CENTRO UNIVERSITÁRIO FEEVALE

FABRÍCIO LUÍS COELHO

CLASSIFICAÇÃO SEMI-AUTOMÁTICA DE MONOGRAFIAS

(Título Provisório)

Anteprojeto de Trabalho de Conclusão

Novo Hamburgo, setembro de 2007.

FABRÍCIO LUÍS COELHO

flcoelho@gmail.com

CLASSIFICAÇÃO SEMI-AUTOMÁTICA DE MONOGRAFIAS

(Título Provisório)

Centro Universitário Feevale
Instituto de Ciências Exatas e Tecnológicas
Curso de Ciência da Computação
Anteprojeto de Trabalho de Conclusão

Professor orientador: Rodrigo Rafael Villareal Goulart

Novo Hamburgo, setembro de 2007.

RESUMO

Com a crescente expansão da Internet, a localização de informações precisas torna-se dia-a-dia mais difícil. Para agilizar o processo de aprendizagem, torna-se necessário o desenvolvimento de ferramentas que facilitem a busca de informações, transferindo a tarefa de classificá-las e ordená-las a aplicações computacionais. Este trabalho foca-se na classificação de monografias, utilizando método de aprendizagem supervisionada para desenvolver um protótipo de classificador semi-automático.

Palavras-chave: Classificação. Inteligência Artificial. Aprendizagem Supervisionada.

SUMÁRIO

MOTIVAÇÃO

A Internet representa uma transição histórica na evolução humana. Sua capacidade de expansão, exponencial e revolucionária, afeta diretamente os indivíduos e suas relações. A rede oferece um universo ilimitado de informações diante de qualquer pessoa com acesso a um computador conectado, admitindo o compartilhamento de idéias entre seus usuários e proporcionando um terreno fértil para a produção e disseminação massiva de novos conhecimentos. Em nenhuma outra época se produziu tanto conhecimento quanto nos dias atuais.

Com uma indústria informativa tão produtiva, nem sempre é possível localizar o que se procura com facilidade. Criatividade e organização parecem, freqüentemente, rumar em sentidos opostos. Para que o processo produtivo não seja interrompido, é necessário um esforço constante na organização e classificação do conhecimento, conferindo maior agilidade na sua transferência entre as pessoas.

O Google[1] classifica seus artigos. Em seu site de pesquisas estão disponíveis trabalhos organizados por categoria. Um sistema que automatiza o processo de classificação de documentos poderia ser empregado na organização dos documentos disponibilizados pelo Google. A quantidade da produção científica da empresa pode não justificar tal sistema, mas poderia servir de apoio à classificação manual.

Num contexto local, as monografias de conclusão do curso de Ciência da Computação podem ser utilizadas como estudo de caso. Na página dos TC's, as monografias estão ordenadas por título. Desta forma, os títulos dos trabalhos devem guardar estreita relação com os conteúdos, de modo que um eventual pesquisador possa localizar o material que procura sem muita dificuldade. Mas isso nem sempre ocorre: freqüentemente, é preciso

[1] - GOOGLE RESEARCH. Disponível em: <http://research.google.com/pubs/papers.html>. Acesso em 13/09/2007.

dedicar tempo à leitura dos resumos e sumários dos trabalhos para encontrar algum que se refira à sub-área desejada.

A classificação – também conhecida como *categorização* – é uma técnica utilizada para distribuir objetos entre determinadas classes. O processo exige conhecimento prévio das propriedades de cada classe, por isso é dividido em dois passos: *aprendizado* e *classificação*. O primeiro passo é identificar e determinar os atributos de cada categoria. O passo seguinte é dividir os elementos entre as categorias (WIVES, 1996). Se fosse necessário realizar a tarefa manualmente, a categorização demandaria um tempo considerável para ser efetuada.

A Inteligência Artificial proporciona modos de aprendizado de máquina capazes de, a partir do estudo do caso, construir uma solução aplicável a problemas similares, de forma eficiente. Árvores de decisão, além de serem extremamente simples, são algoritmos de aprendizagem indutiva que se enquadram nesses requisitos.

Pode-se aplicar uma árvore de decisão a um problema cujo objetivo é definir, a partir de determinadas condições, se interessa ou não esperar por uma mesa de restaurante, por exemplo (RUSSEL, 2004). Para isso, elabora-se uma lista de atributos composta de questões como: se há outros restaurantes adequados nas proximidades; se a espera pode ser confortável; se o dia é de grande movimento; o quanto a fome está incomodando; o tempo de espera estimado, etc. Com base nos atributos selecionados, monta-se a árvore de decisão contendo os testes que seriam aplicados na situação hipotética. Este tipo de algoritmo pode apresentar resultados satisfatórios, com bons níveis de precisão, como se pode verificar no trabalho “Análise de expressões referenciais em corpus anotado da Língua Portuguesa” (ABREU, 2005). Na dissertação, foram desenvolvidos algoritmos para selecionar atributos relevantes à classificação de “descrições definidas como novas no discurso”.

Segundo Aas (1999), alguns dos algoritmos de árvores de decisão mais comumente utilizados para classificação de textos são o C4.5, o CHAID e o CART e Sebastiani (2002) cita os algoritmos ID3, e C5, para a mesma finalidade. Em “Análise...” (ABREU, 2005), por exemplo, foi utilizado o algoritmo j4.8, que é uma implementação em Java do algoritmo c4.5.

Como já abordado, os trabalhos de conclusão de curso da Instituição não são classificados. Cada autor deve determinar palavras-chave que indiquem o assunto do trabalho no interior da monografia, assim como escolhe o título de sua obra. O desafio ora proposto é desenvolver um método para classificar automaticamente esses trabalhos.

