FLAVIO MARGARITO MARTINS DE BARROS

UM SISTEMA DE RECOMENDAÇÃO PARA PÁGINAS WEB SOBRE A CULTURA DA CANA-DE-AÇÚCAR

CAMPINAS

2013

UNIVERSIDADE ESTADUAL DE CAMPINAS

FACULDADE DE ENGENHARIA AGRÍCOLA

FLAVIO MARGARITO MARTINS DE BARROS

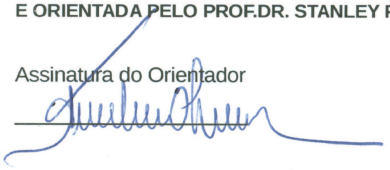
UM SISTEMA DE RECOMENDAÇÃO PARA PÁGINAS WEB SOBRE A CULTURA DA CANA-DE-AÇÚCAR

Orientador: Stanley Robson de Medeiros Oliveira

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Engenharia Agrícola da Faculdade de Engenharia Agrícola da Universidade Estadual de Campinas, para obtenção do Título de Mestre em Engenharia Agrícola, na área de concentração de Planejamento e Desenvolvimento Rural Sustentável.

ESTE EXEMPLAR CORRESPONDE À VERSÃO FINAL DA DISSERTAÇÃO
DEFENDIDA PELO ALUNO FLAVIO MARGARITO MARTINS DE BARROS
E ORIENTADA PELO PROF.DR. STANLEY ROBSON DE MEDEIROS OLIVEIRA

Assinatura do Orientador



CAMPINAS

2013

FICHA CATALOGRÁFICA ELABORADA PELA

BIBLIOTECA DA ÁREA DE ENGENHARIA E ARQUITETURA - BAE - UNICAMP

B278u

Barros, Flavio Margarito Martins de
Um sistema de recomendação de páginas web sobre a cultura da cana-de-açúcar / Flavio Margarito Martins de Barros. --Campinas, SP: [s.n.], 2013.

Orientador: Stanley Robson de Medeiros Oliveira.
Dissertação de Mestrado - Universidade Estadual de Campinas, Faculdade de Engenharia Agrícola.

1. Cana-de-açúcar. 2. Mineração de dados (Computação). 3. Sistemas de recomendação. 4. Serviços Web. I. Oliveira, Stanley Robson de Medeiros. II. Universidade Estadual de Campinas. Faculdade de Engenharia Agrícola. III. Título.

Título em Inglês: A recommender system for web pages regarding sugarcane crop
Palavras-chave em Inglês: Surgarcane, Data Mining (Computing), Recommender systems, Web services
Área de concentração: Planejamento e Desenvolvimento Rural Sustentável
Titulação: Mestre em Engenharia Agrícola
Banca examinadora: Leandro Balby Marinho, Zigomar Menezes de Souza
Data da defesa: 22-02-2013
Programa de Pós Graduação: Engenharia Agrícola

Este exemplar corresponde à redação final da **Dissertação de Mestrado** defendida por **Flavio Margarito Martins de Barros**, aprovada pela Comissão Julgadora em 22 de fevereiro de 2013, na Faculdade de Engenharia Agrícola da Universidade Estadual de Campinas.

Prof. Dr. Stanley Robson de Medeiros Oliveira – Presidente e Orientador
Feagri/Unicamp

Prof. Dr. Leandro Balby Marinho- Membro Titular
UFCG

Prof. Dr. Zigomar Menezes de Souza – Membro Titular
Feagri/Unicamp

FEAGRI
Faculdade de Engenharia Agrícola
Unicamp

Dedico à minha família: minha esposa,
meus pais, irmão e todos os amigos que me
apoiaram nessa empreitada. Agradeço a Deus
pela presença de todos na minha vida.

"O temor do Senhor é o princípio da sabedoria, e
o conhecimento do Santo é entendimento."
Provérbios 9:10

AGRADECIMENTOS

Primeiramente a DEUS que, sempre onipresente, me permitiu chegar até aqui.

Aos orientadores Prof. Dr. Stanley Robson de Medeiros Oliveira e Dr. Leandro Henrique Mendonça de Oliveira pela orientação, suporte, oportunidade de crescimento pessoal e profissional e pelo exemplo do bom trabalho e dedicação à pesquisa.

À CAPES pelo suporte financeiro.

À Embrapa Informática Agropecuária, especialmente ao Laboratório de Tratamento e Organização da Informação Eletrônica, pela oportunidade de utilizar suas dependências físicas e a infraestrutura computacional durante a realização do projeto.

Aos meus pais Marlene e José Martins e ao meu irmão Fábio com os quais sempre tive apoio e incentivo. Em especial a minha mãe, que com suas orações sempre esteve intercedendo por essa conquista.

Ao amor de minha esposa, Ana Paula de Barros, que compartilhou de perto as tristezas e as alegrias dessa empreitada. As noites sem dormir, o stress e as ausências não impediram que ela sempre estivesse ao meu lado me ajudando em tudo e sendo a esposa perfeita que DEUS me deu.

Aos meus sogros e amigos Geralda e Abrão, que também me apoiaram e incentivaram.

Aos meus muitos colegas do IMECC e da FEAGRI com quem pude compartilhar os problemas de pesquisa e obter importantes soluções. Em especial ao meu amigo Wanderson, que estendeu seus momentos de café compartilhando comigo conversas que tanto contribuíram. A todos os meus amigos que compreenderam minhas ausências devido aos estudos. Em especial ao Glauco, com quem sempre se pode contar a qualquer hora.

À minha Igreja, onde sempre recebi apoio dos meus amigos e dos pastores, sempre com boas lições e boas palavras para continuar seguindo no trabalho.

A todos os demais professores, pesquisadores e funcionários da Faculdade de Engenharia Agrícola, da Universidade Estadual de Campinas, que me apoiaram diretamente e indiretamente.

Sumário

LISTA DE TABELAS

LISTA DE FIGURAS

RESUMO

Sistemas de informação web oferecem informações em quantidade elevada, tal que a tarefa de encontrar a informação de interesse torna-se desafiadora. A Agência de Informação Embrapa é um sistema web com o objetivo de organizar, tratar, armazenar e divulgar informações técnicas e conhecimentos gerados pela EMBRAPA (Empresa Brasileira de Pesquisa Agropecuária). O portal está estruturado como uma árvore hierárquica, denominada Árvore de Conhecimento, a qual compreende centenas de páginas web, artigos, planilhas e materiais multimídia. Diariamente o site recebe milhares de acessos tal que os registros dessas visitas são armazenados em um banco de dados. Em domínios onde estão disponíveis informações em quantidade elevada, armazenadas em bancos de dados, as ferramentas de Mineração de Dados são promissoras, pois apresentam recursos para análise e extração de padrões de uso do site para fazer recomendações. Recomendações personalizadas de conteúdo melhoram a usabilidade de sistemas, agregam valor aos serviços, poupam tempo e fidelizam usuários. O objetivo desse trabalho foi projetar, desenvolver e implantar um sistema de recomendação web, baseado em regras de associação, que ofereça recomendações automaticamente de conteúdos da cultura da cana-de-açúcar, de acordo com o perfil da comunidade de usuários. Os dados utilizados nessa pesquisa foram extraídos de um banco de dados de acessos do projeto Agência de Informação Embrapa. A metodologia utilizada na pesquisa compreendeu a preparação dos dados de visitas ao site para uma estrutura de "lista de acessos", onde estão registradas todas as páginas visitadas por cada usuário. A partir destas listas de acesso, regras de associação entre páginas foram geradas por meio do algoritmo Apriori. O conjunto de regras deu origem a uma base de conhecimento que foi armazenada em um banco de dados para fazer recomendações de conteúdo aos usuários. Como suporte à base de conhecimento, para cada página da agência cana-de-açúcar foi criada uma lista de até três das páginas mais visitadas. Essas páginas podem ser oferecidas caso haja ausência de recomendações. O sistema de recomendação foi avaliado com uma métrica denominada taxa de rejeição e, por meio de um questionário aplicado a um conjunto de usuários, foi avaliada a usabilidade da Agência cana-de-açúcar, após a implantação do sistema. A base de conhecimento, gerada na forma de regras de recomendação, também foi avaliada em relação à estrutura de links da Agência, para verificar se a lista de recomendações trouxe conhecimentos sobre a estrutura do portal. De acordo com os resultados da pesquisa, por meio das

recomendações, usuários encontram informações relevantes associadas às suas visitas, aumentam seu tempo de permanência no site e aumentam o uso e visualização dos conteúdos da Agência de Informação Embrapa – Árvore cana-de-açúcar. Em páginas com dezenas de links, a base de conhecimento também atua como uma forma de resumo, apontando os principais links nas páginas.

PALAVRAS-CHAVE: Cana-de-açúcar; Mineração de dados (Computação); Sistemas de recomendação; Serviços Web.

