Aplicando Grafos e *CST (Cross-document Structure Theory)* na Sumarização Automática Multidocumento

Rafael Ribaldo, Ademar T. Akabane, Thiago A. S. Pardo
Instituto de Ciências Matemáticas e de Computação de São Carlos, USP, SP

1. Objetivos

Com a crescente quantidade de informações a disposição do ser humano na rede, a viabilidade de análise destas torna-se quase impossível. Por essa razão, a importância da sumarização automática multidocumento vem crescendo.

O objetivo deste trabalho é criar uma ferramenta que extraia as informações relevantes de textos utilizando técnicas e medidas pré-definidas de grafos e redes complexas (que são um tipo especial de grafo), por exemplo, grau e dinamicidade da rede, aliadas ao conhecimento proveniente da representação semântico-discursiva dos textos fornecida pelo modelo *CST (Cross-document Structure Theory)* (Radev, 2000).

Ao final, a ferramenta criada deverá dar origem a uma interface web acoplável a buscadores, por exemplo, *Google*, contendo opções e métodos de produzir sumários multidocumento. Eventualmente, a ferramenta deve ser um plugin. Além disso, uma ferramenta stand-alone também deve ser produzida. A Figura 1 mostra um esboço dessa ferramenta.

2. Métodos e Procedimentos

Os materiais, a princípio, a serem utilizados para a realização do projeto serão: a) um córpus de textos e resumos; b) um segmentador textual; c) e um thesaurus. Dentre os métodos de sumarização a serem utilizados, prevemos os métodos de Salton et al. (1997) e Antiqueira et al. (2009).

3. Resultados

O resultado obtido, até o momento, foi a estrutura completa de dados para a representação dos textos, utilizando um dígrafo (grafo orientado), além do design da interface. O próximo passo será a exploração dos métodos e dos recursos lingüístico-computacionais previstos.

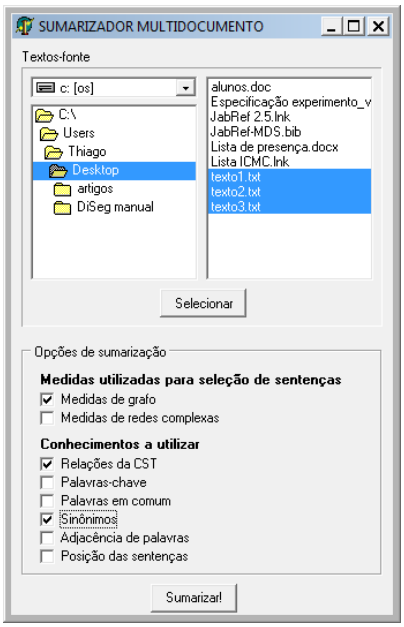

Figura 1: Protótipo da interface mencionada

4. Conclusões Parciais

A partir do que foi estudado é possível notar a grande necessidade de se ter textos concisos. Espera-se que o uso dos conhecimentos e métodos previstos gere sumários de boa qualidade, possibilitando a disponibilização da ferramenta para uso por usuários comuns e também pela comunidade de pesquisa.

5. Referências Bibliográficas

Antiqueira, L.; Oliveira Jr., O.N.; Costa, L.F. Nunes, M.G.V. (2009). A Complex Network Approach to Text Summarization. *Information Sciences*, Vol. 179, N. 5, pp. 584-599.

Radev, D.R. (2000). A common theory of information fusion from multiple text sources, step one: Cross-document Structure. In the *Proceedings of the 1st ACL SIGDIAL Workshop on Discourse and Dialogue*.

Salton, G.; Singhal, A.; Mitra, M.; Buckley, C. (1997). Automatic Text Structuring and Summarization. *Information Processing and Management*, Vol. 33, N. 2, pp. 193-207.