http://dx.doi.org/10.1590/brag.2013.049

# Desenvolvimento e validação de um sistema de recomendação de informações tecnológicas sobre cana-de-açúcar

Flávio Margarito Martins de Barros ([1]*); Stanley Robson de Medeiros Oliveira ([1,2]); Leandro Henrique Mendonça de Oliveira ([2])

*([1]) Universidade Estadual de Campinas (Unicamp), Faculdade de Engenharia Agrícola (Feagri), Caixa Postal 6011, 13083-875 Campinas (SP), Brasil.*
*([2]) Embrapa Informática Agropecuária, Caixa Postal 6041, 13083-886 Campinas (SP), Brasil.*
*(*) Autor correspondente: flavio.barros@feagri.unicamp.br*

*Recebido: 22/abr./2013; Aceito: 30/set./2013*

## Resumo

Sistemas de informações tecnológicas exercem um papel fundamental na agricultura, pois proveem informações de forma rápida e eficiente, dando suporte a decisões estratégicas. No entanto, o excesso de informação disponível pode confundir e dificultar o acesso à informação desejada. Uma alternativa para amenizar esse problema é a adoção de sistemas que ofereçam recomendações automáticas de acordo com o perfil de uma comunidade de usuários. O objetivo deste trabalho foi desenvolver um sistema de recomendação, concebido a partir de técnicas de mineração de dados, para conteúdos relacionados à cultura da cana-de-açúcar. O modelo adotado foi o de listas de recomendação, que são regras de associação entre as páginas *web* visitadas, produzidas a partir dos dados de navegação dos usuários. O sistema foi implantado no portal da Agência de Informação Embrapa, que é um sistema *web,* com o objetivo de organizar, tratar, armazenar e divulgar informações técnicas agrícolas. Dentre os resultados alcançados com a implantação do sistema, destaca-se uma base de conhecimento na forma de regras de associação que descreve o comportamento de uma comunidade de usuários e pode ser utilizada para indicar os *link*s mais importantes em páginas *web* relacionadas à cultura da cana-de-açúcar. Com a adoção desse sistema de recomendação, os principais benefícios para os usuários da agência são: a) acesso a informações detalhadas de etapas de produção, com indicações de material de referência e fontes diferentes de estatísticas confiáveis; b) acesso a informações que favorecem a produção eficiente em termos técnicos, ambientais, sociais e econômicos; c) apoio a agentes públicos e privados no planejamento e tomada de decisão; d) transferência de conhecimento atualizado, em linguagem amigável, para o setor sucroenergético.

Palavras-chave: taxa de rejeição, mineração de dados, regras de associação, informações tecnológicas agrícolas.

## Development and validation of a recommender system for technologial information on sugarcane

## Abstract

Information technology systems play an important role in agriculture, since they provide information in a quick and efficient way to support strategic decisions. However, the excess of information available may confuse and hinder the access to specific information. An alternative is the adoption of systems that provide automatic recommendations according to the profile of a user community. This work aimed to develop a recommender system based on data mining techniques to recommend content regarding the sugarcane crop. The adopted recommendation model relies on recommendation lists that are association rules between web pages, produced from the data users' browsing. The system was deployed on the portal of Embrapa Information Agency, which is a web system that aims to organize process, store and disseminate agricultural technological information. Among the results achieved with the system, it can be highlighted a knowledge base built with association rules, that describes the behavior of a user community and can indicate the most important links in web pages concerning sugarcane. With the adoption of this recommender system, the main benefits to users are: a) detailed information of production stages, with indications of reference material and different sources of reliable statistics; b) information that favors the efficient production considering technical, environmental, social and economic aspects; c) support to public and private stakeholders in planning and decision making; d) transfer of current knowledge with a friendly language for the sugarcane industry.

Key words: bounce rate, data mining, association rules, agricultural technological information.

O Brasil é atualmente o maior produtor mundial de cana-de-açúcar e o maior exportador de produtos dela derivados, sendo a região sudeste a maior produtora. Em particular, o estado de São Paulo é o maior produtor nacional e apresenta mais de 9,6 milhões de hectares de área plantada, de acordo com a União da Indústria de Cana-de-açúcar (UNICA, 2013).