ABSTRACT

Web information systems provide a great amount of information, so that the task of retrieving the information of interest becomes a challenge. Embrapa Information Agency is a web system aimed to organize, treat, store and disseminate technical information and knowledge generated by EMBRAPA (Brazilian Agricultural Research Corporation). The Agency´s portal is structured as a hierarchical tree, called Knowledge Tree, which comprises hundreds of web pages, articles, spreadsheets and multimedia materials. Everyday this site receives thousands of access and the records of these visits are stored in a database. In domains where information is available in high quantity, stored in databases, Data Mining tools are promising, since they have resources for extraction and analysis of usage patterns of the site to make recommendations. Personalized recommendations of content improve the usability of systems, add value to services, save time and retain users. The aim of this work was to design, develop and deploy a web recommendation system based on association rules, which offers automatically recommendations of sugarcane contents, according to the profile of user community. The data used in this study were extracted from a database of accesses from Embrapa Information Agency. The methodology used in the research included a data preparation procedure to transform website visits into a structured access list, in which all page views by each user are stored. From these access lists, association rules between pages were generated by means of the Apriori algorithm. The set of rules has created a knowledge base that was stored in a database to make content recommendations to users. To support the knowledge base, for each page of the sugarcane Agency was created a list of up to three of the most visited pages. These pages can be offered if there are no recommendations. The recommender system was evaluated by using a metric called bounce rate. In addition, through a questionnaire applied to a set of users, the usability of the sugarcane Agency was evaluated, after the system deployment. The knowledge base generated in the form of recommendation rules was also evaluated in relation to link structure of Agency, to verify if the list of recommendations brought knowledge about the structure of the portal. According to the survey results, users find relevant information associated with their visits, increase their time spent on the site and increase the use and the interest of the contents of sugarcane Agency. In

pages with dozens of links, the knowledge base also acts as a form of summarizing them, indicating the main links on the pages.

KEYWORDS: Agricultural recommender systems; Sugarcane; Association rules; Web information systems.

1. INTRODUÇÃO

O Brasil é atualmente o maior produtor de cana-de-açúcar e exportador de açúcar do mundo, sendo a região sudeste a maior produtora. Em particular, o estado de São Paulo é o maior produtor nacional e apresenta grandes extensões de áreas plantadas e muitas usinas instaladas. Além do Estado de São Paulo, também se destacam Paraná, Minas Gerais, Mato Grosso e Goiás (NEVES e CONEJERO, 2007).

Na conjuntura da economia brasileira, a cultura da cana-de-açúcar passou a se destacar 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. Ainda de acordo com o autor, a produção brasileira de etanol a base de cana-de-açúcar é vista como a mais efetiva tecnologia de biocombustíveis no mundo.

Devido à importância capital dessa cultura para agricultura do país, cujo PIB em 2009 foi de R$65,8 bilhões (CEPEA, 2010), 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 segmentos agrícolas, percebe-se a necessidade da boa gestão da informação que auxilie no processo de tomada de decisões no agronegócio (SHEN *et al*., 2011). Ciente dessa demanda, a EMBRAPA (Empresa Brasileira de Pesquisa Agropecuária), que é uma empresa pública com a missão de desenvolver pesquisas e disseminar informações técnicas no âmbito da agricultura, desenvolveu um portal de informações técnicas, a Agência de Informação Embrapa.

A Agência é um sistema web com o objetivo de organizar, tratar, armazenar e divulgar informações técnicas e conhecimentos gerados pela EMBRAPA e outras instituições parceiras. Por meio do endereço eletrônico (http://www.agencia.cnptia.embrapa.br/gestor/cana-de-acucar/Abertura.html), 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 de Informação da cana-de-açúcar 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. De acordo com Bertin *et al*. (2009), a EMBRAPA tem mantido uma política de oferta de informações técnicas para pesquisadores, produtores e a sociedade em geral, com a missão de fazer chegar à sociedade os resultados da pesquisa científica, relativa à toda cadeia produtiva do agronegócio, com o objetivo de criar uma robusta infraestrutura social, técnica e econômica tão necessária ao processo de desenvolvimento.

Sistemas web, como a Agência Embrapa, disponibilizam informações na internet em quantidade elevada. Um portal de informações como esse pode conter milhares de páginas de conteúdo individual, mesmo para uma única cultura, como por exemplo a cana-de-açúcar. Essa oferta de informações em grande quantidade pode confundir e dificultar o acesso pelos usuários (YANG e TANG, 2003).

Sistemas de informação web, em comparação a sistemas de informação convencionais, apresentam características muito distintas. Do ponto de vista do gerenciamento de dados, de acordo com Fraternali (1999), as características dos sistemas web são:

1. Manipulação de dados estruturados e não estruturados.
2. Suporte a acesso exploratório por meio de interfaces de navegação.
3. Customização e adaptação a estrutura de conteúdo.
4. Suporte a comportamento pró-ativo, isto é, recomendação e filtragem[1]

Também Overmyer (2000) ressalta três diferenças:

1. Um foco diferente: o gerenciamento do conteúdo é parecido com o gerenciamento de uma revista.
2. A funcionalidade, usabilidade e o design gráfico são muitos importantes.
3. Ciclos de vida mais curtos.

Nesse contexto, pode-se ver um sistema web como uma plataforma que suporta uma grande variedade de serviços. Como esses sistemas são disponibilizados na web, não é possível saber, a priori, quem são os usuários. Geralmente esses sistemas são desenvolvidos para um conjunto muito variado de tipos de usuários (YANG e TANG, 2003).

Antes do design e da implementação, os desenvolvedores e mantenedores desses sistemas enfrentam muitas dificuldades na definição do perfil dos usuários e suas necessidades. Geralmente, é

[1] Filtragem colaborativa e filtragem baseada em conteúdo. Ambas as técnicas estão apresentadas no capítulo 2.

difícil de se identificar o usuário alvo, que pode se beneficiar mais das informações disponíveis. Mesmo após entrar em operação, devido à grande quantidade de acessos, ainda é difícil determinar o perfil desses usuários. Também as necessidades desses usuários são voláteis, em parte porque com o crescimento esperado de sua quantidade, espera-se uma variação grande de perfis (YANG e TANG, 2003).

Devido às dificuldades explicitadas no parágrafo anterior, uma forma de proceder no caso de os usuários não encontrarem a informação desejada, por causa do volume elevado de informações, ou mesmo por não terem tempo suficiente ou não possuírem as habilidades necessárias para a procura, é a recomendação de conteúdo (KUMAR e THAMBIDURAI, 2010). Geralmente, quando não se tem conhecimento de algum domínio e se precisa de determinada informação ou produto, utiliza-se a recomendação pessoal de terceiros, que pode chegar de forma direta ou indireta, por meio de opiniões e revisões. Sistemas automáticos de recomendação auxiliam nessa tarefa, aumentando a eficiência desse processo de indicação (RESNICK e VARIAN, 1997).

Diante desse cenário, observa-se que a Agência de Informação Embrapa, que compreende centenas de páginas web e registra diariamente milhares de acessos de usuários em um banco de dados, pode se beneficiar da implantação de um sistema de recomendação baseado no perfil de uso da comunidade. Nesse domínio onde há informações de culturas agrícolas em quantidade elevada, armazenadas digitalmente em bancos de dados, para analisar esses dados, as ferramentas de mineração de dados apresentam recursos que fornecem padrões de uso do site para fazer recomendações aos usuários.

Dentre as técnicas utilizadas para identificar o comportamento de uso de sites e oferecer recomendações aos seus usuários pode-se utilizar a mineração de dados, etapa principal do processo de Descoberta de Conhecimento em Bases de Dados (Knowledge Discovery in Data Base – KDD), cujo objetivo é encontrar padrões e tendências nesses dados armazenados (HAN *et al*., 2011). A escolha das técnicas de mineração de dados apresenta-se como uma alternativa promissora, já que essas técnicas podem ser usadas para transformar registros de acessos de páginas Web em recomendações personalizadas para uma comunidade de usuários.

Dessa forma, por meio de ferramentas de mineração de dados, particularmente a geração de regras de associação, tendo como fonte de dados o banco de dados de acessos da Agência de Informação Embrapa, a implantação de um sistema de recomendação de conteúdo agrícola sobre a cana-de-açúcar constitui-se numa forma interessante de melhorar a oferta de informações técnicas

fornecidas pela EMBRAPA aos usuários da Agência. Assim, a proposta desse trabalho é uma das primeiras iniciativas da aplicação de sistemas de recomendação na agricultura.

1.1 Hipótese

É possível diminuir a taxa de rejeição de um sistema de recomendação de páginas Web para a cultura da cana-de-açúcar, desenvolvido por meio da aplicação de técnicas de mineração de dados aos registros dos históricos de navegação de usuários.

1.2 Objetivos

O objetivo geral desta pesquisa é projetar e desenvolver um sistema de recomendação web, baseado em regras de associação, que ofereça recomendações automáticas sobre a cultura da cana-de-açúcar, de acordo com o perfil da comunidade de usuários.

Dentre os objetivos específicos destacam-se:

1) Gerar regras de associação entre páginas, baseadas nos dados de visitas da comunidade de usuários.

2) Implementar, validar e implantar um sistema de recomendação automática, baseado nas regras geradas e que atualiza recomendações automaticamente.

3) Avaliar o impacto das recomendações no uso da Agência de Informação Embrapa por meio do número médio de páginas por sessão de usuário.

1.3 Organização da Dissertação

Para facilitar a compreensão deste trabalho, os capítulos seguintes foram organizados como segue:

O Capítulo 2 apresenta a revisão bibliográfica sobre a cana-de-açúcar, com ênfase na importância econômica da cultura, e a necessidade de sistemas de informações adequados como ferramentas para suporte e tomada de decisões. Foi realizada uma pesquisa extensa sobre sistemas de recomendação, explorando as possibilidades e técnicas mais usadas para construção de sistemas de recomendação no domínio de portais web. Por fim, foi realizada uma revisão das principais técnicas

de mineração de dados, com ênfase em regras de associação e mineração de dados da web.

