Anais 1º Simpósio de Geotecnologias no Pantanal, Campo Grande, Brasil, 11-15 novembro 2006,
Embrapa Informática Agropecuária/INPE, p.566-574.

Aplicação da técnica de *Text Mining* e espacialização de informações sócio-econômicas em sistemas objetivos de previsão de safra para a região da bacia hidrográfica do Pantanal

Laurimar Gonçalves Vendrusculo
Fábio Ricardo Marin
Felipe Gustavo Pilau
Ludmila Roque Ferraz Pacheco

Embrapa Informática Agropecuária - CNPTIA
Av. André Toselo, 209 - Caixa Postal 6041
13083-886 – Campinas, SP, Brasil
laurimar@cntpia.embrapa.br
marin@cnptia.embrapa.br
fgpilau@yahoo.com.br
ludmila@cnptia.embrapa.br

***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 de texto, previsão safras, banco dados, Google Earth.

***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, Database, Google Earth.

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, por meio 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. As geotecnologias utilizam conhecimentos formalizados, por exemplo, por meio de banco de dados climáticos, imagens de satélite, gráficos e outros. Esse conhecimento é também denominado explícito, segundo Nonaka & Takeuchi (1997).

O outro tipo de conhecimento é chamado tácito, o qual não está formalizado e pode ser encontrado com as pessoas e não foi ou não pode ser transformado para representações rigorosas. Nas organizações, os pensamentos, idéias, sentimentos e opiniões das pessoas estão disponíveis na forma de sugestões e reclamações de clientes, manuais, memorandos, notícias, etc. Um desafio constitui-se em transformar, de maneira automática, o conhecimento tácito em explícito. Para tanto, a técnica de *Text Mining* ou mineração de texto, oferta algumas ferramentas, que dentre outras funcionalidades, separam documentos em grupos por assunto ou afinidade (ferramentas de classificação e agrupamento). A dificuldade, neste caso, é lidar com a rica gama de informações, em geral sem estruturada e, assim, de difícil recuperação automática.

Posterior a fase de classificação/agrupamento pode-se ainda combinar este conhecimento com o conhecimento explícito armazenado em bancos de dados estruturados.

Os modelos científicos que apoiam os prognóstico da produção agrícola não dispensam o conhecimento tácito que podem ser encontrados em notícias jornalísticas, as quais constituem fonte alternativa para a localização de outros fatores a serem considerados no processo de estimativa como preconiza Figueiredo (2006). Dentre estes fatores, o autor cita ataque de pragas e doenças; dispersão e variação da dimensão das áreas de cultivo; lavouras consorciadas; rotação de culturas e outros

Este estudo propõe o entendimento de notícias jornalísticas e textos ligados a previsão de safras, utilizando como estudo de caso a região do Pantanal.

A participação total dos estados de Mato Grosso e Mato Grosso do Sul na produção de cereais, leguminosas e oleagionosas, segundo o Levantamento Sistemático da Produção Agrícola - julho de 2006 - IBGE (2006), é de 16,69 % e 6,11% respectivamente. Apesar da contribuição significativa no cômputo da produção nacional, perdendo apenas para o estado de Santa Catarina (17,96%), ressalta-se que grande parte destes produtos não se originam da região do Pantanal.

Especificamente para o cenário pantaneiro, Silva et al. (2001) confirmam que as áreas de lavoura da Região do Pantanal, concentram-se maciçamente na Região de Planalto (área adjacente ao pantanal), a qual sedia 4/5 das áreas de lavoura da região alta. Esta situação reforça a inaptidão agrícola do Pantanal. Em 1985, o Estado do Mato Grosso teve a maior área de lavouras recenseadas (2.666,3 quilômetros quadrados), divididos em 322,6 km^2 no Pantanal e 2.343,7 km^2 no planalto, em relação aos dados censitários de 1975 e 1980. No Mato Grosso do Sul, a maior área recenseada ocorreu também em 1985, com 791,1 km^2 no Pantanal e 3.911,99 km2 no planalto. Itiquira, no Mato Grosso, em 1985, foi o município de maior área de lavoura recenseada (1.280,6 Km^2), representando 27,8% da região do planalto e Pantanal.

Para melhorar e ampliar o entendimento dos dados estruturados tem-se recorrido ao sistemas gerenciadores de banco de dados (SGBD) e ferramentas para visualização de dados georeferenciados.

