Um sistema de recomendação para informações tecnológicas agrícolas

Flavio M. M. de Barros[1], Stanley R. M. Oliveira[1,2], Leandro H. M. Oliveira[2]

[1] Universidade Estadual de Campinas, Brasil
flavio.barros@feagri.unicamp.br
[2] Embrapa Informática Agropecuária
{stanley.oliveira, leandro.oliveira}@embrapa.br

Abstract. The aim of this work was to design, develop and deploy a web recommender system based on association rules to automatically make recommendations of agricultural content, according to the profile of a user community. The data used in this study were extracted from a database of accesses to the site of Embrapa Information Agency. The user visits were stored in a structure of access lists, and from such lists, association rules between pages were generated. The set of rules led to a knowledge base that was used to make content recommendations to users. The system was evaluated using a metric called bounce rate, so that by means of statistical tests it was possible to evaluate the impact of these recommendations. The results revealed that through recommendations, users find relevant information associated with their visits and increase their time spent on the site. The system was validated for the sugarcane crop, but it can be easily extended to provide recommendations regarding other crops.

Categories and Subject Descriptors: H.3.5 [**Online Information Services**]: Web-based services; H.4.0 [**Information Systems Applications**]: General; H.2.8 [**Database Applications**]: Data mining

Keywords: Agricultural Information Systems, Association Rule Mining, Recommender Systems

1. INTRODUÇÃO

Aplicativos modernos, desenvolvidos para o ambiente Web, possuem mecanismos para aprender as preferências de seus usuários. Esses aplicativos são conhecidos como sistemas de recomendação e seus usuários estão cada vez mais acostumados com suas recomendações. Esses sistemas são bem diversos, com diferentes objetivos, desde indicações de produtos até sugestões de leitura, filmes e lugares turísticos. No entanto, no domínio agrícola, nota-se uma carência de tais sistemas já que não se conhece registros na literatura.

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 de cadeias produtivas, como aspectos socioeconômicos e ambientais, planejamento, manejo, colheita, processamento e gestão industrial. O conteúdo foi organizado para atender pesquisadores, produtores rurais, extensionistas e, diariamente, o site recebe milhares de acessos que são registrados em uma base de dados.

O portal da Agência Embrapa disponibiliza milhares de páginas de conteúdo individual, mesmo

para uma única cultura como a cana-de-açúcar. Essa oferta de informações em grande quantidade pode confundir e dificultar o acesso pelos usuários [Yang and Tang 2003]. Assim, a recomendação de conteúdo é uma alternativa viável para auxiliar usuários devido ao volume elevado de informações [Kumar, A; Thambidurai 2010]. A recomendação personalizada de conteúdo aumenta a usabilidade dos sistemas, agrega valor aos serviços, poupa tempo e fideliza usuários.

Uma forma de produzir recomendações é inferir o comportamento dos usuários baseado nos padrões de uso das informações. Em particular, no domínio agrícola, onde há informações em quantidade elevada, armazenadas digitalmente em bancos de dados, as ferramentas de mineração de dados apresentam recursos para análise que podem fornecer padrões de uso do site para fazer recomendações. Esta 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].

Assim, considerando a importância da agricultura para o Brasil e a existência de grandes repositórios de informações agrícolas disponíveis na Web, o objetivo deste trabalho foi projetar e desenvolver um sistema de recomendação web, baseado em regras de associação, que ofereça recomendações sobre a cultura da cana-de-açúcar de acordo com os padrões de acessos de uma comunidade de usuários. No Brasil, essa é uma das primeiras iniciativas de aplicações de sistemas de recomendação no domínio agrícola. Esse sistema pode ser facilmente estendido para outros portais, com outras culturas agrícolas.

2. TRABALHOS RELACIONADOS

Existem experiências bem-sucedidas com sistemas de recomendação de livros, CDs, e outros produtos na Amazon.com [Linden et al. 2003], filmes pela MovieLens [Miller et al. 2003] e notícias pela VERSIFI Technologies [Billsus et al. 2002]. Na implementação desses sistemas, as técnicas mais utilizadas são a filtragem colaborativa e a filtragem baseada em conteúdo [Ricci et al. 2011]. Existem também técnicas híbridas que combinam as duas abordagens [Burke 2000].

