Inferência do impacto causal de um sistema de recomendação na taxa de rejeição de páginas da Agência Embrapa

Flávio Margarito Martins de Barros[1], *Stanley Robson de Medeiros Oliveira*[2]

[1] Faculdade de Engenharia Agrícola, Universidade Estadual de Campinas, Campinas, São Paulo, Brasil, flavio.barros@feagri.unicamp.br

[2] Embrapa Informática Agropecuária, EMBRAPA, Campinas, São Paulo, Brasil, stanley.oliveira@embrapa.br

RESUMO

A Agência Embrapa é um sistema web com o objetivo de organizar, tratar e divulgar informações técnicas e conhecimentos gerados pela EMBRAPA. Diariamente o site recebe milhares de acesso que são registrados em uma base de dados. A partir destes dados foi criado e implantado um sistema de recomendação. Uma das motivações para a implantação do sistema foi o número elevado de sessões de usuário com somente uma visualização de página, uma indicação que o conteúdo do portal era subutilizado pelos usuários. Como a métrica taxa de rejeição, conhecida na literatura como *bounce rate*, pode ser interpretada como a rejeição do usuário em relação ao total de opções de páginas disponíveis, o sistema foi avaliado por meio dessa métrica utilizando técnicas estatísticas tradicionais, como testes de hipótese, antes e depois da implantação do sistema de recomendação, obtendo resultados parciais em relação à queda da métrica em páginas sobre cana-de-açúcar. Os resultados obtidos nesse trabalho indicam que a nova técnica proposta pode ser utilizada para avaliar o impacto de sistemas de recomendação, utilizando um conjunto menor de dados e com maior confiabilidade.

PALAVRAS-CHAVE: Sistema de recomendação, Regras de associação, Séries temporais.

ABSTRACT

Embrapa Agency is a web system designed to organize, process and disseminate technical information and knowledge generated by EMBRAPA. Every day the site receives thousands of access that are registered in a database. From this data was created and implemented a recommender system. One of the motivations for the implementation of this system was the high number of user´s sessions with only one page view, indicating that the portal content was underutilized by users. Considering that the metric bounce rate can be interpreted as the user rejection in relation to the total of pages available, the system was evaluated through this

metric using traditional statistical techniques such as hypothesis tests, before and after the implementation of the recommendation system obtaining partial results regarding the values of the metric dropped in pages on sugarcane. The results of this study indicate that the technique can be used to assess the impact of recommender systems using a smaller data set with greater reliability.
KEYWORDS: Recommender systems, Association rules, Time series.

INTRODUÇÃO

No Brasil, devido a uma demanda por informações técnicas agrícolas de qualidade, a Empresa Brasileira de Pesquisa Agropecuária (EMBRAPA) investiu em um projeto denominado Agência de Informação Embrapa, com o objetivo de organizar, tratar, armazenar e divulgar informações técnicas e conhecimentos gerados ao longo de 40 anos de pesquisa. Por meio do endereço eletrônico (http://www.agencia.cnptia.embrapa.br) o usuário tem acesso a todo o conteúdo do site na forma de textos, artigos, livros, arquivos de imagem, arquivos de som e planilhas eletrônicas. Em particular, a Agência apresenta as principais informações da cadeia produtiva, como aspectos socioeconômicos e ambientais, planejamento, manejo, colheita, processamento e gestão industrial. Todo o conteúdo foi organizado para atender pesquisadores, produtores rurais, profissionais de assistência técnica e extensionistas.

Nesse portal, são disponibilizadas informações em quantidade elevada, tal que com a sobrecarga de informações, muitos usuários podem ter dificuldades em encontrar a informação desejada (YANG e TANG, 2003). Neste cenário, a recomendação de conteúdo é uma alternativa viável para auxiliar usuários (KUMAR e THAMBIDURAI, 2010).

Uma forma de produzir as recomendações é inferir o comportamento dos usuários baseado nos padrões de uso de informações. A mineração de dados, etapa principal do processo de descoberta de descoberta de conhecimento em bases de dados, é uma das melhores alternativas dentre as técnicas utilizadas para identificar o comportamento de uso de sites e oferecer recomendações aos seus usuários (HAN et al., 2011).

