2015-10

Um software para recuperar e analisar artigos open access em agricultura utilizando técnicas de mineração de textos.

Congresso Brasileiro de Agroinformática, X, 2015, Ponta Grossa.
http://www.producao.usp.br/handle/BDPI/49489

Um software para recuperar e analisar artigos *Open Access* em agricultura utilizando técnicas de mineração de textos

Maria Fernanda Moura[1], Gabriel Mari Tararam[1], Ricardo Marcondes Marcacini[2], Luis Eduardo Gonzales[1], Celina Maki Takemura[3], Leandro Eduardo Annibal Silva[1], Fabiano Fernandes dos Santos[4], Solange Oliveira Rezende[4], Sílvio Roberto Medeiros Evangelista[1]

[1] Embrapa Informática Agropecuária, Campinas, São Paulo, Brasil, maria-fernanda.moura@embrapa.br, gabriel.tararam@colaborador.embrapa.br, luis.gonzales@embrapa.br, leandro.annibal@colaborador.embrapa.br, silvio.evangelista@embrapa.br

[2] Fundação Universidade Federal de Mato Grosso do Sul, Três Lagoas, MS, ricardo.marcacini@ufms.br

[3]Embrapa Monitoramento por Satélite, Campinas, São Paulo, Brasil, celina.takemura@embrapa.br

[4]Universidade de São Paulo, São Carlos, São Paulo, Brasil, fabiano.fernandes@icmc.usp.br, solange.rezende@icmc.usp.br

RESUMO

Neste trabalho é apresentado o software CRITIC – Compilação e Recuperação de Informação Técnico-científica e Indução ao Conhecimento, com base em técnicas de mineração de textos sobre artigos científicos. O software usa um provedor de serviços para acesso aos repositórios de referências aos artigos, cujo acesso é aberto. O CRITIC está sendo desenvolvido para ser colocado em repositório Open Source. Na sua versão atual, permite-se realizar uma análise exploratória sobre os resultados das consultas, na qual são automaticamente identificados: tópicos hierárquicos dos temas cobertos na consulta; a distribuição temporal desses temas; e, a distribuição geoespacial dos temas cobertos pelos textos. Discute-se neste, alguns resultados de sua primeira versão, a metodologia de mineração de textos utilizada e a arquitetura do software – projetado para ser facilmente expandido.

PALAVRAS-CHAVE: mineração de textos, revisão bibliográfica, agricultura, acesso aberto.

ABSTRACT

In this paper we present the CRITIC software, which has been developed to the compilation, recovery and induction to knowledge from technical and scientific articles through text

mining techniques. This software has been using a service provider to access the articles references in open access repositories. Additionally, CRITIC is being developed to be an Open Source software. In its current version, it allows to carry out an exploratory analysis of the recovering results, in which: automatically identified topic hierarchies showed some possible topics over the results; the temporal distribution of these topics; and spatial distribution of subjects covered by the texts. Furthermore, some results of the first version are discussed, as well as the text mining methodology and the software architecture - designed to be easily expanded.
KEYWORDS: text mining, literature review, agriculture, Open Access.

INTRODUÇÃO

Em redes de pesquisa, existe a necessidade de se possuir um ferramental de análise da informação que facilite não apenas a identificação de bibliografia e outras fontes de material de divulgação, mas também o trabalho colaborativo entre seus membros. Esse ferramental deve permitir o cruzamento de informações de diversas fontes a fim de avaliar o caminho percorrido pelo tema de pesquisa; por exemplo, tendências, oportunidades, inserção no contexto nacional/internacional e áreas deficitárias em desenvolvimento tecnológico. Para isso, se faz necessário monitorar tanto a produção técnico-científica da própria rede, no passado e presente, como de outras redes ou iniciativas isoladas que visem objetivos semelhantes de PD&I no mesmo domínio de conhecimento.