Na conjuntura da economia brasileira, a cultura da cana-de-açúcar passou a destacar-se a partir do início do ano 2000 como uma opção economicamente viável para a produção de bioenergia em larga escala. De acordo com Gauder et al. (2011), o Brasil ocupa hoje um papel de liderança na produção e distribuição de etanol para o setor automotivo. Devido à relevância dessa cultura para a agricultura do país, que em 2012 já representava aproximadamente 12% do PIB (UNICA, 2012), é muito importante que o país invista em pesquisas e novas tecnologias de manejo, produção, irrigação e escoamento da produção, de forma a manter sua liderança estratégica na produção de cana-de-açúcar. Além disso, é imperativo que essas informações cheguem aos produtores, técnicos e pesquisadores de forma eficiente e eficaz.

No ritmo acelerado das mudanças que vêm impactando os segmentos agrícolas, a necessidade da gestão de informações que auxiliem no processo de tomada de decisões no agronegócio torna-se evidente. Ciente dessa demanda, a EMBRAPA (Empresa Brasileira de Pesquisa Agropecuária) desenvolveu um portal de informações técnicas, a Agência de Informação Embrapa.

A agência é um sistema *web* cujo objetivo é organizar, tratar, armazenar e divulgar informações técnicas e conhecimentos gerados pela EMBRAPA e outras instituições parceiras. Em particular, a Agência de Informação da Cana-de-açúcar (http://www.agencia.cnptia.embrapa.br/gestor/cana-de-acucar/Abertura.html) apresenta as principais informações da sua 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, informações técnicas agrícolas são disponibilizadas em quantidades tão elevadas que, com a sobrecarga de informações, os usuários podem ter dificuldades para encontrar a informação desejada. Assim, a recomendação de conteúdo é uma alternativa viável para auxiliar usuários devido ao volume elevado de informações disponibilizado (Kumar e Thambidurai, 2010).

Uma forma de produzir recomendações é inferir o comportamento dos usuários baseando-se nos padrões de uso de informações. A mineração de dados, etapa principal do processo de descoberta de conhecimento em bases de dados, é uma alternativa 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 cultura da cana-de-açúcar para o Brasil e a existência de grandes repositórios de informações agrícolas disponíveis na *web*, o objetivo deste trabalho foi projetar, desenvolver e avaliar o uso de um sistema de recomendação *web* que ofereça recomendações automáticas sobre a cultura da cana-de-açúcar de acordo com o perfil de uma comunidade de usuários.

Com o objetivo de fornecer recomendações sobre informações tecnológicas referentes à cana-de-açúcar, os dados de uso da base do Portal Agência Embrapa foram extraídos. 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 dessa comunidade.

As regras de associação descrevem a relação entre itens ou produtos que ocorrem com certa frequência em uma base de dados. Uma regra de associação 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) (Agrawal et al., 1993). O Suporte 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 a Confiança 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.

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

em que $X \rightarrow Y$ representa uma regra de associação entre $X$ e $Y$ e $P(X \cup Y)$ representa a probabilidade de encontrar amostras, no conjunto de dados, que contenham $X$ e $Y$.

$$Conf\left(X \rightarrow Y\right) = P\left(X|Y\right) = \frac{Suporte\left(X \rightarrow Y\right)}{Suporte\left(X\right)} \quad (2)$$

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

Para definir os procedimentos metodológicos deste trabalho optou-se por seguir o modelo de processo Cross Industry Standard Process for Data Mining (CRISP-DM). A escolha do CRISP-DM se deu porque essa metodologia é amplamente adotada em projetos de mineração de dados, tanto na academia quanto na indústria, e por ela prover uma ferramenta de suporte rápida, robusta e barata para definir os procedimentos adotados na fase de mineração de dados (Chapman et al., 2000).

De acordo com o modelo CRISP-DM, o processo consiste em seis fases: compreensão do domínio, entendimento dos dados, preparação dos dados, modelagem, avaliação e distribuição.

A primeira etapa refere-se ao entendimento do domínio. Nesse trabalho, o domínio de aplicação é o de sistemas de informações agrícolas, especialmente sistemas de informações tecnológicas sobre a cana-de-açúcar. A próxima etapa refere-se

ao entendimento dos dados, que foram obtidos a partir dos registros dos acessos dos usuários à agência de cana-de-açúcar e, posteriormente, armazenados em um banco de dados.

O banco está estruturado em duas tabelas: a tabela **clientes** e a tabela **tracker.** As informações relativas a cada usuário que visitou o *site* da agência e iniciou uma sessão estão armazenadas na tabela **clientes**, isto é, o usuário requisitou ao servidor o acesso a uma das páginas da agência. Sempre que uma requisição é feita, os seguintes atributos são registrados: **idsessao** (identificador único com 32 caracteres), **ip**, **tempo de permanência**, **latitude**, **longitude**, **cidade**, **país** 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** (nesse 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** (Tabela 1), no banco de dados, possui 2.574.763 linhas, que representam o número de usuários distintos que acessaram conteúdos da cultura da cana-de-açúcar no período compreendido entre outubro de 2010 a janeiro de 2013. A tabela **tracker** (Tabela 2) possui 5.223.003 linhas, cada linha contém a informação de cada requisição individual de uma página do sistema, informação também relativa 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 ele pode ter requisitado e visitado mais de uma página em sua sessão.