Uma vez estruturados, a publicação dos dados, na forma de mapas, utiliza a Internet como meio de divulgação mais freqüente. Para tanto a empresa Google dispõe de serviços que suprem esta necessidade. São eles: Google Map e Google Earth.

O Google Earth (Google Earth, 2006) oferece, uma vasta coleção de imagens aéreas ou de satélites de todo o mundo, atualizadas com freqüência de alguns meses, onde é possível associar informações como fotografias, rotas, dados topográficos e estatísticos e outros. O inconveniente para o uso desse serviço é exigência de conexão em banda larga à Internet, pois a cada interação, o programa-cliente solicita dados de imagens ao site da Google. Para construir uma aplicação personalizada no Google Earth é necessário criar arquivos do tipo KML (Keyhole Markup Language). Esta linguagem compreende a gramática XML e permite a modelagem e armazenamento de aspectos geográficos de pontos, linhas, imagens, polígonos.

Objetivo

Este estudo tem como objetivo a estruturação de informações sócio-econômicas provenientes de notícias jornalísticas na forma de base de dados e respectiva aplicação da técnica de mineração de textos para descoberta de conhecimento. Como estudo de caso, serão utilizadas notícias relacionadas à região do Pantanal MatoGrossense. Para esse estudo a incorporação das informações sócio-econômicas visa melhorar o processo de previsão de safras, onde os resultados experimentais funcionaram como fator de ponderação no cálculo global da produção.

Material e Métodos

Foi realizada a estruturação dos dados utilizando a técnica de modelagem entidade-relacionamento por meio do *software* DBDesigner, conforme ilustra a **Figura 1**. Os dados foram provenientes de matérias jornalísticas de domínio público na Internet.

O sistema gerenciador de banco de dados livre MySQL Server 5.0 (MySQL AB, 2006). implementou a base modelada. A base construída tem por objetivo melhorar a qualificação

das notícias, agregando-as ao veículo jornalístico ou institucional de onde se originaram, quais culturas abrangidas, o fator impactante de quebra ou aumento de safra e outros.

O analista de domínio, interagiu com base de dados relacionando as cidades ou regiões contempladas pelas notícias. Nesta etapa explicitou-se o georeferenciamento da informação, permitindo que uma notícia possa ser relacionada a várias instâncias de cidades ou regiões. Neste caso, o banco de dados consegue responder, textualmente, a questões tais como: quais cidades estão associadas a uma determinada política pública ou quais doenças estão ocorrendo em alguma região de interesse.

Para a captura de notícias realizou-se o Web *clipping* manual. O processo tradicional de *clipping* constitui-se na seleção e extração de notícias, notas e informações que são publicadas na mídia relacionadas a um determinado assunto. A versão *Web* ou virtual do *clipping* é uma forma rápida e econômica de pesquisa aos *sites* de jornais e revistas disponíveis na Internet, onde são acrescidos lista de veículos, portais e agências de notícias.

Analisou-se 113 notícias jornalísticas, escritas em português, de domínio público na Internet, por meio de sites de jornais, que se localizavam fisicamente nos estados de Mato Grosso e Mato Grosso do Sul ou instituições ligadas ao agronegócio, no período de janeiro a setembro 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 nos estados do Mato Grosso e Mato Grosso do Sul.

