Desenvolvimento de Sistemas de Extração de Informações para Ambientes Colaborativos na Web

Douglas Nogueira[1], Vladia Pinheiro[2], Vasco Furtado[1], Tarcisio Pequeno[1]

[1]Mestrado em Informática Aplicada – Universidade de Fortaleza (UNIFOR)
Av. Washington Soares, 1321, Fortaleza, Ceará, Brasil

[2]Departamento de Ciências da Computação – Universidade Federal do Ceará
Campus do Pici-UFC, Fortaleza, Ceará, Brasil.

dougpetcomp@yahoo.com.br, vladia@lia.ufc.br,
(vasco,tarcisio}@unifor.br

Abstract. *This paper describes an architecture for Information Extraction systems, based on Natural Language Processing, for use in web collaborative systems. A web tool to extract information about crimes to the collaborative system WikiCrimes was developed using the architecture. The results of the evaluation of performance, practical usefulness and usability of the tool are presented and discussed.*

Resumo. *Este artigo descreve uma arquitetura de sistemas de Extração de Informação baseados em Processamento de Linguagem Natural para uso em sistemas colaborativos na Web. Uma ferramenta para extração de informação sobre crimes para o sistema colaborativo WikiCrimes foi desenvolvida usando a arquitetura. Os resultados da avaliação dos aspectos de desempenho, utilidade prática e usabilidade da ferramenta são apresentados e discutidos.*

1. Introdução

Cada vez mais cresce o número de sistemas colaborativos na web. Tais sistemas dependem da iniciativa dos usuários da web para geração do conteúdo e de uma inteligência coletiva. De outro lado, a web é uma fonte rica de informações sobre qualquer domínio, seu conteúdo é vasto e, em sua maioria, está na forma não estruturada e em linguagem natural. Se pudéssemos utilizar tal conteúdo para alimentar sistemas colaborativos seria de fundamental importância. Portanto, existe a necessidade crescente de ferramentas que auxiliem a captura rápida, de forma simples, semi-automática e interativa de informações para registro em sistemas colaborativos.

Sistemas de Extração de Informação (EI) visam localizar e extrair, de forma automática, informações relevantes em um documento ou coleção de documentos, contendo textos em língua natural, e estruturar tais informações para os padrões de saída, a fim de facilitar sua manipulação e análise (Grishman, 1997). O Ubiquity[1], por

[1] Em http://labs.mozilla.com/2008/08/introducing-ubiquity/ está disponível uma introdução, exemplos e tutorial sobre o plug-in Ubiquity do navegador Mozilla Firefox.

exemplo, é uma iniciativa na direção de fornecer ao usuário web uma linguagem para realização de *mashups*, que são aplicações web que usam conteúdo de mais de uma fonte para criar um novo serviço completo.

Neste artigo propomos uma arquitetura de sistemas de EI para uso em sistemas colaborativos na Web e apresentamos uma ferramenta – WikiCrimes *Information Extractor (*WikiCrimesIE*)*, desenvolvida segundo a arquitetura proposta. A arquitetura prevê o uso de ferramentas de programação orientada ao usuário e de PLN, que possibilitam melhoria na interação e manipulação de conteúdo em linguagem natural, disponível na Web. Ao final, a avaliação do sistema WikiCrimesIE, trabalhos relacionados e os trabalhos futuros são discutidos.

2. Arquitetura de Sistemas de EI para Ambientes Colaborativos

A arquitetura para sistemas de EI, proposta neste artigo, é apresentada na Figura 1 e é composta por cinco módulos principais: Página Web, Ferramenta de Programação Orientada ao Usuário, Parser Sintático, Analisador Semântico e Sistema Colaborativo.

