Extração de associações entre itens de um portfólio de tecnologias agrícolas

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

Resumo: Este trabalho apresenta uma metodologia para a extração de regras de associação de um portfólio de tecnologias agrícolas, geradas a partir de publicações científicas. Foi necessário semiautomatizar o processo de construção do portfólio, dada a quantidade expressiva de textos que foram selecionados do Sistema Aberto e Integrado de Informação em Agricultura (SABIIA). A partir desse portfólio foram geradas regras de associação para identificar as relações existentes entre atributos como solo, tecnologias, localidade e culturas, a fim de subsidiar especialistas do domínio, especialmente de agricultura irrigada, na verificação de quais tecnologias podem ser adaptadas para os biomas brasileiros.

Palavras-chave: Regras de associação, portfólio, vocábulos tecnológicos, tecnologias agrícolas, mineração de textos.

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

[2] Estatística, doutora em Ciências Matemáticas e da Computação, pesquisadora da Embrapa Informática Agropecuária, Campinas, SP.

[3] Bacharel em Ciência da Computação, Ph.D. em Ciência da Computação, pesquisador da Embrapa Informática Agropecuária, Campinas, SP.

Introdução

Para garantir produções agrícolas mais sustentáveis é importante conhecer as tecnologias empregadas e a sua relação com a região ou o ecossistema relacionado, de acordo com a necessidade da cultura. Para cada cultura, considera-se como fatores intrínsecos a sua natureza aqueles que favorecem o seu desenvolvimento, como o solo e o clima que estão por sua vez relacionados com a localidade.

Uma alternativa para encontrar a relação existente entre a tecnologia e esses fatores é a extração de regras de associação a partir de portfólios tecnológicos, que são planilhas elaboradas a partir das informações coletadas e organizadas a partir de uma coleção delimitada de textos.

Após a geração das regras de associação, estas são avaliadas por especialistas do domínio agrícola, com o objetivo de selecionar quais tecnologias podem ser adaptadas para os biomas brasileiros.

O objetivo deste trabalho foi construir um portfólio semiautomático de tecnologias agrícolas, construído a partir de informações disponíveis em publicações científicas, e extrair regras de associação entre itens desse portfólio.

Materiais e Métodos

A metodologia utilizada neste trabalho é constituída de um processo de mineração de textos retroalimentável, conforme descrito na Figura1 e detalhado a seguir.

1. Busca por textos: os textos foram selecionados do Sistema Integrado e Aberto de Informação em Agricultura (SABIIA) (VACARI et al., 2011) por meio de palavras-chave e expressões de busca escolhidas por especialistas do domínio. O SABIIA coleta metadados de provedores de dados científicos, como artigos científicos e tecnológicos, sendo todos no padrão Open Archives Initiative (OAI). Esses provedores contêm as publicações técnico-científicas da Empresa Brasileira de Pesquisa Agropecuária (Embrapa), considerada uma fonte suficiente de informação para este trabalho.

2. Pré-processamento: nesta etapa utilizou-se a ferramenta I-PreProc (PEREIRA; MOURA, 2015), desenvolvida pela Embrapa Informática Agropecuária, para gerar uma matriz de termos (colunas) por documentos

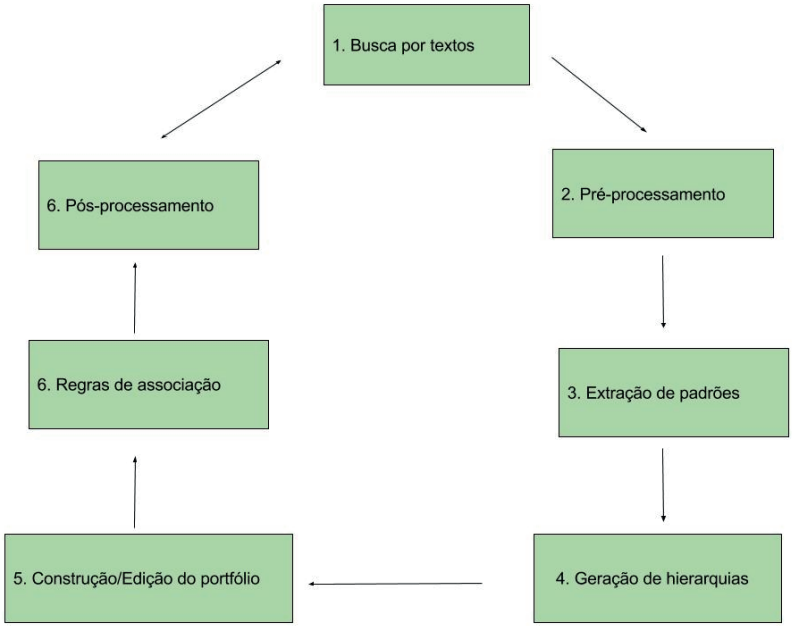

Figura 1. Processo de mineração de textos.