A partir dos dados de uso da Agência Embrapa, por meio de regras de associação (AGRAWAL *et al*., 1993), foi criado e implantado um sistema de recomendação para páginas relativas à cana-de-açúcar (BARROS *et al.*, 2013). A partir das recomendações geradas, o impacto do sistema de recomendação foi avaliado, por meio de testes de hipótese, mostrando que o sistema melhorou a taxa de rejeição para algumas páginas.

Entretanto, ao analisar dados observacionais, a exemplo dos dados de uso do portal Agência Embrapa, é preciso levar em consideração a influência de fatores externos ao efeito que se espera observar (DEVORE, 2006). Por exemplo, no caso da Agência Embrapa, os efeitos do sistema de recomendação podem ser influenciados negativamente ou positivamente por fatores não relacionados ao próprio sistema, tal que a validação do sistema pode ser prejudicada.

De acordo com Brodersen *et. al.* (2014), em áreas como economia e marketing, que também enfrentam o desafio de avaliar o impacto de intervenções por meio de dados observacionais, vem sendo desenvolvidos novos métodos, como modelos bayesianos estruturais de séries temporais, capazes de inferir o impacto causal em dados observacionais. Assim, o objetivo desse trabalho foi avaliar modelos bayesianos estruturais para séries temporais (construídas a partir das taxas de rejeição diárias de páginas da Agência Embrapa) capazes de inferir um impacto causal de uma intervenção de um sistema de recomendação. A partir desses modelos, espera-se prever a resposta das séries temporais na ausência da aplicação de uma intervenção, tal que com os dados reais da série após a intervenção, seja possível comparar a diferença entre as séries observada e predita, permitindo avaliar o impacto causal da intervenção.

MATERIAL E MÉTODOS

Com o objetivo de fornecer recomendações sobre informações tecnológicas agrícolas, foram extraídos dados de uso da base de dados do Portal Agência Embrapa. A técnica de modelagem escolhida foi a geração de regras de associação, por meio do algoritmo Apriori (AGRAWAL *et al*., 1993), captando assim o perfil de acessos da comunidade.

Regras de associação foram introduzidas por Agrawal *et al.* (1993) e descrevem a relação entre itens ou produtos que ocorrem com uma certa frequência em uma base de dados. Uma regra de associação entre itens X e Y tem a forma $X \rightarrow Y$, tal que $X \cap Y = \varnothing$. Para cada regra de associação estão associadas duas medidas tradicionais: confiança (Conf) e suporte (Sup). Sup representa o número de amostras que contêm X e Y. Do ponto de vista conceitual, representa a significância estatística desses itens nas amostras, ao passo que Conf constitui a razão entre o número de amostras que contêm X e Y sobre o número de amostras que contêm X. Do ponto de vista conceitual, a confiança determina a força da regra. Uma regra é considerada interessante quando ela apresenta um suporte e uma confiança, iguais ou superiores, ao mínimo estabelecido pelo usuário. Nas equações (1) e

(2) são apresentadas as expressões para o Suporte e a Confiança de uma regra de associação.

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

onde X → Y representa uma regra de associação entre X e Y e P(X U Y) representa a probabilidade de encontrar amostras no conjunto de dados que contenham X e Y.

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

onde P(X | Y) é a probabilidade condicional de X dado a ocorrência de Y. O Suporte é como definido em (2), sendo que Suporte(X) = P(X).

O banco de dados da Agência Embrapa, que registra a atividade de todos os usuários, está estruturado em duas tabelas: a tabela *clientes* e a tabela *tracker,* tal que cada linha da tabela *clientes* representa um único usuário e cada linha da tabela *tracker* representa um acesso a uma das páginas da Agência. Sempre que uma requisição é feita, são registrados os seguintes atributos do usuário: idsessao (identificador único com 32 caracteres), ip, tempo de permanência, latitude, longitude, cidade, pais e estado. Cada linha na tabela clientes representa um usuário.

Na tabela *tracker* são armazenados dados relativos a cada visualização de página associada a um usuário. Um mesmo usuário pode aparecer em mais de uma linha na tabela. São registrados os seguintes atributos: idtracker (identificador único de cada sessão), idsessão (o mesmo da tabela clientes), página visitada, árvore (neste caso a árvore é a da cana-de-açúcar, mas outras árvores do conhecimento também podem ser acessadas), data do servidor, hora do servidor e tempo da sessão.

