses dados para geração estruturada de resumos (veja a figura (a ser inserido)).

Os mecanismos de busca para informações contidas no domínio de diversos *site* são compostos por menu de navegação, onde o usuário navega entre categorias, ou ainda com uso de palavras-chaves. Independente do método de busca utilizado, esses mecanismos retornam uma lista contendo resumos de páginas internas do *site*, semelhante a página de resposta de uma máquina de busca. Os resumos são compostos com alguns dados extraídos de um banco de dados e compõe um *snippet* estruturado para o documento Web por ela apontado. Neste trabalhos nos referimos a essa lista de documentos formada por resumos como **bloco de índice**, e é de fundamental importância para o arcabouço proposto.

A principal diferença entre os *snippets* presentes nos blocos de índice de um *site* e a página de resposta de uma máquina de busca é a seleção dos dados que o compõe. Em máquinas de busca, como citado anteriormente, métodos automáticos são utilizados para a geração de *snippets* dos documentos. Em *Web sites*, apesar das páginas de respostas serem criadas de maneira dinâmica, os dados que compõe seus

snippets foram previamente selecionados. Isso não implica dizer que em cada solicitação de composição de uma página de resposta uma pessoa ira cria-la em tempo real, mas que selecionou o grupo de dados que deve estar presente. Por exemplo, um artista de música pode ter as seguintes entidades relacionadas: data de nascimento, número de álbuns produzidos, músicas compostas, shows realizados, prêmios ganhos, entre outros. Um pessoa selecionaria quais destas entidades devem fazer parte do *snippet* no bloco de índice quando o documento referente ao artista for retornado como resposta.

A maneira manual como o grupo de dados são selecionados para geração de *snippets* em blocos de índice é o interesse deste trabalho. A intuição parte do principio que os dados presentes nesse *snippet* são os que melhor descrevem o conteúdo do documento *Web* apontado, dado o *feedback* humano. Então, este trabalho propõe a criação de um método automática para construção de *snippets* estruturados a partir dos resumos contidos nas blocos de índice de um *site*.

1.1 Organização do Trabalho

O restante deste trabalho está organizado como segue-se. No capítulo 2 apresenta trabalhos relacionados a geração de *snippets*, tais como, composição, legibilidade e eficiência.

O capítulo 3 oferece detalhes sobre métodos existentes para geração de *snippets* estruturados e define formalmente o problema abordado neste trabalho. Além disso, explora a dificuldade da criação de métodos genéricos para geração de resumos estruturado, apontando limitações e comparando com os métodos atuais.

Capítulo 2

Trabalhos Relacionados

As primeiras abordagens para geração de *snippets* eram independentes das consultas submetidas pelos usuários às maquinas de busca. As sentenças que compunham os resumos eram extraídas a partir de meta-dados, caracteres iniciais do documento, ou alguma abordagem baseada em extração de sentenças, tal como o proposto por Luhn [1958]. Nestas abordagens, os resumos são previamente gerados para cada documento e armazenados para serem exibidos quando necessário.

Tombros & Sanderson [1998] mostraram que a geração de *snippets* independentes da consulta tem pouco impacto na tarefa de auxiliar o usuário a julgar quais documentos são os mais prováveis de conter as informações desejadas, e introduz o conceito de *snippets* direcionados a consulta (do inglês, *query-biased snippets*). Nesta abordagem, as sentenças escolhidas para compor o resumo devem possuir alguma ligação para com a consulta. Desta forma, evidências tais como a localização da ocorrência dos termos da consulta no documento são exploradas. Os autores observaram que esta abordagem permite o usuário julgar a relevância dos documentos com mais precisão e em menor tempo, sem que seja necessário, na maior parte do tempo, acessar todo conteúdo do documento. O Google [1] provavelmente foi a primeira máquina de busca comercial a prover resumos para os documentos utilizando uma abordagem *query-biased*[Turpin et al., 2007].

Sob um aspecto mais geral, a geração de resumos para documentos pode ser dividido em três grandes processos: seleção, condensação e transformação. O primeiro processo, seleção, concerne em selecionar quais informações devem ser incluídas ou excluídas do resumo. O segundo, condensação, está relacionado a substituição de parte do material fonte por ideias de mais alto nível ou conceitos mais específicos. Por fim, transformação, conduz a integração e combinação das ideias para geração do resumo

[1] www.google.com

final [Bando et al., 2010].

