9º Congresso Interinstitucional de Iniciação Científica – CIIC 2015
10 a 12 de agosto de 2015 – Campinas, São Paulo

USO DE FERRAMENTAS DE MINERAÇÃO DE TEXTOS PARA APOIAR A CONSTRUÇÃO DE PORTIFÓLIOS DE TECNOLOGIA AGRÍCOLA.

Carolina Tavares de **Oliveira**[1]**;** Stanley Robson de Medeiros **Oliveira**[2]; Maria Fernanda **Moura**[3]

Nº 15605

RESUMO *– Neste trabalho são relatadas as etapas realizadas para o processo de mineração de textos como auxílio à construção de portfólios tecnológicos. i) os especialistas do domínio selecionaram uma relação de palavras-chaves e jornais de interesse; ii) a busca dos textos foi realizada no sistema SABIIA/Embrapa, dado ao seu acesso ao repositório OAI (Open Archives Initiative) na área de agricultura; iii) criou-se um vocabulário completo de nomes a partir de expressões de busca e de termos mais convenientes relacionados a tecnologia para o setor agrícola; iv) geraram-se hierarquias de tópicos a partir dos documentos e de seu vocabulário; v) geraram-se séries temporais dos tópicos. Mostram-se aqui os principais resultados e limitações dessa análise.*

Palavras-chaves: Mineração de textos, hierarquia de tópicos, portfólios tecnológicos.

ABSTRACT *- This paper reports the steps performed in the process of text mining as an aid to the construction of technological portfolios: i) the domain experts selected a list of keywords and newspapers of interest; ii) the search of the texts was held at SABIIA system / Embrapa, given their access to OAI (Open Archives Initiative) repository in agriculture; iii) it was created a complete vocabulary of names from search expressions and more convenient terms related to technology for the agricultural sector; iv) It was generated hierarchies of topics from the documents and from their vocabulary; v) it was generated time series of topics. This work reports the main findings and limitations of this analysis.*

Key-words: Text mining, Topics hierarchy, technological portfolio

1 Autor, Bolsista CNPq (PIBIC): Graduação em Engenharia Agrícola, Unicamp, Campinas-SP; caroli.aro@gmail.com
2 Orientador:Pesquisador Embrapa Informática Agropecuária,Campinas-SP, stanley.oliveira@embrapa.br
3 Orientadora: Pesquisadora da Embrapa Informática Agropecuária, Campinas-SP; maria-fernanda.moura@embrapa.br

1- INTRODUÇÃO

O processo de mineração de textos visa extrair conhecimento útil de grandes coleções de documentos textuais, buscando padrões e tendências interessantes em textos escritos em língua natural. Envolve a aplicação de algoritmos computacionais que processam textos e identificam informações implícitas que normalmente não poderiam ser recuperadas pelos métodos tradicionais de consulta pois geralmente estes, demandam tempo maior para identificar as informações contidas nos grandes agrupamentos de arquivos. As principais contribuições do processo estão relacionadas a análise quantitativa e qualitativa de grandes volumes de textos e a melhor compreensão do conteúdo disponível neles.

Este trabalho foi elaborado para descrever as etapas do processo de mineração aplicado para auxiliar a construção de portfólios tecnológicos para a agricultura, sendo um dos itens em pauta do projeto pertencente a rede Agrohidro. O objetivo geral desse projeto é avaliar e/ou adaptar tecnologias, utilizáveis em escala de parcela ou superior, que contribuam para a preservação da qualidade da água, para seu uso eficiente, e para o aumento da produtividade da água na agricultura em bacias hidrográficas brasileiras. Com base nisso, portfólios de tecnologias agrícolas serão úteis para avaliar o que tem sido feito para minimizar a degradação dos recursos hídricos, uma vez que o setor agrícola brasileiro é o principal usuário consuntivo destes, permitem análises e discussões para solucionar problemas decorrentes da avaliação e adaptação de tecnologias nos biomas brasileiros. Neste trabalho são relatadas as etapas para mineração de textos .Em materiais e métodos são descritas detalhadamente as etapas e as fontes de dados utilizadas; em resultados e discussão apresentam-se os principais resultados até o momento, perspectivas e possibilidades para futuros trabalhos obtidos. Finalmente, na seção de conclusão mostram-se a utilidade dos resultados,

