Extração de hierarquias de tópicos em textos para apoiar a construção de portfólios tecnológicos

Carolina Tavares de Oliveira[1]
Luisa Miyashiro Tápias[2]
Stanley Robson de Medeiros Oliveira[3]
Maria Fernanda Moura[4]

Resumo: Neste trabalho são relatadas etapas para se extrair hierarquias de tópicos em coleções de documentos, compostas por publicações científicas, com o objetivo de auxiliar a construção de portfólios de tecnologias agrícolas diretamente relacionadas ao uso de recursos hídricos. Com base nas análises ao longo do processo, que é interativo, foi sendo construído um vocabulário controlado, que melhor representa o tema dos portfólios. Com esse vocabulário, que é utilizado para agrupar os documentos, os tópicos encontrados apresentaram um papel relevante na construção de tais portfólios.

Palavras-chave: mineração de textos, agrupamento de textos, recuperação de informação.

Introdução

O foco deste trabalho é construir portfólios de tecnologias para agricultura, nos quais são relatados os recursos tecnológicos empregados para viabili-

[1] Estudante de Engenharia Agrícola da Universidade Estadual de Campinas (Unicamp) bolsista da Embrapa Informática Agropecuária, Campinas, SP.

[2] Estudante de Engenharia Agrícola da Universidade Estadual de Campinas (Unicamp), estagiário da Embrapa Informática Agropecuária, Campinas, SP.

[3] Doutor Ciências da Computação, Pesquisador Embrapa Informática Agropecuária, Campinas, SP.

[4] Doutora em Ciências da Computação, Pesquisadora Embrapa Informática Agropecuária, Campinas, SP.

zar a produção de culturas de forma sustentável, voltando-se a atenção para a questão do uso da água no setor agrícola brasileiro; dado que, este é o principal usuário consuntivo do recurso se comparado ao setor industrial e ao de consumo doméstico. Os portfólios são planilhas elaboradas contendo as informações coletadas e organizadas a partir de uma coleção delimitada de textos. O trabalho aqui proposto visa automatizar ou auxiliar futuros processos de construção de portfólios a partir da literatura, utilizando técnicas de mineração de textos.

O processo de Mineração de Textos pode ser dividido em cinco etapas (MOURA, 2009): a) identificação do problema; b) pré-processamento; c) extração de padrões; d) pós-processamento; e) utilização do conhecimento. Na identificação do problema são definidos os objetivos do processo de Mineração de Textos. Na etapa de pré-processamento, os dados são manipulados a fim de se obter uma representação que possa ser lida por ferramentas de extração de padrões (ferramentas de aprendizado de máquina). Na extração de padrões, técnicas específicas são usadas de acordo com os resultados esperados. No caso dos tópicos, são analisados todos os ramos da hierarquia, procurando identificar os tópicos de maior interesse, que auxiliem a identificação de tecnologias e relações associadas a elas, tais como, tipo de solo, geolocalidade, etc. Ainda, com o arsenal de ferramentas utilizado, também pode-se observar a distribuição temporal de cada tópico em análise, ou seja, em que épocas aquelas tecnologias e suas características associadas estiveram em evidência. No pós-processamento, os resultados encontrados são analisados e validados. Se o processo não resultar em resultados usáveis, repetem-se as etapas, isto é, desde o pré-processamento.

O objetivo deste trabalho é utilizar técnicas de mineração de textos para extrair tópicos de hierarquias em textos visando à construção de portfólios de tecnologias agrícolas relacionadas ao uso de recursos hídricos.

Materiais e Métodos

A metodologia empregada foi composta de quatro fases: a) busca; b) pré-processamento; c) geração de hierarquia; d) pós-processamento (Figura 1), resultando em um processo retroalimentável.

Figura 1. Processo de mineração de textos utilizado.

Busca: Os textos foram selecionados do Sistema Integrado e Aberto de Informação em Agricultura (SABIIA) (VACARI et al., 2011), este é um mecanismo de busca automatizado, que coleta metadados de provedores de dados científicos.

Pré-processamento: utilizou-se a ferramenta I-PreProc, em desenvolvimento na Empresa Brasileira de Pesquisa Agropecuária (Embrapa), para gerar uma matriz de termos (colunas) por documentos (linhas); considerando-se os termos de uma lista de vocábulos previamente fixados. Cada célula da matriz contém a frequência de ocorrência do vocábulo no texto. São gerados dois arquivos: o de extensão DAT com os valores das células (grau de importância de cada termo/palavra em cada documento) e o de extensão HDR com a descrição dos textos (nomes) e vocábulos (termos) utilizados.