Quando estes processos são analisados sob o contexto de resumos para sistemas de recuperação de informação, que precisam ser gerados em microssegundos para exibição dos resultados em tempo real, observa-se um grande custo computacional agregado. Por esta razão, em máquinas de busca para a Web, costuma-se limitar o processo de geração de *snippets* apenas à fase de seleção das sentenças [Bando et al., 2010]. Embora questões como legibilidade, singularidade e compressão das sentenças devem ser consideradas, em um processo conhecido como condensação [Metzler & Kanungo, 2008].

Diversas heurísticas têm sido propostas para geração de *snippets* de maior qualidade. Dentre as abordagens utilizadas, temos o uso de modelos baseados em *language model* [Li & Chen, 2010], modelos baseados em características semânticas para aumentar coerência entre as sentenças selecionadas [Varlamis & Stamou, 2009], e uso de cadeias léxicas [Manabu & Hajime, 2000].

Wang et al. [2007] foi o primeiro trabalho a utilizar aprendizagem de máquina para geração de *snippets*, com os uso dos modelos SVM e *ranking* SVM para mensurar a similaridade entre as sentenças do documento e a consulta processada. Metzler & Kanungo [2008] estendeu o trabalho anterior comparando diversos métodos de aprendizagem em máquina baseados em *ranking* e com inclusão de modelos para seleção do número de sentenças que devem compor o resumo.

Xu et al. [2009] considera que a maioria dos métodos para geração de resumos de documentos *web* não considera realmente o fator humano e utiliza técnicas de *eye-tracking*, juntamente de mineração de dados, para aprender o comportamento de leitura de pessoas e então selecionar sentenças que melhor expressem suas necessidades de informação.

Em um estudo recente, Bando et al. [2010] investigou como humanos constroem *snippets*. Neste estudo foi solicitado que os participante construíssem *snippets* em linguagem natural e *snippets* por extração de texto. Realizando um mapeamento entre quais sentenças foram utilizadas para geração dos *snippets* em linguagem natural, foi observado que em 73% do tempo estas mesmas sentenças eram utilizadas para construção das versões por extração. Em comparação, notou-se que as abordagens automáticas usam estas mesmas sentenças em apenas 22% do tempo. Foi observado também que as métricas de avaliação utilizadas na literatura não refletem a real qualidade do *snippet* gerado, e quando utilizados para comparar as versões geradas por humanos e computadores, as versões automáticas se tornavam bastante competitivas.

Haas et al. [2011] apresenta o conceito de resumos "enriquecidos", que são *snippets* que incorporam informações multimídia, dados estruturados e *metadados*. Esta abor-

dagem fornece elementos que dão aos usuários um melhor entendimento sobre as páginas para os quais apontam. Além disso, os *snippets* gerados a partir dessa técnica tendem a ser mais clicados pelos usuários dos que os *snippets* convencionais.

Outros trabalhos na área atacaram o problema do custo computacional associado à geração de *snippets*. Turpin et al. [2007] provavelmente foi o primeiro a abordar este problema. Os autores apresentaram diversos estudos sobre quais fatores mais influenciam a performance da tarefa de construção de *snippets* e apresentaram a técnica *compressed token system* (CTS), que utiliza compressão e *caching* de documentos em memória para aumentar o desempenho da tarefa de resumo.

Tsegay et al. [2009] apresenta métodos de poda para documentos com o objetivo de aumentar o número de documentos em *cache*, consequentemente aumentando o desempenho da geração de *snippets*. Este trabalho mostra que mesmo quando um documento é reduzido mais que a metade do seu tamanho original, a maior parte dos *snippets* gerados permanece igual.

Ceccarelli et al. [2011] explora o uso de *cache* no processo de geração de *snippets* e apresenta o conceito de *supersnippet*, conjunto de sentenças que é mais provável de compor *snippets* para outros documentos.

Capítulo 3

Geração de *Snippets* Estruturados a partir de Blocos de Índice

3.1 Geração de *Snippets*

Todos os métodos de geração de *snippets* para documentos Web não-estruturados seguem duas etapas: seleção e composição [Metzler & Kanungo, 2008]. Seleciona-se as sentenças que tem maior similaridade para com a consulta (seleção) e as concatena de forma que o texto seja legível (composição). Entretanto, a maioria dos métodos focam apenas na fase de seleção e espera que a simples concatenação das sentenças seja legível.