Para um melhor entendimento do conjunto de dados, uma análise exploratória dos dados também foi feita, de forma a obter uma visão geral que fornecesse características gerais de uso do *site*: páginas mais visitadas, média de páginas visitadas por sessão de usuário, outras agências mais vistas, tecnologias utilizadas, distribuição espacial dos acessos e estatísticas que fornecessem informações sobre perfis de uso.

Após as fases de compreensão do domínio e entendimento dos dados, tem-se a fase de preparação. O objetivo principal dessa fase foi a seleção, seguida da preparação desses dados para um formato adequado à aplicação do algoritmo Apriori. Assim, foi feita a transformação para uma estrutura de transações conforme a Tabela 3.

Por exemplo, na Tabela 3 o usuário 001 acessou a página sobre "praga no colmo" e, em seguida, acessou a página sobre "praga nas raízes". Portanto, cada linha da Tabela 3 representa a sessão de um usuário que acessou uma ou mais páginas com conteúdo relacionado à cana-de-açúcar.

Para definir a importância das regras geradas, foi utilizada a métrica MaxConf, uma vez que cada página pode ter mais de uma regra de associação relacionada. O valor da métrica MaxConf (Wu et al., 2010) é definida como segue:

$$MaxConf(A,B) = máximo\left\{\frac{Suporte(A \cup B)}{Suporte(A)}, \frac{Suporte(A \cup B)}{Suporte(B)}\right\} \quad (3)$$

em que o $Suporte(A \cup B)$ foi definido da Equação 1.

Tabela 1. Descrição e exemplos dos atributos contidos na tabela **clientes**; os exemplos apresentados são de registros reais presentes no banco de dados

| Atributo | Descrição | Exemplo |
|---|---|---|
| Idsessão | Identificador com 32 caracteres | 304af458e75af3c51fa020052d5c6825 |
| ip_cliente | IP do cliente | 192.168.0.1 |
| data_servidor | Data do acesso | Tuesday-31-May-2011 |
| time_stamp | Período do acesso em segundos | 1306855203 |
| cidade_cliente | Cidade de origem do acesso | São Paulo |
| latitude_cliente | Latitude da cidade do acesso | –23.5333 |
| longitude_cliente | Longitude da cidade do acesso | –46.6167 |
| estado_cliente | Estado de origem do acesso | SP |
| país_cliente | País de origem do acesso | Brasil |

Tabela 2. Descrição e exemplos dos atributos contidos na tabela **tracker;** tanto idsessão quanto time_stamp também estão presentes na tabela **clientes**

| Atributo | Descrição | Exemplo |
|---|---|---|
| Idtracker | Identificador da sessão com 32 caracteres | 625bf2356fcb803f8e288de486e9210b |
| Idsesssão | O mesmo da tabela clientes | ---------- |
| local_atual | Página requisitada | Abertura.html |
| árvore_atual | Árvore a qual pertence a página | catalogo20/Abertura.html |
| data_servidor | Data do acesso à página | Thursday-28-October-2010 |
| hora_servidor | Horário do acesso | 11:27:40 |

**Tabela 3.** Exemplo da estrutura de dados presente na tabela **tracker**

| ID usuário | Lista de páginas visitadas |
|---|---|
| 001 | {Praga no colmo, Praga nas raízes} |
| 004 | {Praga no colmo, Praga nas raízes, Produção} |
| 006 | {Produção} |
| 007 | {Produção} |

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 $A \rightarrow B$ significa que, uma vez que um usuário acessa a página *A*, existe alta probabilidade de esse usuário acessar a página *B*.

O algoritmo Apriori inicialmente encontra os grupos de itens ocorrendo frequentemente juntos em amostras que, nesse contexto, são grupos de páginas visitadas em uma sessão de usuário. Esses conjuntos de itens são denominados conjuntos frequentes. Conjuntos frequentes são gerados com base no suporte fornecido como entrada no algoritmo. Assim é importante definir um valor apropriado de suporte tal que o algoritmo possa encontrar padrões de páginas pouco visualizadas e ao mesmo tempo relevantes.

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.