A descrição do material e dos métodos utilizados para o desenvolvimento deste trabalho são apresentados no Capítulo 3. Em particular, a estrutura dos dados é apresentada, com a descrição dos parâmetros e do tamanho do conjunto de dados. São discutidos os softwares utilizados e os programas construídos para essa pesquisa. Considerando a etapa de tratamento, é mostrada a necessidade da transformação da estrutura dos dados e o procedimento utilizado para essa transformação. Por fim, é apresentado o procedimento para geração das regras com o algoritmo Apriori e como os links de recomendação são armazenados no banco de dados e disponibilizados aos usuários.

No Capítulo 4 são apresentadas as análises exploratórias, onde é mostrado o perfil de uso e o perfil dos usuários da Agência Embrapa cana-de-açúcar. Também são apresentadas estatísticas sobre a origem dos acessos e a distribuição dos acessos para cana-de-açúcar e outras culturas na agência. Por fim, é apresentada a arquitetura do sistema de recomendação construído e as análises, tanto da base de conhecimento gerada, quanto dos resultados da validação do sistema junto aos usuários.

No Capítulo 5 são apresentadas as principais contribuições obtidas com a execução do trabalho e são apresentadas sugestões de trabalhos futuros.

2. REVISÃO BIBLIOGRÁFICA

Neste capítulo, será apresentada uma breve discussão sobre a cultura de cana-de-açúcar, em que serão considerados principalmente aspectos relativos à sua importância econômica e estatísticas de produção. Posteriormente, será dada ênfase à conceituação de sistemas de informações agrícolas e a necessidade de um sistema de recomendação de conteúdos para esta cultura. Em seguida, será apresentada a caracterização de um sistema de recomendação e suas principais diferenças com relação aos sistemas de recomendação para informações na web. Além disso, serão apresentados conceitos relativos à mineração de dados com destaque para a geração de regras de associação.

2.1 Cultura da cana-de-açúcar

A cana-de-açúcar é uma planta de origem asiática e foi trazida ao Brasil pelos colonizadores portugueses a partir do seu cultivo em outras colônias. As espécies atualmente cultivadas no Brasil são híbridos derivados do cruzamento das espécies *officinarum* e *spontaneum* (DILLON *et al*., 2007).

Quanto à produção, até o primeiro corte, intervalo que pode demorar de 12 a 18 meses, a cana-de-açúcar recebe o nome de cana-planta. Após este período a rebrota da cana-de-açúcar passa a ter um ciclo normal de 12 meses, quando é denominada cana-soca, de forma que os estágios anuais de corte da cana se repetem até que a lavoura não seja mais rentável economicamente (ANJOS e FIGUEIREDO, 2010).

Esta cultura tem sido relevante para a economia brasileira desde o século XVI. As primeiras mudas vieram da Ilha da Madeira por volta de 1515 e o primeiro engenho estabelecido em 1532. Atualmente o Brasil é maior produtor mundial de cana-de-açúcar, com aproximadamente 9,1 milhões de hectares cultivados (UNICA, 2012). A safra 2011/2012 de cana-de-açúcar no Brasil atingiu 559,215 milhões de toneladas. Do total, 48,44% foram para fabricação de açúcar e 51,56% para a extração de etanol combustível (UNICA, 2012). O Brasil também produz cana-de-açúcar para alimentação animal, para fabricação de cachaça, xarope de cana, dentro outros produtos.

Atualmente o negócio do etanol de cana-de-açúcar brasileiro tem experimentado um grande crescimento, devido ao crescimento dos mercados interno e externo, tendência de aumento dos preços do petróleo, crescimento da frota de carros *flex-fuel* e uma preocupação mundial com o uso de energias

alternativas. Todos esses fatores têm atraído a atenção mundial com relação ao potencial brasileiro de produção do biocombustível, que possui um dos custos mais baixos de produção do mundo (KOHLHEPP, 2010).

Dessa forma, observa-se um crescimento acentuado de áreas plantadas e da produção de cana-de-açúcar ao longo dos anos. A Figura 1 ilustra a expansão das áreas plantadas, onde pode-se perceber um aumento acentuado nos estados de São Paulo, Minas Gerais, Goiás, Mato Grosso do Sul, Mato Grosso e Paraná. Somente a região Centro-Sul concentra 90% da produção de cana-de-açúcar no Brasil, com destaque para o Estado de São Paulo, principal produtor, com mais de 60% da produção nacional de etanol e açúcar, e também mais de 70% das exportações (UNICA, 2012).

No entanto, apesar da importância econômica da cana-de-açúcar para o país, e também da complexidade do sistema produtivo, percebe-se uma lacuna quanto à oferta de informações tecnológicas, voltadas ao atendimento das necessidades do setor, principalmente a partir de 1990, quanto foi extinto o instituto do Açúcar e do Álcool (IAA) (SOUZA *et al*., 2009).

Também, segundo Francisco (2003), tem-se observado um acréscimo na utilização da internet nos setores rurais, principalmente porque, devido à revolução nas comunicações, a internet se transformou em uma ferramenta poderosa para o acesso imediato a informações sobre os preços do mercado mundial, estratégias de negociação, análise do potencial de produtos em diferentes mercados, novas técnicas de produção, novos sistemas de transporte; podendo ainda reduzir custos das transações.

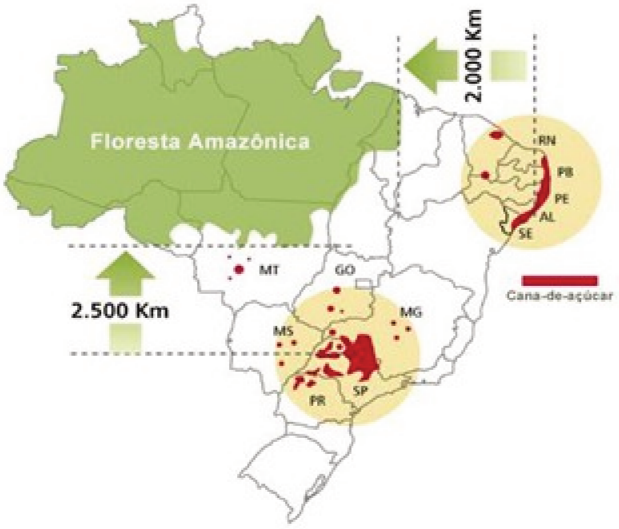


Figura 1: Áreas de plantio de cana-de-açúcar e usinas produtoras de açúcar e etanol. Fonte: Núcleo Interdisciplinar de Planejamento Estratégico (NIPE)/Universidade Estadual de Campinas (Unicamp), Instituto Brasileiro de Geografia e Estatística (IBGE), Centro de Tecnologia Canavieira (CTC) (2008).

Compreendendo esta necessidade de informação, a EMBRAPA, uma das maiores geradoras e detentoras de conhecimento no agronegócio da América Latina, com o objetivo de preencher essa lacuna da oferta de informações, dado o aumento e melhoramento do acesso à internet por parte do meio rural, tem oferecido informações tecnológicas relativas à cultura de cana-de-açúcar, no endereço eletrônico (http://www.agencia.cnptia.embrapa.br/gestor/cana-de-acucar/Abertura.html), onde o usuário tem acesso aos conhecimentos técnicos gerados pela EMBRAPA (OLIVEIRA *et al*., 2009).

Esse portal, denominado Agência de Informação Embrapa, é um sistema de informações *web*, que possibilita a organização, o tratamento, o armazenamento, a divulgação e o acesso à informação tecnológica e ao conhecimento gerado pela Embrapa e outras instituições de pesquisa. Essas informações estão organizadas em uma estrutura ramificada, denominada Árvore do Conhecimento, na qual o conhecimento é organizado de forma hierárquica (EMBRAPA, 2012).

Nesse contexto, considerando a importância da cana-de-açúcar para o país e a oferta de informações tecnológicas disponível na *world wide web*, sobre a cultura, este trabalho apresenta um sistema de recomendação para auxiliar o desenvolvimento do agronegócio da cana-de-açúcar no Brasil.

2.2 Sistemas de recomendação

Quando se necessita de um produto ou serviço relacionado a um assunto que não tem-se domínio é comum buscar a opinião de terceiros para determinação das escolhas (MAES e SHARDANAND, 1995). Esse processo pode ser denominado indicação. A indicação, de forma simples, significa aconselhar, dar sugestão a respeito de algo.

Com a sobrecarga de informações sobre usuários de sistemas web, principalmente aqueles voltados ao comércio digital, muitas organizações têm implantado sistemas de recomendação em seus portais. Existem experiências bem-sucedidas para indicação de livros, CDs, e outros produtos na Amazon.com (LINDEN e SMITH, 2003), filmes pela MovieLens (MILLER *et al*., 2003) e notícias pela VERSIFI Technologies (BILLSUS *et al*., 2002). Atualmente a pesquisa na área vai em direção a melhores representações do comportamento dos usuários e informações dos itens a serem recomendados (RICCI *et al*., 2011).

Em sua forma mais rudimentar, o problema da recomendação nesses sistemas é reduzido à

estimativa de *ratings*[2] para produtos que o usuário desconhece. Esses *ratings* são gerados a partir das opiniões de outros usuários do sistema, de forma que quando um novo usuário aparece, este deve receber a recomendação do produto com os maiores *ratings* (JANNACH *et al*., 2011).

