DESENVOLVIMENTO DO WEBSERVICE EGENT – ESTUDO GRÁFICO DA EVOLUÇÃO DOS N-GRAMS NO TEMPO

GABRIEL L. A. **LIMA**[1]**;** LEANDRO H. M. de **OLIVEIRA**[2]

12616

RESUMO

Este trabalho apresenta um conjunto de ferramentas computacionais para o mapeamento da cobertura temporal de n-grams e termos de um domínio específico do conhecimento agrícola. A partir de um corpus textual formado por textos enriquecidos com metadados anotados automaticamente sobre a cultura de cana-de-açúcar tais ferramentas criam e apresentam a evolução do uso dos termos ao longo tempo. Além dos metadados, e usando o programa REMBRANDT também foram mapeadas a Entidades Mencionadas com classe de Tempo constantes nos conteúdos de texto para futuras análises semânticas a partir destas marcações. Os resultados são bastante satisfatórios e podem ser analisados sob o aspecto estratégico, cuja análise pode ser feita do ponto de vista do alcance dos objetivos propostos e seus trabalhos futuros, e outro mais prático cuja análise foca o ambiente operacional das ferramentas produzidas.

ABSTRACT

This paper presents a set of computational tools for mapping the temporal coverage of n-grams and terms of a specific domais of agricultural knowledge. From a corpus consisting of texts enriched with metadata automatically noted about he culture of cane sugar such tools create and present the use evolution of terms over time. In addition to metadata, and using the software REMBRANDT were also mapped to the Named Entity class Time constants in the text content for future semantic analysis from these markings. The results are quite satisfactory and can be analyzed under the strategic aspect, whose analysis can be done from the standpoint of achieving the proposed objectives and its future work, and a more practical analysis which focuses on the operational environment of the tools produced.

INTRODUÇÃO

O projeto TIENA, Tecnologias Inovadoras para poio à Espacialização de Notícias Agrícolas, tem como objetivo organizar um conjunto de notícias agrícolas sobre a cultura de cana-de-açúcar por meio de técnicas inovadoras em mineração de

[1]Estagiário: Graduação em Ciências de Computação, USP, São Carlos-SP, gabriell@.cnptia.embrapa.br

[2]Orientador: Pesquisador, EMBRAPA CNPTIA, Campinas-SP.

textos em três categorias: tópicos ou assuntos, cobertura temporal e cobertura espacial. A cobertura espacial visa compreender qual é a cobertura geográfica das notícias, podendo ser de caráter geral ou relativa a alguma micro ou macroregião geográfica, e, também futuramente poder-se-á analisar esse fator junto aos modelos de previsão de safra, para avaliar se o fator penaliza os modelos ou não. Já a cobertura temporal visa, em primeiro plano, estudar a evolução do uso dos termos de domínio, tópicos ou assuntos pertencentes a este universo em intervalos e ao longo do tempo, ou seja, como foi a tendência de dispersão destes ao longo do tempo, e, futuramente como tal evolução se relaciona aos modelos de previsão de safra.

Como objetivos específicos a serem atingidos no TIENA podemos enumerar (1) Fornecer uma forma de visualização sobreposta para as três categorias definidas anteriormente – tópico, tempo e espaço - e uma forma de visualização cruzando tais categorias; (2) Desenvolver um classificador semi-automático das notícias sobre uma cultura, com resultados gerados a partir de processos de aprendizagem, que, em tempo real, incrementa gradualmente os tópicos encontrados em presença de novas notícias; (3) Desenvolver um processo semi-automático que identifique a cobertura espacial de notícias sobre uma cultura, sempre em tempo real – atualizando-se os resultados na presença de novas notícias e (4) Desenvolver um processo semi-automático que identifique a cobertura temporal de notícias sobre uma cultura, sempre em tempo real – atualizando-se os resultados na presença de novas notícias.

Neste contexto, este trabalho ocupa-se no estudo e desenvolvimento de ferramentase para a solução informatizada do item 4 acima, e possui o seguintes objetivos específicos: (1) Extrair dados temporais de referência dos textos das notícias coletadas. Tais dados originam essencialmente em dois pontos: Nos metadados da notícia (por meio do acesso a um arquivo XML da notícia) e dentro do texto em si, que apresenta marcas temporais como dia, mês, ano e outras expressões; (2) Utilizar o reconhecimento de entidades mencionadas[3] para se extrair de dentro do texto entidades temporais, as quais se apresentam como menções a um determinado tempo, seja passado, presente ou futuro; (3) Utilizar o pacote NSP[4](BANERGEE, PEDERSEN, 2003) para extração de N-grams do texto. Tais N-grams serão então armazenados para que seja possível pesquisar por eles; (4) Mostrar, graficamente, a

[3]Por “Entidades Mencionadas” entende-se uma entidade referenciada num determinado contexto, podendo assim assumir papéis semânticos diferentes em função desse mesmo contexto.