Para responder questões como essas, existem algumas ferramentas de software, em sua maioria voltadas às análises bibliométricas, colaboração científica e poucos trabalhos em análise de tendências científicas; além, é claro, de ferramentas de mineração de dados (estruturados ou textuais) que podem ser utilizadas no processo. Por exemplo, para recolher referências pode-se utilizar várias máquinas de busca ou recorrer a ferramentas colaborativas, como EndNoteWeb, Mendeley ou softwares livres, como o Zotero e o JabRef . A maioria dessas ferramentas permite armazenar apenas os metadados dos artigos de livre acesso e fornecem estatísticas descritivas básicas sobre a coleção de documentos recuperada. Para realizar uma análise mais detalhada desses dados, sem precisar ler todos os documentos recolhidos, existem poucas ferramentas, especialmente software livre. A ferramenta BibExcel (Pilkington, 2014) foi projetada para analisar dados textuais, desde que formatados de uma maneira similar. Ela gera arquivos que podem ser importados pelo Excel, ou qualquer programa que utilize dados tabelados (por exemplo, no formato CSV); além disso, faz mapeamento de coautoria, alguma classificação de documentos, mapeia fontes e citações,

produz dados para a Google Maps (com a localização dos autores) etc. A Citespace[1] é projetada para a visualização de progressos em um domínio do conhecimento. Dessa forma, ela permite a identificação do rápido crescimento de temas, por meio de *links* entre as citações no mundo das publicações, decompondo uma rede em grupos e encontrando identificadores para nomear os grupos encontrados; e, também permite geoespacializar a colaboração em áreas específicas. A SciTool[2], *Science of Science* (Sci2), muito citada na literatura nos últimos dois anos, é uma ferramenta projetada para o estudo da ciência. Ela provê várias funcionalidades para facilitar o entendimento e interpretação de padrões: pontos temporais de interesse, que determinam períodos de maior atividade em algum tema; identifica nomes de localidades e lhes atribui o correspondente código geográfico, a fim de mapeá-las geograficamente; identifica tópicos de interesse, por meio de agrupamentos independentes; e, também produz redes de coautoria. Todas essas ferramentas trabalham melhor com textos escritos em língua inglesa, cujos dados contenham rótulos (tais como título, resumo, autor, etc), e embora sejam software de uso livre, não são código aberto.

Há mais de vinte anos, a Embrapa desenvolve ferramentas que organizam e permitem buscas sobre os dados que ela utiliza e produz. A partir de 2009, a Embrapa, por meio de seu sistema gestor de bibliotecas, definiu como objetivo estratégico o armazenamento e disseminação de sua produção científica e tecnológica em Acesso Aberto, utilizando repositórios e padrões da *Open Archives Initiative* (OAI)[3]. Para restringir o acesso aos repositórios de interesse em agricultura, e manter uma política de *backup*, foi criado o provedor de serviços Sistema Aberto e Integrado de Informação em Agricultura (Sabiia[4] - (Vacari *et al*, 2011)), caracterizado como sistema responsável pela integração de todos os dados provenientes de repositórios institucionais, periódicos científicos, bibliotecas digitais e outros, tanto internos quanto externos, de interesse da Embrapa. O Sabiia permite buscas simples ou avançadas ao seu repositório e fornece estatísticas básicas, tais como frequência dos artigos em provedores, por autores, por idioma, por assuntos, permite montar um carrinho de artigos, etc; e, ainda, pode-se indicar filtros simples de busca sobre cada item descrito pelas estatísticas básicas - por exemplo, refazer a busca apenas para tais assuntos e autores. Assim, a necessidade de levantamentos bibliográficos e análise dos mesmos por redes de pesquisa vinculadas à Embrapa, passam, quase que total e obrigatoriamente, pelo Sabiia.

---

[1] In: http://cluster.cis.drexel.edu/~cchen/citespace, consultado em maio de 2015.
[2] In: http://sci2.cns.iu.edu/user/, consultado em maio de 2015.
[3] Open Archives Initiative: https://www.openarchives.org/, consultado em maio de 2015.
[4] In: http://www.sabiia.cnptia.embrapa.br/, consultado em maio de 2015.

