Recomendação de dados abertos para solucionar os problemas de comunicação textual : uma análise de métodos para extração de entidades nomeadas

Bianca Munaro Lima, Maria Luiza Machado Campos

Programa de Pós-Graduação em Ciência da Computação NCE/UFRJ
C.P. 68530, CEP 21941-590, Ilha do Fundão, Rio de Janeiro – Brasil

bianca.munaro@nce.ufrj.br, mluiza@nce.ufrj.br

Abstract. Currently, the popularization of Web 2.0 caused an increase in the use of textual communication applications. To complement information exchange and collaboration, data already available on open bases in the Internet can be used to enrich discussions and even help to solve interpretation and ambiguity problems in communication. This paper introduces an approach for recommending linked open data content in textual communication tools and presents the first results of experiments on named entities recognition and extraction, mechanisms that play an essential role in recommendation strategies.

Resumo. Atualmente, a popularização da Web 2.0 provocou um aumento na utilização de aplicações de comunicação textual. Para complementar a troca de informações e a colaboração, dados já disponíveis em bases abertas na Internet podem ser usados para enriquecer discussões e mesmo auxiliar a reduzir problemas de ambiguidade e interpretação nesse tipo de comunicação. Este artigo introduz uma abordagem para a recomendação de conteúdo na forma de dados abertos interligados em ferramentas de comunicação textual, apresentado os primeiros resultados de experimentos para o reconhecimento e extração de entidades nomeadas, mecanismos considerados essenciais em estratégias de recomendação.

1. Introdução

Hoje em dia, para que tantas facilidades sejam oferecidas aos usuários, a Web sofreu grandes mudanças ao longo dos anos. Desde o período de sua criação, além da natural evolução das tecnologias associadas, seus objetivos foram se expandindo, ocorrendo mudanças na natureza de sua utilização.

A chamada Web 1.0, considerada a primeira geração da Web permitia a leitura e a criação de links entre documentos constituídos basicamente por páginas estáticas. Em 2004, a empresa americana O'Reilly Media passa a associar a expressão Web 2.0 ao conceito de uma web colaborativa, onde usuários interagem e geram conteúdo, graças ao surgimento de ferramentas intuitivas e de fácil utilização, como por exemplo, wikis, vídeos, blogs e redes sociais.

Essas informações geradas pelas comunidades associadas as ferramentas da Web 2.0 se encontram fragmentadas, além de frequentemente possuírem inconsistências. Outro desafio, associado à escalabilidade do tratamento desses dados é que, originalmente, estes ambientes interativos são voltados ao entendimento por humanos.

Com isso, há considerável dificuldade para que um sistema processe e interprete essas informações de forma automática. Pra fazer frente a esse desafio, surge a proposta da Web 3.0 e a abordagem de dados abertos interligados (LOD - Linked Open Data), que tem como objetivo organizar, de forma mais inteligente, o conteúdo presente na web, representando os dados de forma que os computadores também possam "entendê-los".

Considerando as possibilidades de evolução das aplicações que os conceitos de LOD podem trazer para a Web, propõe-se este trabalho, tendo como objetivo uma abordagem para o enriquecimento da comunicação textual online. A popularização da Web 2.0 provocou um considerável aumento na utilização dos meios de comunicação textuais. Contudo, nessas ferramentas é comum existirem ambiguidades na interpretação, ocasionando a perda de tempo com buscas, perguntas e esclarecimentos [Horiguchi et al. 2009]. Para complementar essa troca de informações, dados presentes em bases abertas podem ser recomendados, de acordo com a interação em andamento. Assim, novos conhecimentos seriam gerados, a partir da agregação de novas informações, enriquecendo as discussões, solucionando dúvidas e mesmo reduzindo ambigüidades, no próprio ambiente da ferramenta.

Nesse artigo será apresentada uma proposta de abordagem de aperfeiçoamento da comunicação via redes sociais, usando bases da Web de Dados. Em especial, serão discutidos os primeiros resultados da experimentos para identificação e extração de entidades nomeadas, termos relevantes nos conteúdos textuais.

2. Cenário exemplo e abordagem proposta

Devido à grande quantidade de dados sendo gerada na web, os usuários têm dificuldades em localizar e utilizar informações de forma eficiente. Os sistemas de recomendação podem atuar na descoberta de informações interessantes para determinado usuário de acordo com seus interesses, amigos, opiniões, entre outros. Hoje em dia, devido à popularização dos projetos envolvendo dados interligados, já existem bases de dados abertos para diversos domínios, que podem ser utilizadas no processo de recomendação.

Atualmente, existem muitas comunidades virtuais, focadas em diversos contextos, que se utilizam das redes sociais para interação e ações de seu interesse. Um exemplo é o Cirandas[1], criado para apoiar o Fórum Brasileiro de Economia Solidária, que agrega iniciativas de projetos produtivos coletivos. Essas comunidades utilizam cada vez mais as redes sociais para identificar fornecedores de matérias primas e possíveis compradores para seus produtos ou serviços. A interação entre esses grupos poderia ser ainda mais produtiva com a utilização de dados já existentes na web, que possam agregar valor a sua cadeia de atuação. Como exemplo, pode-se citar os empreendimentos de coleta e reciclagem que poderiam utilizar informações de leis, acerca de permissões de uso de áreas públicas, autorização de parcerias com o governo, politicas estaduais de coleta de lixo entre outras.