Essa abordagem parte do princípio que um produto já deva ter sido classificado por outros usuários que já utilizaram aquele produto. Logo, como muitos produtos podem não ter sido classificados, tem-se a necessidade de estimar o *rating* para todos os produtos oferecidos, por meio de algum processo automatizado.

Quanto às técnicas utilizadas para a geração das estimativas dos *ratings*, temos basicamente duas abordagens: a filtragem colaborativa e a filtragem baseada em conteúdo (RICCI *et al*., 2011). Existem também técnicas híbridas que combinam as duas abordagens (BURKE, 2000).

Tanto a filtragem colaborativa, quanto a filtragem baseada em conteúdo são as abordagens mais populares para construir sistemas de recomendação. Sistemas híbridos que combinam as características de ambos e sistemas baseados em *web usage mining*[3] têm sido apresentados como alternativas para contornar os problemas associados às abordagens colaborativas e por conteúdo isoladamente (ITTOO *et al*., 2006).

De acordo com Reategui e Cazella (2005), as estratégias mais utilizadas em sistemas de recomendação são:

1) Listas de Recomendação;
2) Avaliações de Usuários;
3) Suas Recomendações;
4) Usuários que se interessam por X também se interessam por Y;
5) Associação por conteúdo.

As listas de recomendação são uma estratégia que consiste em manter listas de itens organizados por interesse, sem a necessidade de uma análise mais profunda dos dados do usuário. As avaliações dos usuários são informações fornecidas explicitamente pelos próprios usuários, ou na forma de *ratings* ou mesmo com textos descrevendo experiências de uso. Em suas recomendações, os usuários fornecem voluntariamente listas de produtos que consideram interessantes. Na associação por conteúdo itens são oferecidos de acordo com a similaridade a itens sendo visualizados. Das técnicas apresentadas, a mais importante no contexto deste trabalho é a técnica 4, uma vez que o sistema de

[2] *Ratings* são notas ou avaliações fornecidas pelo usuário do sistema web. Geralmente são valores numéricos, que podem ir de 0 a 5 de acordo com a satisfação do usuário.

[3] *Web Usage Mining* é um termo em inglês que designa o campo de estudos que explora aplicações de mineração de dados ao uso de websites. O assunto é explorado no Capítulo 2, subseção 2.6.2.

recomendação implementado foi baseado em listas de recomendação.

De acordo com a técnica 4, a recomendação é obtida a partir da associação entre itens vistos por um usuário, a partir de registros armazenados em uma base de dados. Para determinação desses padrões, há a necessidade da aplicação de técnicas de mineração de dados, uma vez que a base de dados pode conter milhares de registros.

Na Agência de Informação Embrapa, especialmente na árvore de conhecimento[4] relacionada à cana-de-açúcar, os registros de uso podem apresentar as transações como dispostas na Tabela 1. Por meio destes acessos é possível transpor essa estrutura para uma estrutura de "lista de acessos", onde são apresentadas todas as páginas visitadas por cada usuário (Tabela 2). Nesta representação, cada usuário, representado por um ID, acessa uma lista de páginas. Por exemplo, o usuário cujo ID é 001 acessou o conteúdo das páginas relacionadas à praga no colmo e praga nas raízes. Já o usuário cujo ID é 007 acessou o conteúdo da página relacionada à produção.

Tabela 1 – Acessos à Agência Embrapa.

ID	Item (página)	Data do Acesso
001	Praga no Colmo	15/06/2011
001	Praga nas Raízes	15/06/2011
007	Produção	15/06/2011
006	Produção	15/06/2011
004	Praga no Colmo	15/06/2011
004	Praga as Raízes	15/06/2011
004	Produção	15/06/2011

Tabela 2 – Acessos à Agência Embrapa em forma de listas.

ID	Lista de páginas visitadas
001	{Praga no Colmo, Praga nas Raízes}
004	{Praga no Colmo, Praga nas Raízes, Produção}
006	{Produção}
007	{Produção}

Por meio dessa estrutura, com algoritmos de mineração de dados, como o algoritmo Apriori (AGRAWAL *et al*., 1993), é possível inferir regras como a regra (1). Esta regra indica que usuários que acessam o conteúdo da página sobre pragas no colmo podem potencialmente acessar o conteúdo da página pragas nas raízes.

$$\text{Praga no Colmo} \rightarrow \text{Praga nas Raízes} \qquad (1)$$

[4] Árvore de conhecimento é uma estrutura que representa a natureza hierárquica de um assunto de forma gráfica. É denominada "árvore de conhecimento" pois sua representação é semelhante a uma árvore.

De acordo com a técnica de associação por conteúdo, a determinação das recomendações deve se basear no conteúdo dos materiais a serem recomendados para dado usuário. Segundo Mooney e Roy (2000), descobrir os perfis ou gostos, de cada usuário, a partir das características dos produtos, permite caracterizar o perfil de um usuário sem ter de relacionar seus interesses ao de outros usuários. Os itens são recomendados de acordo com a informação sobre o item ao invés das preferências de outros usuários.

A aplicação de uma recomendação automática baseada em conteúdo depende da aplicação de alguma técnica de extração automática de informação do texto, como o *text mining*[5] (ITTOO *et al*., 2006), ou por meio de uma descrição anterior feita por humanos (MOONEY e ROY, 2000).

2.2.1 Filtragem baseada em conteúdo

Segundo Herlocker (2000), há mais de trinta anos cientistas têm tentado resolver o problema da sobrecarga de informações aos usuários por meio de softwares que reconhecem e categorizam automaticamente a informação. Tais softwares geram descrições do conteúdo dos itens oferecidos e então comparam com a descrição da necessidade dos usuários, para determinar se o item é de interesse. A descrição dos interesses dos usuários pode ser determinada de forma explícita, por meio de opiniões ou *ratings*, ou pode ser determinada implicitamente pela observação das ações do usuário. Essas técnicas recebem a designação de filtragem por conteúdo, em virtude da filtragem aplicada pelos softwares estar baseada na análise do conteúdo dos itens. Em uma recomendação baseada em conteúdo, o sistema recomenda itens similares às preferências do usuário no passado (PANAGGIO, 2010).

Ainda de acordo com Herlocker (2000), muitas aplicações que utilizam essa abordagem aplicam técnicas como indexação de frequência de termos, índices de busca booleana, interfaces probabilísticas, interfaces de consulta com linguagem natural e mineração de dados em textos.

Em sua forma mais simples, um sistema baseado em filtragem por conteúdo pode utilizar avaliações dos próprios usuários de produtos de seu interesse, procurar itens similares em conteúdo e oferecê-los como recomendações (REATEGUI e CAZZELLA, 2005). De acordo com Pedronette (2008), a filtragem baseada em conteúdo apresenta alguns inconvenientes como: a dificuldade para análise de conteúdos não textuais tal que a extração e comparação do conteúdo dos objetos é difícil e o

[5] *Text mining* se refere à mineração de dados em textos. Compreende uma ampla gama de técnicas e algoritmos usados para extrair informações de textos automaticamente.

problema do novo usuário onde não se sabe nada dos itens preferidos por certo usuário e por conseguinte do conteúdo desses itens.

Um exemplo de um sistema de recomendação baseado em conteúdo é descrito em Mooney e Roy (2000). Esse sistema é capaz de oferecer recomendações de livros utilizando-se de técnicas de aprendizado de máquina para categorização de textos. Existem outros exemplos de sistemas de recomendação que utilizam aprendizado de máquina para categorização de texto como para páginas web (PAZZANI e BILLSUS, 1997) e grupos de notícias (LANG, 1995). Especificamente, esse sistema de recomendação utiliza um algoritmo de mineração de textos que extrai as palavras mais importantes dos textos. Dessa forma cada livro pode ser representado por um vetor de strings ou palavras, a partir dos quais se pode determinar se o item é adequado como uma recomendação ao usuário, por meio de um cálculo probabilístico de um *score* numérico de cada palavra no vetor.

Ainda segundo Parik *et al*. (2007), sistemas de informações tecnológicas agrícolas têm como extensão natural sistemas de recomendação de conteúdo, uma vez que os produtores podem aprender de seu comportamento e se beneficiarem dessas informações. No entanto, até onde se sabe, não existem registros na literatura sobre sistemas de recomendação baseados em conteúdo para informações tecnológicas agrícolas. Assim a proposta desse trabalho é uma das primeiras iniciativas da aplicação de sistemas de recomendação na agricultura.

2.2.2 Filtragem colaborativa

A abordagem da filtragem colaborativa procura resolver alguns problemas em aberto na filtragem baseada em conteúdo. Na filtragem baseada em conteúdo, o sistema deve obter alguma informação ou representação do conteúdo dos itens que serão recomendados, o que pode ser um problema em sistemas com elevado número de itens com conteúdo não textual.

Na filtragem colaborativa a ideia principal é que os usuários aproveitem dos conhecimentos dos outros usuários para efetuar suas escolhas. Na literatura, o primeiro sistema de recomendação baseado na técnica de filtragem colaborativa foi o Tapestry (GOLDBERG *et al*., 1992) que recomendava memorandos importantes previamente classificados por outros usuários. Em uma recomendação colaborativa, o sistema recomenda itens que foram preferência de usuários com perfis similares ao usuário que está recebendo a recomendação (PANAGGIO, 2010).

