Recuperação Inteligente de Informação

Equipe: André Câmara
Valmir Macário

Descrição do projeto:

O projeto consiste em estender um sistema híbrido para extração de informações de referências bibliográficas. A abordagem utilizada combina técnicas convencionais de classificação de texto com a técnica Hidden Markov Models (HMM).

Este projeto consiste em adicionar mais um classificador ao sistema, o classificador Support Vector Machines (SVM), bem como realizar novos experimentos utilizando apenas HMM para extração de informação de forma isolada.

Arquitetura do sistema:

O processo de extração definido pode ser visto na Figura XXX, tendo sido dividido em duas fases: (1) Extração inicial utilizando as técnicas de classificação para Extração de Informação e (2) Refinamento da saída da fase 1 utilizando um HMM. A primeira fase pode ainda ser subdividida em mais duas etapas intermediárias: divisão do texto em fragmentos e determinação da classe de cada fragmento encontrado. Cada fragmento do texto deve ser representado por um vetor de características, através do qual o classificador estimará a probabilidade de o fragmento apresentado preencher corretamente um dado campo do formulário (Autor, título, conferência, etc). No caso das referências bibliográficas do corpus utilizado neste trabalho, a geração dos fragmentos pode ser feita separando-se o texto de entrada a cada vírgula ou ponto encontrado.

Para a extração de características são utilizadas combinações de três conjuntos de características, sendo dois manualmente definidos em trabalhos anteriores de Nunes (1999) e de Bouckaer et al (2002), e um aprendido automaticamente a partir de uma técnica de seleção de característica utilizada sobre as palavras do texto.

Uma breve descrição destes conjuntos é dada a seguir, maiores informações podem ser obtidas em [referencia a tese]:

1. Manual1: Este conjunto de características foi definido manualmente por Nunes (1999) para o sistema Prodext, sendo o resultado de um trabalho de engenharia de conhecimento. Ele apresenta características específicas do domínio das referências bibliográficas, como a presença de nomes de editoras em um determinado fragmento.

2. *Manual2*: Este conjunto foi definido manualmente por Bouckaert et AL (2002), mas contém características não tão especificas ao domínio das referências 65 bibliográficas como as do conjunto Prodext. Entre as características definidas, estão a presença de números no fragmento e se o fragmento possui palavras que começam com letras maiúsculas.

3. *Automático:* Este conjunto de característica é formado pelas palavras encontradas em um corpus de treinamento, podendo ser aprendido automaticamente. Como o número de palavras distintas existentes no corpus de treinamento é muito grande, utilizamos o método de seleção de características denominado *Information Gain* [Yang & Pedersen,

1997] para escolher as palavras que serão usadas na classificação. Segundo Sebastini (1999) e Aas & Eikvill (1999), os sistemas de classificação de textos normalmente fazem a remoção automática de palavras das *stopwords.* Estas são palavras muito comuns, como alguns verbos, artigos e preposições e em geral não trazem muita informação sobre a categoria a qual pertence um documento. No nosso caso, resolvemos não eliminar essas palavras, pois acreditamos que a presença de algumas *stopwords,* como artigos e preposições, pode trazer informações que auxiliem a diferenciar, por exemplo, um fragmento título de um fragmento que pertença a classe autor ou número.

