Extração de regras de associação em textos para apoiar a construção de portfólios tecnológicos

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

Resumo: O objetivo deste trabalho foi explorar o uso de técnicas para extração de regras de associação em um conjunto de textos para auxiliar a construção de portfólios tecnológicos. A metodologia adotada foi constituída de um processo de busca de documentos na literatura e um processo de mineração de textos, com a intenção de obter regras que indiquem relações entre tecnologias existentes, geolocalidades, tipos de solos, entre outras. Os resultados dos experimentos revelaram a necessidade de mudanças no processo de geração de regras, pois o formato e a quantidade dos dados utilizados resultou em matrizes altamente esparsas, com muitas classes e poucos atributos, o que não nos permitiu obter inferências com um nível de precisão aceitável. Nos próximos passos serão realizados ajustes nos experimentos para que as regras de associação geradas apresentem um nível de precisão aceitável, em termos de suporte e confiança.

Palavras-chave: expressão de busca, vocabulário controlado, Apriori.

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

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

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

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

Introdução

Neste trabalho pretendeu-se construir um portfólio para selecionar tecnologias para o uso sustentável da água na agricultura, a partir da literatura. Os portfólios são planilhas com a relação de tecnologias, locais, época, culturas associadas e possíveis tecnologias associadas. Após a construção dos portfólios, os especialistas de domínio verificarão se essas tecnologias são soluções para problemas decorrentes da avaliação e da adaptação de tecnologias para os biomas brasileiros. A princípio, buscou-se obter, manualmente, informações nos textos, referentes à data, à localidade, às tecnologias presentes e ausentes, e às informações adicionais relevantes, para que seja possível comparar com o método automático. Como o conjunto de textos é muito grande, é interessante automatizar ou semiautomatizar os processos de construção de portfólios utilizando-se técnicas de mineração de textos.

O processo de Mineração de Textos pode ser dividido em cinco etapas: a) identificação do problema; b) pré-processamento; c) extração de padrões; d) pós-processamento; f) utilização do conhecimento. Na identificação do problema são definidos os objetivos do processo de Mineração de Textos. Uma vez definidos o escopo do problema e o objetivo da aplicação, e selecionados os documentos que representem o domínio do problema, pode-se definir a forma de representação dos dados e extração de conhecimento.

Na etapa de pré-processamento, a coleção de textos é manipulada a fim de se obter uma representação que possa ser lida por ferramentas de extração de padrões. Os textos são transformados em uma matriz atributo valor, na qual as linhas correspondem a cada texto e as colunas, às palavras (ou composições de palavras) presentes na coleção de textos, e, cada célula corresponde a uma medida da importância da palavra no texto.

Neste trabalho, na etapa de extrações de padrões, o foco é a experimentação de técnicas de extração de regras de associação, a fim de verificar relações entre tecnologias e outras palavras (composições de palavras) nos textos, tais como condições de aplicação das tecnologias, por exemplo: em presença de seca aplica-se irrigação, seria uma regra do tipo “seca->irrigação”. 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 et al., 1994).

Na última etapa, do processo de mineração de textos, realiza-se o pós-processamento em que é analisado se os dados obtidos nas regras de associação aproximam-se ao esperado. No caso das regras, várias métricas de qualidade podem ser utilizadas para uma avaliação objetiva, bem como a avaliação subjetiva. Se o processo não alcançar resultados usáveis, repete-se o processo, isto é, desde o pré-processamento até que os resultados avaliados sejam compatíveis de acordo com o objetivo do trabalho.

Assim, o objetivo deste trabalho é avaliar o uso de técnicas de mineração de dados para auxiliar a construção de portfólios tecnológicos.

Materiais e Métodos

Foi realizada uma busca de textos no Sistema Aberto e Integrado de Informação em Agricultura (SABIIA) (VACARI et al., 2011) e assim, com as expressões e combinações das palavras fornecidas pelos especialistas, se obteve 2.210 textos. A expressão de busca foi uma combinação de termos fornecidos pelos especialistas do domínio, mais ou menos na forma: *"(bacia hidrográfica) OR (gestão hídrica) OR (modelagem hidrológica) OR (hidrossedimentometria) OR (sistemas de informação geográfica) OR (sensoriamento remoto) OR (disponibilidade hídrica) OR (balanço enérgetico) OR (qualidade da água) OR (indicadores de sustentabilidade) OR (pegada hídrica ecológica) OR (águas cinzas) …..) AND ((videira) OR (bananeira) OR (goiabeira) OR (mangueira) OR (coqueiro) OR (melão) OR (melancia) OR (tomate) OR (cebola)"*.

O resultado foi convertido para uma base de textos, na qual cada resposta de busca corresponde a um texto. Para isso, escreveu-se um script em python, que usa os metadados da busca e baixa os pdfs, quando disponíveis, convertendo-os para textos.