Por fim foi feita a avaliação do modelo. A base de regras foi avaliada para verificar se ela trouxe novas ligações entre as páginas. Também foi avaliado o potencial de resumo e indicação dos *links* principais de uma página por meio da base de regras.

Além das avaliações dos conjuntos de regras também foi realizada uma avaliação com dados *on-line*, feita utilizando-se uma métrica conhecida na literatura como *bounce rate* (Pakkala et al., 2012), termo em inglês que pode ser traduzido por taxa de rejeição. A taxa de rejeição de uma página é dada pela fração de pessoas que entraram no *site* por essa página e abandonaram o *site* logo em seguida, sem visualizar mais nenhuma outra página. Também de acordo com Pakkala et al. (2012), a métrica é um importante indicativo da satisfação dos usuários com um portal, devendo ser monitorada para avaliações.

Para a verificação da significância estatística das variações das taxas de rejeição antes e depois do sistema de recomendação, 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). Um teste paramétrico é um teste estatístico que assume que os dados têm certa distribuição de probabilidade, que nesse estudo é Normal. 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 cana-de-açúcar, no período da pesquisa, ao acessar uma das páginas, tomou uma decisão de forma independente dos demais.

De acordo com Devore (2006), para o teste Z, sob a hipótese de uma distribuição Binomial, utilizando-se a aproximação da distribuição Normal, a estatística de teste pode ser definida como:

$$z = \frac{\hat{p}_1 - \hat{p}_z}{\sqrt{pq\left(\frac{1}{m} + \frac{1}{n}\right)}} \sim Normal(0,1) \quad (4)$$

Para $z > z_\alpha$ em que $\alpha$ é o nível de significância do teste, a hipótese nula é rejeitada.

Também de acordo com Devore (2006), para o teste qui-quadrado, a estatística do teste é dada por:

$$Qp = \frac{\sum_{i=1}^{k}(n_i - np_i)^2}{np_i} \sim \chi^2_{k-1} \quad (5)$$

em que $\chi^2_{k-1}$ é uma variável aleatória com distribuição qui-quadrado com $k - 1$ graus de liberdade e $np_i$ são os valores esperados das proporções.

Especificamente no caso de a hipótese alternativa ser $H_1: p_1 \neq p_2$, os testes são equivalentes. No entanto, nesse trabalho utilizou-se o teste Z para verificar a hipótese $H_1: p_1 \geq p_2$ e o teste qui-quadrado para testar $H_1: p_1 \neq p_2$, pois o teste qui-quadrado só pode ser usado para avaliar diferenças.

Adicionalmente, no banco de dados de acessos da agência cana-de-açúcar é registrada a localização geográfica da origem de cada acesso, dessa forma é possível determinar o país, estado e o município da maioria desses acessos. O estado de São Paulo é a origem do maior volume de acessos, com quase um terço do total (Figura 1).

Por meio da Figura 2, utilizando-se as coordenadas geográficas de latitude e longitude da origem de cada acesso, foi feita uma distribuição espacial das visualizações de páginas de cana-de-açúcar no Brasil. Apesar do portal Agência Embrapa receber acessos do Brasil e de outros países, as regiões sudeste, sul e nordeste (Zona da Mata) são responsáveis pelo grande volume de acessos no país. Essas regiões coincidem com grandes áreas produtoras de cana-de-açúcar.

Como as páginas sobre cana-de-açúcar representam mais de um quarto dos acessos da Agência Embrapa, procurou-se determinar quais são as páginas mais acessadas do portal. A Tabela 4 apresentada as seis páginas mais acessadas, sendo que as outras têm, cada uma, menos de 2,9% dos acessos.

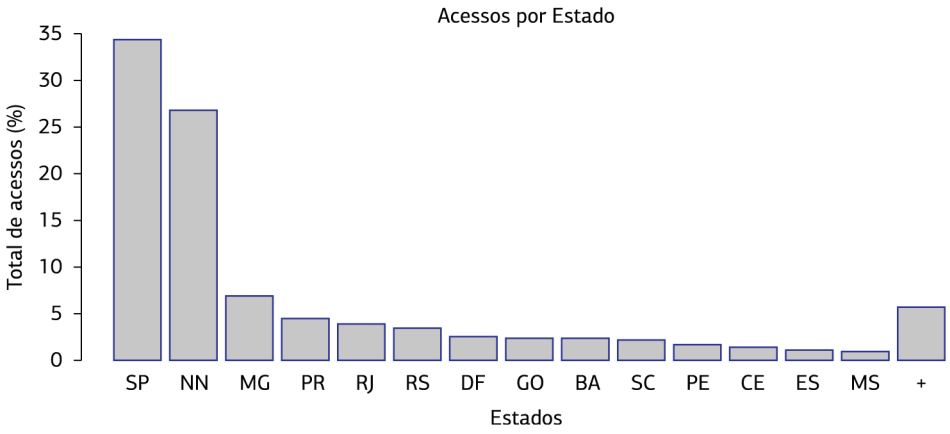


**Figura 1.** Distribuição do total de acessos à árvore de conhecimento de cana-de-açúcar por estado. A sigla NN indica a porcentagem de sessões de localização não identificadaS e o + indica o acumulado dos outros estados.