Como o Sabiia disponibiliza apenas estatísticas descritivas básicas, faz-se interessante incorporar-lhe novas funcionalidades ou, pelo menos, exportar seus dados para outras ferramentas de análise de dados mais sofisticadas. Dessa forma, vem sendo desenvolvido o software CRITIC (Compilação e Recuperação de Informações Técnico-científicas e Indução ao Conhecimento), cujos objetivos são: 1) permitir a formação de uma relação de repositórios de interesse exclusivo de um grupo de pesquisa, cujos artigos podem estar em várias línguas; 2) prover a indexação do repositório de acordo com vocabulário específico; 3) quando os artigos forem de livre acesso, permitir utilizar o texto completo; 4) prover buscas expandidas sobre o repositório, por exemplo, com uso das relações presentes nos *thesaurus* (uma busca por "abacaxi", poderia considerar o termo "abacaxizeiro", bem como, com peso pouco inferior os termo "fruta tropical" e os termos "ananas" e "bromelina" - de acordo com as relações presentes no Thesagro[5]); 5) prover busca simples e avançada; 6) apresentação de estatísticas básicas sobre a consulta; 7) finalizada uma busca, permitir uma análise exploratória mais detalhada dos resultados: i) exibindo os documentos por grupos hierárquicos de assuntos automaticamente identificados, ii) exibir os documentos em mapas, de acordo com sua distribuição geoespacial – automaticamente identificada, iii) exibir os assuntos identificados de acordo com sua distribuição temporal; iv) identificar pontos de interesse, com maior concentração de artigos, temporais e/ou geográficos; v) permitir novos filtros por assuntos, intervalos de tempo e localidades geográficas, para serem graficamente exibidos; 8) permitir a exportação de dados para outras ferramentas de análise; 9) ser modular e reconfigurável a fim de poder incorporar novas análises ou a integração de diferentes ferramentas ou módulos de outras ferramentas; e, 10) ser código aberto. O CRITIC deve ter seu uso compatível com algumas ferramentas de análise bibliométrica e de mineração de dados, incorporando alguns requisitos próprios da área de agricultura, tais como, o uso de repositórios OAI, indexação de acordo com vocabulário de agricultura, geoespacialização de documentos e temas por localidades (regiões, macrorregiões, bacias hidrográficas, rios, etc), organização hierárquica de temas, e outros que venham a surgir. Além disso, o requisito de exportar dados para outros padrões, permite utilizar recursos de várias outras ferramentas.

O CRITIC é um software com base em técnicas de mineração de textos sobre artigos científicos, cuja metodologia de extração de padrões, arquitetura e soluções computacionais, são descritas na seção de Material e Métodos. Uma aplicação, utilizando sua versão 1.0, de abril de 2015, é apresentada em Resultados e Discussões, bem como trabalhos futuros.

[5] Thesaurus brasileiro de agricultura, consultado em maio de 2015, em http://snida.agricultura.gov.br:81 /binagri/html/Cen_Thes1.html.

MATERIAL E MÉTODOS

AA base metodológica do CRITIC é um processo de mineração de textos, para identificação automática de padrões, descrito na próxima subseção. Por outro lado, a escolha de desenvolver um software *open source*, também implicou no uso de software livre e código aberto para todas as soluções de sua arquitetura e ferramental computacional.

PROCESSO DE MINERAÇÃO DE TEXTOS UTILIZADO NO CRITIC

A mineração de textos tradicional implica em obter uma representação dos dados textuais em um padrão que a maior parte das ferramentas de mineração de dados compreende, isto é, uma matriz de descrição da coleção de textos (Aggarwal, 2015). Em geral, as linhas dessa matriz correspondem a cada texto e as colunas a cada palavra de interesse nos textos, e as células a uma medida de importância de cada palavra no texto (frequência, presença ou ausência, etc). A questão de maior importância nesta etapa é a escolha adequada das palavras que devem permanecer na matriz, dado que o número de palavras nos textos é sempre muito grande, mesmo com uso de vocabulário fixado. A partir dessa matriz, pode-se classificar (regressão logística, árvores de decisão, etc) novos textos, realizar uma análise exploratória sobre os textos e suas palavras (agrupamento, análise fatorial, etc), de posse das datas de publicação realizar uma análise temporal da distribuição de palavras ou conjuntos delas, assim como, realizar várias outras análises que dependam de tabulação de dados. O CRITIC transforma automática e estatisticamente o resultado de uma busca, independentemente da língua em que o texto está escrito, isto é, uma coleção de artigos em uma matriz desse tipo, mantendo o conhecimento das datas de publicação. Porém, identificar e extrair os dados de localização geoespacial existentes nos textos implica em usar uma ferramenta linguística com acesso a conhecimento externo. Essas ferramentas reconhecem entidades nomeadas em textos (localidades, nomes próprios, etc) a partir de várias consultas a fontes externas, tais como Wikipedia, bases de localizações geográficas, etc. São ferramentas bastante específicas e dependentes de língua, além de serem computacionalmente bastante complexas e lentas. Após a identificação de nomes de localidades geográficas, ainda é necessário desambiguá-los. Por exemplo, se a localidade “São Francisco” está no texto, precisa-se identificar se é a cidade, um trecho do rio ou a bacia. Para isso é calculado o polígono envolvente, ficando-se com a intersecção dos polígonos em um mesmo texto (Machado *et al*, 2013). Com essas representações - a matriz de incidência de palavras, datas de publicação e localizações geográficas - as demais informações e modelos podem ser inferidos e posteriormente analisados, a partir de ferramentas específicas de visualização.