Aplicou-se um pré-processamento à base textual utilizando a ferramenta I-PreProc (Incremental Pre-Processing), em desenvolvimento no projeto CRITIC@, com um vocabulário controlado. O vocabulário foi determinado por um estudo de resultados anteriores de busca. No vocabulário constam: tipos de solos, geo localidades, todos os termos da expressão de busca e a especificação de tecnologias agrícolas relacionadas ao uso de água, de acordo com o Thesagro (BRASIL, 2015). I-PreProc gera uma representação

de dados com dois arquivos: HDR de descritores e DAT com a matriz atributo valor. Para poder utilizar os resultados como input do algoritmo Apriori (LIU et al., 1998), disponível no ambiente Weka (HALL et al., 2009), foram escritos dois scripts em Python, que leem DAT e HDR e criam: a) um arquivo para o Apriori de Christian Borgelt[5], cujas linhas, que correspondem a cada texto (ou transação), são preenchidas com os vocábulos presentes no texto; e, b) um arquivo do Apriori implementado no Weka, que é uma matriz não esparsa, representada em formato CSV, sendo a primeira linha o nome de todos os vocábulos utilizados na coleção e as demais indicando presença (S) ou ausência (N) do vocábulo no texto, e, a última coluna as variáveis de interesse para o consequente da regra (classes). Sendo que o Apriori de Christian Borgelt apenas calcula as regras de associação. O Apriori do Weka, com matriz não esparsa e classes, gera regras do tipo: *vocábulo x e vocábulo y → tecnologia*.

Uma regra de associação é uma implicação da forma X→Y, onde X e Y são conjuntos frequentes, suporte é a frequência em que X e Y aparecem no conjunto de dados e confiança mede a frequência de itens em Y que aparece nas transações que contém X, equação (2).

Resultados e Discussão

No primeiro experimento, o algoritmo Apriori do Weka foi avaliado para diferentes valores de suporte, mantendo-se a confiança em 90%. A partir da Tabela 1 é possível verificar que para um valor de suporte muito baixo, o número de regras cresce exponencialmente, sendo a maior parte delas regras redundantes, bem como regras que não apresentam nenhuma novidade, como, por exemplo, “aspersão e chuva ==> água”.

Tabela 1. Número de regras geradas pelo Apriori mantendo a confiança de 90% e variação do suporte.

Suporte (%)	Número de regras
1	100.000
3	2.512
5	174
10	13
15	3
20	0

[5] Disponivel em http://www.borgelt.net/apriori.html, consultado em setembro de 2015.

Além disso, a maioria das regras encontradas foram constituídas de poucos vocábulos, pois a matriz de termos submetida ao algoritmo Apriori era muito esparsa (grande número de textos e ausência do vocábulo no texto). Com isso, a frequência de muitos termos era muito baixa (próxima de zero) resultando em regras fracas, ou seja, formadas por termos com um baixo valor de suporte.

Considerando que a análise de regras de associação não apresentou os resultados esperados, outros algoritmos de aprendizado de máquina foram utilizados para tentar identificar a relação entre as palavras-chave e as tecnologias.

O primeiro algoritmo de aprendizado utilizado foi o C.4.5 (QUILAN, 1993) para a indução de regras de classificação por meio de árvores de decisão. A escolha desse algoritmo foi influenciada pela facilidade de uso do algoritmo e pela forma simbólica em que o conhecimento gerado pode ser representado por meio de regras. No entanto, devido à matriz de dados ser tão esparsa, a árvore de decisão não foi gerada completamente, pois a maioria dos atributos possuía valores faltantes.

Em seguida, tentou-se outra abordagem baseada em redes Bayesianas (JOHN; LANGLEY, 1995). Novamente, observou-se que a matriz esparsa como input não favoreceu o uso dessa abordagem. Como consequência, a rede Bayesiana não foi gerada e a relação entre os termos e as tecnologias não foi identificada.

O que pode ser observado é que essa forma de matriz não esparsa, com presença e ausência dos vocábulos e utilizando-se a última coluna com todas as classes de interesse, não permite que se possa inferir algum modelo. Foram utilizados 134 vocábulos, com 114 diferentes classes (tecnologias) em 1.702 textos, sendo que várias classes aparecem nos mesmos textos. Assim, uma opção que se apresenta é fixar uma tecnologia como classe positiva e todas as demais como atributos comuns, ou seja, se a tecnologia em questão não aparecer no texto, a classe será negativa.

Considerações Finais

Os experimentos iniciais mostram que o formato dos dados, como colocado, está levando a matrizes altamente esparsas, com muitas classes e poucos

atributos, o que não nos permite obter inferências com um nível de precisão aceitável. Assim, como trabalho futuro, deve-se procurar diminuir a esparsidade da matriz, tanto considerando as classes como binárias uma a uma, quanto ampliando o vocabulário utilizado.

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**… San Francisco: Morgan Kaufmann, 1994. p. 478-499.

BRASIL. Ministério da Agricultura, Pecuáriia e Abastecimento. **Thesagro**. Disponível em: <http://snida.agricultura.gov.br:81/binagri/html/Cen_Thes1.html>. Acesso em: 15 out. 2015.

HALL, M.; FRANK, E.; HOLMES, G.; PFAHRINGER, B.; REUTEMANN, P.; WITTEN, I. H. The WEKA data mining software: an update. **ACM SIGKDD Explorations Newsletter**, v. 11, n. 1, p. 10-18, June, 2009. DOI: 10.1145/1656274.1656278.

JOHN, G. H.; LANGLEY, P. Estimating continuous distributions in Bayesian classifiers. In: CONFERENCE ON UNCERTAINTY IN ARTIFICIAL INTELLIGENCE,11., San Francisco. **Proceedings**… San Francisco: Morgan Kaufmann, 1995. p. 338-345.

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

QUILAN, R. **C4.5**: programs for machine learning. San Mateo: Morgan Kaufmann, 1993. 302 p. ill.

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.