Extração de Portfólio de Tecnologias de Irrigação a partir de Publicações Científicas

Stanley Robson de M. Oliveira; Maria Fernanda Moura; Celina Maki Takemura; Luísa Miyashiro Tápias; Carolina Tavares de Oliveira; Luis Henrique Bassoi

Resumo

O objetivo do trabalho é apresentar uma metodologia para extração de portfólio de tecnologias de irrigação, a partir de publicações científicas, para melhorar o uso sustentável da água na agricultura. A construção do portfólio foi semiautomatizada por meio de um processo de mineração de textos em documentos selecionados do Sistema Aberto e Integrado de Informação em Agricultura (SABIIA). Com base nesse portfólio, foram geradas regras de associação para identificar a relação entre tecnologias, localidade e culturas, nas regiões do Brasil, com a finalidade de subsidiar especialistas do domínio na verificação de quais tecnologias podem ser adaptadas para os biomas brasileiros.

Termos para indexação: recursos hídricos, mineração de textos, manejo da água.

Extraction of Irrigation Technology Portfolio from Scientific Publications

Abstract

In this work, we present a methodology for extraction of irrigation technology portfolio from scientific publications to improve the sustainable use of water in agriculture. The construction of the portfolio was semi-automated through a text mining process in documents selected from the Integrated and Open Information System in Agriculture (SABIIA). Based on this portfolio, association rules were generated to identify the relationship among technologies, localities and cultures in the Brazilian regions, in order to subsidize domain experts in the verification of which technologies can be adapted to the Brazilian biomes.

Index terms: hydric resources, text mining, management of water.

Introdução

Informações sobre variáveis hídricas são essenciais para o manejo dos recursos hídricos em qualquer ecossistema ou região. Dependendo da região ou ecossistema considerado, podem ser destacadas tecnologias como sistema de produção agrícola irrigado para pequenos produtores, sistema de plantio direto, e o uso da barragem subterrânea e do açude para captação e armazenamento de água. O impacto dessas práticas sobre a recuperação da qualidade e conservação de solo e da água é de fácil percepção, no entanto, dependente da especificidade de cada tecnologia.

Uma das alternativas para colaborar com o manejo dos recursos hídricos na agricultura é a elaboração de um portfólio de tecnologias que contemple diversas escalas (parcela, propriedade agrícola, perímetro irrigado, bacia hidrográfica), cada uma com suas particularidades. O portfólio pode ser representado por uma planilha contendo a relação de tecnologias, locais, época, tipo de solo, culturas e possíveis tecnologias associadas.

Considerando a contribuição da Tecnologia da Informação (TI) no âmbito do Projeto AgroHidro, tanto na geração de soluções como na sua gestão, o objetivo deste trabalho foi construir um portfólio de tecnologias de irrigação a partir de publicações científicas. A partir deste portfólio foram geradas regras de associação para identificar a relação entre tecnologias, localidade e culturas, nas regiões do Brasil, com a finalidade de subsidiar os especialistas do domínio na verificação de quais tecnologias podem ser adaptadas para os biomas brasileiros.

Material e Métodos

A metodologia utilizada neste trabalho é constituída de um processo de mineração de textos, cujas etapas são descritas a seguir:

1) Busca por textos: os textos foram selecionados do Sistema Aberto e Integrado de Informação em Agricultura – SABIIA (VACARI et al., 2011). Este é um mecanismo de busca automatizado que coleta

metadados de provedores de acesso aberto, artigos científicos e tecnológicos, todos no padrão OAI – Open Archives Initiative. Esses provedores contêm a maior parte das publicações utilizadas pelos pesquisadores da Rede AgroHidro, bem como todas as publicações técnico-científicas da Embrapa; o que se mostrou como uma fonte única e suficiente de informação. Primeiramente, buscou-se obter informações nos textos referentes às datas, às localidades, às tecnologias e algumas informações adicionais relevantes. Para isso, foram utilizadas algumas palavras-chave selecionadas por especialistas do domínio.

