RELATÓRIO DE DEFINIÇÃO DO PROJETO

Tópicos Avançados em Processamento de Linguagem Natural

Recuperação Inteligente de Informação - 2007.2

Prof.ª Flávia Barros

1ª Entrega do Projeto

1. Membros da Equipe

Guilherme Alexandre Monteiro [alexandrecordel@gmail.com]

Luciano de Souza Cabral [lscabral@gmail.com]

Marcelo Nunes [ffmasterbr@gmail.com]

Rinaldo José de Lima [rina_lima@yahoo.com.br]

2. Tema do Projeto

Recuperação de Informação usando *Clustering*.

3. Breve Descrição do Projeto

Um sistema de recuperação de informação, com objetivo de seleção e extração de características para posterior *clustering*, utilizando algoritmo Hierárquico e avaliar se o *clustering* está com um nível aceitável usando técnicas de avaliação. Com relação à base, utilizaremos a base DMOZ, do *Open Directory Project* que é o maior e mais abrangente diretório editado da web, construído e mantido por uma vasta comunidade de editores voluntários.

RELATÓRIO DE DEFINIÇÃO DO PROJETO

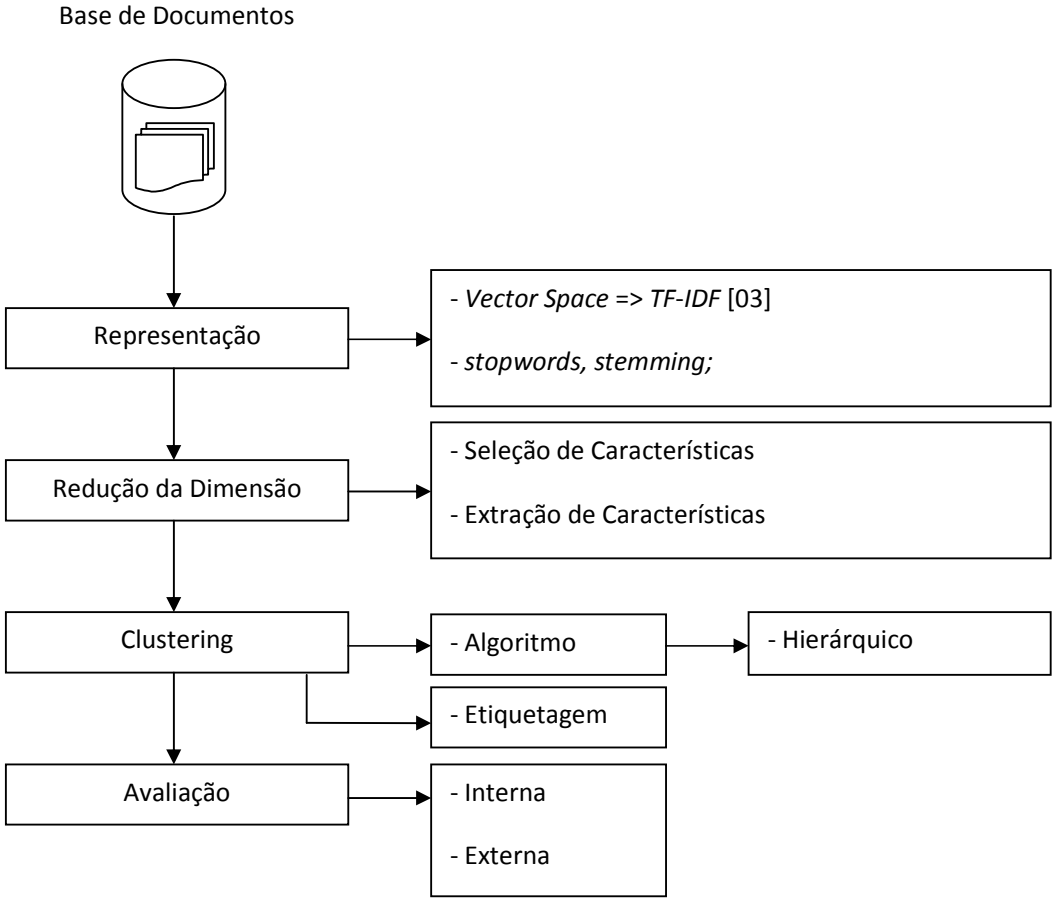
Tópicos Avançados em Processamento de Linguagem Natural

Recuperação Inteligente de Informação - 2007.2

Prof.ª Flávia Barros

2ª Entrega do Projeto

1. Descrição do sistema



Representação

Utilizamos *stopwords*, *stemming* e *TF-IDF*.

Redução da Dimensão

Dividido em seleção e extração de características. Utilizamos *Term Frequency Variance* (TFV) [01] para a seleção, atendido pelo critério:

$$\sum_{j}^{n} tf_j^2 - \frac{1}{n}\left[\sum_{j}^{n} tf_j\right]^2$$

Equação 1 - *Term Frequency Variance* (TFV)

Onde:

O tf_j será calculado conforme mostrado da seção 03, escolhendo cerca de 15% dos termos com o maior critério de qualidade na base. Para extração pretendemos utilizar *Principal Component Analysis* (PCA) *e Independent Component Analysis* (ICA) [07].