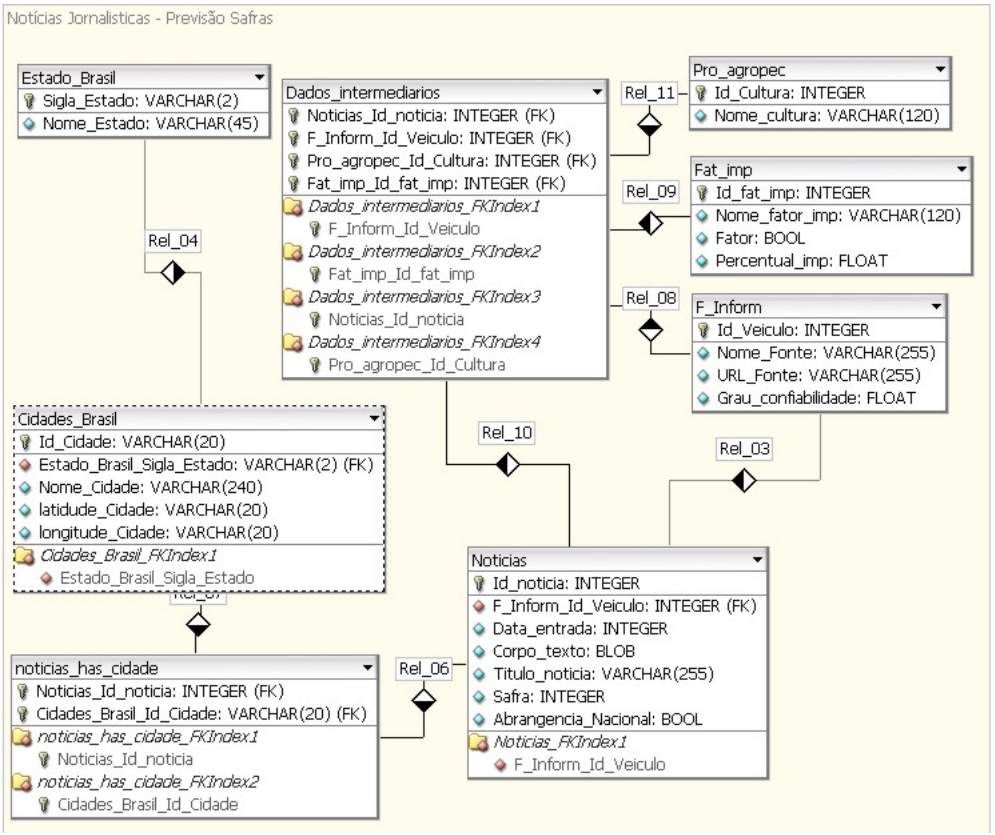

Figura 1 – Modelo de dados para organização das informações sócio-econômicas.

A Figura 2 mostra a arquitetura geral do sistema proposta para este estudo. Os módulos ainda operam de maneira não integrada e aqueles tracejados ainda não foram implementados. Os resultados do módulo de informações sócio-econômicas devem ser quantificados e interagir com aqueles obtidos no modelos agrometeorológicos. Atualmente os dois sistemas funcionam de maneira independente.