**Figura 2.** Distribuição espacial dos acessos às páginas da árvore de cana-de-açúcar.

**Tabela 4.** Contagens e porcentagens das páginas mais vistas sobre cana-de-açúcar

| Páginas | Ocorrências | % |
|---|---|---|
| Processamento da cana-de-açúcar | 117.383 | 8,17 |
| Plantio | 58.902 | 4,10 |
| Cachaça | 58.303 | 4,06 |
| Fabricação do açúcar | 52.837 | 3,68 |
| Fermentação | 48.408 | 3,37 |
| Geração de energia elétrica | 41.210 | 2,87 |

**Tabela 5.** Contagens e porcentagens das páginas mais vistas sobre cana-de-açúcar

| Páginas vistas por sessão | Número de sessões | % |
|---|---|---|
| 1 | 2.122.441 | 83,27 |
| 2 | 237.456 | 9,32 |
| 3 | 72.240 | 2,83 |
| 4 | 35.139 | 1,38 |
| 5 | 21.313 | 0,84 |
| 6 | 14.823 | 0,58 |

Pôde-se perceber que a maioria das páginas mais visitadas está relacionada a produtos feitos a partir da cana-de-açúcar. Somente uma página é relacionada ao plantio e a página relativa ao processamento da cana-de-açúcar recebeu o dobro de acessos em relação à segunda página mais visitada.

Na Tabela 5 pode-se verificar que em mais de 83% das sessões de usuário os visitantes entram em uma página e abandonam o portal. Esse é um indicativo que a maioria dos usuários pode não estar encontrando a informação desejada. De acordo com Kent et al. (2011), a métrica *bounce rate* (taxa de rejeição), que mede o número de sessões de usuário que visualizaram uma página e abandonaram o *site*, provê uma forma de avaliação da satisfação do usuário com páginas ou anúncios clicados.

Com o propósito de conceber uma arquitetura expansível e de fácil alteração, pensou-se em uma estrutura de *software* que fosse capaz de interagir com a estrutura de *software* da Agência Embrapa, criar e atualizar recomendações automaticamente e oferecer as recomendações aos usuários de forma dinâmica. A Figura 3 ilustra a arquitetura geral do sistema de recomendação da agência cana-de-açúcar.

Na Figura 3, à esquerda, tem-se as tarefas que são executadas no servidor e, à direita, as tarefas que são executadas no navegador (cliente). Sempre que um usuário requisita uma das páginas no seu navegador, esse envia uma requisição de página para o servidor Apache. O servidor, por sua vez, ao devolver a página ao navegador, envia também *scripts* desenvolvidos em Java Script, que são responsáveis por enviar ao servidor as informações do usuário, como IP, página acessada, horário do acesso etc. Essas informações são gravadas no servidor por um *script* em PHP e são fonte de dados para o sistema de recomendação.

O *engine* de recomendação foi escrito em R. *Engine,* palavra que pode ser traduzida por motor, é um termo muito utilizado em TI para designar a parte mais importante de um sistema de *software*. Ele é responsável pela aquisição dos dados

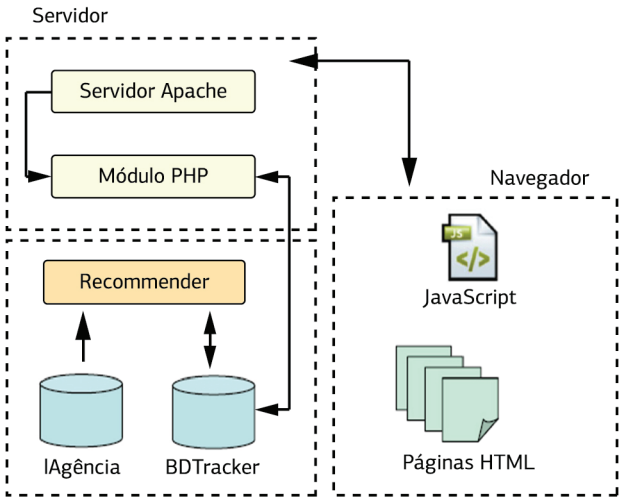


**Figura 3.** Arquitetura do sistema de recomendação da Agência de Informação Embrapa.

no banco de dados, pelo tratamento e transformação dos dados, pela geração das regras e pela gravação das regras no banco. Informações contidas no banco de dados da agência, que contém, dentre outras informações, os identificadores das páginas e também os títulos das páginas, são consultadas pelo *engine* de recomendação. Por fim, as regras, a métrica MaxConf e os títulos compõem as recomendações que são gravadas para consultas posteriores.