[4]O NSP é um software que permite ao usuário identificar e contar N-Grams dentro de um grande corpus, sendo que o tamanho de N é definido pelo usuário. <http://www.d.umn.edu/~tpederse/nsp.html>

distribuição de uma palavra-chave ao longo do tempo, dado que tal palavra-chave corresponde a um N-gram e é fornecida como entrada pelo usuário juntamente com a janela de tempo e a periodicidade; e (5) Exibir, com o auxílio de um concordanciador[5], o trecho nas notícias onde a palavra-chave pesquisada foi encontrada, de forma a permitir ao usuário que veja o contexto onde aquele N-gram ocorreu.

MATERIAL E MÉTODOS

Para o mapeamento da cobertura temporal do termos de domínio foram necessárias a execução das seguintes atividades:

Compilação de um corpus de notícias sobre a cultura de cana-de-açúcar:

Para esta atividade, foi compilado um corpus textual em Português contendo 698 notícias sobre a cultura cana-de-açúcar datado de Janeiro a Abril de 2011. Os textos das notícias que compõem o corpus foram armazenados em dois formatos: (1) em texto puro, usado para as tarefas de processamento e (2) em XML, anotados automaticamente para conter metadados dos textos, como dados de título, fonte, data de publicação e corpo do texto, conforme pode ser visto no exemplo da FIGURA 1.

[5]Ferramenta para apresentação de concordâncias, que são um conjunto de extratos de textos nos quais as ocorrências de uma palavra (chamada de palavra-chave) são mostradas com seu contexto. São apresentadas em 3 formas: KWAC (*Keyword Alongside Context*); KWOC (*Keyword Out of Context*) e KWIC(*Keyword in Context*), a forma mais tradicional.

```xml
<?xml version="1.0" encoding="iso-8859-1"?>
<NOTICIA>
  <DATA>
    13/12/2010
  </DATA>
  <HORA>
    10:37:00
  </HORA>
  <TITULO>
    Petrobras planeja conter "estrangeiros" no alcool
  </TITULO>
  <FONTE>
    Folha de S. Paulo - SP
  </FONTE>
  <CORPO>
    A Petrobras tenta conter...
  </CORPO>
</NOTICIA>
```

FIGURA 1. Notícia no formato XML.

Vale notar que, mais adiante, cada uma das *tags* apresentadas será mapeada para um atributo de uma ou mais tabelas do banco de dados modelado para o *webservice,* de forma que é necessário ter um arquivo bem formado para a manipulação e extração dos metadados.

Estudo da ferramenta REMBRANDT[6]:

O Rembrandt(CARDOSO, 2008) é um sistema de reconhecimento de Entidades Mencionadas (EM) e de detecção de relações entre elas (DRE), em textos escritos em português. Atualmente, o Rembrandt se encontra na versão 1.3 beta e segue o manual do segundo HAREM[7] (MOTA, SANTOS, 2008) para anotar as EM. A ferramenta foi estudada com o com o intuito de avaliar se supriria a necessidade de detecção de referências temporais ocorridas no corpus de notícias. A FIGURA 2, retirada do segundo HAREM, ilustra as categorias, tipos e subtipos de entidades mencionadas.

[6]http://xldb.di.fc.ul.pt/Rembrandt/?do=download-release

[7]O HAREM é uma avaliação conjunta na área do reconhecimento de entidades mencionadas em português. É uma iniciativa que pretende avaliar o sucesso na identificação de classificação automática dos nomes próprios na língua portuguesa (http://www.linguateca.pt/HAREM/).

