Classificação não-supervisionada hierárquica de artigos jornalísticos

Cirillo Ribeiro Ferreira (cirillo.ferreira@usp.br) Orientador: Prof. Dr. Alair Pereira do Lago

Instituto de Matemática e Estatística, Universidade de São Paulo - Trabalho de Formatura Supervisionado

Objetivos

Este trabalho tem como objetivos:

- Criação de uma biblioteca para agrupamento de artigos jornalísticos disponíveis nos meios digitais.
- Proposta e implementação do sistema hVINA (*Hierarchical Viewer of News Articles*) para simplificar a interação entre o usuário e a biblioteca.

Introdução

Com a criação da internet e a popularização de seu uso como ferramenta de comunicação, houve uma explosão de informação que tornou muito difícil a classificação dos documentos produzidos e publicados nela de maneira manual. A área de classificação de documentos é de grande interesse e possui diversas aplicações práticas como classificação de *spam*, identificação de idioma e análise de sentimento. Em especial, a classificação de artigos jornalísticos tem um enorme desafio devido à grande quantidade de novos documentos criados diariamente e a diversidade de temas abordados. Para tal propósito serão utilizados algoritmos de aprendizagem não-supervisionadas, pois não necessitam de um conjunto de treinamento como entrada, permitindo o seu uso em conjuntos de dados bem variados, algo que é bem comum em artigos jornalísticos.

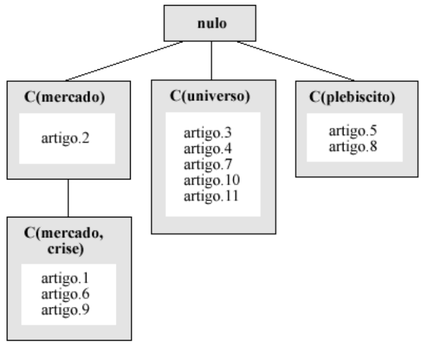

Figura 1: Exemplo de agrupamento feito pelo FIHC

Análise de agrupamento

Análise de agrupamento é uma classificação de padrão que emprega o processo de aprendizagem não-supervisionada e tem como objetivo o particionamento de objetos em grupos cujo membros sejam similares entre si e diferentes dos membros de outros grupos [1].

Algoritmos hierárquicos

Os algoritmos da área de agrupamento são divididos geralmente em duas classes: Algoritmos planos e hierárquicos. Os algoritmos hierárquicos são aqueles que geram uma árvore de grupos (*clusters*). Uma estrutura que fornece mais informação, uma vez que as relações implícitas entre os grupos ficam mais evidentes [2]. Há duas abordagens na criação da árvore, a primeira chamada de divisiva ou *top-down* inicia a construção a partir da raiz até as folhas, a segunda chamada de aglomerativa ou *bottom-up* inicia a construção das folhas à raiz.

FIHC

Foi implementado inicialmente na biblioteca o *Frequent Itemset-based Hierarchical Clustering* (FIHC), que é um algoritmo para agrupamento hierárquico de documentos textuais [3] que utiliza o conceito de conjuntos de itens frequentes (*frequent itemset*). O FIHC está na classe dos algoritmos hierárquicos aglomerativos e baseia-se no seguinte critério de similaridade para a construção da árvore de grupos:

$$Sim(Ci \leftarrow Cj) = \frac{Score(Ci \leftarrow doc(Cj))}{N} + 1 \quad (1)$$

Onde Ci e Cj são os grupos usados na comparação de similaridade, $Score$ é uma medida de relevância de um documento em um grupo;

$$N = \sum_{x} tfidf(x, doc(Cj)) + \sum_{x'} tfidf(x', doc(Cj)) \quad (2)$$

e $tfidf$ é uma medida para avaliação da relevância de uma palavra em um documento.

Conjunto de itens frequentes

Um conceito importante para o entendimento do FIHC é a noção de conjunto de itens frequentes de uma coleção, cuja definição é: um conjunto de palavras que ocorrem em uma quantidade de documentos da coleção acima de um limiar de suporte definido pelo usuário.

Arquitetura da biblioteca

A biblioteca abrange todos os passos de uma solução para agrupamento, desde o pré-processamento dos documentos até os algoritmos de agrupamento propriamente ditos. Além disso, foi arquitetada para trabalhar com diversos idiomas. Na sua primeira versão, a biblioteca prioriza a implementação do algoritmo de agrupamento FIHC. A sua arquitetura é mostrada na figura 2.

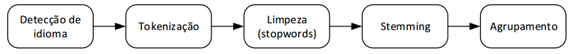

Figura 2: Arquitetura proposta para a biblioteca

O pré-processamento consiste nos seguintes passos:

- Detecção de idioma: O primeiro passo para realizar o agrupamento é identificar o idioma utilizado nos documentos, pois os algoritmos dos passos seguintes necessitam desse conhecimento.
- Tokenização: Segmenta os textos em palavras.
- Limpeza: Remove as palavras que possuem pouca relevância no texto, como as preposições, os artigos e marcações gráficas.
- *Stemming*: Unifica formas variantes de palavras que possuem o mesmo significado, como as palavras "economia" e "econômico".

Já o sistema *hVINA* tem como objetivo criar uma interface amigável para que qualquer usuário possa utilizar a biblioteca, permitindo a ele informar uma coleção de artigos de seu interesse ou utilizar coleções pré-selecionadas pelo sistema.

Conclusão

A biblioteca desenvolvida é modular, permitindo acrescentar o tratamento de outros idiomas ou implementar novos algoritmos de forma não intrusiva. Ademais, outros projetos além do *hVINA* podem utilizá-la facilmente, visto que ela segue as especificações do distribuidor de pacotes do Ruby (RubyGems).

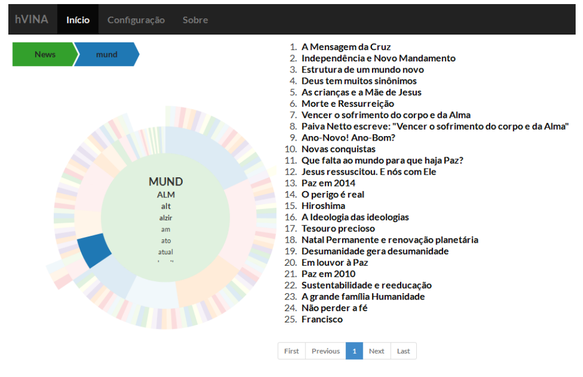

Figura 3: Tela principal do hVINA

O resultado obtido pela biblioteca em conjunto com o sistema *hVINA* demostra a viabilidade de uma solução que tenha boa usabilidade e seja amigável a qualquer usuário, onde não haja a necessidade de treinamento do algoritmo e a configuração seja quase zero. Porém, melhorias na escalabilidade da biblioteca devem ser feitas nos trabalhos futuros.

Referências

[1] A.K. Jain, M.N. Murty, and P.J. Flynn. Data Clustering: A review. *ACM Computing Surveys*, 31(3):264–323, Sept 1999.

[2] C.D. Manning, P. Raghavan, and H. Schütse. *An Introduction to Information Retrival*. Cambridge University Press, 2009.

[3] Benjamim C. M. Fung, Ke Wang, and Martin Ester. Hierarchical Document Clustering Using Frequent Itemsets. 2003.