Nessa abordagem é essencial que o sistema armazene informações a respeito dos itens a partir de usuários que conhecem ou utilizaram esses itens. Essas experiências podem ser registradas como avaliações, históricos de comportamento, históricos de busca entre outros. Segundo Herlocker

(2010), nos primeiros sistemas, o usuário devia informar de modo explícito suas predileções, porém, em seguida esses sistemas automatizaram todo o processo por meio de coleções de pontuações.

Na Tabela 3 pode-se observar um exemplo de como esse processo pode funcionar em um sistema de informações tecnológicas agrícolas:

Tabela 3 – Exemplo de filtragem colaborativa

Usuário	Praga no colmo	Praga nas raízes	Cachaça	Produção
Flávio	x	x		
Stanley		x		
Daniela		x	x	x

Se o usuário Stanley necessita de uma recomendação, o sistema procura outros usuários com hábitos de uso semelhantes. Nesse caso os usuários Daniela e Flávio são utilizados, uma vez que ambos já visualizaram uma mesma página que Stanley: Praga nas raízes. Logo, possíveis recomendações são Praga no colmo, Cachaça ou Produção.

A filtragem colaborativa apresenta também algumas limitações:

- Problema do primeiro avaliador: quando um novo item aparece no sistema não existe forma de fazer uma recomendação desse item até este ser avaliado por um usuário.
- Problema de pontuações esparsas: caso o número de itens seja muito grande e o número de usuários muito pequeno, existe o risco de as pontuações ficarem esparsas.
- Similaridade: usuários com gostos e características singulares, que destoam da média, podem receber recomendações muito pobres.

Segundo Pedronette (2008), dentre os algoritmos mais utilizados nas técnicas de filtragem colaborativa predominam os aqueles baseados em vizinhança. Para Reategui e Cazella (2005), um dos algoritmos mais utilizados é o *k-nearest-neighbor* (k vizinhos mais próximos) que pode ser descrito em três fases:

1. Cálculo da similaridade entre cada usuário e o usuário que vai receber a recomendação;
2. Selecionar um grupo de usuários com mais similaridade (vizinhos) para determinar a predição;
3. Determinar as avaliações dos usuários e fazer a recomendação.

Além das estratégias especificadas, também se encontram na literatura abordagens estatísticas (YU *et al*., 2004), com redes neurais (PAZZANI e BILLSUS, 1997) e ontologias (MIDDLETON *et al*., 2004). No entanto, a filtragem colaborativa e a filtragem baseada em conteúdo são as principais técnicas utilizadas.

Na agricultura, em Kui *et al*. (2011) é descrito um sistema de recomendação baseado em filtragem colaborativa que oferece recomendações de informações tecnológicas agrícolas via internet para fazendeiros. Segundo o autor, como o público-alvo muitas vezes possui conhecimento limitado em tecnologia da informação, geralmente o julgamento e a capacidade de escolha quando há uma quantidade elevada de oferta de informações pode ser muito pobre, de forma que os usuários podem vir a não encontrarem a informação procurada. Esse sistema de recomendação utiliza técnicas de agrupamento, como *K-Means*[6], e a partir dos N itens mais similares, infere os *ratings* dos itens não avaliados pelo usuário alvo, determinando assim quais são os melhores serviços a serem recomendados.

2.2.3 Sistemas de recomendação híbridos

A filtragem híbrida é uma abordagem que visa combinar os pontos positivos da filtragem colaborativa e da filtragem baseada em conteúdo, de forma a melhor atender as necessidades de recomendações dos usuários (HERLOCKER, 2000; ANSARI, 2000).

Como se pode ver na Figura 2, a filtragem híbrida pode combinar as melhores características de ambas as abordagens atenuando seus problemas.

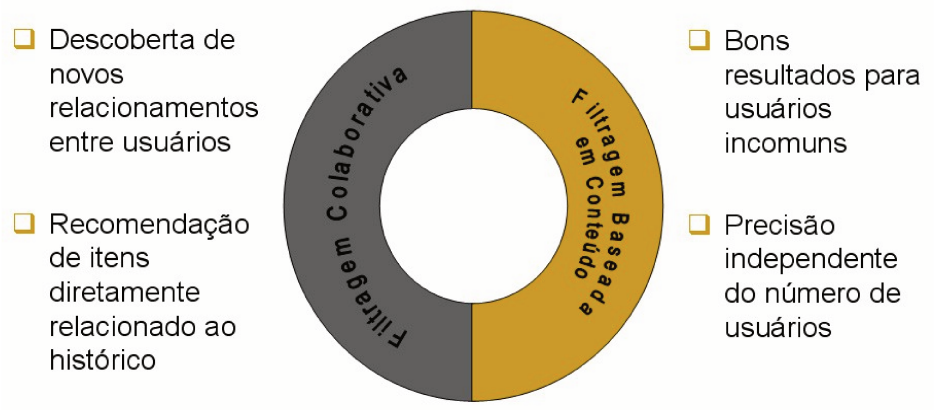


Figura 2: Características da filtragem híbrida (REATGUI e CAZELLA, 2004).

[6] *K-means* é um algoritmo utilizado para agrupamento, baseado em particionamento de acordo com a distância entre elementos (HAN *et al*., 2011).

Segundo Tuzhilin (2005), existem diferentes maneiras de combinar filtragem colaborativa e a filtragem baseada em conteúdo:

1. Implementar métodos colaborativos e baseados em conteúdo separadamente e combinar suas predições;
2. Incorporar algumas características de um método baseado em conteúdo em um método colaborativo;
3. Incorporar algumas características de um método colaborativo em um método baseado em conteúdo;
4. Construir um modelo unificado que incorpora ambas as abordagens;

Todas essas abordagens são encontradas na literatura. Em Claypool *et al*. (1999), o sistema de recomendação realiza uma combinação linear dos *ratings* obtidos por cada módulo, um baseado em conteúdo e outro colaborativo. Em Balabanovic e Shoham (1997) é descrito um sistema baseado na filtragem colaborativa, mas que mantém perfis dos usuários utilizando métodos baseados em conteúdo. Já em Soboroff e Nicholas (1999), o sistema de recomendação usa uma técnica de indexação baseada no significado para criar uma visão colaborativa de uma coleção de perfis de usuários. Finalmente, em Basu *et al*. (1998) é descrito um sistema de recomendação que utiliza uma abordagem unificada utilizando características baseadas em conteúdo e características colaborativas.

As técnicas utilizadas em sistemas de recomendação, vistas nesta e em seções anteriores, abarcam abordagens tão diversas como filtragem colaborativa, filtragem baseada em conteúdo, técnicas híbridas, ontologias, mineração de dados, dentre outras. O campo tem se desenvolvido nos últimos anos, principalmente devido à abundância de informações que circula na internet. O comércio eletrônico tem se beneficiado dessas tecnologias há anos. Ainda assim, dentro do contexto da agricultura, praticamente não são encontradas aplicações dessas tecnologias.

2.3 Mineração de dados

Mineração de Dados (*Datamining - DM*) e Descoberta de Conhecimento em Base de Dados (*Knowledge Discovery in Databases – KDD*) são termos usados para se referir à pesquisa, técnicas e ferramentas utilizadas para extrair informações de grandes volumes de dados. Segundo Piatetsky-Shaphiro e Frawley (1991), Fayyad *et al*. (1996), KDD é o processo completo de extração de conhecimento em bases de dados. Mineração de dados é somente uma etapa de todo o processo.

Em geral, o termo mineração de dados é utilizado por muitos pesquisadores como sinônimo do processo de KDD (CABENA *et al*., 1997; CHAPMAN *et al*., 2000; KURGAN e MUSILEK, 2006). Portanto, neste trabalho será utilizado somente o termo mineração de dados.

Nas últimas décadas, a pesquisa na área tem avançado com o advento dos computadores e a internet pois a geração de dados brutos é muito maior que a capacidade humana para analisá-los. Como resultado, no meio acadêmico e na indústria, pesquisadores têm buscado alternativas para trabalhar com um volume elevado de dados que impossibilita sua análise manual (KRIEGEL *et al*., 2007).

Em resposta às necessidades comuns em projetos de mineração de dados nos anos 90, um grupo de organizações envolvidas com a mineração de dados (Terradata, SPSS, -ISL-, Daimler-Chryslere OHRA) propuseram um guia de referência para o desenvolvimento de projetos de mineração de dados (MARISCAL *et al*., 2010).

De acordo com Chapman *et al*. (2000), o processo de mineração de dados é composto por seis fases do modelo CRISP-DM (Cross Industry Standard Process for Data Mining).

Abaixo uma breve descrição das fases do processo:

- **Compreensão do domínio**: o foco da fase inicial é o entendimento dos objetivos e necessidades do projeto e então converter esse conhecimento em um problema de mineração de dados propriamente dito. Compreende também o planejamento inicial para alcançar esses objetivos.

- **Entendimento dos dados**: essa fase se inicia com a coleta inicial dos dados e prossegue com atividades associadas ao entendimento dos dados, com objetivo de identificar problemas, compreender as especificidades dos dados e formar hipóteses.

- **Preparação dos dados**: esta fase compreende todas as atividades relacionadas à construção do conjunto final de dados, que será utilizado para análise em algum software estatístico ou de mineração de dados. De forma a utilizar os dados como entrada para esses softwares, a estrutura dos dados pode precisar ser alterada, ou mesmo alguma transformação seja necessária como inclusão de novas variáveis, mudança de escala dos valores das variáveis, mudança na estrutura dos dados ou seleção de um conjunto menor de atributos. É também nesta fase que ocorre a limpeza

do conjunto de dados, onde são verificados erros de digitação, dados faltantes e inconsistências no banco de dados.