ARQUITETURA E FERRAMENTAL DO CRITIC

Nesta subseção encontra-se a arquitetura da versão 1.0 do CRITIC, na qual cada módulo pode ser substituído ou novos módulos podem ser acoplados. Ele foi projetado para ser acesso Web, pois seus usuários estão habituados ao Sabiia. Dessa forma, alguns de seus processos precisam ser disparados, mensalmente, "*off line*". Esses processos compreendem a atualização da base de dados para busca, pré-processamento dos dados e obtenção das localidades geográficas, que são processos computacionalmente pesados. Os processos "*on line*", via Web, são basicamente os de busca e visualização – que disparam a seleções de submatrizes, cálculos de tópicos (análise exploratória de temas), distribuição temporal dos temas e seleções de novas submatrizes. A seguir, de acordo com a ilustração na Figura 1, os módulos da versão 1.0 do CRITIC, que correspondem aos seus processos, são descritos:

. Coletor Sabiia: esse processo (módulo) faz parte da ferramenta Sabiia e foi mantido sem alterações, isto é, todo o código e processo foram reutilizados. Ele consulta os provedores OAI e atualiza a base de dados do Sabiia, indexando-a na sequência.

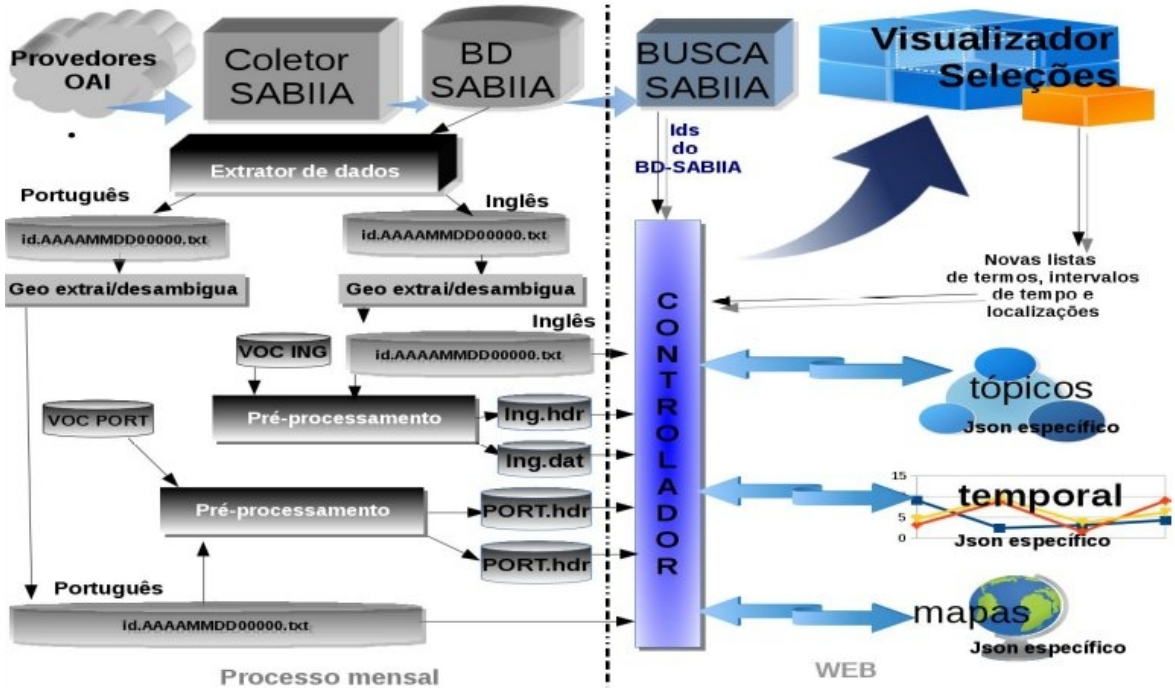
. Extrator de dados: este processo (um aplicativo JAVA) encarrega-se de recuperar todos os textos da base de dados da Sabiia, colocando-os em um padrão XML que permite a identificação da data de publicação e das partes do texto que devem ser utilizadas no processo de mineração (título, resumo, palavras-chaves), além de incluir a informação geoespacial – que não é presente no padrão atualmente armazenado pelo Sabiia. O processo divide os dados em duas bases de textos: uma em português e outra em inglês. Nesta versão, o CRITIC trabalha apenas com essas duas línguas – podendo-se acoplar processos para o tratamento de outras línguas.