2- MATERIAIS E MÉTODOS

Cada etapa gerou um resultado que era pré-requisito para a etapa subsequente. Com base nisso o processo deu-se da seguinte maneira: i) os especialistas do domínio selecionaram uma relação de palavras-chaves e jornais de interesse; ii) a busca dos textos foi realizada no sistema

SABIIA/Embrapa[1], e, sempre que possível ou disponível, buscou-se por textos completos; iii) criou-se um vocabulário de nomes a partir de expressões de busca e de termos mais convenientes relacionados a tecnologia na agricultura; iv) geraram-se hierarquias de tópicos a partir dos documentos e de seu vocabulário; e, v) geraram-se séries temporais dos tópicos

Nesta seção são descritas detalhadamente as principais fontes de dados para o processo e suas etapas de coleta e análise.

2.1 Definição de palavras chaves pelos especialistas

Para a seleção dos documentos textuais, houve a necessidade de se obter palavras ou termos técnicos que seriam utilizados para restringir a busca dos documentos no sistema SABIIA., uma vez que os termos deveriam ser escolhidos para gerar buscas de documentos relacionados a tecnologia na agricultura, portanto as palavras foram selecionadas pelos especialistas na área e integrantes do projeto da rede Agrohidro. No total foram reunidos 469 termos,sendo 178 em inglês 291 em português. Alguns deles são: Jurisdictional lands;Riparian zones; Cattle fencing; Scream-crossing guidelines, sistemas de informação geográfica; sensoriamento remoto; pegada hídrica ecológica; águas cinzas; hidrodinâmica do solo; dimensionamento do bulbo úmido; termometria; programação da irrigação; uso eficiente da água; produtividade da água; proteção de nascentes; recuperação de áreas degradadas; modelos hidrológicos, entre outros.

2.2 Seleção de textos e seus respectivos metadados

Os textos foram selecionados do Sistema Integrado e Aberto de Informação em Agricultura (SABIIA), este é um mecanismo de busca automatizado, que coleta metadados de provedores de dados científicos de acesso aberto, no padrão OAI (Open Archives Initiative), previamente selecionados. No total foram reunidos em um diretório 643 documentos (em pdf) e seus metadados (armazenados em um diretório denominado descritores).O SABIIA acessa os principais repositórios OAI com literatura de interesse da agricultura, veja: http://www.sabiia.cnptia.embrapa.br/.

2.3 Conversão dos PDFS

Foi escrito um programa em Python nomeado de CRIAtextbase.py, para converter os pdfs no formato txt(formato adequado para o processamento),depois de convertidos os arquivos foram

1

armazenados na pasta "textbase", e os não convertidos mas com metadados na pasta "texbaseSoMeta". O resultado do processo foi:

número de pdfs convertidos: 579

textbase: 462(documentos completos, convertidos e com metadados)

textbaseSoMeta: 56 arquivos só com metadados.

2.4 Criação do vocabulário

Para que se possa obter tópicos mais significativos na área agrícola e ligados à palavras-chaves sobre tecnologias, criou-se um vocabulário. Neste trabalho, o vocabulário é considerado como uma lista de palavras que será utilizada no processo de mineração de textos. Para isso partiu-se de um vocabulário inicial, obtido do Thesagro[2]. A questão da similaridade foi lexicalmente tratada dado que removem-se inflexões dos termos na composição do vocabulário. Assim para encontrar termos lexicalmente similares utilizou-se a medida de "edit distance" superior a 0.60(função do Python). Em termos do uso de ferramentas, definiu-se um filtro para a criação do vocabulário: para unigramas foi considerado aqueles no intervalo 5<=tf=>250; gerou-se stoplist nova; e na sequência, palavras de até três gramas. Utilizou-se a ferramenta TaxtEdit para preprocessar o vocabulário. Criou-se um diretório nomeado discover.names com nomes stemmizados dos atributos e suas estatísticas de frequências. A stemização (do inglês, *stemming*) é o processo de reduzir palavras flexionadas (ou às vezes derivadas) ao seu tronco (*stem*), base ou raiz, geralmente uma forma da palavra escrita. Para criar um vocabulário ordenado e com stems mapeados às palavras mais frequentes, fez-se um programa em Python nomeado mapeiaVOC.py. Com entrada VOC_PT.txt (lista de vocábulos de interesse) e stemWdtF_PT.all (frequências dos stems e palavras que lhes originam) e o resultado foi: MAPA_PT.txt (lista final de vocábulos) e VOC_PT_o.txt (vocábulos ordenados).

