APLICAÇÃO DO *TEXT MINING* PARA INCORPORAÇÃO DE INFORMAÇÕES SÓCIO-ECONÔMICAS EM SISTEMAS OBJETIVOS DE PREVISÃO DE SAFRA

LAURIMAR G. VENDRUSCULO[1], FABIO R. MARIN[2], BERNARD BARBARISI[3], FELIPE G. PILAU[4]

[1] Engª. Eletricista, Pesquisador II, Embrapa Informática Agropecuária, Caixa Postal: 6041 - CEP: 13083-970 - Campinas – SP, Fone (0xx19) 3789-5733, *laurimar@cnptia.embrapa.br*
[2] Engº. Agrônomo, Pesquisador III, Embrapa Informática Agropecuária, Campinas – SP
[3] Graduando em Engenharia Ambiental , Embrapa Informática Agropecuária, Campinas – SP
[4] Engº. Agrônomo, Bolsista CNPq/CONAB, Embrapa Informática Agropecuária, Campinas – SP

Escrito para apresentação no

XXXV Congresso Brasileiro de Engenharia Agrícola

31 de julho a 4 de agosto de 2006 - João Pessoa – PB

RESUMO: Vários estudos acadêmicos e esforços governamentais têm sido empreendidos para predizer, com confiança, a área plantada e a produtividade, no intuito de estimar oficialmente as safras agrícolas brasileiras. A estimativa oficial é baseada em levantamentos sistemáticos, por município, com informação colhida através de entrevistas em estabelecimentos rurais e outros setores organizados da sociedade. É importante, contudo, que outros fatores sejam considerados para a consolidação dos números regionais, estaduais e nacionais, especialmente, os fenômenos climáticos, condições para o manejo das lavouras, ocorrência generalizada de pragas e doenças. Sob esta ótica, o presente estudo apresenta a técnica de mineração de textos para incorporação de fatores sócio-econômicos no processo de previsão de safras. Estes fatores foram analisados no contexto de notícias jornalísticas por meio do software Eurekha, que possibilitou formar agrupamentos com índice de similaridades aceitáveis.

PALAVRAS-CHAVE: INFORMAÇÃO NÃO ESTRUTURADA, MINERAÇÃO TEXTO, PREVISÃO SAFRAS

TEXT MINING APPLICATION FOR INCORPORATION SOCIOECONOMIC INFORMATION IN OBJECTIVE SYSTEMS OF HARVEST FORECAST

ABSTRACT: Several academic studies and governmental efforts have been undertaken to predict, with trust, the planted area and the productivity, in the intention of esteem the Brazilian agricultural harvests officially. The official estimate is based on systematic data, for municipal district, with information picked through interviews by rural establishments. However factor as public politics, climatic phenomena, crop management, diseases, pests, and others have to be considered in the global calculation of the harvests. Under this optics, the present study presents the technique of text mining for incorporation economic factors in the process of harvests forecast. These factors were analyzed in the context of journalistic news through the software Eurekha. This tool formed groupings with index of acceptable similarities

KEYWORDS: NON-STRUCTURED INFORMATION, TEXT MINING, HARVEST FORECAST

INTRODUÇÃO: Continuados esforços acadêmicos e governamentais têm sido empreendidos para predizer a área plantada e a produtividade (quantidade produzida por unidade de área), no intuito de estimar oficialmente as safras agrícolas brasileiras. No Brasil, o Instituto Brasileiro de Geografia e

Estatística (IBGE) e a Companhia Nacional de Abastecimento – CONAB, são os responsáveis governamentais pela aferição e divulgação de informação sobre predição de safras atualmente. O IBGE ao longo de seus anos de existência tem mantido um levantamento sistemático da produção agrícola, com dados obtidos de forma subjetiva, por meio de consulta a especialistas, por município, com um censo agropecuário, de periodicidade variável, com informação colhida, através de entrevistas por estabelecimento rural. Com o objetivo de aprimorar o sistema de estimativas das safras agrícolas brasileiras, foi instituído, em 2003, o projeto GeoSafras, sob coordenação da CONAB e participação de uma vasta rede multi-institucional (Figueiredo, 2006). Está previsto no GeoSafras o uso de geotecnologias bem como a aplicação de modelos agrometeorológicos e espectrais para prognósticos de rendimento e estimativa de área plantada. No entanto, Figueiredo (2006) preconiza o uso que outros fatores a serem considerados no processo de estimativa de safras, tais como: ataque de pragas e doenças; dispersão e variação da dimensão das áreas de cultivo; lavouras consorciadas; rotação de culturas e outros. Sob esta ótica, o presente estudo apresenta a técnica de mineração de textos para incorporação de fatores sócio-econômicos no processo de previsão de safras. Estes fatores serão analisados no contexto de notícias jornalísticas e podem ser utilizados como fator de ponderação no cálculo global da produção.