No Brasil, a Secretaria Especial de Informática (PRODASEN) desenvolveu um portal especializado em informação jurídica e legislativa, chamado LeXml[2]. Ele utiliza

[1] FBES - Fórum Brasileiro de Economia Solidária: O que é o Cirandas. Disponível em: http://cirandas.net/fbes/o-que-e-o-cirandas; acesso em: 29/04/2012

[2] LEXML: Rede de informação legislativa e jurídica. Disponível em: http://www.lexml.gov.br/; acesso em 29/04/2012

um modelo de referência e a metalinguagem XML para estruturar seus dados. Em nosso cenário exemplo, informações presentes no LeXml podem enriquecer as discussões e a interação entre comunidades, como a de reciclagem e coleta presente no Cirandas. Por outro lado, informações presentes em outras bases de dados também podem complementar a informação da legislação.

Portanto, em nossa abordagem, é possível recomendar conteúdo em ferramentas de comunicação textual utilizando informações da WoD. Com isso espera-se facilitar a interação entre usuários, permitindo que dados que complementem a discussão sejam visualizados. Para alcançar esse objetivo a abordagem contempla os seguintes passos:

- Avaliar as bases de dados existentes e selecionar as que podem ser utilizadas para recomendação e estabelecer ligações entre elas;
- Analisar os algoritmos de extração de entidades nomeadas para português;
- Desenvolver módulo de análise textual para descoberta de palavras importantes;
- Utilizar um algoritmo de recomendação para calcular a distância semântica entre as entidades para recomendar informações relacionadas;
- Apresentar as informações encontradas para o usuário;

No momento, o foco de nosso trabalho tem sido a análise de softwares voltados para a extração de entidades nomeadas e a avaliação dos resultados para o idioma Português. Os primeiros resultados obtidos são discutidos na próxima seção.

3. Análise de softwares para extração de entidades nomeadas

De acordo com os objetivos apresentados, foram analisados dois softwares para extração de entidades nomeadas para a língua portuguesa. São eles: Ltasks e AlchemyApi. Existem também os extratores de entidades mencionadas, que levam em consideração o contexto no qual a entidade está inserida. Portanto, também foi analisado um software para extração de entidades mencionadas em português, o Rembrandt.

O AlchemyAPI [3] é uma plataforma que disponibiliza um conjunto de funcionalidades de processamento de linguagem natural e utiliza dados de LOD como fonte de conhecimento. Na extração de entidades nomeadas é possível obter os resultados como URIs correspondentes a recursos em bases da WoD. O LTasks[4] é uma ferramenta baseada no Apache OpenNLP, que disponibiliza um web service para acessar o serviço de extração de entidades nomeadas. O Rembrandt[Cardoso 2008] é um sistema de reconhecimento de entidades mencionadas (REM) para a língua portuguesa. Ele usa a Wilkipédia como fonte de conhecimento e aplica um conjunto de regras gramaticais que definem o conceito de determinada palavra de acordo com o contexto.

Os 3 softwares foram utilizados para extrair entidades de 18 páginas da Wikipédia sobre reciclagem, totalizando 8.697 palavras. Para cada software foi medida a acurácia (número de entidades corretas/ total de entidades classificadas) e a taxa de

---

[3] AlchemyAPI: Transforming text into knowledge. Disponível em: http://www.alchemyapi.com/; acesso em: 26/04/2012

[4] Language Tasks. Disponível em: http://ltasks.com/; acesso em 28/04/2012

erro (número de entidades erradas/ total de entidades classificadas), de acordo com os totais representados na Tabela 1.

Tabela 1. Valores de acurácia e taxa de erro para cada software utilizado

	LTasks	AlchemyAPI	Rembrandt
Acurácia	141/205 = 68%	41/50 = 82%	126/175 = 72%
Taxa de erro	64/205 = 32%	9/50 = 18%	49/175 = 28%

O LTasks reconheceu muitas entidades, porém errou muito nos textos que falavam de elementos químicos e em nomes de organizações em inglês. O AlchemyAPI reconheceu poucas entidades na sua versão para língua portuguesa, porém teve menos erros e classificou doenças muito bem. O Rembrandt reconheceu muitas entidades e teve poucos erros, a maioria deles em siglas de elementos químicos.

De acordo com os resultados, pode-se perceber que a eficácia dos softwares é bastante dependente de domínio, classificando melhor certos tipos de texto. Isso ocorre porque cada um deles tem suas bases de treinamento e categorias para a classificação das entidades.

Dentre os softwares analisados, o Rembrandt é o único em que a licença permite a modificação e acesso ao código. Portanto, podem ser feitos ajustes e adaptações para que ele reconheça entidades de domínios mais específicos como a legislação brasileira. Para isso podem ser acrescentadas novas bases de treinamento. Existem também outras ferramentas de código aberto que permitem a extração de entidades nomeadas, a utilização de bases de treinamento e a adaptação para língua portuguesa como o Lingpipe, o extrator de Stanford e o Apache OpenNLP.

4. Conclusão

Neste artigo foram analisados softwares para a extração de entidades nomeadas, fundamentais a esta proposta. Os resultados dos primeiros experimentos mostram que eles não são satisfatórios para domínios específicos e para o idioma portugues. Portanto, na continuidade de nosso trabalho, será necessário incorporar modificações e desenvolver mehorias para que sejam reconhecidas entidades de domínios mais abrangentes e que sejam adequados às peculiaridades de nosso idioma.

Ainda como trabalhos futuros, já estamos trabalhando nos mecanismos para interligação das informações das bases de LOD aos conteúdos das aplicações de comunicação textual.

Referências

Horiguchi, S., Hoshi, T., Inoue, A., Okada, K. (2009). GaChat: A chat system that displays online retrieval information in dialogue text. VISSW. IUI2009.

Cardoso, N. (2008). REMBRANDT - Reconhecimento de Entidades Mencionadas Baseado em Relações e ANálise Detalhada do Texto. Em Cristina Mota & Diana Santos (eds.), Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM. Linguateca, 2008, pp. 195-211.