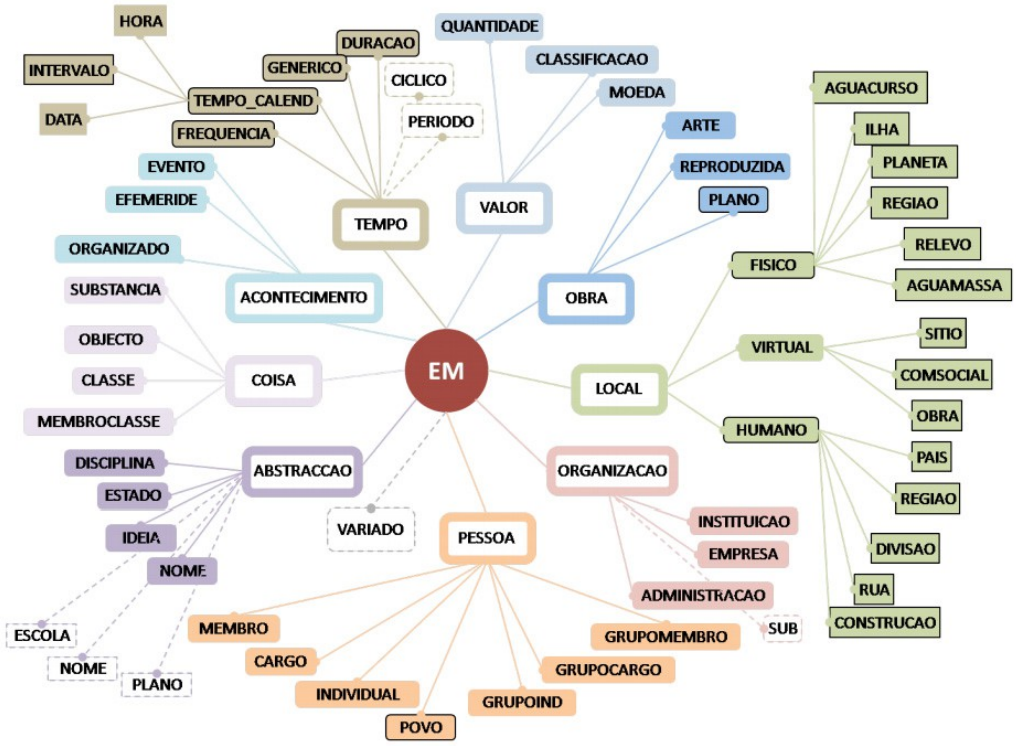

FIGURA 2. Classificação das entidades mencionadas de acordo com o segundo HAREM.

Com o conhecimento dos tipos de entidades que poderiam ser encontradas no texto, foi executado então o processo de detecção das entidades mencionadas em todo o corpus através do REMBRANDT. Para isso, após a execução do REMBRANDT sobre o coprus, foi escrito um *script* Groovy para separar, por notícia, as entidades mencionadas com a categoria TEMPO, tornando possível comparar as entidades da categoria TEMPO extraídas com a presença de referências temporais no texto, de forma a determinar o quão preciso é o REMBRANDT e o quanto as entidades temporais influenciariam no desenvolvimento do *webservice.*

Modelagem do banco de dados:

Levando em consideração que o sistema final deveria usar uma interface web, optou-se pelo uso de um banco de dados para armazenar os dados pré-processados. Para garantir a fácil interoperabilidade entre as diversas partes do projeto TIENA, o modelo do banco desenvolvido cobre não apenas as necessidades específicas deste trabalho, indo além e cobrindo as necessidades de todos os demais trabalhos que integram o projeto TIENA.

Extração dos N-grams:

O Pacote NSP foi usado para a extração dos N-grams, cuja execução mais simples é feita pelo fornecimento de um arquivo de entrada contendo texto (corpus),

dois arquivos com definições de *stoplists* e um terceiro arquivo com a regra de formação dos *tokens*[8].

Pode-se dividir a extração dos N-grams em duas etapas:

- Extração das Unigramas aplicando o uso das regras das lexias complexas(POTTIER, 1972) em adição com todos os Unigramas encontrados pela aplicação das regras de formação de *tokens* do português brasileiro.
- Extração dos demais N-grams apenas utilizando as regras de formação de *tokens* do português brasileiro.

O processo de extração foi aplicado em todo o corpus, sendo que para cada notícia foi gerado um arquivo em separado para cada tamanho N do N-gram, com N variando de 1 até 10. Na FIGURA 3 pode-se ver um exemplo de um arquivo gerado pelo NSP para N-grams de tamanho 2 (bigramas):

```
11
line<>of<>2 3 2
of<>text<>2 2 2
second<>line<>1 1 3
line<>and<>1 3 1
and<>a<>1 1 1
a<>third<>1 1 1
first<>line<>1 1 3
third<>line<>1 1 3
text<>second<>1 1 1
```

FIGURA 3. Exemplo de saída do NSP.