***Clustering e* Avaliação**

Pretendemos utilizar o algoritmo *Bisection K-Means* [02] para agrupar os documentos. As etapas de avaliação interna e externa ocorrem respectivamente sem e com informação de classes dos documentos. Para a função de etiquetagem utilizaremos também o *Term Frequency Variance* (TFV), indicado na seção anterior.

A interna é utilizada como um critério para o agrupamento, para maximizar a coesão dos clusters. A externa serve para avaliar a qualidade do agrupamento baseado em informação das classes dos arquivos, onde pretendemos utilizar uma adaptação do *F-Measure* [02] para problemas de agrupamento.

2. Documentos

Com relação à base, utilizamos a base DMOZ, do *Open Directory Project* que é o maior e mais abrangente diretório editado da web, construído e mantido por uma vasta comunidade de editores voluntários. Restrito ao subconjunto de Inteligência Artificial do ODP, definidos na linguagem Inglesa, que utiliza apenas a descrição de cada documento. [08]

3. Preparação dos documentos

Stopwords

Utilizamos como Lista de *StopWords,* uma disponível no site da *Pagex* [09], como a maioria dos motores de busca não consideram extremamente palavras comuns, a fim de economizar espaço em disco ou para acelerar os resultados da pesquisa, estas palavras são filtradas.

Stemming

Utilizamos a biblioteca *stemming Oleander* [10], que é uma implementação em C++ do algoritmo *stemming* de Porter, e suporta grande parte dos idiomas da Europa Ocidental.

TF-IDF

Term Frequency (TF) – Trata-se da freqüência do termo no documento, ou seja, quanto maior, mais relevante é o termo para descrever o documento. [11]

Inverse Document Frequency (IDF) - Inverso da freqüência do termo entre os documentos da coleção, ou seja, Termo que aparece em muitos documentos não é útil para distinguir relevância. [11]

$$tfidf_j = tf_j \log \frac{|T_r|}{DF_t}$$

$$tf_j = \begin{cases} 1 + \log t_j & \text{if } t_j > 0 \\ 0 & \text{otherwise} \end{cases}$$

Equação 2 - *TF-IDF* [03]

4. Base de documentos pronta para consulta:

- Não utilizamos persistência com os índices invertidos. Criamos na inicialização do sistema um mapeamento *hash* em memória, onde a *string* que indica o termo é o valor e o *list* é uma lista com informações referentes aos termos nos documentos onde o mesmo ocorre. Ex: "*map<string*, *list<type*> >".
- Este type pode indicar qualquer informação sobre os termos nos documentos onde ele ocorre, tais como: posição, freqüência. O *type* indica a freqüência do termo nos documentos.

5. Tipos de consultas que o sistema consegue/pretende processar

Nosso foco é o processo de agrupamento de documentos. Podemos efetuar uma consulta posterior ao agrupamento, para saber se a qual grupo de documentos o documento desejado pertence, por exemplo.

6. Próximos passos

a) Vamos fazer o *F-Measure* [02], para o corpus de teste;
b) Alterar certas variáveis como o *bias* da seleção de atributos, critério de escolha na divisão dos clusters, para melhorar o *F-Measure*;
c) Escrita do relatório final.

7. Referências

[01] Liu, Tao; Liu, Shengping; Chen, Zheng; Ma, Wei-Ying. *An Evaluation on Feature Selection for Text Clustering*. 2003.

[02] Steinbach, Michael; Karypis, George; Kumar, Vipin. *A Comparison of Document Clustering Techniques*. 2000.

[03] Ribeiro-Neto, Berthier & Baeza-Yates, Ricardo. *Modern Information Retrieval*. Addison Wesley. ACM Press. 1999.

[04] Geraci, Filippo; Pellegrini, Marco; Maggini, Marco; Sebastiani, Fabrizio. *Cluster Generation and Cluster Labelling for Web Snippets: a Fast and Accurate Hierarchical Solution*. 2006.

[05] Larsen, Bjornar & Aone, Chinatsu. *Fast and Effective Text Mining Using Linear-time Document Clustering*. 2002.

[06] Tang, Bin; Shepherd, Michael; Milios, Evangelos. *Comparing and Combining Dimension Reduction Techniques for Efficient Text Clustering*. 2005.

[07] Hyvärinen, A & Oja, E. *Independent component analysis: algorithms and applications*. Neural Networks 13 (2000) 411–430.

[08] DMOZ. *Open Directory Project.* <www.dmoz.org>. Acesso em: 25 out 07.

[09] Stop Words List. *Pagex*. <www.pagex.com/webtools/stopwords.cfm>. Acesso em: 11 nov 07.

[10] The Oleander Stemming Library. *Oleander Solutions*. <www.oleandersolutions.com/stemming.html>. Acesso em: 11 nov 07.

[11] Barros, Flavia Almeida. Slide de Aula. *cap2-1.ppt*. *Recuperação Inteligente de Informação*. CIn-UFPE.