A metodologia acima descrita também é utilizada, ou pelo menos adaptada a documentos estruturados. Haas et al. [2011] procura informações de Web semântica para identificar dados estruturados na página, mas também fornece uma interface para que o proprietário do *site* decida qual a melhor forma de se apresentar os *snippets*, selecionando *template* e dados irá compor o resumo. Poucos domínios possuem métodos automáticos de extração de dados. Zhang et al. [2012] não trata do problema de extração de dados, há a suposição de que esses dados já foram extraídos e rotulados. Em ambos os trabalhos, incluindo Haas et al. [2011], o problema de extração não é um fator importante, há formas de se fornecer os *metadados* e dados estruturados de maneira não automática, restando apenas a fase de apresentação ao usuário.

O arcabouço proposto é um modelo de geração de *snippets* estruturados, abordando desde a fase de extração dos dados estruturados e *metadados* do documento Web a apresentação do resumo. O uso de bloco de índice ajuda a reduzir a dificuldade do problema de extração, apesar da abordagem *open* Web.

3.2 Blocos de Índice

Como explicado na introdução, os mecanismos de busca de informação de *sites* retornam uma lista de resumos para documentos Web, onde chamamos essa lista de bloco de índice. Os blocos de índice formam uma região regular na página Web e os resumos nele contidos são estruturados (veja a figura 3.1).

Figura 3.1. Exemplo de uma resposta para uma consulta submetida a um site. A região regular de registro é o bloco de índice.

Seja um site S formado por um conjunto de documentos Web $(D_1, D_2, ..., D_{n_1})$ para algum n_1 não nulo. Generalizamos o conceito de bloco de índice, B, como sendo qualquer região regular, $R \subset D_{i_1}$, que contenha objetos ricos em dados, d_j, i.e., $R = (d_1, d_2, ..., d_{n_2})$, para $n_2 \geq 2$, e possua as características:

- Cada dado d_i é uma tupla formada de dois ou mais elementos (atributo dos objetos ricos em dado). Logo, $d_i =< a_1, ..., a_m >$ para $m \geq 2$. Sendo $d_i[j]$ é j-ésimo atributo de d_i.

- Existe um $d_i[j]$ com a função de apontador para outro documento Web, D_{i_2}, pertencente a S, onde $D_{i_1} \not\subseteq D_{i_2}$ e $d_i \subset D_{i_2}$.

Os blocos de índice podem ser amplamente utilizados para geração de resumos pelas máquinas de busca. Se um documento retornado como resposta possuir como *backlink* um registro pertencente a um bloco de índice, este pode ser utilizado em substituição ou complemento ao *snippet* tradicional. Contudo, os *snippets* são tão completos quanto as informações providas no bloco de índice, mas a adição de dados como *metadescription* ou *query-biased snippet* do documento Web pode ajudar a suprir uma possível falta de informação. Esse problema também é explorado neste trabalho.

3.3 Arcabouço para Geração de Snippets Estruturados

O arcabouço proposto para utilização de blocos de índice para geração de *snippets* estruturados é composta de três fases:

- Identificação dos blocos de índice e extração dos dados estruturados.
- Rotulação dos atributos
- Composição do *snippet*.

3.3.1 Identificação dos Blocos de Índice e Extração dos Dados Estruturados

O processo de identificação de blocos de índice consiste em encontrar as regiões regulares do documento *Web* e verificar quais destas regiões possuem as características citadas na seção 3.2. O problema principal consiste em localizar as regiões regulares.

A abordagem utilizada é similar aos processo de extração de dados que utilizam alinhamento de árvore. Sejam duas árvores $T1$ e $T2$, o objetivo é estabelecer uma correspondência entre sub-árvores e nodos de $T1$ e $T2$ que são equivalentes. Por equivalência entendemos nodos que representam os mesmo atributos e tipos de entidades.

O método para realizar a equivalência entre as subárvores é uma variação da ideia proposta em Fernandes et al. [2011]. Nesse trabalho é apresentado uma estrutura hierárquica chamada de árvore SOM (acrônimo para *Site Objetc Model*) que resume as árvores DOM (acrônimo para *Domain Object Model*) de todas as páginas de um Web *site*. Ela agrupa nodos similares, formando um tipo de *cluster*, e permite a identificação de regularidades entre os documentos *Web* do site e a segmentação do deles.

Utilizando o mesmo princípio da SOM, mas ao invés de agrupar as DOM de um Web *site*, agrupa-se as subárvores presentes em uma única DOM de um documento Web e pode-se segmentar o documento em regiões regularidades, em consequência, identificar os blocos de índice.