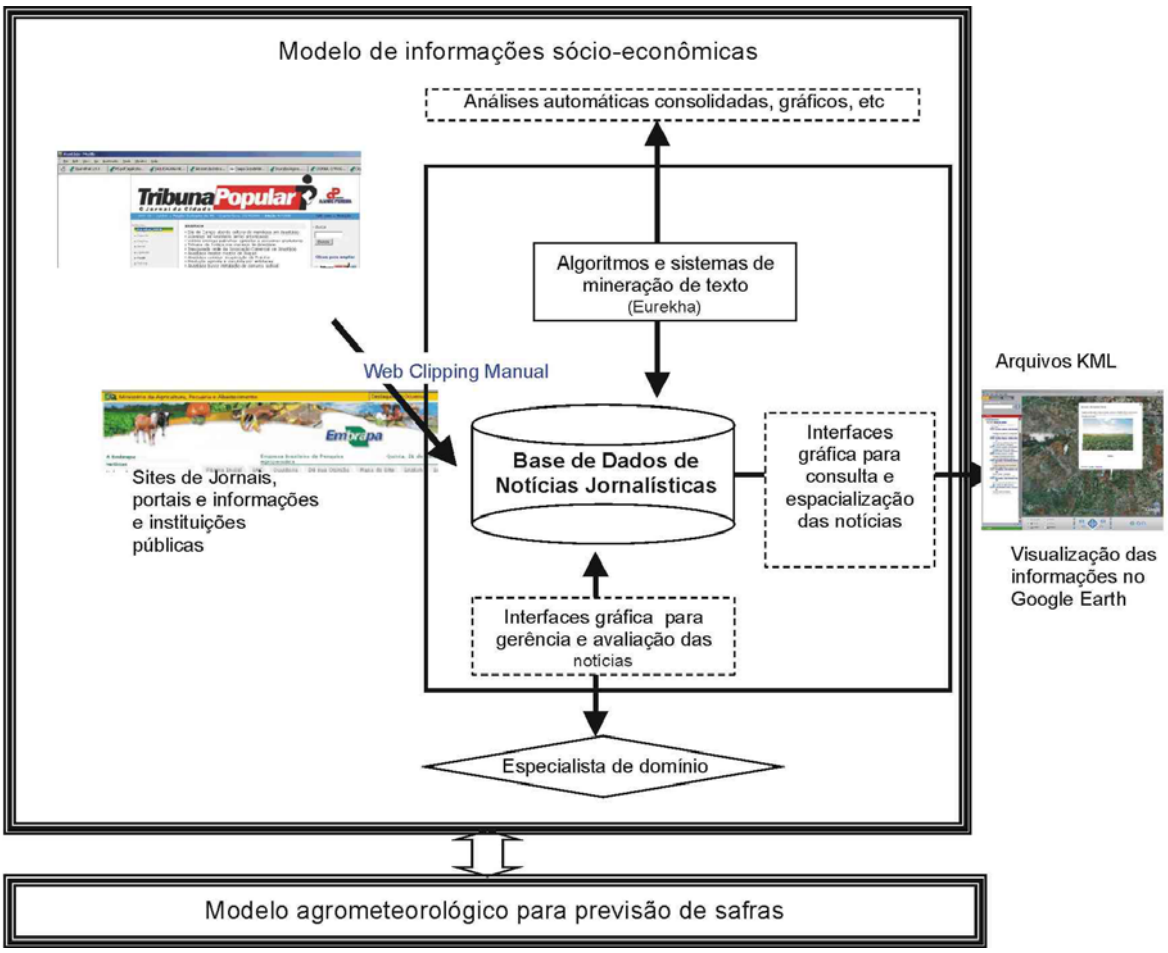

Figura 2 – Arquitetura proposta para o sistema de avaliação de informações sócio-econômicas.

Parte da técnica CRISP-DM - CRoss-Industry Standard Process for Data Mining - (Shearer, 2000) foi aplicada visando encontrar padrões de texto que auxiliassem 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, proposta por Wives em (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. Para cada notícia, inicialmente armazenada no SGBD, foi gerado um arquivo do tipo texto, conforme ilustra a Figura 3.

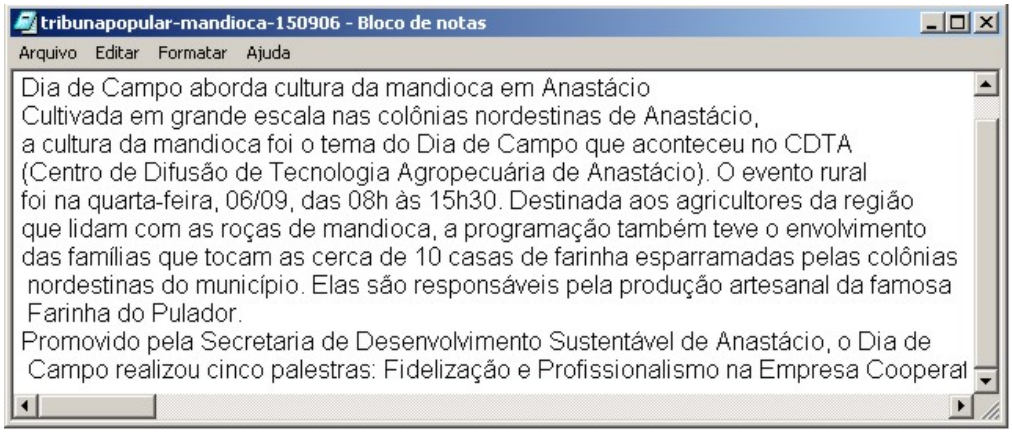

Figura 3 – Formato padrão dos conteúdos textos.

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 programa gera os chamados centróides ou palavras-chave. Estes centróides dão uma idéia sobre o assunto discutido por determinado conjunto de notícias. Exemplo: *Cluster* [4] – ANASTÁCIO MANDIOCA MUNICÍPIO PRODUTORES. Nesta situação, os textos agrupados versaram, em sua maioria, sobre notícias relativas à produção familiar de mandioca em colônias agrícolas do município de Anastácio, localizada no Mato Grosso do Sul. Duas notícias, desse *cluster*, descrevem ações de apoio aos agricultores familiares como a entrega de patrulhas agrícolas e evento para difusão de tecnologias. As quatro notícias relativas ao agrupamento [4] são sumarizadas pela Tabela 1.

