MinTI: Mineração de Texto e Imagem para Recuperação Inteligente de Documentos do CPDOC-FGV

Leonardo Silva Kury[1], Aristófanes Corrêa[1], Anselmo Paiva[1,] Asla M. Sá[2], Moacyr Silva[2], Paulo Cezar Carvalho[2].

[1]Universidade Federal do Maranhão (UFMA)
Núcleo de Computação Aplicada (NCA) - São Luís – MA – Brasil
[2]Fundação Getúlio Vargas (FGV)
Centro de Matemática Aplicada (CMA) - Rio de Janeiro – RJ – Brasil

leokury@gmail.com, ari@dee.ufma.br, paiva@deinf.ufma.br, {asla.sa, moacyr.silva, pcezar}@fgv.br

Abstract. *CPDOC-FGV (Centro de Pesquisa e Documentação de História Contemporânea do Brasil) presently hosts a large set of texts and images, totalizing 1,8 million documents, mostly digitalized. Traditional types of search are inefficient and moreover, the data base grows faster than the human capacity to categorize and analyse the information. In this context, CMA-FGV (Centro de Matemática Aplicada), in partnership with NCA-UFMA (Núcleo de Computação Aplicada), agreed to develop a software tool called MinTI, supported by recent advances in text and image mining. Our goal is to create a tool for intelligent information retrieval from CPDOC-FGV database. By automatizing the process of information extraction we hope to improve the potentialities of the database.*

Resumo. *O Centro de Pesquisa e Documentação de História Contemporânea do Brasil (CPDOC-FGV) tem atualmente conjuntos documentais que totalizam cerca de 1,8 milhões de documentos. As formas tradicionais de busca em acervo se tornam ineficientes à medida que a base de dados cresce numa velocidade que ultrapassa a capacidade humana de catalogação e análise da informação. Neste contexto, o Centro de Matemática Aplicada (CMA-FGV), em parceria com o Núcleo de Computação Aplicada (NCA-UFMA), estão pesquisando e desenvolvendo a ferramenta computacional MinTI, baseada em tecnologia recente desenvolvida nas áreas de mineração de textos e imagens, para recuperação inteligente de informações da base de dados do CPDOC-FGV, visando acelerar e automatizar o processo de extração de informações relevantes para melhorar o aproveitamento do potencial da base de dados.*

1. Introdução

O Centro de Pesquisa e Documentação de História Contemporânea do Brasil (CPDOC-FGV) tem como um de seus objetivos abrigar conjuntos documentais relevantes para a história recente do país. Os conjuntos documentais do CPDOC-FGV hoje totalizam cerca de 1,8 milhões de documentos. Com o passar dos anos, o CPDOC-FGV tem

armazenado cada vez mais informações (texto, áudio, imagem e vídeo) em suas bases de dados. A quantidade de informação armazenada aumenta diariamente e ultrapassa a habilidade técnica e a capacidade humana de interpretação dessa informação. Apesar do enorme valor desses dados, a sua organização atual não consegue aproveitar totalmente o material que está armazenado em sua base. A maior parte dessa informação encontra-se na forma textual descrita em linguagem natural e é atualmente recuperada utilizando o sistema de computação *Accessus*. Este sistema ainda não é dotado de recursos inteligentes de agrupamento e classificação de informações.

Nesse contexto, propusemos como colaboração do Centro de Matemática Aplicada (CMA-FGV), em parceria com o Núcleo de Computação Aplicada (NCA-UFMA), o desenvolvimento de uma ferramenta computacional baseada em mineração de textos e imagens para recuperação inteligente de informações da base de dados do CPDOC-FGV, visando a extração de informações relevantes para acelerar e automatizar o aproveitamento do potencial da base de dados.

Neste artigo, abordaremos brevemente o problema de Mineração de Textos na Seção 2. A Seção 3 mostra o protótipo inicial da ferramenta de mineração de texto e imagem, denominado *MinTI*, que está sendo desenvolvida em parceria pelo CMA-FGV e pelo NCA-UFMA. Na Seção 4 descreveremos o problema de mineração de imagens, no contexto do CPDOC-FGV, como trabalho futuro.

2. Mineração de Texto

A Mineração de Texto é o processo de obtenção de informações relevantes a partir de textos descritos em linguagem natural. Inspirado originalmente na mineração de dados, que consiste em extrair informação de banco de dados estruturados, difere desta pela natureza dos dados processados que são não-estruturados ou semi-estruturados (Hotho and Nürnberger, 2005). A Mineração de Textos combina técnicas de Banco de Dados, Inteligência Artificial, Aprendizado de Máquina e Processamento de Linguagem Natural entre outros. Alguns exemplos de aplicações da área de mineração de textos são: sumarização, classificação, agrupamento, tradução automática e extração de informações de textos.

O processo de mineração de textos é bastante complexo e pode ser subdividido nas seguintes etapas (Aranha, 2007):

- **Coleta** é a etapa inicial do processo em que é feita a aquisição dos dados. É de fundamental importância para o desenvolvimento que se obtenha dados de qualidade. Ao fim dessa etapa, espera-se que tenha formado uma base de dados textual, conhecido na literatura como *corpus*.
- No **pré-processamento**, os dados são preparados para as etapas seguintes com a finalidade de serem formatados e representados para o processamento das etapas seguintes. Dependendo dos dados, pode ser uma etapa demorada e consumir boa parte do cronograma do processo.
- A **indexação** é responsável por organizar os termos adquiridos de forma a facilitar o acesso e recuperação dos dados. Uma boa estrutura de indexação garante agilidade e rapidez ao processo de mineração.