Existem também casos de aplicação de regras de associação em sistemas de recomendação para descobrir relações entre páginas visitadas. Por exemplo, em [Kazienko 2009] foram determinadas regras de associação entre páginas para criar listas de recomendações. Em [Yang and Parthasarathy 2003] e [Jorge et al. 2002] foram implantados sistemas de recomendação baseados exclusivamente em listas de recomendação, obtidas por meio de mineração de regras. Mais recentemente regras de associação foram utilizadas no contexto de classificação em um sistema de recomendação aplicado ao turismo [Lucas et al. 2013] e em [Paranjape-Voditel and Deshpande 2013] foram utilizadas regras de associação para a criação de um sistema de recomendação de portfólios no mercado de ações.

Esse trabalho se diferencia dos demais em três aspectos fundamentais: a) escopo da aplicação: no domínio agrícola, aplicações de sistemas de recomendação são raramente encontradas; b) extensibilidade: o sistema proposto pode ser facilmente estendido para diversas culturas agrícolas; c) método de avaliação das recomendações: a avaliação do sistema proposto foi realizada do ponto de vista da usabilidade do site, por meio da taxa de rejeição. Conforme [Sculley et al. 2009], a taxa de rejeição pode ser utilizada para avaliar se os usuários encontram a informação desejada em um site.

3. ARQUITETURA DO SISTEMA DE RECOMENDAÇÃO

A arquitetura do sistema é basicamente dividida em duas partes: servidor e navegador, conforme Figura 1. O servidor compreende o servidor apache, o módulo PHP, o banco de dados e o módulo Recommender, implementado em R. Quando um usuário acessa uma das páginas da Agência Embrapa, o servidor apache responde enviando dois arquivos: uma página em HTML e o script responsável pelas consultas às recomendações e, depois, atualiza a página visitada. Assim que o navegador carrega a página, o script é inicializado e envia uma requisição de consulta ao servidor com o endereço da

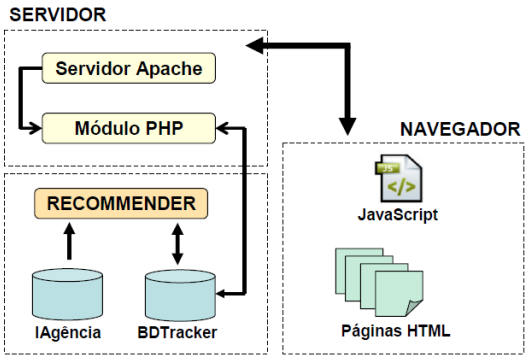

Fig. 1. Arquitetura do sistema de recomendação para informações tecnológicos agrícolas.

página visitada. Em seguida, é acionado o módulo PHP que realiza a consulta ao banco de dados "BDTtacker". Caso haja recomendações, o servidor envia as informações ao script que atualiza a página com os links (sugestões de páginas a ser visitadas). Caso contrário, a página fica inalterada.

A inferência das regras 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 uma vez por semana, assim os novos acessos dos usuários, incluindo os cliques em recomendações, atualizam o banco de dados contribuindo para o aparecimento de novas regras. Um dos resultados mais importantes desse trabalho é a base de conhecimento gerada. Essa base, representada pelas regras de associação entre páginas, além de ser utilizada para recomendações também pode ser utilizada para adaptar o site de acordo com o perfil de uso dos usuários [Perkowitz and Etzioni 2000].

4. RESULTADOS

O conteúdo do portal da Agência Embrapa está estruturado em árvores de conhecimento. Como os acessos à árvore do conhecimento da cana-de-açúcar representam quase metade do total de acessos à Agência Embrapa, o sistema proposto foi validado para essa cultura agrícola. As visitas dos usuários à árvore da cana-de-açúcar foram estruturadas em duas tabelas: clientes e tracker, sendo que na tabela clientes são armazenadas informações de cada usuário que entrou na Agência e iniciou uma sessão e na tabela tracker são armazenados dados relativos a cada visualização de página associada a um usuário.

Sempre que uma requisição é feita, são registrados os seguintes atributos do usuário na tabela clientes: 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 registrados os seguintes atributos: idtracker (identificador único de cada sessão), idsessão (o mesmo da tabela clientes), página visitada, árvore (neste caso a árvore do conhecimento é a da cana-de-açúcar), data do servidor, hora do servidor e tempo da sessão.