MATERIAL E MÉTODOS: Foram analisadas 139 notícias jornalísticas, escritas em português, de domínio público na Internet, por meio de sites de jornais ou instituições ligadas ao agronegócio, no período de dezembro de 2005 a fevereiro de 2006. As notícias foram adquiridas semanalmente, de maneira manual, e selecionadas aquelas que narravam fatos que tiveram impacto na produção agrícola em função de fatores climáticos, políticas públicas, ocorrência de pragas e doenças e outros. Em um primeiro estudo, as notícias foram associadas, uma a uma, por um analista de domínio, ao estado de abrangência da notícia, nome da cidade, e respectivas coordenadas geográficas, nome da cultura e classificação do fator impactante na produção. Como fatores positivos à produção utilizou-se: investimentos e clima favorável, e como fatores negativos: falta de água, granizo, doença e praga. Todos os atributos foram consolidados em uma planilha eletrônica e posteriormente importados por um sistema de informação geográfica. Por meio desta ferramenta a notícia foi espacializada permitindo que uma análise mais acurada fosse feita. Utilizou-se parte da técnica CRISP-DM (CRoss-Industry Standard Process for Data Mining) (SHEARER, 2000) para descoberta de padrões de texto que auxiliariam na descoberta de conhecimento implícito das notícias jornalísticas. As etapas que compõem o CRISP-DM são: (*i*) compreensão do negócio; (*ii*) compreensão dos dados; (*iii*) preparação dos dados; (*iv*) modelagem; (*v*) avaliação; e (*vi*) aplicação. O estudo utilizou o software livre Eurekha 2.0 (Personal Edition), disponível em: http://www.leandro.wives.nom.br/eurekha/eurekha.htm, desenvolvido por WIVES (1999). O Eurekha permite a submissão de textos à análise de algoritmos de *clustering* (*best-star, cliques, full-star, stars*), que corresponde a etapa de modelagem do CRISP-DM. A ferramenta proporcionou a obtenção do conhecimento (padrões, relacionamentos) com base no agrupamento de notícias com as mesmas características de similaridade. Cada notícia foi separada em arquivo do tipo texto, conforme ilustra a Figura 1.

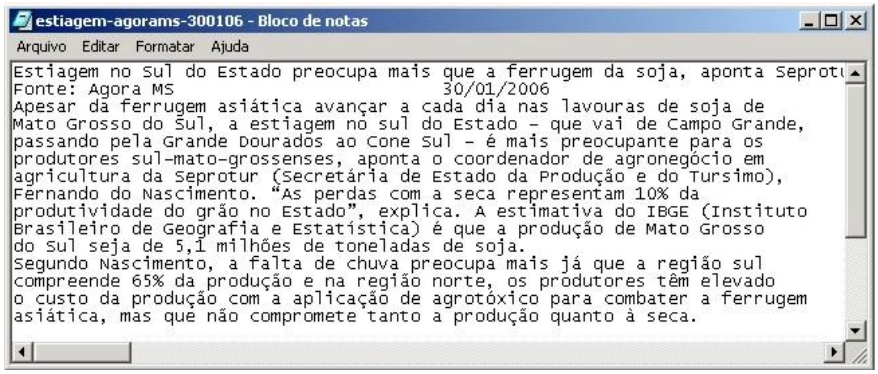

Figura 1 – Formato padrão dos conteúdos textos.

O uso do Eurekha permitiu o agrupamento distinto de notícias que tinham mesma ocorrência de palavras. Para tanto, o software calcula uma matriz com os índices de similaridades entre os arquivos-

textos. Para cada grupo, o software gera os chamados centróides ou palavras-chave. Estes centróides dão uma idéia sobre o assunto discutido por determinado conjunto de notícias. Exemplo: Cluster [24] – sul soja mato grosso. Nesta situação, os textos agrupados versaram, em sua maioria, sobre notícias relativas à produção de soja nos estados do Mato Grosso e Mato Grosso do Sul. As 4 notícias relativas ao agrupamento 24 são sumarizadas na Tabela I.

Tabela I – Meta-informação sobre as notícias do agrupamento 24

Nº. Notícia	Título	Fonte	Data publicação
1	Estiagem no Sul do Estado preocupa mais que a ferrugem da soja, aponta Seprotur.	Agora MS	30/01/2006
2	Ferrugem asiática da soja chega a 638 focos no país	Folha do Estado	31/01/2006
3	Perda na safra de soja já pode chegar a 15% devido à estiagem	Campo Grande News	16/01/2006
4	Chuvas dão trégua em MT, colheita da soja precoce é retomada.	A Notícia Digital	18/01/2006