- Na etapa de **mineração** são executados algoritmos, cálculos estatísticos e inferências com o objetivo de extrair automaticamente informações relevantes da base de dados.
- Finalmente, a **análise** dos dados é realizada para interpretação e visualização dos resultados obtidos.

No presente contexto, efetuaremos as etapas de coleta, pré-processamento e indexação sobre a base de dados do CPDOC-FGV. O foco de pesquisa deste trabalho está no estudo comparativo e aplicação de algoritmos de aprendizado de máquina e de outras áreas relacionadas durante a etapa de mineração.

3. *MinTI*

O protótipo inicial por nós desenvolvido, chamado *MinTI*, implementa nesta fase somente o módulo de mineração de texto. A fase atual o protótipo aceita dois tipos de entrada de texto: texto inserido diretamente na caixa de texto do programa, ou texto obtido a partir de um arquivo especificado pelo usuário (ver Figura 1).

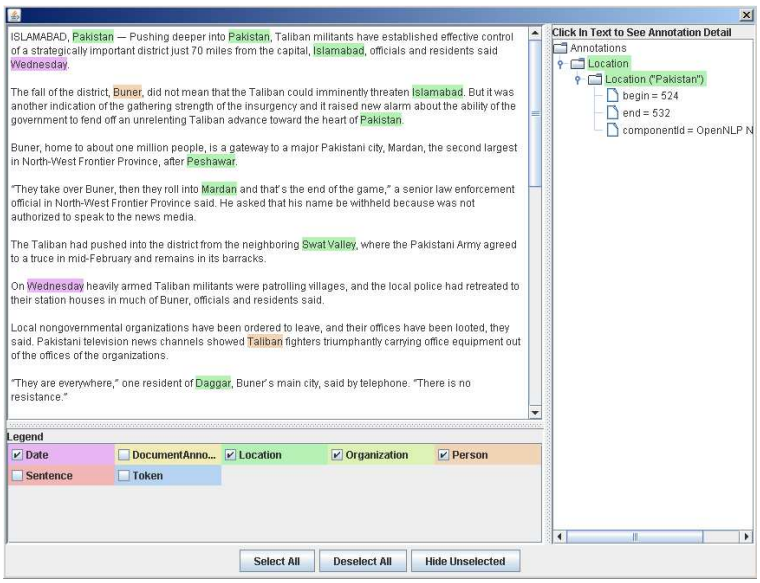

Figura 1 - Tela do *MinTI* exibindo o resultado da tarefa de NER para um artigo do *The New York Times*

O *MinTI* é baseado na arquitetura do padrão UIMA (*Unstructured Information Management Architecture*) da OASIS (*Organization for the Advancement of Structured Information Standards*). A implementação desta arquitetura utiliza o Apache UIMA - *framework* de código aberto desenvolvido em linguagem JAVA - que suporta o idioma português e o inglês. Para o idioma português o protótipo executa a tarefa de etiquetação de classes gramaticais (*Part-of-Speech Tagging*, ou simplesmente POS Tagging). Esta tarefa resume-se a atribuir uma classe gramatical da linguagem para cada *token* do texto. A tarefa executada no idioma inglês é o reconhecimento de entidades nomeadas (*Named Entity Recognition*, ou NER). Essa é uma subtarefa da extração de informações que visa localizar e classificar elementos atômicos do texto em categorias pré-definidas como nome de pessoas, organizações, locais, datas, etc. Essa diferenciação entre as tarefas dos

idiomas deve-se a falta de *corpus* anotado em português para o treinamento dos algoritmos.

A próxima fase de desenvolvimento do *MinTI* implementará a etapa de mineração propriamente dita. Em seguida passaremos a abordar a mineração de imagens.

4. Trabalhos Futuros: Mineração de Imagem

Um problema comum enfrentado pelos grandes arquivos de dados é a concentração de conhecimento nas pessoas que manipulam e/ou alimentam a base de dados. Uma vez que essa pessoa não está mais disponível perde-se a *expertise* de busca às informações arquivadas. No arquivamento de imagens de personalidades históricas este fato se torna especialmente crítico uma vez que uma determinada pessoa se "familiariza" aos personagens de um dado período e passa a reconhecê-los em fotografias ainda não catalogadas, tarefa que requer a habilidade particular deste indivíduo que outra pessoa não familiarizada não seria capaz de executar.

Para o CPDOC-FGV um procedimento de reconhecimento automático de personalidades em bancos de fotografias e vídeos seria de grande utilidade. O reconhecimento de face a partir de imagens fotográficas e imagens de vídeo está emergindo como uma atividade na área de pesquisa com numerosas aplicações comerciais. Estas aplicações requerem algoritmos robustos para reconhecimento de faces humanas sob diferentes condições de iluminação, expressões faciais e orientações. A mineração de imagens agrupa novos conceitos e tecnologias que englobam as áreas de processamento de imagens e vídeos e conceitos de aprendizado de máquinas. A mineração de imagens ocorre sobre os dados extraídos da imagem, essas características são utilizadas por modelos de mineração.

Referências

A Hotho, A Nürnberger (2005) A brief survey of text mining, LDV Forum-GLDV Journal for Computational Linguistics and Language Technology

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

MW Berry (2004) Survey of Text Mining: Clustering, Classification, and Retrieval, Springer

MW Berry, M Castellanos (2007) Survey of Text Mining II: Clustering, Classification, and Retrieval, Springer

CPDOC. Centro de Pesquisa e Documentação de História Contemporânea do Brasil, Fundação Getúlio Vargas. Sítio: http://www.cpdoc.fgv.br

ARANHA, C.N. (2007) Uma abordagem de pré-processamento automático para mineração de textos em português: sob o enfoque da inteligência computacional, Tese de Doutorado, Departamento de Engenharia Elétrica, PUC-Rio