. Geo extrai/desambigua: este processo, desenvolvido em Python, é o responsável por disparar um reconhecedor de entidades nomeadas e associá-las a um objeto geográfico. Posteriormente, aplica um processo de desambiguação de topônimos que usa a distância entre esses objetos. Como os reconhecedores de entidades nomeadas são dependentes de língua, para textos em inglês, foi utilizado o Open Calais (Reuters, 2008) e, para textos em português, por hora, optou-se por fazer uma tradução automática dos textos para inglês e utilizar o mesmo arcabouço.

. Pré-processamento: é utilizada a ferramenta PreProc, que utiliza a eTMLib (Dias *et al*, 2012), um aplicativo JAVA, que é o responsável pela construção das matrizes de incidência

de palavras nos textos. Nesta versão optou-se pelo uso de um vocabulário de agricultura (VOC-PORT e VOC-ING), com base no Thesagro e AgroVoc[6].

Figura 1: Arquitetura da Versão 1.0 do CRITIC



. Busca Sabiia: esse processo, aplicativo JAVA/Web, permite busca simples e avançada sobre a base de dados do Sabiia (Vacari et al, 2011). Foi totalmente reutilizado, porém, sofreu pequenos ajustes a fim de disparar os processos do CRITIC. Realizada uma consulta é possível disparar a análise dos resultados da consulta a partir de um botão na interface de consulta, então o Controlador dispara os demais processos.

. Controlador: é o coração do CRITIC/Web, um aplicativo JAVA que a partir de uma lista de identificadores dos resultados de busca provê processo para extrair as submatrizes de interesse e: dispara o processo de identificação de tópicos e de construção de arquivos intermediários de visualização (formato JSON JavaScript Object Notation[7]) para biblioteca D3 (Data-Driven Documents[8]); dispara o processo de identificação das distribuições temporais de cada tópico identificado e constrói um arquivo de representação dos gráficos; recupera a informação geoespacial de cada texto resultado da busca e cria a representação dos mapas para a biblioteca do *Google Maps*.

[6] Em http://aims.fao.org/vest-registry/vocabularies/agrovoc-multilingual-agriculture-thesaurus, consultado em maio de 2015.
[7]Em http://json.org, consultado em maio de 2015.
[8]Em http://d3js.org/, consultado em maio de 2015.

. Tópicos: tópicos são os descritores de agrupamentos de textos que, idealmente, devem corresponder a algum tema de interesse sobre a pesquisa realizada. Os agrupamentos, na versão 1.0 do CRITIC foram obtidos por agrupamento divisivo, com base no algoritmo "*bissecting kmeans*" implementado na ferramenta TORCH[9] (Marcacini e Rezende, 2010). A TORCH fornece seus próprios identificadores de grupos (palavras que estatisticamente melhor discriminam os grupo); e foi acoplada como um módulo da versão 1.0 do CRITIC.

. Temporal: neste módulo, é possível definir a evolução temporal de tópicos extraídos dos metadados de produção científica, de forma não supervisionada. Além disso, o usuário pode definir algumas palavras-chave para apoiar a definição de um tópico mais específico de seu interesse; e monitorar a evolução deste tópico ao longo do tempo.