2) Pré-processamento: nesta etapa utilizou-se a ferramenta I-PreProc (PEREIRA; MOURA, 2015), em desenvolvimento pela Embrapa Informática Agropecuária, para gerar uma matriz de termos (colunas) por documentos (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 e regras de associação sobre os textos já pré-processados.

   a) Hierarquias de tópicos: a extração de uma hierarquia de tópicos visa facilitar a navegação e exploração da coleção de textos, que é hierarquicamente agrupada 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 palavras 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; o que auxiliou a construção do portfólio.

b) Regras de associação: as regras de associação buscam encontrar o relacionamento entre itens de dados que ocorram com uma certa frequência, ou seja, identificar padrões em dados históricos (AGRAWAL; SRIKANT, 1994). Para a geração das regras de associação foi utilizado o algoritmo Apriori (LIU et al., 1998). As regras de associação podem ser representadas da forma $X \rightarrow Y$, em que X e Y são conjuntos de atributos tais como tecnologia, tipo de solo, local, cultura, tal que $X \cap Y = \emptyset$. Para cada regra estão associadas duas medidas tradicionais: confiança (Conf) e suporte (Sup). Sup representa o número de tuplas que contêm X e Y, ao passo que Conf constitui a razão entre o número de tuplas que contêm X e Y sobre o número de tuplas que contêm X. Uma regra é considerada interessante quando ela apresenta um suporte e uma confiança iguais ou superiores ao mínimo estabelecido pelo usuário.

4) Pós-processamento: com base no o portfólio, regras de associação foram geradas para identificar a relação entre tecnologias, localidade e culturas, nas regiões do Brasil. O objetivo dessas regras é obter relações tais como "região nordeste e irrigação por gotejamento implica que a cultura é uva de mesa", ou seja, que possam auxiliar a identificação das relações de tecnologias com cultura, e, consequentemente, a classificação dessas tecnologias. Ainda no pós-processamento, avaliou-se a qualidade das regras obtidas e da própria metodologia empregada.

Resultados e Discussão

A partir da SABIIA foram reunidos 2.209 documentos e metadados, originados das expressões de busca fornecidas pelos especialistas do domínio. O conjunto de resultados da SABIIA foi utilizado para acessar os textos completos de todos os papers de acesso livre e, para aqueles cujo acesso não é livre, foram utilizados os metadados; criando-se uma base de textos. Esta base foi pré-processada com o uso da ferramenta

I-PreProc e um vocabulário controlado. O vocabulário foi criado a partir da junção de quatro glossários da área de recursos hídricos e dois tesaurus (Thesagro e Agrovoc). Após o pré-processamento, gerou-se a hierarquia de tópicos. A partir da hierarquia realizou-se uma análise exploratória nos tópicos identificados, para extrair localidades, tipos de solos, etc, e documentos repetidos. Com a eliminação dos documentos repetidos, o portfólio ficou constituído por uma planilha com 1.490 linhas e sete atributos (Tecnologia, Tecnologia Associada, Local, UF, Região, Tipo de Solo e Cultura Agrícola).

Para a geração das regras, foram conduzidos vários experimentos em que a Cultura foi fixada como o consequente da regra (Y) e os demais atributos foram combinados no antecedente da regra (X). Desta forma, as regras identificaram quais tecnologias, locais, tipos de solo ou UF estão associados a uma determinada cultura agrícola. As 1.490 linhas do portfólio foram segmentadas por região para facilitar a geração de regras nesta 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.