População do banco de dados:

Após o processamento do corpus, tanto pelo o REMBRANDT quanto pelo o NSP, foi criado um *script* Groovy para a população do banco de dados, podendo ser usado como tanto para a carga inicial dos dados como para futuros incrementos. Além disso, um arquivo de conFIGURAção determina quais etapas da população do banco devem ser ou não feitas, dizendo se existem novos textos para serem inseridos no banco, se será feita a detecção de entidades mencionadas e se serão extraídos os N-grams dos novos textos. Além disso, o *script* criado se encarrega, também, de extrair os dados contidos nos metadados de cada notícia, fazendo as transformações necessárias para o armazenamento de tais informações no banco de dados. Uma vez populado o banco, é possível incrementá-lo a qualquer instante, dado que os *scripts* responsáveis por processar os textos e inseri-los no banco foram feitos de forma a

[8]Entende-se por *token* um conjunto de caracteres com um significado coletivo.

evitarem duplicatas e também evitar processar o mesmo conteúdo mais de uma vez, otimizando o desempenho.

Criação do *Webservice* EGENT:

O Webservice criado foi batizado de EGENT que representa o acrônimo de *Estudo Gráfico da Evolução de N-Grams no Tempo*. O EGENT segue uma arquitetura *Model-View-Controller*[9] e foi desenvolvido utilizando a tecnologia Java Server Faces (JSF[10])e a biblioteca PrimeFaces[11], além do Hibernate[12] para persistência dos dados.

***Webservice* EGENT:**

A ferramenta criada permite ao usuário verificar a distribuição de ocorrências de um certo N-gram ao longo do tempo. Para tal análise, o usuário deve entrar com uma entrada composta por um intervalo de tempo, uma escala e uma palavra-chave. A FIGURA 4 exemplifica a entrada:

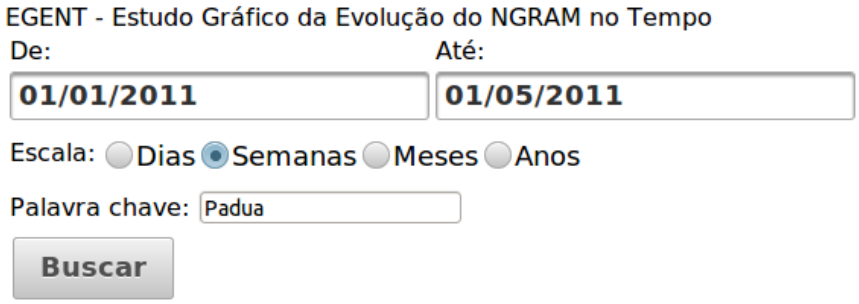

FIGURA 4. Interface de entrada do EGENT.

As FIGURAS 5, 6 e 7 mostram o exemplo de saída correspondente ao exemplo de entrada da FIGURA 4. As saídas apresentadas são, respectivamente, o gráfico da distribuição no tempo, a tabela contendo as possíveis classificações feitas pelo REMBRANDT e a tabela contendo os contextos onde a palavra-chave foi encontrada.

[9] http://www.macoratti.net/vbn_mvc.htm

[10] http://www.oracle.com/technetwork/java/javaee/javaserverfaces-139869.html

[11] http://primefaces.org/

[12] http://www.hibernate.org/

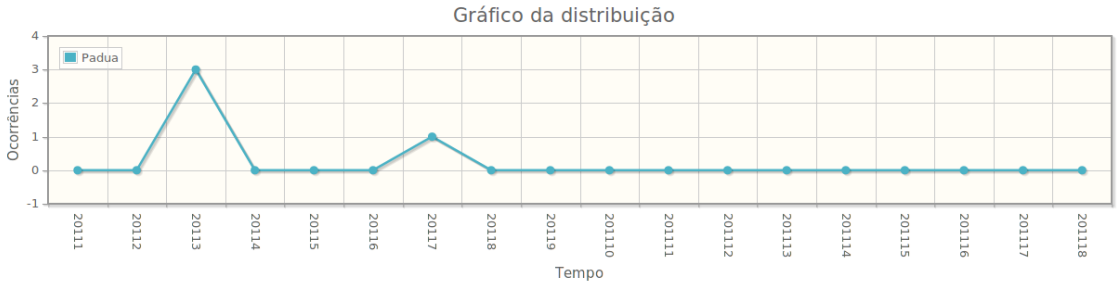

FIGURA 5. Primeira saída, gráfico da distribuição.

Categorização pelo Rembrandt

Categoria	Tipo	Subtipo
LOCAL	HUMANO	DIVISÃO

FIGURA 6. Segunda saída, tabela de entidades mencionadas.

Concordanciador

Texto
O setor sucroalcooleiro do centro-sul brasileiro deve terminar a safra 2010/11 com uma receita recorde de R$ 50 bilhões. O volume é 25% superior aos R$ 40 bilhões verificados na safra 2009/10, de acordo com o diretor técnico da União da Indústria de Cana-de-Açúcar (Unica), Antonio de Padua Rodrigues. Segundo ele, em relação à safra 2008/09, período no qual os efeitos da crise financeira mundial foram mais expressivos, com ganhos de R$ 33 bilhões, o avanço foi de 51,15% Os números não incluem impostos.