RESULTADOS E DISCUSSÃO: Com o valor 2% para o coeficiente de sensibilidade, e a escolha do algoritmo *best-star*, foram encontrados 48 grupos de notícias. Em função da importância da produção da commoditie soja na economia nacional, este estudo analisou o percentual de notícias que versavam sobre esta cultura e as diversas facetas da sua produção. Foram encontradas 59 notícias, onde um dos centróides foi a palavra soja. Este valor correspondeu a 42 % da população de notícias analisadas. Procurou-se enfatizar as informações naqueles estados brasileiros cuja produção desta cultura é significativa em relação a produção nacional e optou-se pelo cluster 24. Este agrupamento reuniu 4 notícias, conforme ilustra a Figura 2, e apesar dos centróides indicarem apenas a localização da produção, estava implícito nas notícias, os motivos da quebra da produção, quais sejam: a influência da ferrugem asiática, o impacto da estiagem no Mato Grosso do Sul e excesso de chuvas no norte do Mato Grosso durante o período analisado. Apesar da matriz de similaridade apresentar valores poucos significativos (valores ideais seriam próximo a 1), as notícias do cluster 24 possuem valores razoáveis pelo número de notícias avaliadas. A Figura 2, mostra o valor de 0,1098 da notícia 1 em relação a notícia 2. O maior valor de similaridade neste grupo foi de 0,1205 entre a notícia 1 e 3.

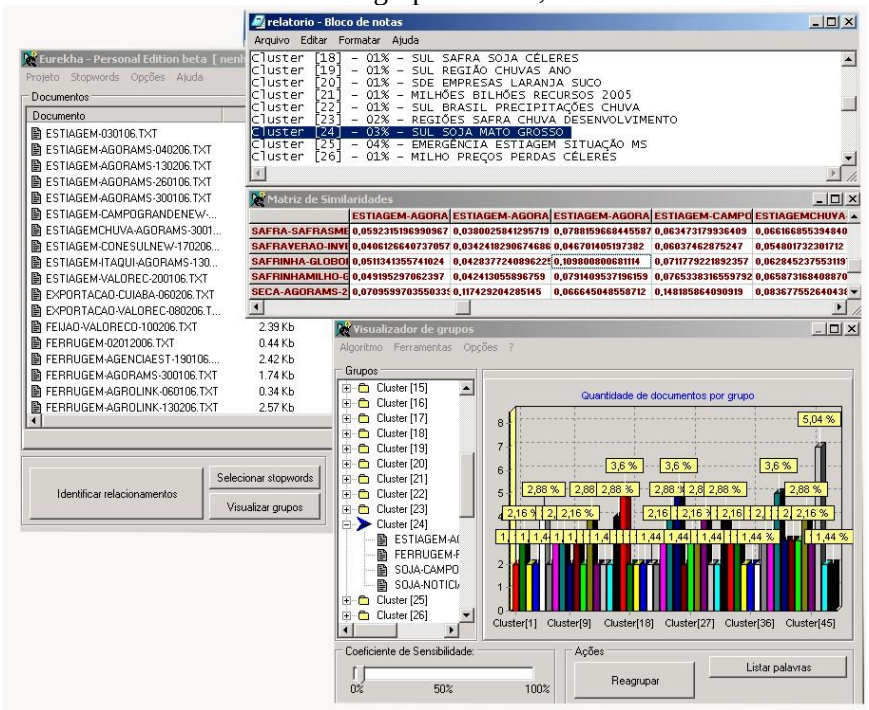

Figura 2 – Interface do software Eurekha enfatizando informações sobre a produção de soja nas notícias analisadas.

CONCLUSÕES: Os agrupamentos encontrados pelo software Eurekha mostraram-se coerentes em termos de conteúdo semântico. Em relação ao agrupamento que descreve a produção de soja nos Estados do Mato Grosso e Mato Grosso do Sul, as notícias mostram forte indicação da tendência de penalização da produção em função da estiagem, ferrugem asiática e chuvas em excesso. A análise também revelou o grande destaque que a cultura de soja teve na mídia no período analisado. Porém, ainda não foi possível inserir regras mais complexas para o cômputo do percentual de penalização da produção em escala regional. Para tanto, novos algoritmos serão testados e integrados com um sistema computacional para estimativa de produtividade que utiliza o método agrometeorológico, em desenvolvimento na Embrapa Informática Agropecuária. O método agrometeorológico enfatiza o grau de penalização sobre o rendimento da cultura face às condições climáticas nos períodos críticos do desenvolvimento vegetativo da planta e seu modelo pode ser acrescido de fatores (deflacionários ou inflacionários) que contemplem informações sôcio-econômicas. Um sistema mais completo que analise este tipo de informações deve também considerar outros indicadores como manejo da cultura, infestação de pragas, controle fito-sanitários, comercialização de adubos e outros.

REFERÊNCIAS BIBLIOGRÁFICAS

FIGUEIREDO, D. S. **Projeto *GeoSafras* – Aperfeiçoamento do Sistema de Previsão de Safras da Conab** Disponível em: http://www.conab.gov.br/download/GeoSafras/Manuais/projetogeosafras.pdf Consultado em 14 mar. 2006.

SHEARER, COLIN. **The CRISP-DM Model: The Blueprint for data mining**. Journal of Data Warehousing. V. 5, No. 4, p. 13-22, 2000.

WIVES, L.K. **Um estudo sobre agrupamento de documentos textuais em processamento de informações não estruturadas usando técnica de clustering**. Dissertação de Mestrado, PPGC/UFRGS, Porto Alegre (RS), 1999.