A tabela clientes, no banco de dados, possui 2.574.763 linhas, que representam o número de usuários distintos que acessaram conteúdos sobre a cana-de-açúcar no período compreendido entre outubro de 2010 a janeiro de 2013. A tabela tracker possui 5.223.003 linhas, onde cada linha contém a informação de cada requisição individual de uma página do sistema, também relativo ao mesmo período. 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.

A base de conhecimento gerada para a árvore da cana-de-açúcar é composta por vinte e oito regras de associação entre as páginas (Tabela I), relacionando páginas de conteúdo textual e páginas com recursos eletrônicos, como arquivos em pdf e vídeos. Mesmo com o suporte e a confiança baixos

($sup = 0.0005$ e $conf = 0,51$) e com o volume elevado de sessões de usuário (mais de 2 milhões), não emergiram padrões de acesso que relacionassem páginas que já não estavam ligadas entre si. Conforme Tabela II, somente seis páginas possuem mais de uma recomendação. Além das páginas apresentarem muitos links (não são recomendações, mas hiperlinks para outras páginas), estas apresentam de 1 a 4 recomendações. Esse resultado é particularmente importante, pois mostra que a base de conhecimento tem um potencial de resumo para direcionar o usuário aos links mais importantes.

Table I. Base de conhecimento com 28 regras de associação entre páginas da cultura da cana-de-açúcar.

Regra	Antecedente	Consequente	Sup	Conf
1	Fabricação do açúcar	Produção dos açúcares líquido	0,007%	83%
2	Extração	Moendas	0,015%	83%
3	Processamento da cana	Açúcar e álcool: o combustível do Brasil [vídeo]	0,010%	80%
4	Variedades	3ª geração de variedades CTC	0,007%	79%
5	Custos e rentabilidade	Planilha geral de custos e rentabilidade	0,013%	77%
6	Cachaça	Fábrica de aguardente de cana-de-açúcar	0,007%	75%
7	Cachaça	O perfil da cachaça	0,005%	72%
8	Processamento da cana	Açúcar e álcool: a tecnologia sucroalcooleira [vídeo]	0,007%	72%
9	Queima	Exigências	0,012%	70%
10	Variedades	Variedades RB de cana-de-açúcar	0,007%	68%
11	Processamento da cana	Modelo de otimização para o planejamento em usinas	0,010%	64%
12	Processamento da cana	Açúcar e álcool: a produção do álcool [vídeo]	0,012%	63%
13	Plantio	Mudas	0,119%	63%
14	Açúcar	Mercado	0,018%	62%
15	Correção e adubação	Adubação e calagem em cana-de-açúcar	0,006%	62%
16	Doenças	Outras doenças	0,042%	60%
17	Qualidade de matéria-prima	Produção de etanol de cana-de-açúcar	0,007%	59%
18	Abertura	Cana-de-açúcar	0,006%	59%
19	Cachaça	A arte de produzir cachaça [vídeo]	0,005%	57%
20	Diagnose nutricional	Expectativa da produtividade	0,006%	56%
21	Plantio	Recomendações técnicas em Rondônia	0,008%	55%
22	Análise de solo	Interpretação da análise	0,032%	54%
23	Preparo do solo	Plantio direto	0,035%	53%
24	Implicações	Exigências	0,009%	53%
25	Abertura	Pré-produção	0,147%	52%
26	Doenças fúngicas	Outras doenças	0,036%	51%
27	Meio ambiente	Diagnóstico agroambiental	0,010%	51%
28	Meio ambiente	Impactos	0,017%	51%

De acordo com [Sculley et al. 2009], quando a taxa de rejeição de um site (conjunto de páginas) ou de uma única página é alta, isto pode ter dois significados: os usuários encontram exatamente o que estavam procurando ou eles não encontraram a informação desejada e não acham o site atrativo para explorá-lo. Assim, para verificar o efeito do sistema de recomendação foram calculadas as taxas de rejeição para as páginas da cana-de-açúcar que receberam recomendações. Todas as sessões foram consideradas no período de 25 de novembro de 2012 até 16 de janeiro de 2013.