A geração das recomendações é executada no servidor uma vez por semana. Assim, os novos acessos dos usuários, inclusive os cliques em recomendações, atualizam o banco de dados de clientes, que também contribuem para o aparecimento de novas regras de recomendação automaticamente.

Essas regras, que são geradas e atualizadas semanalmente, podem mudar de acordo com novos padrões de acesso que eventualmente surjam. Entretanto, um dos resultados mais importantes desse trabalho é a base de conhecimento gerada. Essa base, representada pelas regras de associação entre as 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 e Etzioni, 2000). Na Tabela 6 é apresentada a base de conhecimento de regras que está em operação no sistema de recomendação.

A base de conhecimento, gerada para a agência de cana-de-açúcar, contém 28 regras de associação entre as páginas. Essas regras relacionam páginas de conteúdo textual e páginas com recursos eletrônicos, como arquivos em pdf e vídeos.

Além das regras geradas especificamente para a agência cana-de-açúcar, o sistema de recomendação foi construído de tal forma que durante o processo de recálculo das regras de recomendação também são geradas regras de associação para páginas de todas as outras agências de informação disponíveis (agronegócio do leite, milho, feijão, manga, gado de corte, entre outros). No entanto, essas regras de associação relativas às outras árvores de conhecimento, apesar de estarem na base, ainda não são oferecidas aos usuários em forma de recomendação.

A base de conhecimento completa, incluindo todas as árvores de conhecimento, compreende um total de 684 regras de associação, com Suporte mínimo de 0,0005 e onfiança mínima de 0,51.

Na análise dos *links* de todas as páginas antecedentes, nas 28 regras, um dos resultados mais importantes desse trabalho é que todas as recomendações eram para páginas que já estavam ligadas por um *link*. Esse resultado mostra que, mesmo com o Suporte e a Confiança baixos 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.

Outro ponto a se destacar na Tabela 6 é o fato de a maioria dos padrões emergentes de regras de associação estar ligada a páginas sobre subprodutos da cana-de-açúcar. Na Agência de Informação Embrapa não existe nenhum mecanismo que permita identificar o usuário, no entanto, pelos padrões observados, há evidências de que uma grande parte da clientela do portal componha-se de pesquisadores, gestores e outros profissionais ligados à indústria da cana-de-açúcar.

Como se pode observar na Tabela 7, somente seis páginas possuem mais de uma recomendação. Além de as páginas apresentarem muitos *links* (não são recomendações, mas *hiperlinks* para outras páginas), elas apresentam somente de uma a quatro recomendações. Esse resultado mostra que a base de conhecimento tem um potencial de resumo e de direcionar o usuário aos *links* mais importantes.

Nos testes com o restante das páginas da agência, o resultado é distinto, pois das 684 regras, 263 ligam páginas para as quais não havia *links*. Isso equivale a mais de 38% das regras. Em relação a toda a Agência Embrapa, não considerando somente cana-de-açúcar, a base de conhecimento gerada pode contribuir diretamente para a reestruturação do portal.

Por fim, para verificar o efeito do sistema de recomendação, as taxas de rejeição para toda a Agência Embrapa foram calculadas para as páginas da cana-de-açúcar e para aquelas que receberam regras de recomendação. Todas as sessões foram consideradas no período de 25 de novembro de 2012 até 16 de janeiro de 2013, pois foi durante esses dias que as páginas consideradas continham recomendações. Os dados são apresentados na Tabela 8.

Nas linhas destacadas têm-se as páginas que apresentaram variação na taxa de rejeição e em que, em pelo menos um dos testes estatísticos, essa variação foi significativa. As diferenças entre as proporções de rejeição foram testadas utilizando-se um teste não paramétrico (teste qui-quadrado) e um teste paramétrico (teste Z).

Observando-se os dados da Tabela 8 vê-se que em algumas páginas o sistema de recomendação não teve impacto na

**Tabela 6.** Base de conhecimento com 28 regras de associação 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 | 3.ª 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 ambiente | Diagnóstico agroambiental | 0,00009567 | 0,51 |
| Meio ambiente | Impactos | 0,00017338 | 0,51 |

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 nos testes. Para as páginas Diagnose das necessidades nutricionais, Queima e Implicações, houve um número pequeno de sessões em que essas foram a primeira página visitada.