A tabela *clientes*, no banco de dados, possuía 2.574.763 linhas, que representavam o número de usuários distintos que acessaram conteúdos no período compreendido entre outubro de 2010 a janeiro de 2013. A tabela *tracker* possuía 5.223.003 linhas, onde cada linha contém a informação de cada requisição individual de uma página do sistema, também relativo ao período de outubro de 2010 a janeiro de 2013. É importante notar que cada linha da tabela *tracker* representa uma requisição de página. Logo um mesmo usuário pode aparecer em várias linhas, pois este pode ter requisitado mais de uma página em sua sessão.

Para a etapa de modelagem, com o conjunto final de dados já tratados, foram

determinadas as regras de associação mais relevantes entre as páginas de conteúdo da Agência cana-de-açúcar, de forma a oferecer recomendações de conteúdo, baseadas no perfil da comunidade de usuários. Cada regra de associação relaciona somente duas páginas, o antecedente e o consequente. Assim, uma regra de associação entre duas páginas A → B significa que, uma vez que um usuário acessa a página A, existe alta probabilidade deste usuário acessar a página B. Neste trabalho, foi utilizado um suporte baixo o suficiente (sup = 0.0005), tal que regras representando páginas com poucos acessos fossem encontradas. Essas regras foram ordenadas pela confiança e armazenadas no banco de dados da Agência de Informação Embrapa.

A arquitetura do sistema foi dividida em duas partes: servidor e navegador, conforme a Figura 1. O servidor compreende o Apache, um software com a função de servir páginas web, o módulo PHP, responsável pelo processamento das páginas, e banco de dados e o módulo de recomendação (*Recommender*). A inferência das regras de associação envolve a interação entre o *Recommender* e os bancos "BDTracker" e "IAgência". No primeiro, o *Recommender* consulta as informações sobre os históricos de navegação dos usuários e no segundo os títulos das páginas que serão apresentados nas recomendações. As regras são recalculadas semanalmente. Na Tabela 1 são apresentadas 28 regras e as respectivas páginas onde o impacto do sistema de recomendação foi avaliado.

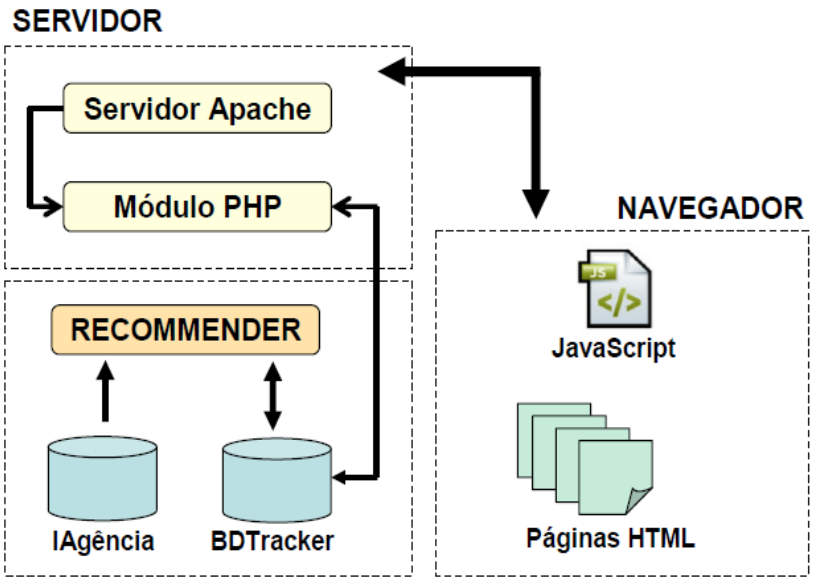

Figura 1 – Arquitetura do sistema de recomendação.

Tabela 1: Base de conhecimento com 28 regras de associação entre páginas sobre cana-de-açúcar.