Para a verificação da significância estatística das variações das taxas de rejeição, antes e depois da implantação do sistema, foram utilizados dois testes estatísticos: teste Z e o teste qui-quadrado para diferenças entre proporções. Os dois testes são paramétricos, sendo que o segundo é o teste estatístico mais comumente utilizado para testes de homogeneidade [Devore 2006]. A hipótese para utilização dos testes paramétricos foi que os dados tinham uma distribuição Binomial, pois foi considerado nesse trabalho que cada usuário que entrou na Agência Embrapa no período da pesquisa, ao acessar uma das páginas, tomou uma decisão de forma independente dos demais. É importante salientar que no caso de a hipótese alternativa ser a simples diferença entre as taxas de rejeição, isto é $H_1 : p_1 \neq p_2$, os dois testes são equivalentes. No entanto, o teste Z foi utilizado pois é possível testar a hipótese $H_1 : p_1 > p_2$. Assim, nas linhas destacadas da Tabela III, tem-se as páginas que apresentaram variação na taxa de rejeição e que, em pelo menos um dos testes estatísticos, essa variação foi significativa.

Table II. Regras com suas respectivas páginas antecedentes, número total de recomendações e o total de links na página.

Regras	Antecedente	Recomendações	Total de Links
1	Fabricação do açúcar	1	44
2	Extração	1	41
3,8,11,12	Processamento da cana-de-açúcar	4	49
4,1	Variedades	2	45
5	Custos e rentabilidade	1	39
6,7,19	Cachaça	3	49
9	Queima	1	46
13,21	Plantio	2	44
14	Açúcar	1	37
15	Correção e adubação	1	52
16	Doenças	1	49
17	Qualidade de matéria-prima	1	41
18,25	Abertura	2	25
20	Diagnose das necessidades nutricionais	1	49
22	Análise de solo	1	48
23	Preparo do solo	1	43
24	Implicações	1	44
26	Doenças fúngicas	1	49
27,28	Meio ambiente	2	42

Observando-se os dados da Tabela III, vê-se que em algumas páginas o sistema de recomendação não teve impacto na taxa de rejeição. Para as páginas de Extração, Processamento da cana-de-açúcar, Cachaça, Queima, Plantio, Doenças, Análise do solo, Preparo do solo e Meio ambiente, não houve variação estatisticamente significativa para ambos os testes. Para o caso de algumas páginas, como a de "Diagnose das necessidades nutricionais", "Queima" e "Implicações", houve um número pequeno de sessões onde estas eram a primeira página visitada.

Table III. Estatísticas sobre a resposta ao sistema de recomendação para as recomendações.

Página	Taxa Rejeição (antes)	Taxa Rejeição (depois)	Qui-Quadrado (p-valor)	Z (p-valor)
Fabricação do açúcar	**0.6780**	**0.7757**	**0.0000**	**0.0000**
Extração	0.8984	0.9153	0.2922	0.1461
Processamento da cana-de-açúcar	0.7747	0.7743	0.9775	0.4887
Variedades	**0.9392**	**0.9003**	**0.0002**	**0.0001**
Custos e rentabilidade	**0.7463**	**0.4832**	**0.0000**	**0.0000**
Cachaça	0.9437	0.9414	0.7278	0.3639
Queima	0.8949	0.8727	0.6026	0.3013
Plantio	0.8421	0.8492	0.4904	0.2452
Açúcar	**0.8455**	**0.9078**	**0.0013**	**0.0007**
Correção e adubação	**0.9580**	**0.9250**	**0.0010**	**0.0005**
Doenças	0.5497	0.5776	0.3602	0.1801
Qualidade de matéria-prima	**0.9624**	**0.9152**	**0.0000**	**0.0000**
Abertura	**0.2849**	**0.2373**	**0.0000**	**0.0000**
Diagnose das necessidades nutricionais	0.6268	0.5000	0.3005	0.1503
Análise de solo	0.9257	0.8849	0.0160	0.0080
Preparo do solo	0.7031	0.7537	0.0113	0.0056
Implicações	**0.9552**	**0.8571**	**0.0911**	**0.0456**
Doenças fúngicas	**0.9718**	**0.9514**	**0.0674**	**0.0337**
Meio ambiente	0.9638	0.9810	0.3484	0.1742

Por outro lado, aproximadamente metade das páginas apresentaram variações estatisticamente significativas (p-valor $< 0{,}05$, em pelo menos um dos testes). Destas, a maioria teve a taxa de rejeição diminuída com exceção das páginas "Fabricação do açúcar" e "Açúcar". Especialmente, no caso das páginas de "Fabricação do açúcar" e "Açúcar", a diferença positiva pode ser interpretada à luz do volume de acessos: como estas são duas das páginas mais acessadas e com altas taxas de rejeição, isso

pode indicar que os usuários já encontraram as informações desejadas nas próprias páginas. Outra hipótese é que a quantidade de dados usada antes da disponibilização do sistema foi muito maior do que a quantidade de dados utilizada após a implantação do sistema, para essas páginas. Assim, essa variação pode ser resultado dessa desproporção, que pode diminuir com a coleta de mais dados.