Tabela 1 – Meta-informação sobre as notícias do agrupamento 4

Nº. Notícia	Título	Fonte	Data publicação
1	Valério entrega patrulhas agrícolas a pequenos produtores	Assessoria da Prefeitura de Anastácio	15/09/2006
2	Prefeitura de Guia Lopes vai ampliar Patrulha Agrícola	Diário MS	04/05/2006
3	Produção agrícola é discutida por entidades	Tribuna popular	11/08/2006
4	Dia de Campo aborda cultura da mandioca em Anastácio	Tribuna popular	15/09/2006

O *software* Google Earth 3.0.076 - sistema operacional Microsoft Windows 2000, foi utilizado para exercitar a visualização das notícias. Os atributos mapeados pela base em MySQL foram traduzidos em arquivos do tipo KML (Keyhole Markup Language), como aquele mostrada na Figura 4.

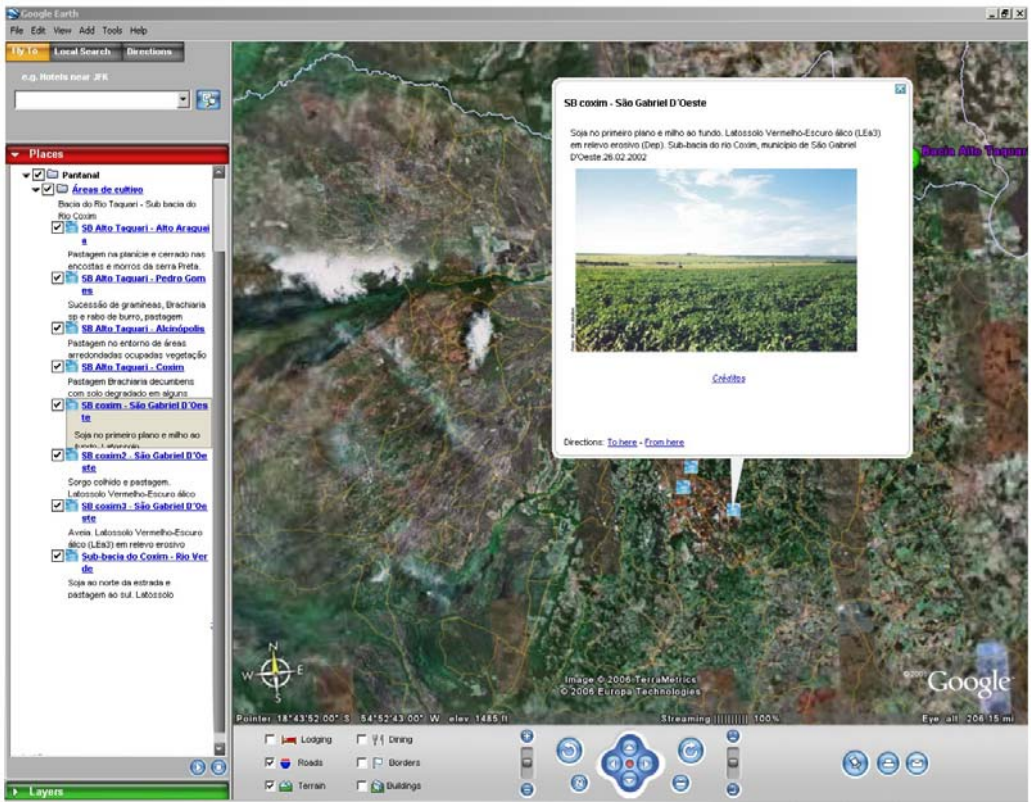

Figura 4 – Espacialização de informações sobre áreas cultivadas no Pantanal.

Resultados e Discussão

Com o valor de 10 % para o coeficiente de sensibilidade, e a escolha do algoritmo *best-star*, foram encontrados vinte e seis grupos de notícias, no entanto, quarenta e quatro notícias não foram agrupadas pelo *software*. O sistema considerou que as notícias não agrupadas tratavam de um assunto distinto.

