1. Introdução

O crescimento exponencial do conhecimento, maximizado nos anos recentes pelo advento da Internet e outros avanços da tecnologia da informação e comunicação, configura-se um grande paradoxo para o desenvolvimento do aprendizado humano, posto que, ao mesmo tempo em que se disponibiliza um vasto universo de informações pertinentes sobre praticamente quaisquer áreas de conhecimento, também se dificulta o acesso, pesquisa e processamento deste "excesso" de informação.

Esse paradoxo levanta questões importantes quanto à utilização e transferência desse conhecimento acumulado tais como: como organizar e integrar conhecimento a partir de um volume tão grande de informação desestruturada? Como gerir este conhecimento permitindo o seu reuso e aplicação de forma ágil e flexível? Como acessá-lo de forma rápida e prática?

Entendemos que boa parte das questões mencionadas anteriormente, como desafios importantes à gestão eficaz do conhecimento, são análogas aos problemas abordados e adequadamente endereçados durante anos de evolução constante da tecnologia de banco de dados. Sistemas Gerenciadores de Bancos de Dados (SGBDs) nada mais são que um conjunto de tecnologias que suportam armazenamento, manipulação, consulta e integração de bases de informação e, portanto, acreditamos que as mesmas abordagens utilizadas em bancos de dados são propensas a serem usadas no gerenciamento de conhecimento.

Todavia, algo que distancia bastante as técnicas hoje existentes para gerenciamento de bancos de dados daquelas necessárias à área da gestão de conhecimento é que as primeiras pressupõem que os dados estejam previamente estruturados (ex. Datawarehouse, DataMining etc), ou seja, os dados devem estar armazenados dentro de um padrão ou modelo.

Por outro lado, documentos existentes em bibliotecas encontram-se em sua maioria em formato de texto (ex. pdfs, docs etc), os quais constituem-se em uma base desestruturada e caótica, disponibilizados em uma forma totalmente despadronizada. Estes documentos possuem diversos elementos que poderiam ser utilizados na área de gestão de conhecimento, mas estes elementos precisam ser extraídos destes documentos e armazenados de uma forma estruturada, a fim de poder-se aplicar sobre os mesmos as técnicas de integração, consulta, e armazenamento que os bancos de dados disponibilizam.

Este trabalho faz parte do projeto Partnership in Global Learning (PGL) da PUC-RJ aonde um grupo de trabalho vem desenvolvendo estudos para usar a tecnologia de armazenamento de objetos de aprendizagem no desenvolvimento dos seus cursos em diversas áreas do conhecimento. Esta abordagem de armazenamento de LOs (objetos de aprendizagem) vem evoluindo na área de bancos de dados para e-learning (Melo e Baruque, 2003). Diversos trabalhos já foram desenvolvidos neste projeto todos voltados para a gerência de LOs como, por exemplo: (Baruque e Melo, 2005), (Gomes et al., 2006) e (Leal e Melo, 2006) .

O LO é descrito por elementos de metadados, a fim de proporcionar melhores buscas e reuso e eles podem ser constituídos de diversos ALOs (Atomic Learning Object) que são a menor unidade de aprendizado (Pereira, Porto e Melo, 2003). É preciso que os metadados possam ser definidos a partir do conteúdo existente no texto, permitindo o uso indiscriminado de qualquer conteúdo digitalizado em forma de ALO.

Propomos aqui a criação dos ALOs baseados em textos extraídos de documentos digitais. Sugerimos um processo semi-automático com este objetivo que faz uma seleção e classificação de partes destes documentos. Ao final deste processo um gestor de conhecimento ou o Instructional Designer pode fazer uma pesquisa definindo palavras de interesse e o tipo de ALO que deseja. A pesquisa retorna as partes dos textos, que possuem as palavras definidas além de uma classificação. O resultado é apresentado ordenado pela classificação feita de acordo com os parâmetros da consulta solicitada. Após a triagem final executada pelo gestor e baseado nos dados obtidos, fica mais simples definir um novo ALO que pode ser armazenado em um repositório de objetos de aprendizado de uma forma estruturada.

1.1.Motivação

Esta dissertação visa investigar métodos e técnicas para estruturação de conhecimento armazenado em textos. Quando consideramos a qualidade e a diversidade do conteúdo hoje existente em documentos digitais, antevemos um grande benefício na conjunção de técnicas de Bancos de Dados, Mineração de Textos, Aprendizado de Máquina e Gestão de Conhecimento para suporte a objetivos educacionais. Contribuir para o avanço, ainda que em parte, da extração e estruturação do conhecimento existente em milhares de documentos monolíticos existentes, e disponibilizá-los para utilização amigável por parte da comunidade científica e da sociedade em geral é um desafio muito motivante e objeto principal desta dissertação.

1.2.Objetivo

O objetivo principal deste trabalho é desenvolver um processo que permita a obtenção de partes de textos contidos em documentos que possam ser utilizados como Objetos de Aprendizagem. Investigamos métodos e técnicas que possibilitem e apóiem este processo.

Queremos que um usuário tenha como opção o tipo de texto desejado, o que especifica a proposta instrucional do texto que ele procura, e palavras chaves. Desta forma ele objetiva encontrar partes dos documentos que contenham um assunto de interesse em um formato específico, o que lhe resulta em um ALO. Como estas informações o gestor pode criar o ALO e preencher os dados existentes em seus metadados o que permite que eles sejam armazenados em repositórios de ALOs e utilizados na composição de LOs complexos e nas consultas integradas desenvolvidas em outros trabalhos para este tipo de repositório.

Utilizamos técnicas de análise da linguagem natural e Aprendizado de Máquina para gerar um modelo que permita a classificação de textos. Este modelo é então usado para ordenar os textos selecionados em um *ranking*, apresentando para o usuário no topo da lista os textos mais “parecidos” com o que foi solicitado.

Investigamos a utilidade e desempenho do Aprendizado Semi-Supervisionado para a execução da tarefa de classificação. Utilizamos o algoritmo proposto por (Nigam, Maccallum, Thrun e Mitchell, 2000) que, baseado na combinação de Expectation-Maximization (EM) e um classificador naive Bayes, aprende a classificação a partir de uma quantidade pequena de textos já classificados ou seja, exemplos que de alguma forma já foram etiquetados, acrescida de textos não etiquetados, que são mais fácil de serem obtidos.

1.3.Organização

No capitulo dois fazemos uma revisão sobre Objetos de Aprendizagem que é o que procuramos obter no processo de mineração de texto. Especificamos as classes que serão usadas no processo de classificação de textos e como elas foram definidas. No terceiro capitulo descrevemos as técnicas de Aprendizado de Maquina e todos os fundamentos teóricos utilizados no desenvolvimento do algoritmo de aprendizado e classificação e a justificativa para a utilização deste tipo de algoritmo de aprendizado. Definimos também as métricas utilizadas para medir a eficiência do classificador desenvolvido. No capítulo quatro apresentamos o Sistema SQLLOMining que foi desenvolvido para apoiar o processo de obtenção de LOs. No capítulo cinco apresentamos o Estudo de Caso detalhando os experimentos que foram feitos e os resultados alcançados. No capítulo seis apresentamos alguns trabalhos relacionados e faremos uma comparação. E finalmente no capítulo sete apresentamos as conclusões.