- A ferramenta de programação orientada ao usuário permite o desenvolvimento de comandos para captura de informações contidas em páginas web, facilitando sua manipulação. Este módulo é responsável por ler textos em páginas Web e enviá-lo ao módulo seguinte (*parser* sintático) junto com objetivos de extração de informação.
- O *parser* sintático recebe o texto selecionado de uma página web e realiza a análise sintática do texto e uma resolução restrita de seus referentes.
- O analisador semântico recebe a árvore sintática gerada no módulo anterior e é responsável por inferir o significado das sentenças do texto da página Web e responder aos objetivos enviados pelo usuário.

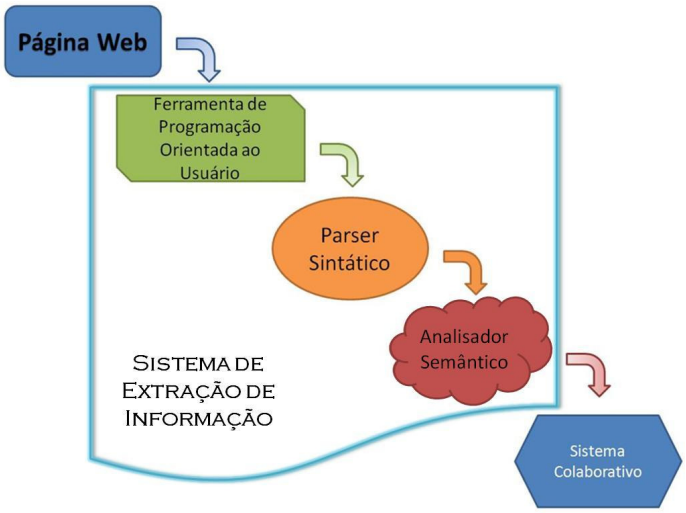

Figura 1. Arquitetura de Sistemas de EI para Ambientes Colaborativos.

O caráter de inovação na arquitetura proposta reside em prover, a usuários web, facilidades de interação e extração de informações de textos, em linguagem natural, que

contenham informações de interesse para sistemas colaborativos em que sejam participantes.

3. *WikiCrimes Information Extractor*

Sistemas colaborativos permitem a comunicação de idéias, compartilhamento de recursos e coordenação dos esforços de trabalho. O sistema WikiCrimes, por exemplo, permite a seus usuários acessarem e realizarem registros de ocorrências criminais via web. Uma das necessidades do projeto WikiCrimes é fornecer aos usuários uma ferramenta que facilite o registro de crimes a partir de notícias sobre crimes veiculadas na web.

Para atender a esta necessidade, desenvolvemos a ferramenta WikiCrimesIE (WikiCrimes *Information Extractor*) para extração de informações sobre crimes e registro no banco de dados do sistema, instanciando os módulos da arquitetura proposta:

- Ferramenta de programação orientada ao usuário: foi desenvolvido um comando na ferramenta Ubiquity - *mapcrimes*, para enviar ao framework de PLN o texto selecionado e os objetivos de extração (quais informações são necessárias para o WikiCrimes). A ferramenta Ubiquity é um plug-in do Mozilla Firefox que disponibiliza uma coleção de comandos derivados de uma linguagem rápida, fácil e natural, que permite aos usuários o acesso à informação e a modificação dela.
- Parser Sintático: foi utilizado o parser morfossintático PALAVRAS (Bick, 2000) que realiza a análise morfológica e sintática de sentenças em língua portuguesa.
- Analisador Semântico: foi utilizado o *Semantic Inferentialism Analyser* (SIA), componente principal do *framework* SIM (Semantic Inferentialism Model) (Pinheiro et AL, 2008).

A justificativa para uso do framework SIM para o módulo de análise semântica é que muitas vezes as informações sobre crimes (tipo do crime, arma utilizada, causas/motivos do crime, etc) estão implícitas no texto jornalístico e algumas inferências mais complexas, oriundas dos usos dos conceitos na prática lingüística, precisam ser realizadas.