2.5 Pré-processamento da coleção de textos

[2] Thesagro Agrícola Nacional em http://snida.agricultura.gov.br:81/binagri/html/Cen_Thes1.html.

Preprocessar a coleção de textos, neste trabalho, é convertê-los para uma matriz. Nesta matriz cada linha corresponde a um documento e cada coluna a um vocábulo do vocabulário. Nas células foram colocadas frequências de cada vocábulo em cada texto.

Assim para esta etapa criou-se um diretório denominado preProcessamentoColecao no qual foi armazenado a ferramenta de pré-processamento preProc e a pasta texbase (com todos os arquivos convertidos em texto e os respectivos descritores). Dentro do diretório preProc foi armazenado na pasta vocabulário o arquivo VOC_PT_o.txt. Para pré-processar precisou-se do arquivo de configuração dadosTech.xml armazenado no diretório preProc. Para executar no Java deu-se o comando no terminal java -jar preprocessMain.jar dadosTech.xml, sob o diretório preProc.O Resultado foi Matriz1.dat (as células da matriz) e Matriz.1.hdr (a descrição dos documentos e vocábulos utilizados).

2.6 Mapeamento da matriz criada

O objetivo da stemmização é reduzir o número de vocábulos. Como os vocábulos stemmizados nem sempre são bem compreendidos ao se mostrar o resultado, nesta etapa eles são trocados pelos termos mais frequentes correspondentes a eles a esse processo denomina-se mapear.

Assim, foram transferidas para o diretório MapeiaMatriz/Mapeamento, a Matriz1.dat e Matriz.1.hdr criadas conforme 2.5. Para realizar o mapeamento executou-se:

java -jar Mapeamento.jar MAPA_PT.txt Matriz_1.hdr TechMapeado.hdr

Sendo que o MAPA_PT.txt é o mapa criado conforme descrito na etapa 2.4. O resultado foi: TechMapeado.hdr. Colocou-se o arquivo hdr junto com ao seu correspondente arquivo dat no diretório Hierarquias.

2.7 Geração de hierarquias e séries temporais

O diretório Hierarquias continha os resultados da etapa descrita em 2.6, ou seja, TechMapeado.hdr e TechMapeado.dat. Além disso continha TopicVisAnterior e websensor-cnptia. Com a finalidade de gerar séries temporais. Para executar esse processo faz-se

java -Xmx2G -cp torch/dist/torch.jar torch.websensors.InitialModelLearning ./confTech.ini

O resultado final fica em Hierarquias/TechDendro.xml.

Para visualizar os resultados da hierarquia e séries, cria-se com a ferramenta TopicVis alguns arquivos que podem ser navegados, da seguinte forma:

```
java -jar TopicVis.jar TechDendro.xml textbase/ ./ teste
```

Resultando em dendogram.json dendogran.html com a data em que foi criado, a ser aberto no navegador.