- **Modelagem**: nesta fase, várias técnicas de modelagem, como por exemplo análise de regressão, árvores de decisão, redes neurais, regras de associação, dentre outras, são selecionadas e aplicadas aos dados. Nesta fase, mais de um modelo pode ser gerado. Dependendo da técnica utilizada, podem ser necessários novos ajustes, novas transformações no conjunto de dados ou ajustes de parâmetros do modelo. Os parâmetros que podem ser ajustados nos modelos dependem das técnicas utilizadas: podem ser a confiança e o suporte para regras de associação, número de níveis em uma árvore de decisão ou a inclusão de novos termos da equação de uma análise de regressão.

- **Avaliação**: nesta fase do projeto, os modelos já foram desenvolvidos. Antes de passar a fase final de desenvolvimento dos modelos finais, é importante revisar todos os passos executados, para verificar que os objetivos foram alcançados. No fim dessa fase, uma decisão a respeito do uso dos resultados da análise deve ser tomada.

- **Distribuição**: geralmente a construção dos modelos não é a fase final de um projeto de mineração de dados. Mesmo que o fim da pesquisa seja aumentar o conhecimento sobre os dados, será necessário organizar o conhecimento extraído bem como apresentá-lo de forma palatável ao seu usuário final. Dependendo do projeto, a distribuição pode ser simplesmente a preparação de um relatório ou muitas vezes pode ser uma tarefa mais complexa. A tarefa da distribuição nem sempre é responsabilidade do analista, mas é importante que mesmo que essa tarefa seja desempenhada por terceiros, que estes tenham algum conhecimento da forma como podem utilizar os modelos.

O ciclo externo, na Figura 3, simboliza a natureza cíclica da mineração de dados. O processo não termina uma vez que uma solução é encontrada. As lições aprendidas durante o processo podem gerar novos questionamentos, geralmente mais pertinentes ao assunto. Assim, o

analista ao final da fase de avaliação, por meio de novos conhecimentos adquiridos no próprio processo de mineração de dados, pode voltar a fase inicial e construir um modelo ainda melhor.

Em mineração de dados, de acordo com Han *et al*. (2011), as tarefas têm a função de especificar o tipo de padrão a ser encontrado nas bases de dados. Pode-se caracterizar as tarefas em duas grandes categorias: tarefas descritivas e tarefas preditivas, conforme Figura 4.

Tarefas descritivas têm a finalidade de caracterizar propriedades gerais dos dados na base de dados e tarefas preditivas têm por objetivo fazer inferências a partir dos dados de forma a fazer predições. As tarefas preditivas têm por objetivo a construção de modelos, a partir de um conjunto de dados, de forma a fazer inferências em novos dados. As principais tarefas de predição são classificação e regressão.

A classificação é o processo de encontrar um modelo com objetivo de distinguir classes, de forma que a partir desse modelo possa-se predizer a classe de objetos desconhecidos (HAN *et al*., 2011). Os modelos obtidos podem ser apresentados como regras, árvores de decisão, fórmulas matemáticas ou redes neurais.

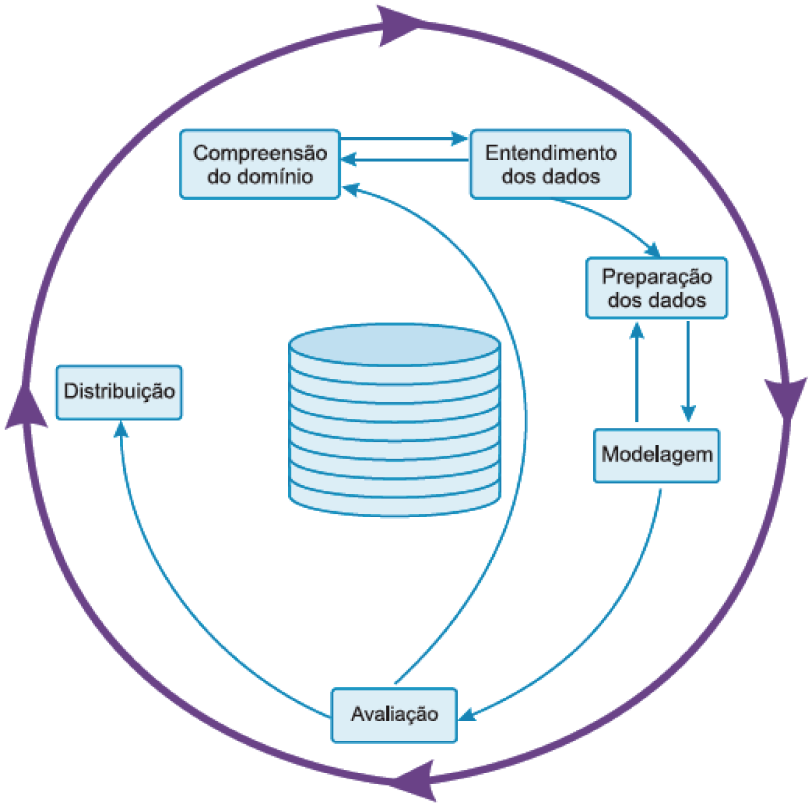


Figura 3: Fases da metodologia CRISP-DM (CHAPMAN *et al*., 2000)

Figura 4: Tarefas de Mineração de Dados (adaptado de REZENDE *et al*., 2005).

Árvores de decisão são amplamente utilizadas e são uma das técnicas de mineração de dados mais utilizadas (HAN *et al*., 2011). Em uma árvore de decisão o nó localizado no extremo superior da árvore representa seu início e é denominado nó raiz. Nós localizados na extremidade inferior são denominados nós folha e representam o valor de predição do atributo meta (variável dependente) e os nós internos denotam um teste em um atributo, de forma que cada ramo da árvore é o resultado desse teste.

Após a construção da árvore, esta pode ser utilizada para classificação de exemplos cuja classe é desconhecida. Para isso, um caminho é traçado a partir do nó raiz, descendo pelos ramos, até atingir um nó folha, que representa a classe de predição do exemplo em questão (KRIEGEL *et al*., 2007). O número de regras está associado ao número de caminhos da raiz até as folhas da árvore.

A regressão é uma metodologia estatística muito utilizada para fazer predições (HILL *et al*., 2003). Essas predições se caracterizam pela determinação de tendências de variações nos dados em função das variáveis existentes. Conceitualmente é similar à classificação, porém se aplica na predição de valores contínuos.

As tarefas descritivas consistem na identificação de padrões inerentes a um banco de

dados. Os dados desse banco não possuem classe especificada. Entre essas tarefas, destacam-se as regras de associação e clusterização. Especialmente regras de associação, foram introduzidas por Agrawal *et al*. (1993) e descrevem a relação entre itens ou produtos de uma base de dados. Existem diversas aplicações dessa técnica, como por exemplo, análise de informação médica, estudo do acesso a computadores e análise de perfil de compras de clientes.

A estrutura dos dados para aplicação do algoritmo pode ser descrita com uma estrutura de "cesta" (AGRAWAL *et al*., 1993). Nessa estrutura cada transação é representada como uma tupla, um vetor binário, tal que o valor 1 indica a presença do item na transação, enquanto o valor 0 representa a ausência do item. As regras de associação podem ser representadas da forma $X \rightarrow Y$, onde X e Y são conjuntos disjuntos de atributos, isto é, $X \cap Y = \varnothing$. Nessas regras, X representa o antecedente e Y o consequente.

Para cada regra de associação estão associadas duas medidas tradicionais: confiança (Conf) e suporte (Sup). Sup representa o número de tuplas que contêm X e Y. Do ponto de vista conceitual, representa a significância estatística desses itens nas tuplas, ao passo que Conf constitui a razão entre o número de tuplas que contêm X e Y sobre o número de tuplas que contêm X. 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.

$$Suporte(X \rightarrow Y) = P(X \cup Y) \tag{2}$$

onde $X \rightarrow Y$ representa uma regra de associação entre X e Y e P(X *U* Y) representa a probabilidade de encontrar transações no conjunto de dados que contenham X e Y.

$$Confiança(X \rightarrow Y) = P(X / Y) = \frac{Suporte(X \rightarrow Y)}{Suporte(X)} \tag{3}$$

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).

Para exemplificar o processo de geração de regras de associação, serão utilizados os dados na Tabela 4. A primeira coluna representa a identificação da transação e as outras colunas indicam se um

determinado artigo está presente na transação.

Tabela 4 – Exemplo de transações em um banco de dados.

ID	Artigo 1	Artigo 2	Artigo 3	Artigo 4	Artigo 5	Artigo 6	Artigo 7
1	Não	Sim	Não	Sim	Sim	Não	Não
2	Sim	Não	Sim	Sim	Sim	Não	Não
3	Não	Sim	Não	Sim	Sim	Não	Não
4	Sim	Sim	Não	Sim	Sim	Não	Não
5	Não	Não	Sim	Não	Não	Não	Não
6	Não	Não	Não	Não	Sim	Não	Não
7	Não	Não	Não	Sim	Não	Não	Não
8	Não	Não	Não	Não	Não	Não	Sim
9	Não	Não	Não	Não	Não	Sim	Sim
10	Não	Não	Não	Não	Não	Sim	Não

Primeiramente, o usuário deve especificar os níveis de suporte e confiança. Para exemplificar, seja Sup = 0,3 e Conf = 0,8. Por meio de algum pacote de software, pode-se gerar regras como a seguinte:

Conjuntos de itens frequentes: Artigo 2, Artigo 4.	**Sup = 0,3**
Regra: Se Artigo 2 → Artigo 4	**Conf = 1**

Inicialmente o algoritmo determina o conjunto de itens frequentes, isto é, aquele conjunto de itens que aparecem juntos em uma fração do total maior ou igual ao suporte. A partir desses conjuntos, também denominados *conjuntos frequentes* (AGRAWAL *et al*., 1994), o algoritmo procura regras que satisfaçam o requisito mínimo de confiança, isto é, tais que a confiança calculada seja maior ou igual ao nível mínimo determinado.