Antecedente	Consequente	Sup	Conf
Fabricação do açúcar	A diferenciação de produtos na cadeia produtiva do açúcar: o processo de produção dos açúcares líquido e líquido invertido	0.00007146	0,83
Extração	Moendas	0.00014799	0,83
Processamento da cana-de-açúcar	Açúcar e álcool: o combustível do Brasil [vídeo]	0.00010036	0,80
Variedades	3ª geração de variedades CTC	0.00007185	0,79
Custos e rentabilidade	[Planilha geral de custos e rentabilidade: sem os coeficientes técnicos]	0.00013433	0,77
Cachaça	Fábrica de aguardente de cana-de-açúcar	0.00006677	0,75
Cachaça	O perfil da cachaça	0.00005467	0,72
Processamento da cana-de-açúcar	Açúcar e álcool: a tecnologia sucroalcooleira [vídeo]	0.00007380	0,72
Queima	Exigências	0.00012027	0,70
Variedades	Variedades RB de cana-de-açúcar	0.00006951	0,68
Processamento da cana-de-açúcar	Um modelo de otimização para o planejamento agregado da produção em usinas de açúcar e álcool	0.00010075	0,64
Processamento da cana-de-açúcar	Açúcar e álcool: a produção do álcool [vídeo]	0.00012183	0,63
Plantio	Mudas	0.00118747	0,63
Açúcar	Mercado	0.00018158	0,62
Correção e adubação	Adubação e calagem em cana-de-açúcar	0.00005818	0,62
Doenças	Outras doenças	0.00042134	0,60
Qualidade de matéria-prima	Produção de etanol de cana-de-açúcar: qualidade da matéria-prima	0.00007458	0,59
Abertura	Cana-de-açúcar	0.00006326	0,59
Cachaça	A arte de produzir cachaça: visita a um produtor rural artesanal [vídeo]	0.00005037	0,57
Diagnose das necessidades nutricionais	Expectativa da produtividade	0.00006287	0,56
Plantio	Recomendações técnicas para o cultivo da cana-de-açúcar forrageira em Rondônia	0.00008122	0,55
Análise de solo	Interpretação da análise	0.00031825	0,54
Preparo do solo	Plantio direto	0.00034910	0,53
Implicações	Exigências	0.00008981	0,53
Abertura	Pré-produção	0.00146511	0,52
Doenças fúngicas	Outras doenças	0.00036081	0,51
Meio ambienteasmor	Impactos	0.00017338	0,51

A metodologia utilizada na pesquisa também incluiu a preparação de dados de 1.450.484 sessões de usuários, relativas ao período de novembro de 2010 a setembro de 2013, onde a intervenção do sistema de recomendação ocorreu a partir do dia 01 de novembro de 2012. Foram calculadas séries temporais de taxas de rejeição, como em Sculley *et. al.,* (2009), para 20 páginas da Agência Embrapa cana-de-açúcar.

A partir dos dados das séries temporais, foram ajustados modelos bayesianos estruturais, definidos nas equações (3) e (4), para cada série relativa a cada página.

$$y_t = Z_t^T \alpha_t + \varepsilon_t \tag{3}$$

$$\alpha_{t+1} = T_t \alpha_t + R_t \eta_t \tag{4}$$

onde $\varepsilon_t \sim N\left(0, \sigma_t^2\right)$ e $\eta_t \sim N\left(0, Q_t\right)$ são independentes das outras variáveis. A equação (3) é a equação das observações, que relaciona os dados observados y_t a um vetor de estados d-dimensional α_t , e a equação (4) é a equação de estado que governa a evolução do vetor de estados α_t ao longo do tempo. Modelos bayesianos estruturais, como definidos em (3) e (4), são flexíveis e modulares, abarcando uma série de outros modelos, como por exemplo, os modelos ARIMA, que podem ser reescritos dessa forma (BRODERSEN *et. al.*, 2014).

Na Figura 2 é mostrada uma aplicação do modelo em uma série temporal fictícia e a interpretação do resultado. No primeiro painel, baseado nos dados da série antes da intervenção (em preto) o modelo foi ajustado e a série predita após a intervenção (em azul, a partir do 70 no eixo das abscissas), sob hipótese de que o regime não mudou. No segundo painel é mostrada a diferença entre a série predita e a observada, isto é, o efeito causal para cada ponto no tempo. Por fim, no terceiro painel são mostrados os efeitos causais acumulados no tempo. De acordo com a Figura 2, claramente há um efeito na série devido à intervenção.