Geração de hierarquias: geram-se dendrogramas, que são representações de agrupamentos hierárquicos, por meio do software torch[5] que transforma a matriz em uma hierarquia (arquivo xml). Cada nó corresponde a um tópico com seus descritores, e, associados a cada tópico, os gráficos de distribuição temporal destes. Na sequência, utiliza-se a ferramenta topicVis (em desenvolvimento na Embrapa), que lê o xml gerado pela torch e cria arquivos nos padrões jsons e html, que podem ser visualizados utilizando-se um navegador.

Pós-processamento: analisam-se as hierarquias, verificando se os resultados são descartáveis ou utilizáveis, neste caso, na construção dos portfólios. Se o resultado ainda for insuficiente, dada esta validação subjetiva, então verifica-se a qual etapa retornar.

[5] Disponível em: <http://sites.labic.icmc.usp.br/torch/>. Acesso em: 20 out. 2015.

A partir do uso deste processo, são relatados alguns resultados inicias (portfólios tecnológicos) a partir dos quais fez-se uma primeira análise. Depois dessa primeira análise, verificou-se a necessidade de se buscar por novas fontes de dados (selecionar uma nova base de textos e melhorar os vocábulos e expressões de busca) para que fossem geradas novas hierarquias.

Resultados e Discussão

A partir da SABIIA foram reunidos 643 documentos e metadados, gerou-se dendrogramas a partir da coleção conforme a literatura. A partir das hierarquias realizou-se uma análise exploratória nos tópicos identificados, utilizando as informações apresentadas para a elaboração manual de uma primeira aproximação do portfólio de tecnologias, mostrado na Tabela 1.

Tabela 1. Representação de portfólios tecnológicos.

Portfólio de tecnologias								
Gráfico	**Documento**	**Tópico**	**Localidade**	**Ano**	**Titulo**	**Cultura**	**Tecnologia**	**Tecnologia associada**
Entre 09/05/1995 a 08/09/2001	0083.190.txt	Condutividade	Sete Lagoas MG	2013	Irrigação de pastagens	Forrageira	Irrigação por micro-aspersão	Fertirrigação
		Hidráulica						Localizada
		Censo						Aspersão
		Crédito						Capsula porosa
		Aviaria						Gotejamento
		Cevada						Infiltração
		Contrato						
...	...	...	...	...	...	...	...	...

Além das hierarquias, verificavam-se os conteúdos dos textos em cada tópico e a distribuição temporal do tópico – picos desta. Uma parte da hierarquia é mostrada na Figura 2.

Depois da análise dos dados, verificou-se que poderia ser mais interessante ampliar o vocabulário do domínio, para que os descritores dos tópicos fossem mais significativos e abrangessem temas relacionados às tecnologias agrícolas associadas a recursos hídricos. Selecionou-se terminologias nas subáreas de geo localidades, tipos de solo, e tecnologias. Esses termos foram obtidos da planilha e expandidos com as relações do

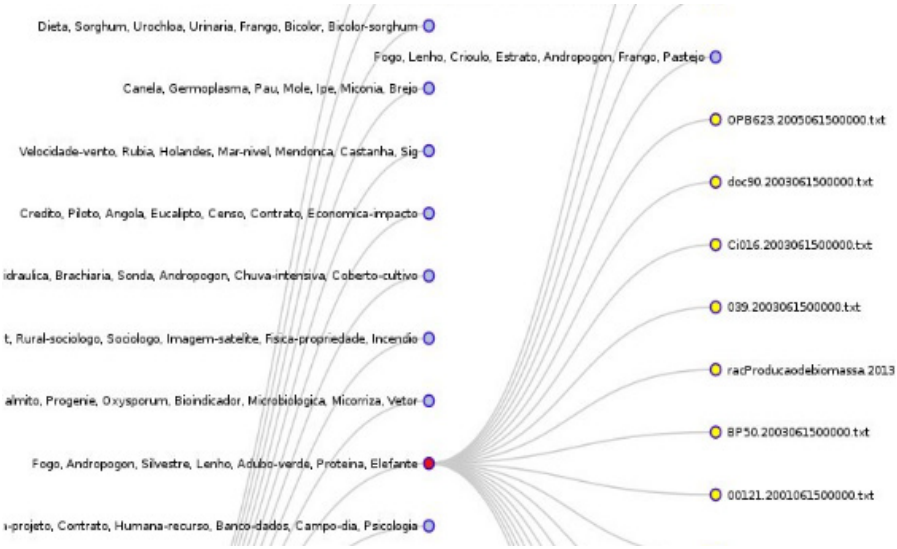

Figura 2. Tópicos construídos para os primeiros 643 documentos.