(linhas); considerando-se a intersecção entre os termos presentes nos documentos e uma lista de vocábulos previamente fixados. Cada célula da matriz contém a frequência de ocorrência do vocábulo no documento. 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) presentes nos textos.

3. Extração de padrões: como a base de textos não é pré-categorizada, nesta etapa utilizaram-se algumas técnicas de aprendizado de máquina não supervisionado, tais como a obtenção de hierarquias de tópicos sobre os textos já pré-processados.

4. Hierarquias de tópicos: a extração de uma hierarquia de tópicos tem como objetivo facilitar a navegação e exploração da coleção de textos, que é agrupada hierarquicamente de acordo com a similaridade entre os documentos – descritos como vetores de frequência de termos. Cada tópico é descrito por uma relação de palavras (ou termos); a relação contém as pa-

lavras estatisticamente mais significativas no grupo, dado algum critério. A função dessa relação de palavras é ajudar a identificar a que tópico (tema) o grupo de documentos se refere.

5. Construção/Edição do portfólio: o portfólio foi construído manualmente a partir da elaboração de planilhas contendo as seguintes informações: cultura, tecnologia, tecnologia associada, solo, localidade, região do Brasil (Norte, Nordeste, Sul, Sudeste ou Centro-Oeste) e UF para cada texto. Nesta etapa também foram descartados os textos repetidos encontrados e arquivos com textos ilegíveis.

6. Regras de associação: buscam encontrar as relações entre itens de dados que ocorram com uma certa frequência, ou seja, identificar padrões em dados históricos (AGRAWAL; SRIKANT, 1994). Essas regras foram obtidas com a utilização do algoritmo Apriori (LIU et al., 1998). Sendo X e Y conjuntos de atributos tais como tecnologia, tipo de solo, local, cultura, tal que $X \cap Y = \Phi$, as regras podem ser representadas da forma $X \rightarrow Y$. Para cada regra estão associadas as métricas: suporte (Sup) e confiança (Conf) conforme equações (1) e (2), respectivamente.

$$Suporte(X \rightarrow Y) = P(X \cup Y) \quad (1)$$

$$Confiança(X \rightarrow Y) = P(X/Y) = \frac{Suporte(X \rightarrow Y)}{Suporte(X)} \quad (2)$$

Do ponto de vista conceitual, o suporte representa a significância estatística dos itens (termos) nas tuplas, ao passo que a confiança determina a força da regra.

7. Pós-processamento: nesta etapa, analisam-se se as regras obtidas fornecem resultados descartáveis ou utilizáveis por especialistas do domínio. Caso o resultado tenha sido insuficiente, se verifica qual etapa deve-se retornar para iniciar o processo novamente até obter resultados favoráveis.

Resultados e Discussão

Foram reunidos 2.209 documentos e metadados a aprtir da base de dados SABIIA, utilizando-se as expressões de busca construídas pelos especialistas do domínio. Esse conjunto de resultados consiste de textos completos

de todos os artigos com acesso livre e, no caso de textos sem acesso livre, foram utilizados metadados; criando-se uma base de textos. Essa base primeiramente foi pré-processada com a ferramenta I-PreProc e um vocabulário controlado, criado a partir da junção de quatro glossários da área de recursos hídricos e dois tesaurus (Thesagro e Agrovoc).

Depois foi gerada a hierarquia de tópicos e a partir dela realizou-se uma análise exploratória para a construção do portfólio. Foram eliminados os textos repetidos ou ilegíveis. Em seguida, foram extraídos os atributos: Tecnologia, Tecnologia associada, Localidade, UF, Região, Tipo do solo e Cultura, para cada texto. A versão final do portfólio ficou constituída por uma planilha com 1.490 linhas. A Tabela 1 apresenta parte do portfólio gerado.

Para gerar as regras, fixou-se a Cultura como o consequente de regra (Y)

Tabela 1. Visão parcial do portfólio de tecnologias.

Tecnologia	Tecnologia associada	Localidade	UF	Regiao	Solo	Cultura
irrigação	manejo de água	lago cujubim grande	RO	N		feijão-caupi
irrigação						melão
irrigação	manejo de água	petrolina	PE	NE	latossolo	uva
irrigação	adubação mineral	Vale do São Francisco	MG	NE		
irrigação		cruz das almas	BA	NE		banana
irrigação	irrigação por aspersão	petrolina	PE	NE	latossolo	uva
irrigação	irrigação por aspersão	belém	PA	N		açaí
irrigação	irrigação por aspersão	amazônia	AM	N		açaí
irrigação	irrigação por gotejamento	paty dos alferes	RJ	SE	latossolo	tomate
irrigação	irrigação por gotejamento	lavras	MG	SE	latossolo	tomate
irrigação	manejo de água	petrolina	PE	NE	latossolo	uva
irrigação	irrigação por gotejamento	petrolina	PE	NE	latossolo	uva
irrigação	irrigação por gotejamento	urussanga	SC	S		uva
manejo	manejo de irrigação	urussanga	SC	S		uva
irrigação	irrigação por aspersão	brasília	DF	CO		tomate
adubação	irrigação por gotejamento	brasília	DF	CO		tomate
		sul do rio grande do sul	RS	S		tomate
irrigação	irrigação localizada	cansanção	BA	NE	planossolo	banana
irrigação		mossoró	RN	NE		tomate