2.3.1 Mineração de dados na agricultura

Mineração de dados tem sido aplicada em diversos campos da economia e tem desempenhado um papel cada vez mais importante na agricultura. O conhecimento em mineração de dados permite à agricultura, assim como permitiu a outros setores, utilizar a informação de forma mais eficiente e eficaz (CHINCHULUUN e XANTHOPOULOS, 2010).

O campo de aplicações de mineração de dados na agricultura é ainda relativamente novo (MUCHERINO *et al*., 2009). Ainda assim existem aplicações na área de solos, clima, qualidade de frutos, classificação de doenças, classificação de imagens de satélite e etc.

Em Wu *et al*., (2008) são apresentados os dez algoritmos de mineração de dados mais utilizados em todos os domínios de aplicação. Embora não exista nenhuma técnica específica para resolver problemas exclusivos da agricultura, é possível encontrar na literatura aplicações em agricultura de quase todos estes algoritmos.

Uma das áreas onde a mineração de dados tem sido mais utilizada é no estudo dos solos. A maior parte das aplicações encontra-se na área de classificação de solos, onde as técnicas de mineração podem ser utilizadas para classificação de grandes conjuntos de dados (KUMAR e KANNATHASAN, 2010). Por exemplo, em Vibah *et al*. (2007) foi desenvolvido um modelo de classificação de solos, baseado em clusterização e classificação: inicialmente foi aplicado o algoritmo *K-means* para agrupamento de dados de solo e, em seguida, foi utilizado o algoritmo *random forest*[7] para classificação final dos solos dentro dos clusters.

Técnicas de mineração de dados também têm recebido atenção na área de climatologia, uma vez que modelos preditivos podem subsidiar decisões estratégicas com relação à época de plantio. Por exemplo, em Romani *et al*., (2010) foi desenvolvido um algoritmo para detecção de padrões de associação em séries temporais de dados climáticos e séries temporais de índices obtidos em imagens de satélite. Técnicas como esta, podem permitir predições de dados climáticos a partir de grandezas que podem ser facilmente extraídas de imagens de satélite.

Ainda em climatologia, também pode-se citar Boschi (2010), que analisou o comportamento espaço temporal da precipitação pluvial e dos veranicos no Estado do Rio Grande do Sul, em quatro zonas homogêneas, num período de 20 anos. Para análise das zonas homogêneas foram utilizadas técnicas de clusterização. Jan *et al*. (2009) desenvolveram um sistema que utiliza dados climáticos e outros atributos para previsão do clima, utilizando o algoritmo de classificação k-NN.

No segmento de análise de imagens na agricultura, Nonato (2010) desenvolveu modelos preditivos para identificar áreas cultivadas com cana-de-açúcar em imagens de sensoriamento remoto, no Estado de São Paulo. Nesse trabalho, foram utilizadas técnicas de seleção de atributos e árvore de decisão binária na identificação de áreas cultivadas com cana-de-açúcar.

Também existem aplicações na agricultura com a análise de dados mercadológicos. Em geral,

[7] *Random forests*, que pode ser traduzido por florestas aleatórias, são algoritmos que combinam técnicas de amostragem e árvores de decisão (HAN *et al*., 2011) para obter modelos de classificação com maior acurácia.

essas aplicações podem ser utilizadas para gestão estratégica de propriedades rurais. Por exemplo, Denobile (2005) desenvolveu um modelo de gestão estratégica, com o foco no cliente, para a comercialização de produtos orgânicos por meio de venda direta. Por meio de regras de associação e árvores de decisão foi elaborado um modelo de segmentação de clientes. Esse modelo foi utilizado também para identificar hábitos de compras dos clientes da propriedade.

Ainda relacionado à gestão de processos e produção agrícola, em Mucherino *et al*. (2009), foi utilizada a técnica de clusterização para prever a qualidade do processo de fermentação de vinhos. Para monitorar o processo de fermentação, mais de 22 mil observações foram selecionadas. Em seguida, o algoritmo *k-means* foi aplicado sobre estas amostras com o propósito de agrupá-las em clusters.

Além das aplicações já mencionadas, técnicas de mineração de dados também podem ser utilizadas para criar modelos de predição na agricultura. As técnicas de mineração de dados podem ser usadas, por exemplo, na construção de sistemas de alerta de doenças. Meira (2008) desenvolveu um sistema de alerta de doenças em culturas agrícolas, com aplicações na ferrugem do cafeeiro. Apesar de sistemas de alerta de doenças permitirem racionalizar o uso de agrotóxicos, são pouco utilizados na prática devido à complexidade dos modelos, dificuldade de obtenção dos dados e custos para o agricultor. Entretanto, por meio de estações meteorológicas automáticas, banco de dados e monitoramento agrometeorológico na web, foram gerados dados com os quais foi desenvolvido o sistema. Nesse sistema foram utilizadas árvores de decisão.

Sistemas de informações agrícolas também começam a receber atenção da comunidade científica. A Agência de Informação Embrapa (BERTIN *et al*., 2009) e também outros sistemas de informação (PARIK, 2007) são exemplos de soluções amplamente utilizadas para oferta de informações técnicas e para o processo de tomada de decisões. Nesses sistemas, com o aumento cada vez maior do volume de informação, técnicas de mineração de dados podem ser usadas para filtragem da informação.

2.3.2 Mineração de dados na web

A quantidade de dados armazenados em arquivos de computador e bases de dados tem crescido exponencialmente ao longo dos anos. Com o crescimento explosivo da oferta de dados na *World Wide Web* faz-se necessário que os usuários utilizem ferramentas automáticas para a obtenção da informação desejada. Com o objetivo de criar ferramentas capazes de auxiliar o usuário nessa tarefa,

pode-se aplicar técnicas de mineração de dados. Mineração de dados na web pode ser classificada em três tipos: *web structure mining*, *web content mining* e *web usage mining* (SINGH e SINGH, 2010).

Web mining (WM) não tem uma definição precisa, mas em termos gerais podemos dizer que WM é a aplicação de técnicas de mineração de dados em dados da web com o objetivo de obter conhecimento potencialmente útil. As fontes desses dados podem ser arquivos de log[8], a estrutura de *hiperlinks*[9] ou mesmo o próprio conteúdo das páginas (SINGH e SINGH, 2010).

O *web content mining* (WCM) trata da descoberta de informações a partir do conteúdo das páginas. Pode ser pensado como uma extensão da tarefa realizada pelos mecanismos de busca da internet (DUHAN *et al*., 2009). O *web usage mining* (WSM) trata da descoberta da estrutura de ligações entre as páginas em contraste com WCM que trata do conteúdo das páginas. Por fim *web usage mining* (WUM) é utilizado para descobrir os padrões de navegação do usuário por meio dos arquivos de log nos servidores, obtidos no processo de interação do usuário com o site (DUHAN *et al*., 2009). A taxonomia considerada está representada na Figura 5.

As tarefas de mineração web, segundo Pierrakos *et al*. (2003), têm sido utilizadas para solucionar diversos problemas como:

- Encontrar informação relevante;
- Criar conhecimentos a partir da web;
- Personalização da informação;
- Aprendizado sobre os consumidores.

[8] Arquivos de log são arquivos de texto onde aplicativos podem armazenar várias informações, como por exemplo informações do tráfego de um servidor.

[9] *Hiperlinks* são referências à outras páginas contidas em uma página web. Sites complexos compreendem dezenas, centenas e até milhares de páginas interligadas por *hyperlinks.*

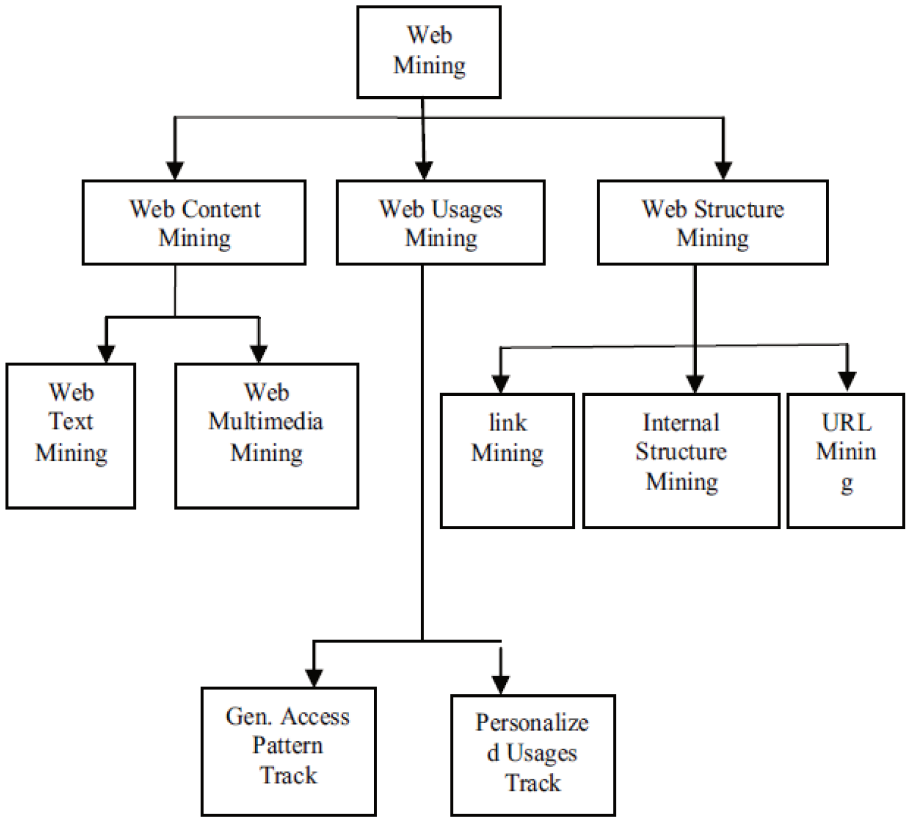


Figura 5: Taxonomia do Web Mining (SINGH e SINGH, 2010)