Os *clusters* que com maior número de notícias foram o [13] e [19], e seus respectivos centróides versaram sobre:

Cluster [13] - 09% - ALGODÃO SAFRA PRODUÇÃO PRODUTORES
Cluster [19] - 09% - SAFRA ANO PRODUTOR DOURADOS

Das seis notícias reunidas no *cluster* [13], quatro discutiam o atraso na produção do algodão em função do excesso de chuva no Estado do Mato Grosso. Ressalta-se que as quatro notícias descreveram este fato no período de 12 de julho a 25 de julho de 2006.

No cluster [19], das seis notícias agrupadas, três discutiram, de maneira sumária, a safra de milho na região de Dourados. Dois deles falavam da produção de soja em Sinop, e uma última notícia versava sobre o uso de agrotóxico ilegal no Estado do Mato Grosso do Sul. Como os municípios não pertencem a Região do Pantanal, o sistema de previsão de safras não deverá computar as contribuições oriundas destas notícias.

As interfaces do *software* Eurekha, são apresentadas pela Figura 5, com ênfase nas informações do *cluster* [4].

A Figura 5 enfatiza o índice da matriz de similaridades entre as notícias três e quatro, que alcança o valor de 0,0966. Apesar de pouco significativo (o valor ideal é igual a um), esse valor é maior que o alcançado entre as notícias quatro e um (0,0962) e aquele entre as notícias um e três (0,1424). Ou seja, a discussão dos problemas enfrentados pelos pequenos produtores e a apresentação de modelos rentáveis de desenvolvimento está intimamente ligados ao Dia de Campo para difusão de novas tecnologias para a mandioca.

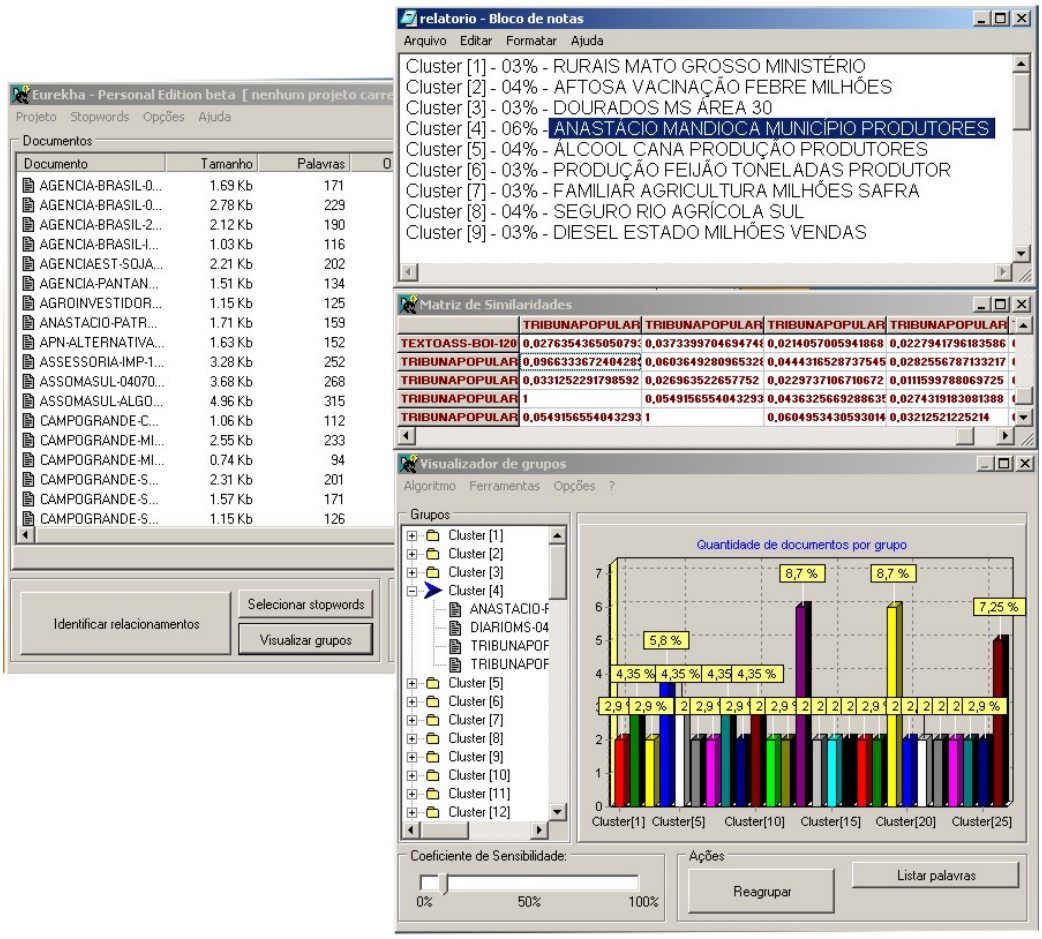

Figura 5 – Interface do programa Eurekha enfatizando informações sobre a produção de mandioca nas notícias analisadas.