. Mapas: este módulo é uma classe JAVA do Controlador, que recupera os identificadores dos textos em cada grupo de tópico e gera o mapa (um mapa Google) daquele grupo apresentando o retângulo envolvente das localizações.

. Visualizador/Seleções: este é um JSP (JAVA *Server Pages*) que controla as exibições dos agrupamentos, mapas, textos e gráficos temporais, de acordo com os tópicos selecionados. Como esse processo de visualização fornece elementos para uma análise exploratória de dados, optou-se por permitir a seleção de filtros. Na versão 1.0, os filtros podem ser conjuntos de palavras que aparecem nos tópicos apresentados e intervalos de tempo.

RESULTADOS

Nesta seção exemplifica-se o uso da versão 1.0 do CRITIC sobre resultados de uma busca e discute-se semelhanças entre suas funcionalidades e de outras ferramentas que permitem fazer o mesmo tipo de análise de dados. Assim:

. Busca: na Figura 2 é ilustrado o resultado de uma busca pela tema "cana-de-açúcar" em uma base de dados de testes do Sabiia, que possui 990 textos no tema. Deve-se notar que além das palavras-chaves cana-de-açúcar e suas variações morfossintáticas, outras palavras têm peso nos resultados como "solos", "controle biológico", etc. Pode-se filtrar a busca, selecionando repositórios ou palavras-chaves na parte esquerda da tela e depois ativar o botão "Atualizar" - que atualiza os resultados de busca. Considerando-se a busca satisfatória, pode-se ativar o botão "Analisar Busca".

. **Dendrograma**: na Figura 3 é ilustrado o resultado do agrupamento hierárquico dos textos resultantes da busca, com seus respectivos descritores (tópicos ou temas). Entre os possíveis tópicos de interesse, encontram-se temas como cevada, palmito e outras culturas,

[9]Disponível em http://sites.labic.icmc.usp.br/torch/, consultado em maio de 2015.

não apenas cana, solos, clima, etc. Esses outros tópicos podem ser navegados a fim de explorar o porquê de terem surgido; por exemplo, com a soja ocorre uma rotação de culturas e há influência na produção de ambas. O usuário pode navegar entre todos esses tópicos, na árvore em português ou inglês, e ir abrindo os nós de seu interesse. Associado a cada tópico é apresentada uma relação dos textos sob aquele tópico, gráfico do tópico e mapa; além de que todo texto lido pelo usuário é relacionado no histórico.

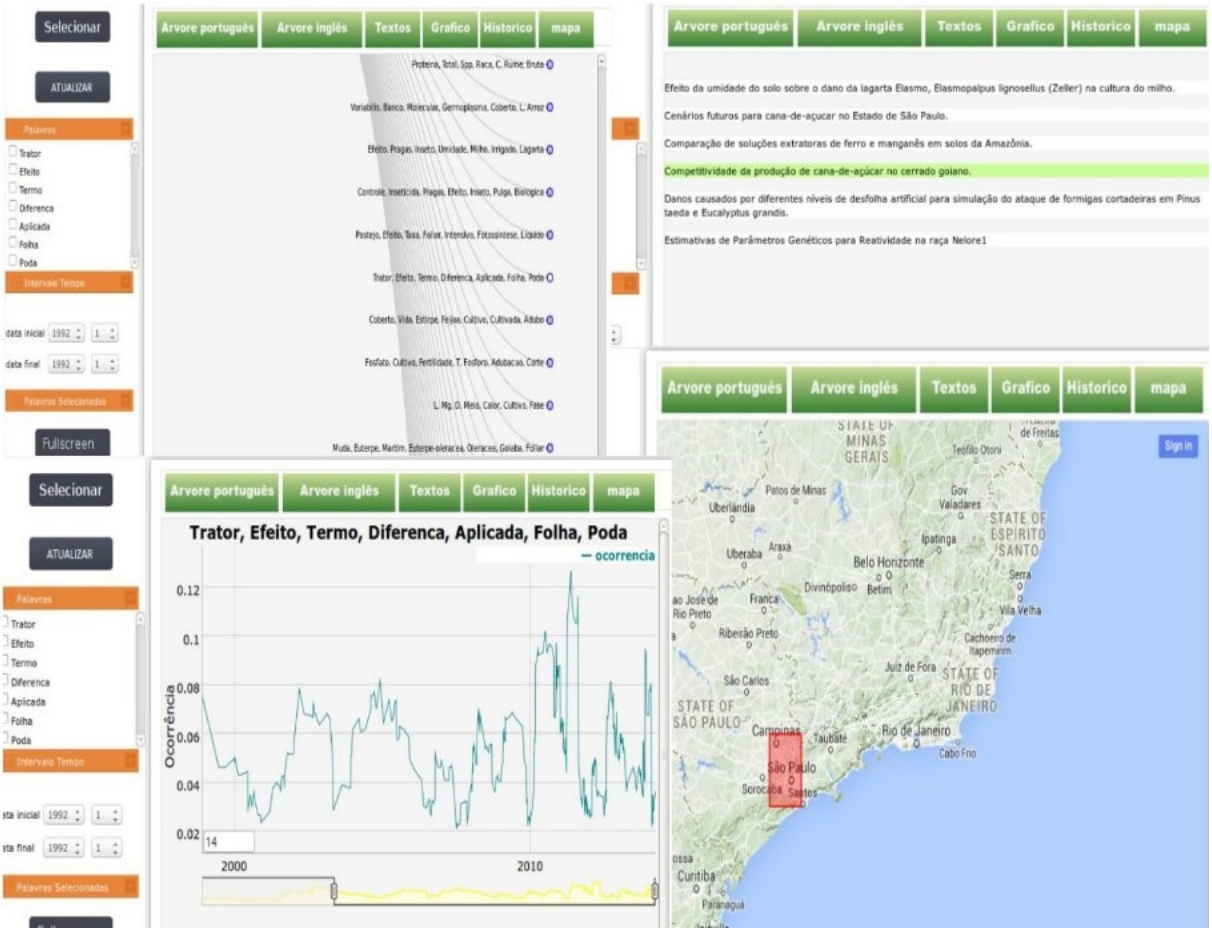
Figura 2: Busca por cana-de-açúcar e seus 990 resultados