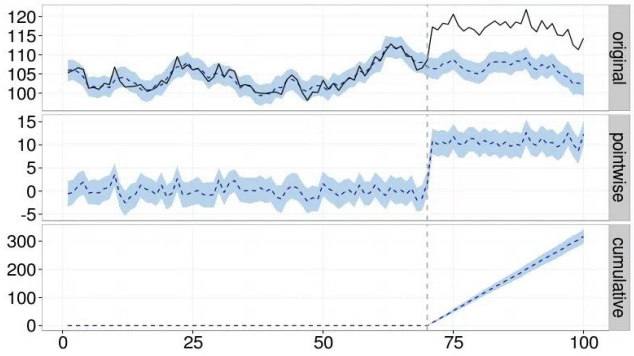

Figura 2 – Exemplo de aplicação da técnica em uma série temporal.

Os modelos bayesianos foram ajustados para as séries entre o período de novembro de 2010 a outubro de 2012, tal que a projeção deste modelo foi comparada aos dados reais de taxa de rejeição para as mesmas páginas durante 60 dias após a intervenção. Pela comparação da diferença entre a projeção e a série real, foi avaliada a probabilidade de obtenção dos valores efetivamente observados sob a hipótese do efeito nulo de intervenção.

RESULTADOS E DISCUSSÃO

Na Figura 3, observa-se o resultado da modelagem para a série temporal das taxas de rejeição diárias para uma das páginas da Agência. No painel 1 da Figura 3 é apresentada a série original e as previsões após a intervenção para a série de taxas diárias de rejeição da página de Cachaça. No painel 2, onde são mostradas as diferenças entre a série original e a série predita, nos 60 dias após intervenção do sistema de recomendação observa-se uma ligeira queda na taxa de rejeição, mas somente observando a série não é possível verificar se a diferença é estatisticamente significativa. Assim, na Tabela 2, são apresentadas as probabilidades de ocorrência da série efetivamente observada caso estas séries evoluíssem de acordo com o modelo ajustado antes da intervenção, para as 11 das 20 páginas analisadas nesse trabalho.

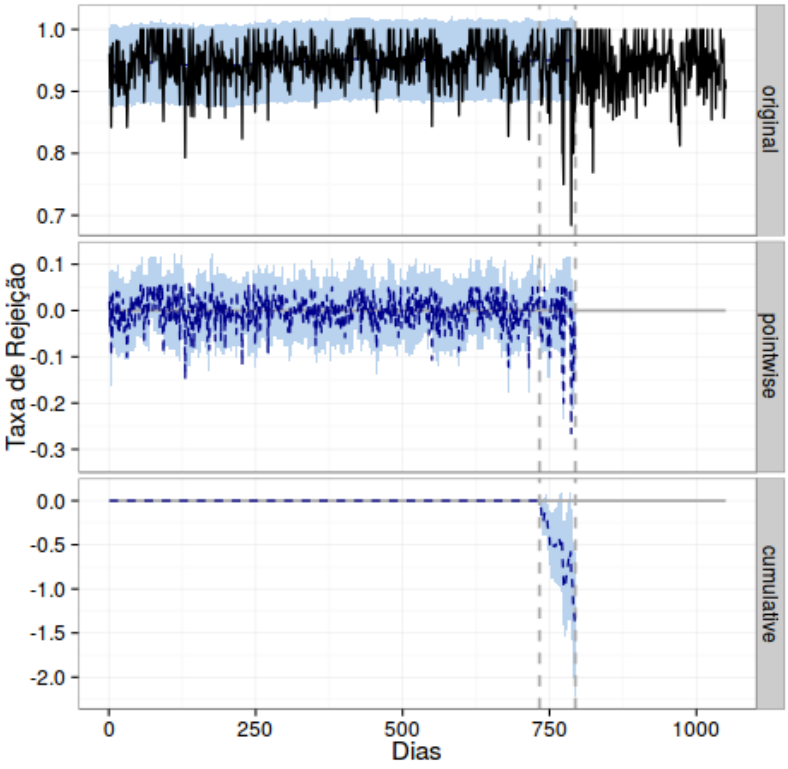

Figura 3 – Modelo para a série da página da cachaça.