Conclusões

Apesar do percentual significativo (38,9%) de notícias analisadas, em relação ao número total, não terem sido agrupadas pelo *software* Eurekha, os vinte e seis agrupamentos encontrados mostraram-se coerentes em termos de conteúdo semântico.

Em relação ao agrupamento, analisado nesse trabalho, a relevância das notícias sobre a cultura da mandioca no município de Anastácio é confirmada pelo Relatório de Produção Agrícola Municipal 2003, organizado pelo IBGE. Nesse relatório a mandioca figura como a cultura de maior rendimento de produção e área plantada. Novas políticas e ações públicas

devem favorecer ainda mais o cultivo e comercialização da mandioca, visto ser uma vocação natural do município aliada a cultura das colônias nordestinas presentes em Anastácio.

Em função dos resultados obtidos com os algoritmos de mineração de texto, o sistema proposto prevê a associação de índices de deflação ou inflação desses resultados aos valores de produção agrícola, calculados por meio de modelos agrometeorológicos. O método agrometeorológico, de uma maneira geral, enfatiza o grau de penalização sobre o rendimento da cultura face às condições climáticas nos períodos críticos do desenvolvimento vegetativo da planta.

É imperativo, na nova abordagem, a inserção de regras mais complexas para o cômputo do percentual de penalização da produção em escala regional.

O sistema constitui-se em ferramenta importante também para manejo dos recursos hídricos, pois Azevedo & Monteiro (2006), consideram a atividade agropecuária como grande responsável pela degradação intensa das águas, sejam superficiais ou subterrâneas.

A infra-estrutura e metodologia propostas para este estudo podem colaborar para a previsão de safras em outras regiões do Brasil.

6. Referências Bibliográficas

Azevedo, Andréa Aguiar; Monteiro, Jorge L. Gomes, **Análise dos impactos ambientais da atividade agropecuária no cerrado e suas inter-relações com os recursos hídricos na região do pantanal**. Disponível em: < http://assets.wwf.org.br/downloads/wwf_brasil_impactos_atividade_agropecuaria_cerrado_pantanal.pdf#search=%22itiquira%20agropecu%C3%A1ria%22> Acesso em: 21 set. 2006.

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> Acesso em: 14 mar. 2006.

Google Earth. **A 3D interface to the planet** . Disponível em: < http://earth.google.com/ > . Acesso em: 13 mar. 2006.

IBGE, **LSPA – Levantamento Sistemático da produção agrícola – Pesquisa mensal de previsão e acompanhamento das safras agrícolas no ano civil**. Disponível em: < ftp://ftp.ibge.gov.br/Producao_Agricola/Levantamento_Sistematico_da_Producao_Agricola_%5Bmensal%5D/Fasciculo/ > Acesso em: 20 set. 2006.

Nonaka, I.; Takeuchi, H. ***Criação de conhecimento na empresa*: como as empresas japonesas geram a dinâmica da inovação**. Rio de Janeiro: Campus, 1997.

Nonaka, Ikujiro; Takeuchi, Hirotaka. ***Criação de conhecimento na empresa***. Rio de Janeiro: Campus, 1997.

MySQL AB **Developer Zone**. Disponível em: <http://dev.mysql.com> Acesso em: 19 set. 2006.

Silva, J. dos S. V. da; Moraes, A S.; Seidl, A F. **Evolução da agropecuária do pantanal brasileiro 1975-1985**. Corumbá: Embrapa Pantanal, 2001.

Shearer, C.. *The* CRISP-DM Model: The blueprint for data mining. **Journal of Data Warehousin**g. 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***. Porto Alegre: CPGCC da UFRGS, 1999. 101p. (Dissertação de Mestrado).