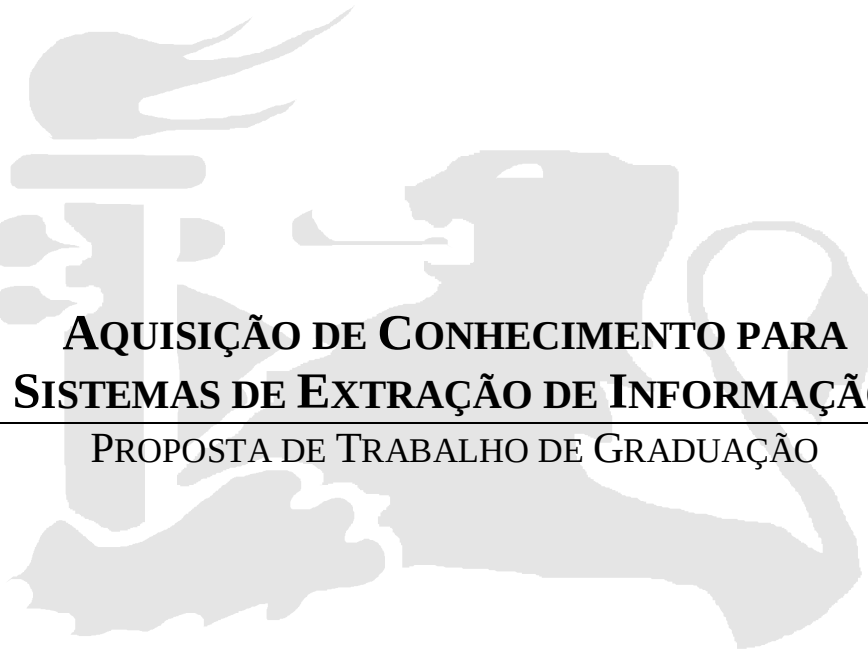
UNIVERSIDADE FEDERAL DE PERNAMBUCO
GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO
CENTRO DE INFORMÁTICA



AQUISIÇÃO DE CONHECIMENTO PARA SISTEMAS DE EXTRAÇÃO DE INFORMAÇÃO

PROPOSTA DE TRABALHO DE GRADUAÇÃO

Aluno: Mozart Vasconcelos Silva
Orientadora: Flávia de Almeida Barros

Recife, 24 de Maio de 2005

Contexto

A quantidade de informação existente na internet vem aumentando cada vez mais, tornando-se mais difícil a localização da informação desejada [Grishman 1997] [Arvind 2003]. Mesmo quando conseguimos recuperar documentos de interesse, o seu conteúdo nem sempre pode ser tratado automaticamente pelo computador, requerendo um trabalho manual extenso. Assim, as informações contidas nesses documentos não ficam facilmente disponíveis ao usuário, pois não é possível tratar tanta informação manualmente. Além disso, as informações estão apresentadas de forma não estruturada em documentos textuais. Para resolver esse problema são utilizados sistemas de extração de informação [Freitag 1998] [Gaizauskas 1998].

Extração de Informação (EI) envolve a criação de uma representação estruturada da informação selecionada de um documento estruturado ou não [Smith 1997], tornando a informação não estruturada em uma estrutura tabular a partir de templates de saídas. Esses sistemas têm como objetivo extrair informações relevantes para o usuário, que não necessita assim ler grandes quantidades de texto a fim de localizar a informação desejada.

Os sistemas de EI são utilizados principalmente para a criação de bases de dados específicas para um domínio de informação. Por exemplo, se um sistema de EI tem como finalidade extrair preços de livros de diferentes páginas na internet, essa base poderá ser utilizada para um outro sistema que faça comparações entre os preços dos livros.

Wrappers são programas especiais para usados para a extração de dados estruturados ou semi-estruturados, explorando as regularidades da aparência dos textos, usando padrões que ajudam a identificar os dados relevantes.

Objetivos

O objetivo deste Trabalho de Graduação é a criação de um módulo de aquisição de conhecimento que irá ajudar o engenheiro de conhecimento na construção de templates de saída e um conjunto de regras bases para uma determinada aplicação. O processo de aquisição é baseado em um "training corpus" (corpo de treinamento) de documentos fontes, no caso abordado o "corpus" é composto por web pages com os dados a serem coletados pelo engenheiro de conhecimento.

Cronograma

	Mês															
Atividade	*Maio*				*Junho*				*Julho*				*Agosto*			
Pesquisa sobre extração de informação																
Arquitetura e funcionalidades																
Implementação do sistema																
Realização de experimentos																
Refinamento																
Elaboração do relatório final																

Referências Bibliográficas

[Arvind 2003] Arasu, Arvind & Garcia-Molina, Hector Extracting Structured Data from Web Pages. SIGMOD Conference, 2003

[Freitag 1998] Freitag, D., Information extraction from HTML: Application of a general machine learning approach. Proc. the Fifteenth National Conference on Artificial Intelligence, pp. 517-523, AAAI press, Madison, Wisconsin, 1998.

[Gaizauskas 1998] Gaizauskas, Robert & Wilks, Yorick, Information Extraction: Beyond Document Retrieval. Computational Linguistics and Chinese Language Processing vol 3 no 2, pp 17-60, 1998

[Grishman 1997] Grishman, Ralph, Information extraction techniques and challenges. International Summer School SCIE-97. Springer-Verlag, 1997

[Smith 1997] Smith, Dan & Lopez, Mauricio. Information Extraction for semi-structured documents. Proceedings of the Workshop on Management of Semistructured Data, 1997

Datas e Assinaturas

_____	_____
Mozart Vasconcelos Silva	Flávia de Almeida Barros

24 de Maio de 2005.