FIGURA 7. Terceira saída, contexto.

RESULTADOS E DISCUSSÃO

Os resultados deste trabalho podem ser analisados sob dois aspectos: um estratégico, cuja análise pode ser feita do ponto de vista do alcance dos objetivos propostos e seus trabalhos futuros, e outra mais prática cuja análise foca o ambiente operacional das ferramentas produzidas.

Sob o aspecto estratégico, este trabalho serviu como base para análise inicial quantitativa e qualitativa da cobertura temporal dos n-grams e termos de domínio sobre uma determinada área do conhecimento, principalmente no sentido que a mesma pode ser empregada para outras análises mais elaboradas que consideram esse tipo de cobertura como ponto de partida. Isso porque a partir dos resultados produzidos, foi possível verificar o comportamento temporal de um determinado n-gram de maneira muito flexível, inclusive com a possibilidade alteração na escala de tempo desejada e o tamanho do n-gram. Portanto, este é o principal resultado deste trabalho visto que preenche a lacuna da análise temporal para a previsão de safras pretendida pelo TIENA.

Sob o aspecto prático, este trabalho produziu bastante resultados do ponto de vista operacional, já que implementou diversas ferramentas computacionais de base que tornaram possível tanto a análise temporal quanto a análise espacial pretendidas.

Tais ferramentas são, por exemplo, (1) a que identifica e extrai os metadados do corpus e faz a carga do banco de dados; 2) a que executa o REMBRANDT sobre o corpus e armazena as EM extraídas para futuras referências; 3) a que extrai os n-grams do corpus em seus diversos tamanhos e faz sua associação com seu tempo de ocorrência e 4) o webservice EGENT, que utiliza todas essas informações para a análise da cobertura temporal. Juntas, e desde que se tenha um corpus anotado com anotações de tempo, tais ferramentas podem utilizadas para empreender a análise de cobertura temporal dos n-grams de qualquer área do conhecimento.

CONCLUSÃO

A partir do formato de arquivos proposto para os textos que compõem o corpus de notícias e com a extração dos metadados foi possível criar diversas formas de se exibir os dados contidos em uma notícia, atingindo com exito o objetivo facilitar o estudo de como um determinado termo ou n-gram se desenvolve ao longo do tempo e de que formas ele influencia os modelos de previsão de safra. A combinação das 3 saídas escolhidas permite ao usuário não apenas ver a distribuição dos n-grams ao longo do tempo, mas também ver de que forma tais n-grams ocorrem no texto, adicionando possibilidades de pesquisa as quais o usuário poderia não ter em mente previamente e, consequentemente, gerando maior conhecimento sobre um determinado assunto ou palavra-chave. Vale notar que os resultados obtidos foram satisfatórios mesmo utilizando um corpus relativamente pequeno, logo, espera-se obter resultados ainda melhores com um corpus mais abrangente.

Além disso, a exibição das entidades mencionadas detectadas pelo REMBRANDT auxiliará nos trabalhos futuros, uma vez que permitirá, mais facilmente, o estudo das classificações feitas pelo REMBRANDT dado uma palavra-chave de interesse. Isso porque uma notícia sobre a queda na safra de cana-de-açúcar, por exemplo, pode mencionar que houve quedas em um período diferente no ano passado, entretanto, não é um indício de que o N-gram dado como entrada pelo usuário tenha ocorrido em notícias no período indicado pela menção temporal, ficando para trabalhos futuros a análise mais a fundo a possível influência das entidades temporais no indício da ocorrência de um N-gram.

AGRADECIMENTOS

A Embrapa Informática Agropecuária (CNPTIA), pela oportunidade de estágio.

REFERÊNCIAS

BANERJEE, S.; PEDERSEN, T.; *The Design, Implementation and Use of the N-gram Statistics Package*. Carnegie Mellon University, Pittsburgh, USA, 2003 13p. Disponível em: <http://cpansearch.perl.org/src/TPEDERSE/Text-NSP-1.13/doc/cicling2003.pdf >

CARDOSO, N.; REMBRANDT – Reconhecimento de Entidades Mencionadas Baseado em Relações e Análise Detalhada do Texto. In: MOTA, C.; SANTOS, D.; Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo Harem, 2008, Cap. 11, p. 195-211.

MOTA, C.; SANTOS, D.; Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM. Linguateca, 2008. Disponível em: <http://www.linguateca.pt/HAREM/actas/LivroSegundoHAREM.html>

POTTIER, B.; *Grammaire de l'espagnol*, 1972, França.