5. CONCLUSÃO

Neste trabalho, foi apresentado um sistema de recomendação para informações tecnológicas agrícolas. Sua validação se deu por meio de um estudo de caso relacionado à cultura da cana-de-açúcar. Os resultados demonstraram que a partir da implantação desse sistema, houve um impacto positivo na usabilidade do portal da Agência Embrapa. O sistema implantado permite aos usuários acessar mais informações sem a necessidade de realizar novas buscas, o que é muito interessante para usuários não especialistas e menos experientes (ex.: produtores rurais e extensionistas).

Vinte páginas relacionadas à cultura da cana-de-açúcar receberam recomendações, e destas, mais da metade tiveram a taxa de rejeição diminuída com significância estatística. A base de conhecimento, composta de 28 regras, atua como uma forma de resumo, indicando quais são os links mais importantes nas páginas do portal sobre a cana-de-açúcar.

REFERENCES

Billsus, D., Brunk, C., and Evans, C. Adaptive interfaces for ubiquitous web access. *Communications of the ACM - The Adaptive Web* 45 (5): 34–38, 2002.

Burke, R. Knowledge-based recommender systems. *Encyclopaedia of Library and Information Systems* 69 (32), 2000.

Devore, J. L. *Probabilidade e Estatística para Engenharia e Ciências*. Thomson Learning, São Paulo, SP, Brasil, 2006.

Han, J., Kamber, M., and Pei, J. *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2011.

Jorge, A., Alves, M. A., and Azevedo, P. Recommendation with association rules: A web mining application. In *in Proceedings of Data Mining and Warehousing, Conference of Information Society 2002, Eds. D. Mladenic and M. Grobelnik, Josef Stefan Institute*, 2002.

Kazienko, P. Mining indirect association rules for web recommendation. *Int. J. Appl. Math. Comput. Sci.* 19 (1): 165–186, Mar., 2009.

Kumar, A; Thambidurai, P. Collaborative Web Recommendation Systems -A Survey Approach. *Global Journal of Computer Science and Technology* 9 (5), 2010.

Linden, G., Smith, B., and York, J. Amazon.com recommendations: Item-to-item collaborative filtering. *IEEE Internet Computing* 7 (1): 76–80, Jan., 2003.

Lucas, J. P., Luz, N., Moreno, M. N., Anacleto, R., Figueiredo, A. A., and Martins, C. A hybrid recommendation approach for a tourism system. *Expert Systems with Applications* 40 (9): 3532 – 3550, 2013.

Miller, B. N., Albert, I., Lam, S. K., Konstan, J. A., and Riedl, J. MovieLens unplugged. In *Proceedings of the 8th international conference on Intelligent user interfaces - IUI '03*. ACM Press, New York, New York, USA, pp. 263, 2003.

Paranjape-Voditel, P. and Deshpande, U. A stock market portfolio recommender system based on association rule mining. *Applied Soft Computing* 13 (2): 1055 – 1063, 2013.

Perkowitz, M. and Etzioni, O. Towards adaptive web sites: Conceptual framework and case study. *Artificial Intelligence* 118 (1–2): 245 – 275, 2000.

Ricci, F., Rokach, L., Shapira, B., and Kantor, P. B., editors. *Recommender Systems Handbook*. Springer, 2011.

Sculley, D., Malkin, R. G., Basu, S., and Bayardo, R. J. Predicting bounce rates in sponsored search advertisements. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '09*. ACM Press, New York, New York, USA, pp. 1325, 2009.

Yang, H. and Parthasarathy, S. On the use of constrained associations for web log mining. In *WEBKDD 2002 - Mining Web Data for Discovering Usage Patterns and Profiles*, O. Zaïane, J. Srivastava, M. Spiliopoulou, and B. Masand (Eds.). Lecture Notes in Computer Science, vol. 2703. Springer Berlin Heidelberg, pp. 100–118, 2003.

Yang, H.-L. and Tang, J.-H. A three-stage model of requirements elicitation for Web-based information systems. *Industrial Management & Data Systems* 103 (6): 398–409, 2003.