Classificação e agrupamento de textos para processamento multidocumento

Felipe Tassario Gomes, Thiago Alexandre Salgueiro Pardo

NILC - Núcleo Interinstitucional de Lingüística Computacional – Instituto de Ciências Matemáticas e Computação – Universidade de São Paulo – São Carlos – SP – Brasil

felipc@grad.icmc.usp, taspardo@icmc.usp.br

Abstract. *This paper describes the analysis and development of an online tool to assist searching news, which is capable of finding news about an specific topic and later clustering the results into subtopics. Here we present the algorithms and evaluation methods considered, the developed architecture of the tool, as well as the results obtained from the evaluation process.*

Resumo. *Este artigo descreve o estudo e o desenvolvimento de uma ferramenta online de auxílio à busca de notícias, que realiza uma busca sobre notícias de um mesmo tópico e agrupa os resultados em subtópicos. Aqui são apresentados o estudo dos algoritmos e seus métodos de avaliação, o desenvolvimento da arquitetura da ferramenta, bem como a análise e conclusão de seus resultados.*

1. Introdução

Para muitas tarefas de processamento de textos, faz-se necessário o agrupamento e classificação dos textos em função dos tópicos de que tratam e em seu desenvolvimento no decorrer do tempo. Por exemplo, na área conhecida como sumarização multidocumento, busca-se por um resumo de um conjunto de notícias que versam sobre um mesmo assunto. A leitura do resumo, que agrega as principais informações do evento e seu desenvolvimento, dispensa a leitura das várias notícias relacionadas.

No cenário atual, onde um grande conjunto de notícias de diversas fontes estão rapidamente acessíveis na *web* através dos *websites* de jornais ou de portais agregadores de notícias, existe uma vasta quantidade de informação disponível sobre um mesmo tópico, bem como uma grande quantidade de temas novos surgindo a cada momento. Tal fluxo de informação torna difícil acompanhar o desenvolvimento de um tópico, bem como obter o panorama geral de um assunto sem previamente conhecer os detalhes e a sequência de acontecimentos do mesmo.

Neste projeto, propôs-se o desenvolvimento de uma ferramenta online de auxílio à busca textual de notícias, capaz de buscar notícias sobre um mesmo tópico, agrupá-las em função de seus subtópicos e exibi-las para o usuário.

2. Objetivos

Neste trabalho de iniciação científica, visou-se um estudo de técnicas relacionadas ao tema de classificação e agrupamento de textos, bem como seus critérios de avaliação; a escolha a adaptação de alguma técnica de agrupamento para o domínio de textos jornalísticos publicados em português brasileiro; o desenvolvimento de um ambiente

online para integrar esta técnica; além da apresentação e familiarização do aluno à área de mineração de dados e processamento de línguas naturais.

3. Material e método

Nesta seção, serão descritas as bases teóricas, as técnicas relacionadas a esta pesquisa e as metodologias utilizadas para implementá-la e construir a arquitetura da ferramenta.

3.1. Conjunto de Dados

Um conjunto de dados é uma coleção de entradas (cada uma denominada Instância), geralmente apresentada em forma de tabela, que possuem um mesmo conjunto de variáveis (denominadas Atributos) utilizadas para descrevê-las. Tais atributos podem ser valores numéricos ou atributos nominais. Dos vários atributos presentes num conjunto de dados, existe um atributo especial denominado Classe, que representa uma relação direta ou indireta de seu valor com os outros atributos de sua instância.

As instâncias e os seus atributos geralmente vem de medidas de observações naturais ou métricas previamente estabelecidas. Um exemplo de conjunto de dados com este aspecto é o publicado em [Fischer, 1936], em que foram obtidas medidas de 4 atributos diferentes em 50 flores do gênero Íris, além do atributo classe tomado como a espécie da flor.

3.2. Classificação e agrupamento de dados

A classificação de um conjunto de dados consiste em determinar a classe de cada instância deste conjunto com base nos valores de seus atributos [Witten e Frank, 2005]. Esta classificação pode se dar através de modelos matemáticos simbólicos/estatísticos como o classificador naïve-Bayes, o algoritmo KNN [Mitchell, 1997], ou o algoritmo sIB [Slonim et. Al, 2002].

O agrupamento de dados consiste em aplicar as mesmas técnicas de classificação num conjunto de dados em que não se conhece previamente as suas possíveis classes [Witten e Frank, 2005].

3.3. Agrupamento de textos

Para agrupar um conjunto de textos, é preciso determinar os atributos de cada instância do conjunto. Estes atributos podem ser em função de diversas métricas, como a quantidade de palavras ou parágrafos do texto, palavras presentes e suas frequências, sequências de palavras (n-gramas), conjunto de palavras (*bag of words*), entre outros.