A figura 2 apresenta o resultado da extração de informação do seguinte texto "*Mais um crime com características de execução sumária foi registrado em Fortaleza. Na noite de terça-feira, o jovem Marcelo dos Santos Vasconcelos, 29, foi fuzilado na porta de casa. O crime ocorreu na Rua Casimiro de Abreu, em Parangaba*". A interface da ferramenta WikiCrimesIE apresenta o texto recebido pelo comando *mapcrimes* do Ubiquity e as informações extraídas pelo SIA. No exemplo apresentado, o local do crime "Rua Casimiro de Abreu, Parangaba" (*A*) foi extraído da sentença *"O crime ocorreu na Rua Casimiro de Abreu, em Parangaba"*, e foi localizado no mapa digital. O tipo do crime "homicídio" (*B*) foi extraído da sentença "*o jovem Marcelo dos Santos Vasconcelos, 29, foi fuzilado na porta de casa"*.

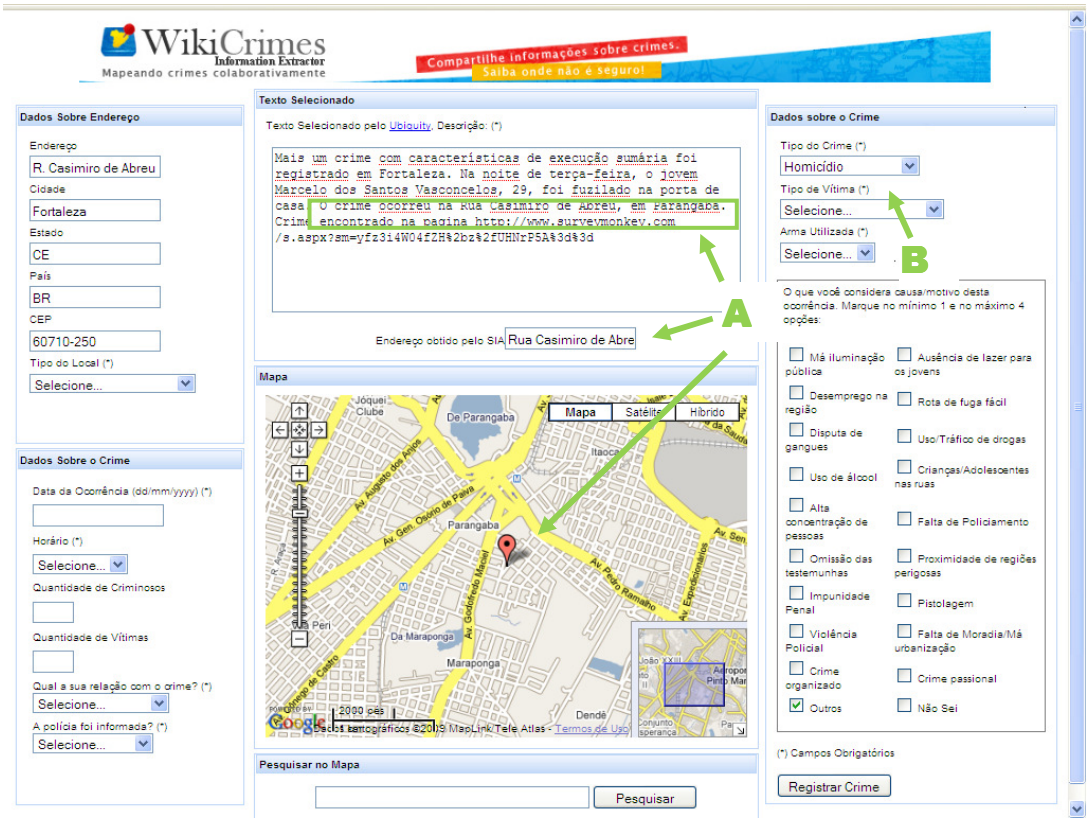

Figura 2. Interface do sistema WikiCrimesIE onde foi extraída informação sobre o local do crime e tipo do crime do texto selecionado pelo usuário. O endereço foi localizado no mapa geoprocessado.