Identificado a região regular correspondente a um bloco de índice é iniciado o processo de extração. Diversos métodos na literatura abordam o problema de se extrair dados de um documento Web, como Reis et al. [2004] e Zhai & Liu [2005]. Não é foco deste trabalho propor um novo método para extração de dados, será utilizado métodos já existente na literatura.

3.3.2 Rotulação dos Atributos

Como resultado do processo de extração obteremos um *dataset* anônimo, i.e., um conjunto de dados que não possui rotulamento. Como exemplo, para os registros pertencentes a Figura 3.1 poderíamos ter o seguinte *dataset* como resultado:
$D_1 = \langle$"Lord of the Rings Part 1: The Fellowship of the Ring", "To celebrate the release of the first [...]", "$9.74"$\rangle$,
$D_2 = \langle$"The Lord of The Rings Part 2: The Two Towers", "To celebrate the release of the first [...]", "$9.74"$\rangle$,
$D_3 = \langle$"The Hobbit", "The prelude to The Lord of the Rings, [...]", "$9.74"$\rangle$

Para que os dados acima possam ser utilizados na construção de *snippet*, eles precisam serem adicionados a um *template* para que então possam compor a página de resposta da máquina de busca. O *template* a ser utilizado está diretamente relacionado ao domínio que pertencem os dados extraídos. Diferentes domínios possuem diferentes atributos, e essa heterogeneidade precisa ser considerada na geração de *snippets*, havendo necessidade da rotulação do *dataset* criado.

O problema de rotulamento pode ser definido como encontrar rótulos descritivos para um conjunto de dados relacionais anônimos. Diversos trabalhos na literatura abordaram esse problema. Da Silva et al. [2007] propôs um método totalmente automático e independente para as tarefas de seleção e atribuição de rótulos. O método utiliza a máquina de busca *Web* para encontrar o melhor rotulo para os atributos presentes em cada registro do *dataset*. Dado a variedade de domínios que podem estar contidos nos *dataset* gerados, este é a provavelmente a melhor opção para resolver o problema.

3.3.3 Composição do *Snippet*

Supondo a correta execução das etapas anteriores é possível então a composição do *snippet* estruturado para apresentação dos resultados da máquina de busca. Dado um *dataset* rotulado e um *template*, a tarefa deveria se resumir apenas em preencher o *template*. Entretanto, para o correto posicionamento das informações, o domínio dos dados deve ser considerado para a geração *template* a ser utilizado. A característica *open* Web do método, impede a princípio o uso de *template* genérico, a geração de um *template* de forma automática não é foco do trabalho.

A quantidade de dados a ser apresentado no *snippet* também é um fator a ser considerado, o resumo é tão completo quanto os dados providos no bloco de índice. Este fato pode tornar os *snippets* pouco informativos e consideramos esta hipótese no processo de composição.

Consideramos três tipos de *snippets* estruturados:

- Apenas dados extraídos dos blocos de índice.
- Dados extraídos dos blocos de índice e *metadescription* do documento Web.
- Dados extraídos dos blocos de índice e *query-biased snippet* do documento Web.

Para fins de comparação e avaliação de qualidade também são considerados *snippets* formados por:

- Apenas dados extraídos dos blocos de índice sem a utilização de *template*.
- Apenas o *metadescription* do documento Web.
- Apenas o *query-biased snippet* do documento Web.

3.4 Metodologia de Avaliação

Para avaliar o arcabouço apresentado é necessário verificar três itens:

Qualidade dos *snippets* gerados: avaliação do usuário.

Para fazer a avaliação de qualidade dos *snippets* estruturados será utilizada a estratégia proposta por Zhang et al. [2012]. Utilizando uma plataforma de *crowdsourcing*, o experimento de avaliação segue o modelo de um jogo. Os usuários são convidados a julgar a relevância de cada documento baseado no seu *snippet* O ganho do usuário é proporcional a qualidade da avaliação realizada. Julgamentos aleatórios conduzirão a um lucro baixo, enquanto um julgamento cuidadoso irá aumentar os lucros. Zhang et al. [2012] mostrou que esta forma de condução de experimentos é bastante efetiva.

Os dois primeiros itens estão relacionados sobre a viabilidade de utilização do método. É necessário que ele cubra a maior variedade possível de respostas de busca para justificar sua implementação. Sendo um processo *offline*, o tempo entre a coleta das páginas e construção dos *snippets* precisa ser razoável para ter uso prático.

Abrangência do método proposto: presença de páginas de índice nos Web *sites*.

