Extração de informação de senso comum da Wikipédia

Daniel Feitosa, Thiago A. S. Pardo
Instituto de Ciências Matemáticas e de Computação de São Carlos

1. Objetivos

A utilização dos avanços na área computacional para prover suporte ao processo de ensino-aprendizagem e atribuir a ele um caráter flexível e portável, em abordagens que combinam Aprendizagem Eletrônica (AE) e Educação à Distância (EaD) tem se tornado cada vez mais comum [1]. Dentre estes avanços está o projeto Open Mind Commom Sense (OMCS), que possui o objetivo de obter uma grande base de senso comum, que possa ser utlizada em diversas aplicações. A extensão deste projeto no Brasil (OMCS-Br) possui um sistema de aprendizado por meio de interação humana. Neste contexto, este projeto tem como meta acelerar, de forma eficiente, o crescimento desta base.

2. Material e Métodos

Para cumprir com o objetivo deste projeto, será preciso utilizar uma base de senso comum em potencial. A internet seria esta base de busca, porém precisa-se restringir a textos bem escritos. Tentando alcançar os dois ideais, opta-se pela Wikipédia, que se acredita ser uma forte base de compartilhamento de informação, pois já foi utilizada com sucesso em projetos semelhantes [2]. Precisa-se, também, desenvolver um método eficiente para extração de informação desta base. Propõe-se, portanto, o uso da Wikipédia para extração de informações candidatas a senso comum.

3. Resultados e Discussão

Dois problemas a serem resolvidos são a forma de acesso à Wikipédia e qual método de extração utilizar. Buscando um acesso rápido à base da Wikipédia, optou-se por manter em servidor uma cópia em formato SQL, o que causa problemas de atualização, que são menos importantes do ponto de visto do projeto. Para extração de informação, existem diversos métodos já explorados com sucesso, tais como expressão regular e tipos de autômatos [3], precisando-se agora juntá-los a métodos de exploração de conteúdo web já testados com sucesso [4].

Com base nas informações acima, foi criada uma arquitetura (Figura 1), que servirá para o desenvolvimento do sistema de extração.

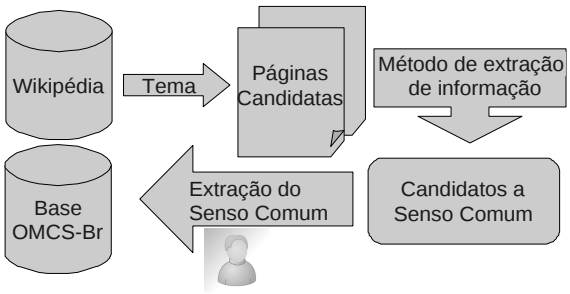

Figura 1: Arquitetura do sistema de extração

4. Conclusões

Até o desenvolvimento atual, estão claras as diversas formas de extração dos candidatos a senso comum, bem como acessar a base da Wikipédia. Mesmo com a forma de acesso à base definida, este projeto pretende, até seu término, encontrar a forma mais eficiente de extrair os candidatos. Como conclusões, pretende-se acelerar o crescimento da base do OMCS-Br, bem como afirmar a Wikipédia como uma base forte para obtenção de informação.

5. Referências Bibliográficas

[1] ZHANG D.; ZHAO, J. L.; ZHOU, L.; NUNAMAKER, J. F. (2004). Can e-learning replace classroom learning? **Communications of the ACM,** US (New York, NY): ACM Press, v. 47, n. 5, mai. 2004. p. 75-79.

[2] R. M. Casado, E. Alfonseca, P. Castells, *Automatic Extraction of Semantic Relationships for WordNet by Means of Pattern Learning from Wikipedia, in NLP and IS: 10th International Conference on Applications of NL to IS* (2005).

[3] Jackson, P.; Moulinier, I. Natural language processing for online applications: Text retrieval, extraction, and categorization, cáp. 4 John Benjamins Publishing Co (2002).

[4] S. F. Adafre; V. Jijkoun; M. de Rijke; In: C. Peters et al, editors, *Evaluation of Multilingual and Multi-modal Information Retrieval. 7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006, Revised Selected Papers*, LNCS 4730, pages 537-540, September 2007.