3- RESULTADOS E DISCUSSÃO

O resultado final foi a geração dendrogramas (Figura 1 e 2), que podem ser abertos no navegador. Dendrogramas são representações icônicas que organizam determinados fatores e variáveis. Neste trabalho são considerados hierárquicas de tópicos. Nesses diagramas cada ramo da árvore corresponde a palavras que funcionam como descritores dos tópicos, agrupando um conjunto de textos relacionados á estas expressões. Um possível tópico é apresentado na figura 2, observa-se que para as palavras “Mosca, Fruta, Larva, Infestans, Esterco, Fêmea, Disco”,há um grupo de textos associados a esse tópico, além disso o dendrograma gera uma série temporal que é um gráfico com a distribuição no tempo das publicações dos textos. Desta forma, pode-se explorar os resultados a partir dessa visualização. Porém, a projeção das árvores demonstrou um resultado com termos específicos e técnicos e os textos aparentemente não se enquadravam diretamente nas categorias de tecnologia agrícola buscadas no trabalho. Estes resultados ajudam a vislumbrar, principalmente a distribuição de temas vinculados á área agrícola, Porém, não foi possível obter correlações explícitas entre tecnologias e a distribuição geográfica delas, bem como uma aplicação de um processo para identificar toponímias nos textos (localizações geográficas). Para isso, prevê-se um trabalho futuro que utilize regras de associação entre as categorias de tecnologias agrícolas e os termos mais frequentes nos textos, além de suas localidades geográficas disponíveis no território brasileiro. Além disso não foi obtido o portfólio tecnológico, este é o objetivo incluído no plano de trabalho que será realizado como continuação deste trabalho.


Grafico
Adubo-verde, Oxido, Mucuna, Mucuna-preto, Condutividade-eletrica, Guandu, Espontanea
Clorado, Curva-nivel, Corte-gado, Carro, Piper, Demanda, Encosta
Piper, Caule, Colheita-epoca, Sodio, Oxigenio, Pimenta, Longo-pimenta
Aspecto-economica, Milho-verde, Credito, Poder, Capitulo, Dispar, Intensiva-terra
Pimenta, Longo-pimenta, Piper, Corte-epoca, Desenvolvimento-projeto, Broto, Triturador
056.2009061500000.txt
37.2011061500000.txt
059.2012061500000.txt
025.2002061500000.txt
0071.2009061500000.txt
BoletimPD28AMAZORIENTAL.2004061500000.txt
doc231.2007061500000.txt
00173.2008061500000.txt
0019.2000061500000.txt
bolpd17.2001061500000.txt
0050.2000061500000.txt
057.2012061500000.txt
0073.2006061500000.txt
047.2008061500000.txt
Producaobiomassa.1998061500000.txt
025.2006061500000.txt
040.2008061500000.txt
Bico, Gota, Colmo, Campo-dia, Espiga, Efeito-estufa, Pau
Capitulo, Pimenta, Oxido, Longo-pimenta, Adubo-verde, Piper, Mucuna
Cebola, Bulbo, Fenol, Abelha, Mellifera, Bioindicador, Allium
Mosca, Fruta-mosca, Larva, Infestans, Esterco, Femea, Disco
dade, Brachiaria, Araras, Indicador-sustentavel, Pousio, Pesca, Bioteste
ndicador-sustentavel, Sig, Landsat, Socorros, Banco-dados, Cartografia