Figura 3: Agrupamento hierárquico com os tópicos encontrados

. **Abrindo um nó específico:** abrindo-se o nó "Trator, Efeito, Termo, Diferença, Aplicada, Folha, Poda", ilustração da Figura 4, pode-se observar os textos sob o tema (e abrir os textos que são disponíveis na internet, a partir da URI), verificar a distribuição do tópico ao longo das publicações que cobrem a busca no tempo (os picos do gráfico são os pontos de interesse), e verificar o mapa de distribuição dos textos.

. **Filtrando a análise realizada**: de forma similar ao filtro da busca do Sabiia, pode-se escolher quais palavras e intervalos de tempo devem ser mantidos nas análises seguintes. Para isso, selecionam-se as palavras na parte esquerda das telas, apresentadas na Figura 4, e então, vai-se selecionando cada qual em cada novo nó navegado. Por fim, atualiza-se a análise de dados com as palavras e intervalos de tempos selecionados - botão Atualizar.

Figura 4: Visão geral dos resultados com tópicos, textos, gráfico e mapa



DISCUSSÕES E TRABALHOS FUTUROS

A proposta do CRITIC é trabalhar com repositórios OAI, funcionalidade que em sua primeira versão é proporcionada pelo Sabiia. Porém, o Sabiia recupera apenas os metadados dos repositórios, tais como título, nome de autores, resumo, etc. Em versão futura, a indexação dos textos dos repositórios OAI deverá ser executada sobre artigos completos, quando

disponíveis gratuitamente, ou metadados, caso contrário; sempre com base em vocabulário agrícola. O uso de textos integrais, não apenas metadados, deve melhorar a identificação de localidades geográficas e a extração de tópicos, pois só resumos e palavras-chaves não refletem por completo os conteúdos dos textos. Além disso, a busca deve poder ser expandida a partir das relações das palavras em um *thesaurus*. Não há ferramenta, citada na literatura até o momento, que possua diretamente essas funções e possa ser acoplada ao CRITIC.