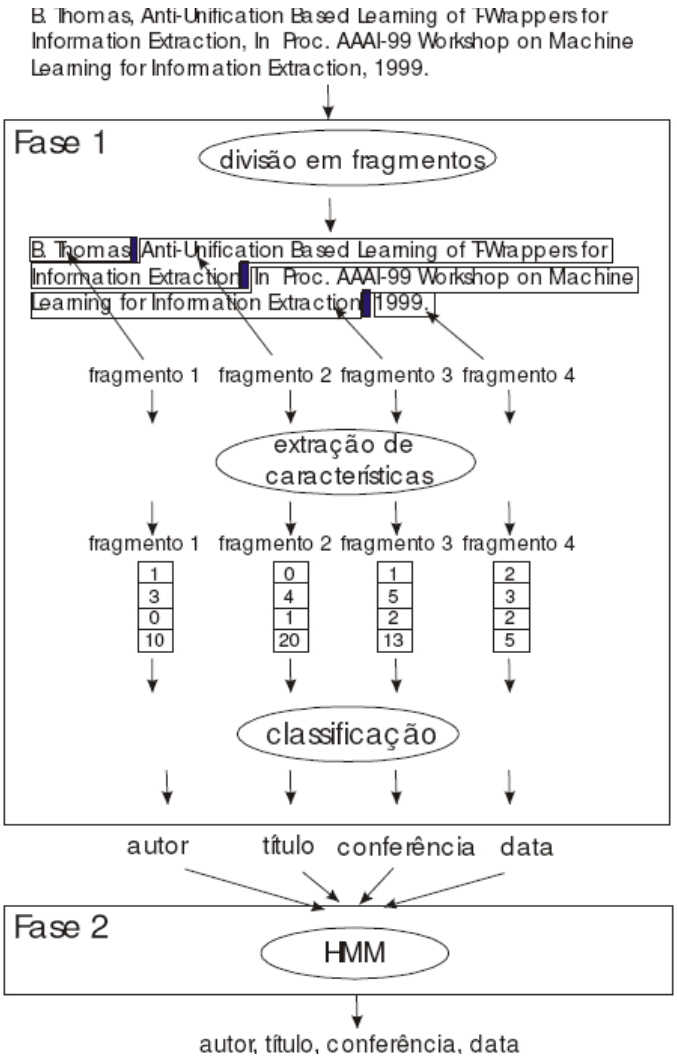

Figura 1. Etapas realizadas pelo sistema (Fonte: [SILVA, 2004])

Diversos classificadores são usados na tarefa de classificação de textos. São eles: o *Naive Bayes* [DUDA et al, 2001], as regras de classificação geradas pelo algoritmo PART [DUDA et al, 2001] e um classificador baseado em instâncias, o k-NN [DUDA et al, 2001]. A API disponibilizada pela ferramenta WEKA [DUDA et al, 2001] foi utilizada para a implementação desses classificadores.

Neste projeto será adicionado mais um classificador ao sistema, o Support Vector Machines (SVM), também utilizando-se a API disponibilizada pelo WEKA.

A fase de refinamento dos resultados obtidos na fase 1 é realizada através de um HMM que recebe o resultado da classificação individual de todos os fragmentos do texto e realiza uma nova classificação simultânea para todos eles. A Figura 4.6 ilustra como funciona esta fase. Este HMM foi treinado exclusivamente a partir da saída do classificador da fase anterior.

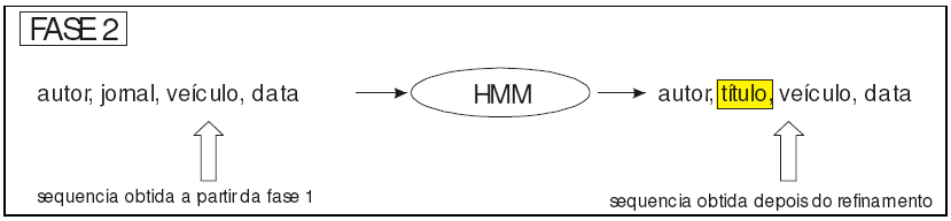

Figura 2. Processo de refinamento da fase 1 (Fonte: [SILVA, 2004]).

Base de dados:

O corpus de documentos a ser utilizado nos experimentos é o *Bibliography on computational linguistics, systemic and functional linguistics, artificial intelligence and general linguistics,* composto de 6000 referências bibliográficas no formato Bibtex. As bibliografias neste formato possuem tags que indicam a classe a qual cada subcadeia do texto pertence.

Referências:

[Aas & Eikvil, 1999] Aas, K. & Eikvil, L., Text Categorization: a survey. Technical Report #941, Norwegian Computing Center, 1999.

[Bouckaert, 2002] Bouckaert, R. R., Low level information extraction: a Bayesian network based approach. TextML 2002.

[DUDA et al, 2001] Duda, R.O.; Hart, P.E.; Stork, D.G. Pattern Classification. John Wiley & Sons, 2001.

[NUNES, 2000] Nunes, C. C. R., ProdExt: Um Wrapper para extração de produção técnica e científica de páginas eletrônicas. Dissertação mestrado, Centro de Informática da Universidade Federal de Pernambuco, Recife, Brasil, 2000.

[Sebastini, 1999] Sebastiani, F., Machine learning in automated text categorization, Tech. Rep. IEI-B4-31-1999, Consiglio Nazionale delle Ricerche, Pisa, Italy, 1999

[SILVA, 2004] Silva, E. F. A., Um Sistema para Extração de Informação em Referências Bibliográficas Baseado em Aprendizagem de Máquina. Dissertação mestrado, Centro de Informática da Universidade Federal de Pernambuco, Recife, Brasil, 2004.

[YANG & PEDERSEN, 1997] Yang, Y., Pedersen, J., O., A comparative study on feature selection methods in text categorization, in Proc. of the 14th International Conference on Machine Learning, ICML97, 1997.