e os outros atributos foram combinados no antecedente de regra (X), ou seja, para uma cultura agrícola verificou-se quais tecnologias, locais, tipos de solo ou UF estão associados. O portfólio foi segmentado por região para obter uma melhor visualização do resultado e facilitar a geração de regras na seguinte escala: a) Norte (21 instâncias); b) Nordeste (773 instâncias); c) Sudeste (199 instâncias); d) Centro-Oeste (72 instâncias); e) Sul (65 instâncias); f) 360 instâncias sem a definição de UF e Região foram descartadas para não influenciar os resultados.

Dessa forma gerou-se as regras de associação para cada região do Brasil com especificações mínimas de suporte e confiança. A seguir são exempli-

ficadas algumas das regras obtidas a fim de obter a validação dos especialistas.

Norte (22 regras, suporte de 6% e confiança de 90%): (**Se Localidade** = Capitão Poço **& UF**=PA **& Solo**=latossolo ==> Cultura=banana). Outro exemplo foi (**Se Tecnologia_associada**=manejo de água **& UF**=RO ==> **Cultura**=feijão).

Nordeste (21 regras, suporte de 1% e confiança de 80%): (**Se Tecnologia_associada**=manejo de irrigação **& Localidade**=Cruz das Almas **& UF**=BA ==> **Cultura**=banana) e (**Se Tecnologia_associada**=irrigação por aspersão & UF=PE ==> **Cultura**=uva).

Sudeste (21 regras, suporte de 2% e confiança de 80%): (**Se Tecnologia_associada**=irrigação por gotejamento **& Localidade** = viçosa ==> **Cultura**=tomate) e (**Se Tecnologia_associada**= manejo de cobertura de solo **& Localidade**=Paty dos Alferes **& UF**=RJ ==> **Cultura**=tomate).

Centro-Oeste (48 regras, suporte de 4% e confiança de 90%): (**Se Tecnologia_associada**=manejo de água **& Localidade**=Brasília ==> **Cultura**=tomate) e (**Se Tecnologia_associada**=variabilidade melhoramento genético & UF=DF ==> **Cultura**= batata-doce).

Sul (31 regras, suporte de 4% e confiança de 80%): (**Se Tecnologia_associada**=irrigação por gotejamento **& Localidade**=Santa Maria ==> **Cultura**=tomate) e (**Se Localidade**=Santana do Livramento **& UF**=RS ==> **Cultura**=uva).

Nota-se que apenas alguns atributos estiveram presentes nas regras, como localidade, tecnologia associada e cultura, embora o portfólio seja constituído de mais atributos e boa parte das regras foi redundante ou não apresentou novidades.

Considerações Finais

A construção do portfólio envolveu muito trabalho manual para uma grande quantidade de textos com a análise exploratória de hierarquias, o que revela a necessidade de processos semiautomatizados para futuros portfólios.

Ao realizar a construção de tópicos, baseou-se em técnicas que envolvem filtros estatísticos, e não em processamento de língua natural, obtendo-se

resultados puramente estatísticos. Outro problema encontrado foi o fato de a maioria das ferramentas para lidar com o processamento de língua serem concebidas para a língua inglesa, exigindo a tradução do português para o inglês, o que pode ocasionar erros de sintaxe, de semântica e perdas de informações importantes.

O portfólio apresentou uma alta quantidade de dados esparsos, já que muitos textos não apresentaram informações para preencher todos os atributos (tipo de solo, localidade, tecnologia associada, entre outras), por isso os suportes foram muito baixos em quase todos os casos analisados.

Uma solução em andamento é a utilização de ferramentas que reconhecem entidades nomeadas em textos, como localidades, termos industriais, tecnologias, etc.

Referências

AGRAWAL, R.; SRIKANT, R. Fast algorithms for mining Association Rules in Large Databases. In: INTERNATIONAL CONFERENCE ON VERY LARGE DATA BASES, 20., 1994, Santiago. **Proceedings**... Santiago: Morgan Kaufmann, 1994. p. 478-499. VLDB.

LIU, B.; HSU, W.; MA,Y. Integrating classification and Association Rule Mining. In: INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 4.,1998, New York. **Proceedings**… Menlo Park: AAAI, 1998. p. 80-86.

PEREIRA, R. G.; MOURA, M. F. I-Preproc: uma ferramenta para pré-processamento e indexação incremental de documentos. In: MOSTRA DE ESTAGIÁRIOS E BOLSISTAS DA EMBRAPA INFORMÁTICA AGROPECUÁRIA, 11., 2015, Campinas. **Resumos expandidos**... Brasília, DF: Embrapa, 2015. p. 17-23.

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.