Por outro lado, aproximadamente metade das páginas apresentou variações estatisticamente significativas ($p<0,05$, em pelo menos um dos testes). Dessas, 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 sob o aspecto do volume de acessos. Como essas 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 explicação é que a quantidade de visitas antes da disponibilização do sistema de recomendação foi muito maior do que a quantidade de dados utilizados 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.

Algumas páginas destacaram-se na diminuição da taxa de rejeição após a disponibilização do sistema de recomendação, como, por exemplo, Abertura, Custos e rentabilidade e Implicações. A página Abertura já apresentava um valor baixo de taxa de rejeição, o que era esperado para a página principal, mas o destaque foi a diminuição significativa das taxas das páginas de Custos e de Implicações.

A página Implicações, no seu final, não apresenta opções de *links* para o usuário, assim as recomendações adicionadas podem ter atuado como um suporte ao usuário, pois antes do sistema de recomendação mais de 95% dos usuários abandonavam o *site* após acessar essa página. Em particular, no caso da página Custos e rentabilidade houve uma queda

**Tabela 7.** Regras com as respectivas páginas antecedentes e o número total de recomendações junto ao 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, 10 | 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 |

**Tabela 8.** Respostas ao sistema de recomendação

| Página | Taxa rejeição (recomendação) | Taxa rejeição (sem recomendação) | Qui-quadrado (p-valor) | Z (p-valor) |
|---|---|---|---|---|
| Fabricação do açúcar | 0,7757 | 0,6780 | 0,0000 | 0,0000 |
| Extração | 0,9153 | 0,8984 | 0,2922 | 0,1461 |
| Processamento da cana-de-açúcar | 0,7743 | 0,7747 | 0,9775 | 0,4887 |
| Variedades | 0,9003 | 0,9392 | 0,0002 | 0,0001 |
| Custos e rentabilidade | 0,4832 | 0,7463 | 0,0000 | 0,0000 |
| Cachaça | 0,9414 | 0,9437 | 0,7278 | 0,3639 |
| Queima | 0,8727 | 0,8949 | 0,6026 | 0,3013 |
| Plantio | 0,8492 | 0,8421 | 0,4904 | 0,2452 |
| Açúcar | 0,9078 | 0,8455 | 0,0013 | 0,0007 |
| Correção e adubação | 0,9250 | 0,9580 | 0,0010 | 0,0005 |
| Doenças | 0,5776 | 0,5497 | 0,3602 | 0,1801 |
| Qualidade de matéria-prima | 0,9152 | 0,9624 | 0,0000 | 0,0000 |
| Abertura | 0,2373 | 0,2849 | 0,0000 | 0,0000 |
| Diagnose das necessidades nutricionais | 0,5000 | 0,6268 | 0,3005 | 0,1503 |
| Análise de solo | 0,8849 | 0,9257 | 0,0160 | 0,0080 |
| Preparo do solo | 0,7537 | 0,7031 | 0,0113 | 0,0056 |
| Implicações | 0,8571 | 0,9552 | 0,0911 | 0,0456 |
| Doenças fúngicas | 0,9514 | 0,9718 | 0,0674 | 0,0337 |
| Meio ambiente | 0,9810 | 0,9638 | 0,3484 | 0,1742 |

muito acentuada. Essa queda pode estar também associada à visibilidade das recomendações, pois de todas as 20 páginas presentes, essa é a página em que as recomendações estão mais visíveis.

No geral, as taxas de rejeição apresentaram queda em quase metade das páginas. Para as páginas em que não houve queda da taxa de rejeição, três possibilidades podem ter ocorrido: a) o resultado pode dever-se à baixa exposição dos *links* de recomendação; b) o conteúdo de algumas delas já solucionara as necessidades de muitos usuários; c) falta de observações em algumas páginas que são menos visitadas.

Baseado na diminuição da taxa de rejeição de quase metade das páginas da agência cana-de-açúcar, verificou-se uma associação entre a implantação do sistema de recomendação e a melhora na utilização desse conjunto de páginas da agência cana-de-açúcar. Experiências semelhantes com sistemas de recomendação baseados em regras de associação são encontradas em Changchien et al. (2004) e Kazienko (2009). Em Kazienko (2009), o sistema de recomendação foi avaliado em relação à estrutura de *hyperlinks* de um portal, mostrando que listas de recomendação podem ser usadas também como forma de determinar os *links* mais

importantes de uma página. Já em Changchien et al. (2004), um sistema de recomendação com um dos módulos baseado em regras de associação foi avaliado em termos de retorno financeiro comparado às promoções tradicionais, obtendo ganhos superiores a 10% em venda de produtos.