Verificar o quão comum é a presença de documentos Web referenciados por um bloco de índice. A ideia é descobrir a aplicabilidade do arcabouço em máquinas de busca e alternativa aos *snippets* tradicionais quando aplicável. Uma vez que blocos de índice são o *template* natural de uma resposta submetida a um sistema de busca, e como a maioria dos *sites* atuais oferecem esse recurso, a intuição é uma alta aplicabilidade do arcabouço.

Custo computacional do processo: custo total desde a identificação dos blocos de índice a construção dos *snippets*.

Query-biased snippets possuem elevado custo computacional para uma máquina de busca. Os resumos são gerados sob demanda, as sentenças do documento Web são extraídas de acordo com a consulta submetida e possuem restrição de tempo, o resultado tem que ser apresentado o mais rápido possível.

O arcabouço proposto utiliza uma abordagem independente da consulta, e portanto o custo de geração do *snippet* se restringe em ler as informações do *dataset*. Apesar da abordagem "independente da consulta"não ser considerada uma boa alternativa, espera-se que os resultados dos experimentos de qualidade torne o arcabouço uma exceção a regra.

Outros custos computacionais a serem considerados para a aplicação do arcabouço são os de identificação, extração e rotulamento. É necessário que todo o sistema seja escalável para ter aplicação prática.

Referências Bibliográficas

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

Ceccarelli, D.; Lucchese, C.; Orlando, S.; Perego, R. & Silvestri, F. (2011). Caching query-biased snippets for efficient retrieval. Em *Proceedings of the 14th International Conference on Extending Database Technology*, EDBT/ICDT '11, pp. 93--104, New York, NY, USA. ACM.

Da Silva, A. S.; Barbosa, D.; Cavalcanti, J. a. M. B. & Sevalho, M. A. S. (2007). Labeling data extracted from the web. Em *Proceedings of the 2007 OTM Confederated international conference on On the move to meaningful internet systems: CoopIS, DOA, ODBASE, GADA, and IS - Volume Part I*, pp. 1099--1116.

Fernandes, D.; de Moura, E. S.; da Silva, A. S.; Ribeiro-Neto, B. & Braga, E. (2011). A site oriented method for segmenting web pages. Em *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, SIGIR '11, pp. 215--224, New York, NY, USA. ACM.

Haas, K.; Mika, P.; Tarjan, P. & Blanco, R. (2011). Enhanced results for web search. Em *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.

Li, Q. & Chen, Y. P. (2010). Personalized text snippet extraction using statistical language models. *Pattern Recogn.*, 43:378--386.

Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM J. Res. Dev.*, 2:159--165.

Manabu, O. & Hajime, M. (2000). Query-biased summarization based on lexical chaining. *Computational Intelligence*, 16(4):578--585.

Metzler, D. & Kanungo, T. (2008). Machine learned sentence selection strategies for query-biased summarization. Em *SIGIR Learning to Rank Workshop*.

Reis, D. C.; Golgher, P. B.; Silva, A. S. & Laender, A. F. (2004). Automatic web news extraction using tree edit distance. Em *Proceedings of the 13th international conference on World Wide Web*, WWW '04, pp. 502--511, New York, NY, USA. ACM.

Tombros, A. & Sanderson, M. (1998). Advantages of query biased summaries in information retrieval. Em *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '98, pp. 2--10, New York, NY, USA. ACM.

Tsegay, Y.; Puglisi, S. J.; Turpin, A. & Zobel, J. (2009). Document compaction for efficient query biased snippet generation. Em *Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval*, ECIR '09, pp. 509--520, Berlin, Heidelberg. Springer-Verlag.

Turpin, A.; Tsegay, Y.; Hawking, D. & Williams, H. E. (2007). Fast generation of result snippets in web search. Em *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '07, pp. 127--134, New York, NY, USA. ACM.

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

Wang, C.; Jing, F.; Zhang, L. & Zhang, H.-J. (2007). Learning query-biased web page summarization. Em *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, CIKM '07, pp. 555--562, New York, NY, USA. ACM.

Xu, S.; Jiang, H. & Lau, F. C. (2009). User-oriented document summarization through vision-based eye-tracking. Em *Proceedings of the 14th international conference on Intelligent user interfaces*, IUI '09, pp. 7--16, New York, NY, USA. ACM.

Zhai, Y. & Liu, B. (2005). Web data extraction based on partial tree alignment. Em *Proceedings of the 14th international conference on World Wide Web*, WWW '05, pp. 76--85.

Zhang, L.; Zhang, Y. & Chen, Y. (2012). Summarizing highly structured documents for effective search interaction. Em *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '12, pp. 145--154.