3.1. Avaliação e Análise dos Resultados

A tabela 1 apresenta os resultados do SIA quanto a sua precisão em extrair o local e tipo do crime, de uma centena de textos da web descritivos de crimes. Os resultados para o atributo "tipo do crime" são ainda mais motivadores.

Tabela 1. Precisão do SIA na extração do "Local do Crime" e "Tipo do Crime"

	Local do crime	Tipo do crime
Precision	52%	48%

Em outro experimento realizado com 12 usuários do sítio WikiCrimes, eles foram solicitados a ler 20 textos simples descritivos de crime, identificar o local do crime e registrá-lo no WikiCrimes diretamente (sem o uso do WikiCrimesIE). Foi cronometrado o tempo que cada usuário levava para realizar esta tarefa e este foi comparado com o tempo que a WikiCrimesIE levava para extrair as informações (local do crime e tipo do crime): em média, a solução WikiCrimesIE/SIA foi 49,84% mais rápida. Em todos os casos, o tempo médio da extração com o uso da ferramenta foi bem menor do que sem o uso da ferramenta.

Além disso, foi aplicado questionário para os usuários e realizada uma análise qualitativa, que abordou aspectos de usabilidade, desempenho e reuso. Todas as respostas do questionário corroboraram nossas hipóteses quanto à utilidade prática, desempenho e usabilidade de ferramentas como o WikiCrimesIE para sistemas colaborativos.

4. Trabalhos Relacionados

Em geral, métodos de extração de informação utilizam técnicas estatísticas de aprendizado de máquina, por exemplo, Hidden Markov Model (HMM), estão sendo aplicadas para criação de regras de acordo com o tipo de texto analisado, visando minimizar a participação humana (Glickman & Jones, 1999). Sistemas baseados em HMM como o DATAMOLD (Borkar, Deshmukh e Sarawagi, 2001) e AUTOBIB (Geng e Yang, 2003), são sistemas determinísticos de aprendizagem de regras que extraem informações de textos não estruturados e criam um registro estruturado. A precisão alcançada na extração de informações é geralmente alta. Contudo, sistemas que usam HMM costumam consumir muito tempo de processamento, o que é inviável para ambientes colaborativos e interativos na web. Enfim, existem diversas abordagens e sistemas para EI, porém, em sua maioria, dedicados a extração de informações em lote e sem preocupação com a interação e usabilidade, aspectos críticos em ambiente web.

5. Conclusão

Neste trabalho apresentamos um framework para desenvolvimento de ferramentas de extração de informação para ambientes colaborativos na web. A ferramenta para extração de informações sobre crimes – WikiCrimesIE – foi desenvolvida usando a arquitetura proposta e foram avaliados os aspectos de desempenho, utilidade prática e usabilidade da ferramenta. Os resultados comprovaram as vantagens do uso da ferramenta em relação aos usuários humanos, principalmente quando eles são solicitados para ler uma quantidade de textos e a extrair uma série de informações, repetidamente. Trabalhos futuros envolvem a melhoria do analisador semântico usado.

Referências

Bick, E. (2000) 'The Parsing System "Palavras"'. Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. Aarhus University Press.

Borkar, V., Deshmukh, K., Sarawagi, S. (2001) Automatic segmentation of text into structured records. Proceedings of the 2001 ACM SIGMOD International Conference on Management of Data, California, p. 175–186.

Geng, J., Yang, J. (2003) AUTOBIB: Automatic extraction and integration of bibliographic information on the web. 29th VLDB Conference Berlin, Germany.

Glickman, O., Jones, R. (1999) Examining machine learning for adaptable end-to-end information extraction systems. AAAI 1999 Workshop on Machine Learning for Information Extraction.

Grishman, R. (1997). Information Extraction: Techniques and Challenges. SCIE'97: International Summer School on Information Extraction. Springer-Verlag, p.10–27.

Pinheiro et AL. (2008) SIM: Um Modelo Semântico-Inferencialista para Sistemas de Linguagem Natural. VI Workshop em Tecnologia da Informação e da Linguagem Humana (TIL 2008), WebMedia, Brasil.