Para a região Norte, foram encontradas 22 regras, considerando o suporte de 6% e a confiança de 90%, sendo boa parte delas redundante ou não apresentava novidade. Um dos exemplos dessas regras foi (*Se Localidade* = Capitão Poço & UF=PA & *Solo*=latossolo ==> *Cultura*=banana). Outro exemplo foi (*Se Tecnologia_associada*=manejo de água & UF=RO ==> *Cultura*=feijão). Para dados disponíveis das demais regiões do Brasil foram geradas regras similares. Por exemplo, para a região Nordeste, com suporte de 1% e confiança de 80%, foram geradas 21 regras. Exemplos dessas regras incluem (*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). Na análise para a região Sudeste, foram geradas 21 regras com suporte igual a 2% e confiança igual a 80%, como por exemplo, (*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). Para a região Centro-Oeste, foram geradas 48 regras, considerando suporte igual a 4% e confiança igual a 90%. Exemplos dessas regras incluem: (*Se Tecnologia_associada*=manejo de água & *Localidade*=Brasília ==> *Cultura*=tomate) e (*Se Tecnologia_associada*=variabilidade melhoramento genético & UF=DF ==> *Cultura*= batata-doce). Por fim, para a região Sul, foram geradas 31 regras de associação para o suporte igual a 4% e a confiança igual a 80%, com destaque para as regras (*Se Tecnologia_associada*=irrigação por gotejamento & *Localidade*=Santa Maria ==> *Cultura*=tomate) e (*Se Localidade*=Santana do Livramento & UF =RS ==> *Cultura*=uva).

Embora o portfólio seja constituído dos atributos: tecnologias, locais, época, tipo de solo, culturas e tecnologias associadas, nota-se que apenas alguns destes estiveram presentes nas regras. Por exemplo, localidade, tecnologia associada e cultura apresentaram uma frequência mais expressiva nas informações extraídas dos textos. Esses resultados reforçam a necessidade de: (a) avaliar a inclusão de outros atributos de interesse dos especialistas, resultando na geração de regras com mais específicas (com mais atributos) ou (b) selecionar outros repositórios de informações agrícolas para a geração do portfólio de tecnologias.

Conclusões

A construção dessa primeira versão do portfólio ainda envolveu muito trabalho manual, com base na análise exploratória das hierarquias. Essa experiência revela a necessidade de revisão do processo semiautomatizado para a criação de futuros portfólios. Um dos problemas encontrados foi na construção dos tópicos. Ela foi baseada em técnicas que não envolvem processamento de língua natural, logo obtém resultados puramente estatísticos. A maioria das ferramentas para lidar com processamento de língua natural foram concebidas para a língua inglesa, exigindo uma tradução do português para o inglês das publicações científicas, nos experimentos realizados. A tradução dos textos para o inglês pode gerar erros de sintaxe e de semântica, de forma que infor-

mações importantes nos documentos podem ser perdidas; por isso não foi utilizado na metodologia proposta. E, outro problema encontrado foi a grande quantidade de dados esparsos na constituição do portfólio. Muitos documentos não apresentavam tipo de solo, localidade, tecnologia associada, entre outras. Por essa razão, os valores do suporte foram muito baixos em quase todos os casos analisados, variando de 1% a 4%, ao passo que a confiança variou de 80% a 90%, sendo estes resultados, no entanto, compatíveis com aqueles apresentados na literatura. Dessa forma, o processo interativo está sendo revisado para que se obtenham melhores resultados e mais especificidade na construção do portfólio, à medida que novas buscas são realizadas e outras ferramentas para processamento de língua natural são avaliadas. A versão atual do portfólio encontra-se em análise por especialistas do domínio. O objetivo é ter, na versão final do portfólio, uma avaliação das tecnologias encontradas, com apontamentos sobre sua contribuição para o melhor uso de água e solo, delimitadas espacialmente, por bioma e região.

Referências

AGRAWAL, R.; SRIKANT, R. Fast algorithms for mining association rules in large databases. In: BOCCA, J. B.; JARKE, M.; ZANIOLO, C. (Ed.). **Proceeding of the 20th International Conference on Very Large Data Bases**. San Francisco: Morgan Kaufmann Publishers, 1994. p. 478-499.

LIU, B.; HSU, W.; MA,Y. Integrating classification and association rule mining. In: INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 4., 1998, Palo Alto. **Proceedings**... Menlo Park: AAAI Press, 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.