Das ferramentas existentes a que mais se aproxima do CRITIC é a Sci2. Em comum elas realizam o mesmo tipo de análise temporal, identificando pontos de interesse e encontram as localizações geográficas citadas nos textos. A Sci2 também procura tópicos entre os textos, porém, não são tópicos hierárquicos, ela divide a coleção de textos em grupos planos independentes e atribui-lhes identificadores. A vantagem de tópicos hierárquicos, além da própria organização em si, é que, por vezes, assuntos mais específicos, e talvez de maior interesse, podem estar próximos às folhas da árvore, mesmo quando os mais genéricos (nós mais próximos à raiz) parecem incompreensíveis ou sem nexo. Esses tópicos específicos podem estar diluídos e até escondidos nos grupos quando a divisão é plana. A Sci2 também encontra redes de autoria, funcionalidade ainda não presente na CRITIC. Porém, a Sci2 não é código aberto, o que inviabiliza reusar seus módulos ou mesmo contribuir com seu desenvolvimento, com novos módulos. Outra ferramenta com alguma similaridade, é a Google Ngram Viewer[10], cuja funcionalidade é localizar termos em coleções indexadas pelo Google, mostrar sua distribuição temporal e permitir a recuperação dos textos, por termos e/ou intervalo de tempo. Porém, a distribuição temporal apresentada é para cada termo, não a distribuição conjunta de termos (ou palavras) como nos tópicos da CRITIC.

Como trabalhos futuros, tem-se a questão da indexação da busca por vocabulário fixado e de algumas poucas melhorias na versão – como seleção de localidades, etc; só após a liberação dessa nova versão em beta teste e a fixação de *bugs* que venham a surgir, o CRITIC será colocado em repositório *Open Source*. Pretende-se expandir as funcionalidades do CRITIC para outras análises de dados, tais como correlações entre temas, expandir a cobertura de regiões geográficas e línguas utilizadas, permitir tarefas de classificação.

CONCLUSÕES

O objetivo do CRITIC é ser um software de mineração de textos para a área de agricultura, utilizando repositórios OAI, indexando-os de acordo com vocabulário de agricultura, realizando a geoespacialização de documentos e temas, explorando a organização hierárquica

[10]Em https://books.google.com/ngrams, consultado em maio de 2015.

de temas, e outras funcionalidades que venham a ser requisitadas. As soluções aqui apresentadas, bem como a própria versão 1.0 do ambiente, mostram que grande parte desses objetivos já foram alcançados. Incorporando-se novas funcionalidades de exportação de dados para vários padrões, permitir-se-á utilizar recursos de várias outras ferramentas. E, após sua colocação em repositório *Open Source*, espera-se que ele atinja um público maior e que receba contribuições ao seu desenvolvimento.

REFERÊNCIAS

AGGARWAL, C. Mining Text Data. In: AGGARWAL, C. Data Mining. 2015, ISBN: 978-3-319-14141-1, p 429-455.

DIAS, V.F. MOURA, M.F. CRUZ, S.A.B. HIGA, R.H. Evolução da eTMLib - Embrapa's Text Mining Library para pré-processamento de dados textuais. In: Mostra de Estagiários e Bolsistas da Embrapa Informática Agropecuária, 8., 2012, Campinas. Resumos... Brasília, DF: Embrapa, 2012.

MACHADO, L.S. MOURA, M.F. TAKEMURA, C.M. Modelagem e desenvolvimento de reconhecedor geoespacial para documentos textuais. In: Mostra de Estagiários e Bolsistas da EMBRAPA Informática Agropecuária, 9., 2013, Campinas. Resumos... Brasília, DF: Embrapa, 2013, p. 35-38.

MARCACINI, R. M. REZENDE, S. O. Torch: a tool for building topic hierarchies from growing text collections. In WFA'2010: IX Workshop on Tools and Applications. Belo Horizonte - MG, Brasil. Webmedia'2011: Brazilian Symposium on Multimedia and the Web, pages 1-3, 2010.

PILKINGTON, Alan. Technology management: A comprehensive bibliometric analysis. In: Technology Management Conference (ITMC), 2014 IEEE International. IEEE, 2014. p. 1-4.

REUTERS, Thomson. OpenCalais. Retrieved June, v. 16, 2008.

VACARI, I. VISOLI, M.C. GONZALES, L.E. Acesso aberto a informação científica agropecuária na internet: caso do sistema aberto e integrado de informação em agricultura (Sabiia). In: Congresso Brasileiro de Agroinformática, 8., 2011, Bento Gonçalves. Anais... Florianópolis: UFSC; Pelotas: UFPel, 2011.