Para isso, selecionar-se-á um grupo de trabalhos já concluídos para efetuar um processo estatístico que encontre as cinco palavras mais comumente usadas, que servirão como atributos de um texto. Determinar-se-á, ainda, um conjunto de categorias, a partir das quais cada monografia será classificada manualmente. Deve-se, então, induzir a aprendizagem de uma árvore de decisão, a partir de um novo conjunto de trabalhos de exemplo, fornecidos como entrada. A árvore gerada será utilizada para desenvolver um protótipo de aplicação que classifique automaticamente novos textos.

O propósito do software é viabilizar a seleção de monografias de acordo com a área de interesse, através da busca por palavras-chave. A aplicação poderá ser integrada à página do curso de Ciência da Computação, para gerar automaticamente uma página que ordene as monografias por categoria. O software poderá ser integrado às páginas dos demais cursos oferecidos pela Feevale.

O aplicativo proposto não tem utilidade restrita limitada ao ambiente acadêmico. Seu campo de utilização é vasto, podendo ser aplicado ainda na classificação de artigos em seminários e conferências, ou em áreas profissionais, onde o grande volume de documentos retarda a localização de informações, como o direito e a biologia, entre outras.

OBJETIVOS

Objetivo geral

Desenvolver um protótipo de aplicação para classificar semi-automaticamente monografias, a partir de técnicas de aprendizagem supervisionada.

Objetivos específicos

- Estudar técnicas de aprendizagem supervisionada;
- Estudar árvores de decisão;
- Estudar métodos de classificação de textos;
- Estudar o framework Weka;
- Desenvolver uma taxonomia das áreas de trabalhos de conclusão de curso em Ciência da Computação;
- Estudar o comportamento prático de algoritmos e técnicas de árvores de decisão;
- Determinar os atributos para aprendizagem de um conjunto de monografias já realizadas;
- Efetuar a classificação manual de trabalhos de conclusão já finalizados;
- Desenvolver um protótipo de aplicação para classificação automática de textos;
- Avaliar o protótipo desenvolvido;
- Analisar os resultados da avaliação.

METODOLOGIA

O desenvolvimento deste trabalho segmentar-se-á em duas partes: Trabalho de Conclusão 1 (TC1) e Trabalho de Conclusão 2 (TC2).

A primeira parte do trabalho compreende o estudo de métodos de aprendizagem supervisionada, de metodologias de classificação de textos e documentos, o desenvolvimento de uma taxonomia das áreas de trabalhos de conclusão de curso em Ciência da Computação e a redação do primeiro relatório de atividades.

O principal método para a realização desta fase será a realização de pesquisas bibliográficas em livros, monografias e na Internet. A taxonomia das áreas de trabalhos de conclusão será desenvolvida através de entrevistas com professores do curso.

Compreendem a segunda parte do trabalho, atividades práticas com algoritmos e técnicas de árvores de decisão; a seleção de trabalhos para classificação manual - que posteriormente servirão para testar os algoritmos; a classificação manual dos trabalhos de conclusão, selecionados a partir da taxonomia desenvolvida na fase anterior; o desenvolvimento do protótipo de aplicação; a avaliação do protótipo; a análise dos resultados da avaliação do software.

Os recursos que serão empregados na realização da fase final do trabalho são: monografias de conclusão, para o aprendizado da árvore de decisão; o framework weka, para construção da árvore de decisão; e ferramenta para o desenvolvimento do software, a ser selecionada oportunamente.

CRONOGRAMA

Trabalho de Conclusão I

Etapa	Meses			
	Ago	Set	Out	Nov
Elaborar anteprojeto				
Estudar métodos de aprendizagem supervisionada				
Estudar metodologias de classificação de textos e documentos				
Desenvolver uma taxonomia das áreas de trabalhos de conclusão de curso em Ciência da Computação				
Redigir o relatório do TC1				

Trabalho de Conclusão II

Etapa	Meses					
	Jan	Fev	Mar	Abr	Mai	Jun
Estudar o comportamento prático de algoritmos e técnicas de árvores de decisão						
Selecionar trabalhos para classificação manual e posteriores testes dos algoritmos						
Classificar manualmente trabalhos de conclusão						
Desenvolver o protótipo de aplicação para classificação automática de textos						
Realizar avaliação do protótipo						
Analisar os resultados						
Redigir o relatório do TC2						

BIBLIOGRAFIA

AAS, Kjersti; EIKVIL, Line. **Text Categorisation: A Survey**. Norwegian Computing Center, 1999. Noruega, 1999. 37p.

ABREU, Sandra Collovini de. **Análise de expressões referenciais em corpus anotado da Língua Portuguesa**. UNISINOS, 2005. Dissertação de Mestrado (pós-graduação em computação aplicada). Ciências Exatas e Tecnológicas, Universidade do Vale do Rio dos Sinos, 2005. 105 p.

GOOGLE RESEARCH. Disponível em: <http://research.google.com/pubs/papers.html>. Acesso em 13/09/2007.

RUSSEL, Stuart et al. **Inteligência Artificial**. 2.ed. Traduzido por: Vandenberg D. de Souza. Rio de Janeiro: Elsevier, 2004. 1021 p. Tradução de Artificial Intelligence.

SEBASTIANI, Fabrizio. **Machine learning in automated text categorization**. Consiglio Nazionale delle Ricerce, 2002. Itália, 2002. 47p.

WIVES, Leandro Krug. **Um estudo sobre agrupamento de documentos textuais em processamento de informações não estruturadas usando técnicas de "clustering"**. UFRGS, 1999. Dissertação de Mestrado (pós-graduação em Ciência da Computação). Instituto de Informática, Universidade Federal do Rio Grande do Sul, 1999. 102 p.