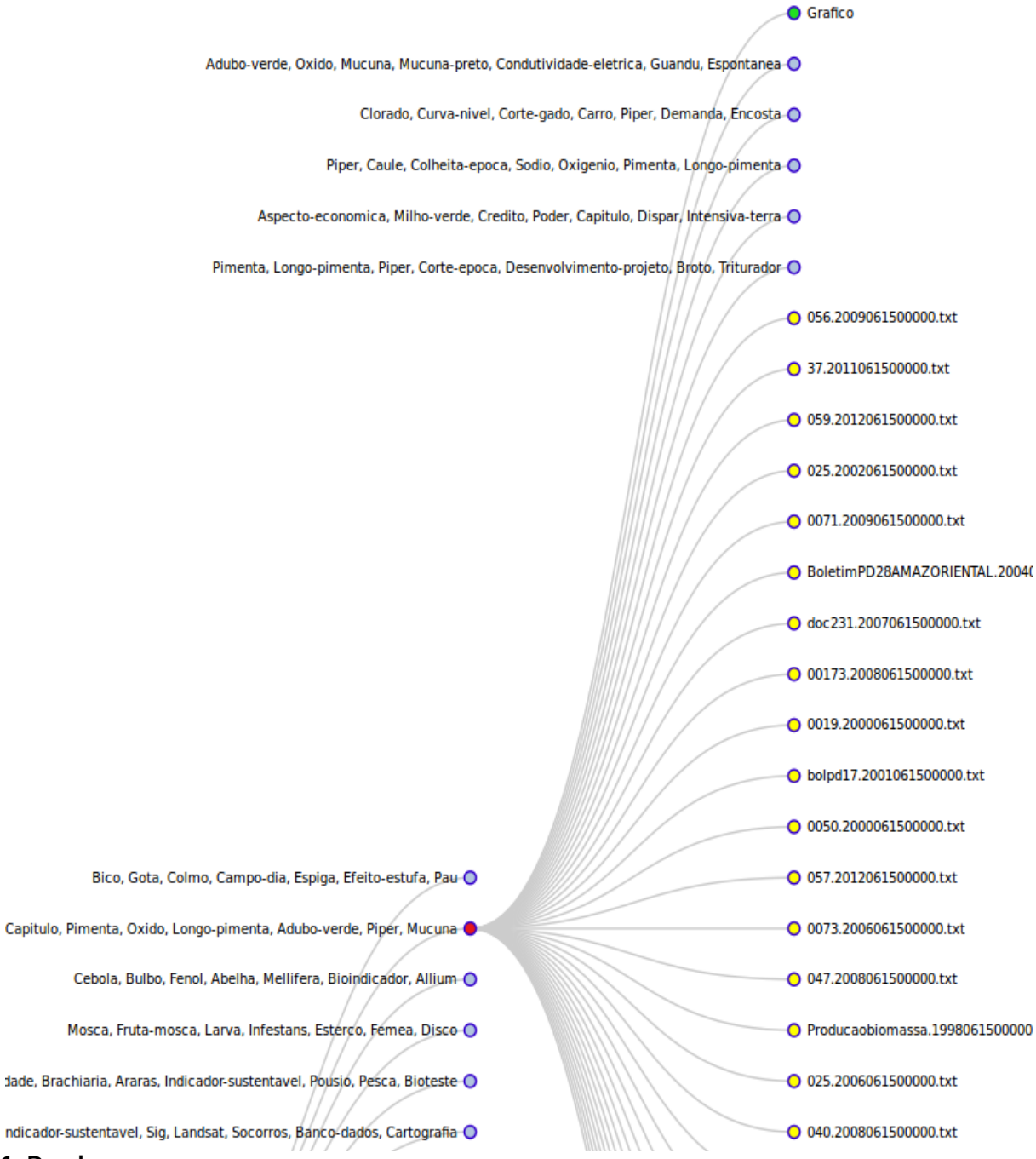

Figura 1: Dendrograma

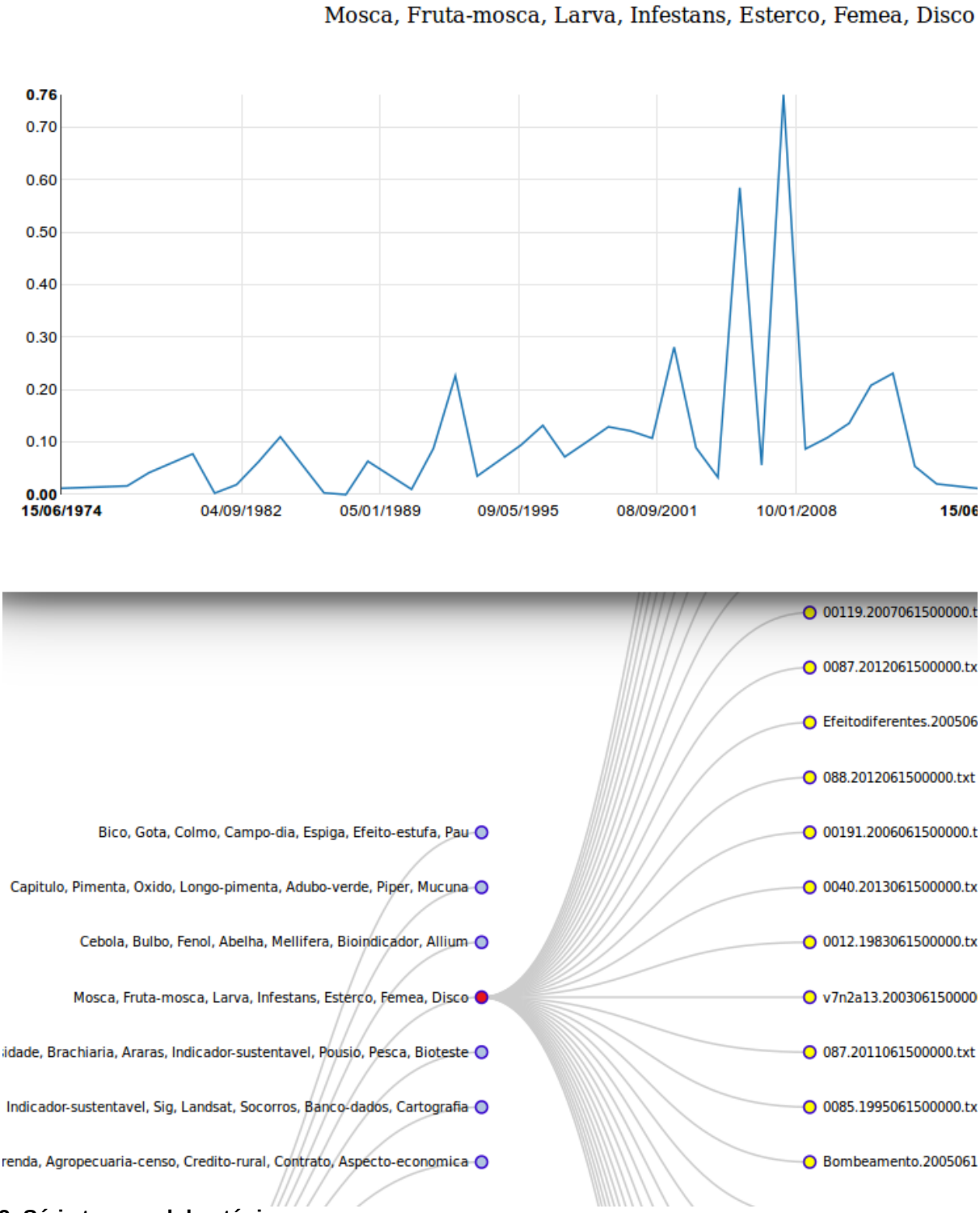

Figura 2: Série temporal dos tópicos

4- CONCLUSÃO

Este trabalho permitiu mostrar como a mineração de textos pode auxiliar a questão da reunião de informações necessárias para se obter um portfólio que condiz com os objetivos do projeto. No entanto, não foi obtido um portfólio, este é o objetivo incluído no plano de continuação do trabalho.

Neste trabalho consta a construção de dendrogramas que facilitam a visualização das informações de grandes volumes textuais. A perspectiva futura, e a proposta de continuação deste envolverá extração de padrões utilizando técnicas como regras de associação e extração de toponímias dos textos.

5- AGRADECIMENTOS

Agradeço a Embrapa Informática Agropecuária pela oportunidade de estágio e ao CNPq pela bolsa concedida. Agradeço aos meus orientadores Maria Fernanda e Stanley.

6- REFERÊNCIAS BIBLIOGRÁFICAS

AMO, S. Técnicas **de mineração de textos. Universidade** Federal de Uberlândia. Faculdade de Computação, disponível em http:// www.deamo.prof.ufu.br/,2003

BASTOS, V. M. **Ambiente de descoberta de conhecimento na Web para a língua Portuguesa. Phd thesis,** Universidade Federal do Rio de Janeiro,COPPE,2006.