Thesagro[6]. Alguns termos obtidos são Irrigação por aspersão, Irrigação por microaspersão, Irrigação pivô central, Irrigação subterrânea, Irrigação por sulco, Irrigação por pote de barro, pivô central etc. De geo localidades, foram utilizadas as presentes nos textos, como Teresina, Salvador, Sete Lagoas, Campinas, etc. E, também foram incluídos só tipos de solos, como Latossolo, Neossolo, Organossolo, etc. Construiu-se um novo vocabulário, buscaram-se por novos textos e metadados no SABIIA, novas hierarquias foram geradas a partir da mudança do vocabulário. O resultado é mostrado na Figura 3.

Nesses diagramas cada ramo da árvore corresponde às palavras que funcionam como descritores dos tópicos, agrupando um conjunto de textos similares. Um possível tópico é apresentado na Figura 3. Na Figura 3 está em evidência o tópico descrito pelas palavras Melancia, Irrigação, Fertirrigação, Água, Adubo, Parnaíba; ao qual está associada uma série temporal, que indica a distribuição no temporal desse tópico em relação aos textos agrupados Os resultados apresentam uma melhora subjetiva, como

[5] Disponível em: <http://snida.agricultura.gov.br:81/binagri/html/Cen_Thes1.html>. Acesso em: 20 out. 2015.

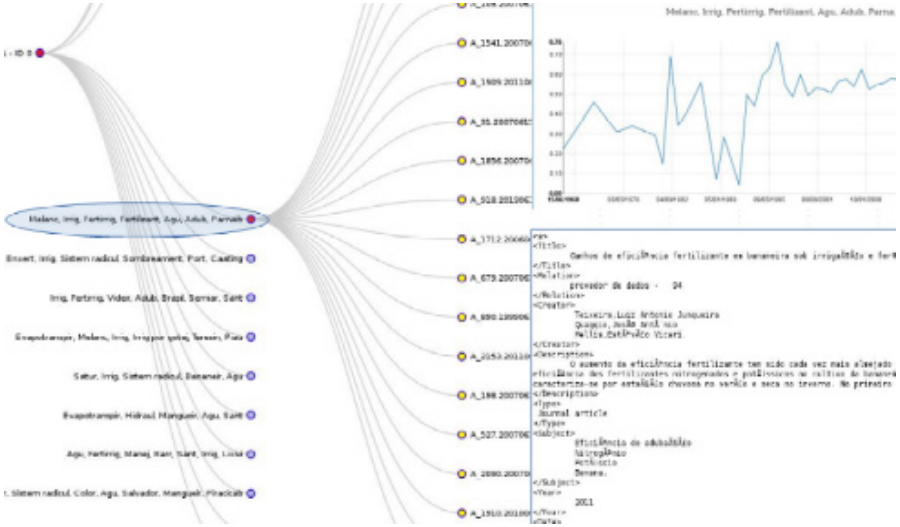

Figura 3. Distribuição temporal de um tópico e exemplo de documento.

se pode observar, na Figura 2, os tópicos contém termos, tais como nomes de tecnologias utilizadas no setor agrícola brasileiro, diretamente associadas a recursos hídricos.

Considerações Finais

O processo interativo utilizado está possibilitando que se obtenham melhores resultados, e mais especificidade na construção dos portfólios, isso foi observado repetindo o processo para se gerar novas hierarquias de tópicos.

Referências

MOURA, M. F. **Contribuições para a construção de taxonomias de tópicos em domínios restritos utilizando aprendizado estatístico**. 2009. 137 f. Tese (Doutorado em Ciências de Computação e Matemática Computacional) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, SP. Disponível em: <http://ainfo.cnptia.embrapa.br/digital/bitstream/item/17875/1/MFM_Tese_5318963-2.pdf). Acesso em: 20 out. 2015.

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. Não paginado. Disponível em: <http://ainfo.cnptia.embrapa.br/digital/bitstream/item/46253/1/89811-1.pdf >. Acesso em: 20 out. 2015.