Conclui-se ser possível desenvolver e implantar um sistema de recomendação capaz de melhorar a usabilidade de um sistema de informações agrícolas, especialmente sobre cana-de-açúcar. Os resultados revelaram que a presença das recomendações melhora a usabilidade do portal e permite que os usuários acessem mais informações, sem a necessidade de novas buscas. Como os padrões foram extraídos a partir de sessões de usuário que visitaram várias páginas, os padrões de associação refletem os hábitos dessa comunidade de usuários, de forma que com essas recomendações os usuários menos experientes podem ter acesso a esses padrões de forma automática.

Do total de páginas na árvore de cana-de-açúcar, 20 delas receberam recomendações, e dessas, mais da metade tiveram a taxa de rejeição diminuída com significância estatística.

A partir dos históricos de acesso dos usuários foram extraídas 28 regras de associação entre páginas da cultura de cana-de-açúcar, sendo que a maioria das páginas apresentava uma ou, no máximo, quatro recomendações.

## REFERÊNCIAS


AGRAWAL, R.; IMIELINSKI, T.; SWAMI, A.N. Mining Association Rules between Sets of Items in Large Databases. SIGMOD, v.22, p.207-216, 1993. http://dx.doi.org/10.1145/170036.170072

CHANGCHIEN, S.W.; LEE, C.F.; HSU, Y.J. On-line personalized sales promotion in electronic commerce. Expert Systems with Applications, v.27, p.35-52, 2004. http://dx.doi.org/10.1016/j.eswa.2003.12.017

CHAPMAN, P.; CLINTON, J.; KERBER, R.; KHABAZA, T.; REINARTZ, T.; SHEARER, C.; WIRTH, R. CRISP-DM1.0: step-by-step data mining guide. Illinois: SPSS, 2000. 78p.

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

GAUDER, M.; GRAEFF-HOENNINGER, S.; CLAUPEIN, W. The impact of a growing bioethanol industry on food production in Brazil. Applied Energy, v.88, p.672-679, 2011. http://dx.doi.org/10.1016/j.apenergy.2010.08.020

HAN, J.; KAMBER, M.; PEI, J. Data mining: concepts and techniques. 3rd ed. Morgan Kaufmann Publishers, 2011. p.703.

KAZIENKO, P. Mining indirect association rules for web recommendation. International Journal of Applied Mathematics and Computer Science, v.19, p.165-186, 2009. http://dx.doi.org/10.2478/v10006-009-0015-5

KENT, M.L.; BRYAN, J.C.; REBEKAH, A.H.; REBECCA, A.P. Learning web analytics: A tool for strategic communication. Public Relations Review, v.37, n.5, 2011. http://dx.doi.org/10.1016/j.pubrev.2011.09.011

KUMAR, A.; THAMBIDURAI, P. Collaborative Web Recommendation Systems - A Survey Approach. Global Journal of Computer Science and Technology, v.9, p.30-35, 2010.

PAKKALA, H.; PRESSER, K.; CHRISTENSEN, T. Using Google Analytics to measure visitor statistics: The case of food composition websites. International Journal of Information Management, v.32, p.504-512, 2012. http://dx.doi.org/10.1016/j.ijinfomgt.2012.04.008

PERKOWITZ, M.; ETZIONI, O. Towards adaptive Web sites: Conceptual framework and case study. Artificial Intelligence, v.118, p.245-275, 2000. http://dx.doi.org/10.1016/S0004-3702(99)00098-3

UNIÃO DA AGROINDÚSTRIA CANAVIEIRA DE SÃO PAULO - UNICA. Disponível em: <http://www.unica.com.br/noticia/871626920312979436/liderancas-do-setor-sucroenergetico-cobram-politicas-publicas-em-audiencia-no-senado/>. Acesso: 5 dez. 2012.

UNIÃO DA AGROINDÚSTRIA CANAVIEIRA DE SÃO PAULO – UNICA. Área total por estado. Disponível em: <http://www.unicadata.com.br>. Acesso: 22 ago. 2013.

WU, T.; CHEN, Y.; HAN, J. Re-examination of interestingness measures in pattern mining: A unified framework. Data Mining and Knowledge Discovery, v.21, p.371-397, 2010. http://dx.doi.org/10.1007/s10618-009-0161-2