Neste trabalho, os atributos do texto foram determinados como as palavras (uni-gramas) e sequências de duas e três palavras consecutivas no texto (bi-gramas e tri-gramas). Para medir os valores dos atributos, o texto tem toda a sua pontuação removida e é então dividido palavra a palavra. Os atributos de todos os textos são determinados e colocados numa tabela, e por fim é feita a contagem da frequência de cada atributo em cada instância. Previamente à contagem, o texto tem um conjunto especifico de palavras removidas, chamadas de *stop words*, que consistem de palavras muito frequentes na língua, porém de baixo valor semântico (exemplo: “de”, “como”, “para”), que podem influenciar negativamente os resultados.

3.4. Avaliação de agrupamentos

Existem diversas métricas para avaliar a qualidade de um algoritmo de agrupamento, como descritas em [Manning et. Al, 2008]. As métricas consistem em aplicar o algoritmo de agrupamento num conjunto de dados de teste em que se conhece previamente as classes, e analisar os acertos e erros do algoritmo. Em particular, a métrica de pureza consiste em penalizar a quantidade de instâncias incorretamente classificadas. A pureza P pode ser dada por:

P = Instâncias corretamenta classificadas / Número de instâncias

3.5. Arquitetura desenvolvida

A ferramenta desenvolvida consiste num sistema online composto por diversos módulos responsáveis por cada etapa do funcionamento do sistema. Em particular, os módulos principais do sistema podem ser vistos como:

- Módulo de Busca de Notícias: Este módulo é responsável por perguntar um tópico de notícias desejado pelo usuário e buscar um conjunto de notícias sobre este assunto num motor de busca, como o Yahoo! Notícias;

- Módulo de Processamento de Texto: Este módulo recupera os textos encontrados no motor de busca e processa cada instância do texto, identificando seus atributos;

- Módulo de Agrupamento: Este módulo recebe a tabela de instâncias e atributos gerada pelo módulo anterior e aplica um algoritmo de agrupamento no conjunto de dados. Os algoritmos utilizados estão presentes no ambiente de mineração de dados Weka [Witten e Frank, 2005].

4. Resultados

De um conjunto de textos jornalísticos do português brasileiro disponível no corpus CSTNews [Aleixo e Pardo, 2008], foram utilizados 21 textos de 5 tópicos de notícias diferentes, que foram aplicados na ferramenta utilizando 3 algoritmos de agrupamento diferentes no ambiente Weka: Simple K Means, Expectation Maximization (EM) [Dempster et. al, 1977], e sIB [Slonim et Al., 2002]. A medida da pureza resultante do agrupamento é dado na Tabela 1.

Tabela 1. Medida da pureza num conjunto de 25 textos

Algoritmo	Pureza
Simple K Means	0.428
EM	0.571
sIB	0.952

5. Conclusões

Neste trabalho, foi desenvolvido uma ferramenta online para visualização de notícias jornalísticas utilizando técnicas de mineração de dados aplicadas a textos. Tal ferramenta estPelos resultados obtidos, acredita-se que o algoritmo sIB tem melhores aplicações no domínio de textos com o uso de atributos derivados dos n-gramas do texto. Entretanto, medir a qualidade dos resultados em um conjunto de textos desconhecido (obtidos de um motor de busca) é difícil pois não se conhece as classes

dos textos (como acontece no corpus preparado), estando sujeitos à medidas subjetivas. A ferramenta aqui desenvolvida, bem como seu código fonte, está disponível para uso através do site do endereço http://www.nilc.icmc.usp.br/nilc/tools/newshead/.

Referências

Aleixo, P., Pardo, T.A.S. (2008) CSTNews: Um córpus de Textos Jornalísticos Anotados Segundo a Teoria Discursiva Multidocumento CST. In *Séries de Relatórios Técnicos do ICMC, USP*, nº 236.

Dempster, A., Laird, N., and Rubin, D. (1977) Maximum likelihood from incomplete data via the EM algorithm. In *Journal of the Royal Statistical Society, Series B*, pp. 1-38

Fisher, R. A. (1936) The Use of Multiple Measurements in Taxonomic Problems. In *Annals of Eugenics 7,* pp. 179-188.

MacQueen, J.B. (1967): Some Methods for classification and Analysis of Multivariate Observations. In *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*", Berkeley, University of California Press, pp. 281-297

Manning, C. D., Raghavan, P., Schütze, H. (2008) Introduction to Information Retrieval. Cambridge University Press, 2008.

Mitchell, T. (1997) Machine Learning. McGraw Hill.

Slonim, N., Friedman, N. Tishby, N. (2002). Unsupervised document classification using sequential information maximization. In *25th International ACM SIGIR Conference on Research and Development in Inforation Retriavel,* pp. 129-136.

Witten, I.H, Frank, E. (2005) Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, 2ª edição.