Da Tabela 2, observa-se que os resultados obtidos indicam que o sistema de recomendação teve impacto na diminuição da taxa de rejeição dessas páginas. Em 11 das 20 páginas analisadas observou-se queda significativa na métrica, isto é, quando a probabilidade de observação da série, supondo um regime inalterado, foi inferior a 0.1, mostrando um impacto significativo. Salienta-se que a taxa de rejeição de uma página pode estar associada a outros fatores, além da falta de interesse do usuário, como no caso de conteúdos específicos que são um fim em si mesmos. Nesses casos o usuário pode se sentir satisfeito com a informação obtida e abandonar o portal com uma única visualização. Algumas páginas apresentam conteúdos que podem demandar informações adicionais, tal que nesses casos a presença do sistema de recomendação pode ter um impacto positivo.

Os resultados obtidos a partir da Tabela 2 estão de acordo com os resultados discutidos em Barros *et. al.*, (2013), onde foram utilizados testes de hipótese para proporções. Em ambos os casos, tanto por meio de testes de hipótese para proporções, quando por meio das séries temporais, foi possível avaliar o impacto da presença das recomendações. Ainda assim, a abordagem apresentada neste trabalho tem diversas vantagens, como por exemplo a quantidade reduzida de dados necessários e também o fato de o modelo fornecer uma medida mais confiável do impacto causal.

Tabela 2: Título das páginas e as probabilidades de ocorrência da série observada sob hipótese de que o sistema de recomendação não teve efeito.

Página	Probabilidade
Cachaça	0,002
Fabricação do Açúcar	0,036
Qualidade da matéria-prima	0,006
Custos e rentabilidade	0,001
Variedades	0,059
Agricultura de precisão	0,030
Monitoramento ambiental	0,055
Análise de solo	0,071
Queima	0,082
Corte	0,052
Carregamento	0,022

CONCLUSÕES

O método proposto foi bem sucedido em avaliar o impacto da intervenção de um sistema de recomendação, de acordo com resultados da literatura. Também, a metodologia proposta neste trabalho é inovadora pois foi uma das primeiras aplicações de sistemas de recomendação na agricultura e uma das primeiras aplicações da avaliação do impacto causal de sistemas de recomendação em dados de uso online, com resultados efetivos.

REFERÊNCIAS

AGRAWAL R., IMIELINSKI T., SWAMI A. N. Mining Association Rules between Sets of Items in Large Databases. **SIGMOD**, Washington, v.22, n.2, p.207-216, 1993.

BARROS F. M. M.; OLIVEIRA, S. R. M.; OLIVEIRA, L. H. M. Desenvolvimento e validação de um sistema de recomendação de informações tecnológicas agrícolas sobre cana-de-açúcar. **Bragantia**, Campinas, v.72, n.4, p.387-395, 2013.

BRODERSEN, K. H.; GALLUSSER, F.; KOEHLER, N. R.; SCOTT, S. L. Inferring causal impact using Bayesian structural time-series models. **Annals of Applied Statistics,** 2014.

DEVORE, J. L. Probabilidade e Estatística para Engenharia e Ciências, 6 ed. São Paulo: Thompson, p. 692, 2006.

HAN, J.; KAMBER, M.; PEI, J. **Data Mining:** Concepts and Techniques, 3ed. San Francisco: Morgan Kaufmann Publishers, 2011. p.703.

KUMAR, A.; THAMBIDURAI, P. Collaborative Web Recommendation Systems - A Survey Approach. **Global Journal Of Computer Science And Technology**, Chennai, v. 9, n. 5, p.30-35, Jan. 2010.

SCULLEY, D.; MALKIN, R.; BASU, S.; BAYARDO, R. Predicting bounce rates in sponsored search advertisements. **Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining – KDD09,** 2009.

YANG, H.; TANG, J. A three-stage model of requirements elicitation for web-based information systems. **Industrial Management And Data Systems**, Taipei, v. 103, n. 5-6, p.398-409, 2003.