As fontes de dados para a realização do WUM são:

- *Web Server Logs*: contém os logs das requisições das páginas. Geralmente contém o IP do usuário, tempo, data, página requisitada, código HTTP dentre outras informações. Estas informações podem ser armazenadas em um único arquivo ou separadas em logs de acesso, logs de erro, e *referrer logs*[10]. Em geral, somente os administradores dos sites têm acesso a essas informações.

- *Proxy Server Logs:* um *web proxy* é um processo de armazenamento de informações do tráfego entre o *browser*[11] e o servidor. Ele diminui a quantidade de informação redundante que trafega na rede, servindo também como fonte de dados para descoberta de padrões.

[10] São registros extras nos logs de acesso que informam o site anterior do visitante de uma página.
[11] São aplicativos com a função de processamento e visualização de páginas web. Também conhecidos como navegadores de internet. São exemplos: Internet Explorer, Google Chrome e Mozilla Firefox.

- *Browser Logs:* os próprios *browsers* como o Internet Explorer, Mozilla ou Chrome podem ser alterados, ou mesmo *scripts* rodando no lado do cliente podem ser utilizados para reunir informações dos usuários. Esses logs são outra fonte de dados.

As tarefas de *web mining* envolvem três fases: pré-processamento, descoberta de padrões e análise dos padrões (THANGAMANI e THANGARAH, 2011). O estágio inicial do WUM é o pré-processamento. É nessa fase que dados irrelevantes e incompletos são retirados dos logs. O arquivo de saída do pré-processamento pode ser utilizado para a descoberta de padrões como: caminhos de navegação, regras de associação, mineração de dados sequenciais e etc. Por fim, tem-se a análise dos padrões identificados que pode ser feita com a ajuda de ferramentas gráficas.

2.3.3 Sistemas de recomendação com regras de associação

Em um ambiente web, usualmente regras de associação podem ser utilizadas para definir ocorrências associadas de páginas em conjuntos de sessões de usuário. As sessões atuam como as "cestas de compras" tal que por meio da mineração de regras de associação é possível determinar a relação entre páginas vistas frequentemente juntas (KAZIENKO, 2009).

Regras de associação que revelam similaridades entre páginas web, derivadas dos comportamentos de uso de um site, podem ser utilizadas como listas de recomendação (ADOMAVICIUS e TUZHILIN, 2001; MOBASHER *et al*., 2000; NAKAGAWA e MOBASHER, 2003; YANG e PARTHASARATHY, 2003).

Tipicamente sistemas de recomendação utilizam a técnica de regras de associação para descobrir relações entre páginas a partir dos históricos de navegação. Geralmente esse histórico é retirado dos arquivos de *log* do servidor, transformado em um conjunto de sessões de usuário, e a partir destas sessões são extraídas regras de associação entre as páginas. Existem trabalhos onde as regras de associação tradicionais são acompanhadas dos padrões sequenciais (Nakagawa e Mobasher, 2003).

Estas listas de recomendação podem fornecer padrões de uso do site que estendem a estrutura de *hyperlinks* já presente, ou mesmo atuar como um mecanismo de filtragem e ranqueamento dos *hyperlinks* mais importantes (KAZIENKO, 2009).

A avaliação de sistemas de recomendação baseados em regras de associação pode ser feita do ponto de vista da técnica de recomendação, ou sob o ponto de vista da usabilidade do site. Em

Jorge *et al*., (2002), uma proposta de sistema de recomendação, muito semelhante a este trabalho, foi avaliada do ponto de vista do algoritmo enquanto nesse trabalho buscou-se avaliar o sistema de recomendação do ponto de vista do usuário, como em Putman (2010).

3. MATERIAL E MÉTODOS

A abordagem desta pesquisa baseia-se na aplicação de algoritmos de mineração de dados nos registros de uso do portal Agência Embrapa - cana-de-açúcar, armazenados digitalmente em um banco de dados.

Com o objetivo de fornecer recomendações de leitura para os usuários, os padrões emergentes de uso da Agência cana-de-açúcar foram extraídos da base de dados. A técnica de modelagem escolhida foi a geração de regras de associação, por meio do algoritmo Apriori, captando assim o perfil de acessos da comunidade. Como o portal tem um volume de tráfego elevado, as regras de associação foram geradas em períodos de menor acesso no servidor e foram armazenadas no banco de dados para futuras recomendações.

Para dar suporte aos procedimentos realizados neste trabalho, optou-se por seguir um modelo de processo conhecido como 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 prover uma ferramenta de suporte rápida, robusta e barata para definir os procedimentos adotados na fase de mineração de dados (CHAMPMAN *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. Nas seções seguintes deste capítulo, são apresentadas estas fases e seus significados.

3.1 Contexto

Segundo Chapman *et al*. (2000), antes de definir as fases do ciclo de vida de um projeto de mineração de dados, deve-se definir o contexto do processo. O contexto é uma avaliação preliminar, mas uma análise detalhada do problema será apresentada posteriormente na descrição das fases da metodologia CRISP-DM. As quatro dimensões do contexto são:

- **Domínio da aplicação:** a área de estudo é a de sistemas de recomendação de páginas

Web para conteúdos de cana-de-açúcar.

- **Tipo do problema de mineração de dados:** o problema envolve a geração de regras de associação.

- **Aspectos Técnicos:** devido à estrutura dos dados que estão armazenados em um banco de dados relacional, deve-se efetuar transformações em sua estrutura. As transformações utilizadas são descritas, em detalhe, na fase de preparação dos dados na Seção 3.4. Após a geração das regras, estas devem ser armazenadas em uma estrutura apropriada para a consulta no banco de dados da agência.

- **Ferramentas e Técnicas:** Nesse trabalho os softwares utilizados foram:

  - Sistema operacional Linux para instalação dos aplicativos utilizados.

  - Sistema gerenciador de banco de dados PostgreSQL 8.4 (http://www.postgresql.org.br/) para armazenamento e consulta dos dados.

  - Softwares R (http://www.r-project.org/) para execução dos algoritmos apropriados à mineração de dados.

Todas as ferramentas utilizadas nesse trabalho são *open source*[12] e não apresentam nenhum custo para sua utilização. Esses softwares foram escolhidos em virtude de sua potencialidade, escalabilidade, ampla utilização no meio acadêmico e custo nulo.

3.2 Compreensão do domínio

A fase de compreensão do domínio resultou na elaboração do Capítulo 2, onde foi feita uma

[12] *Open source* é o termo em inglês para código aberto, criado pela OSI *(Open Source Initiative)* e refere-se a *software* também conhecido por software livre.

revisão da literatura. Conforme discutido na introdução, o objetivo principal desse trabalho foi a construção, implementação e implantação de um sistema de recomendação. Assim, realizou-se uma pesquisa em busca de conhecimentos atualizados sobre a oferta de informações tecnológicas relacionadas à cultura da cana-de-açúcar, sistemas de recomendação e mineração de dados.

3.3 Entendimento dos dados

3.3.1 Coleção de dados e descrição

Nesta fase do trabalho, foi realizado um levantamento dos dados para avaliar possíveis problemas de qualidade e determinar as informações necessárias para alcançar os objetivos traçados. Assim, o conjunto de dados selecionado foi o conjunto de acessos às páginas da Agência de Informação Embrapa – cana-de-açúcar.

O conjunto de dados estava armazenado em um banco de dados, resultante das informações de acesso ao portal Agência de Informação Embrapa. Os acessos a cada página, vídeo e outros materiais foram automaticamente transferidos para um formato estruturado em um banco de dados relacional[13], nesse caso o PostgreSQL 8.4.

O banco está estruturado em duas tabelas: a tabela *clientes* e a tabela *tracker*. Na tabela *clientes* estavam armazenadas informações relativas a cada usuário que entrou na Agência e iniciou uma sessão, isto é, requisitou ao servidor o 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.

Para determinação do fim de uma sessão, o sistema utiliza uma heurística baseada no seu tempo de duração: sempre que um usuário requisita uma página e cessa a atividade, após trinta minutos a sessão é encerrada. Dessa forma um mesmo usuário que faz duas requisições separadas no tempo em mais de trinta minutos será tratado como dois usuários diferentes. Ainda assim, neste trabalho cada linha da tabela clientes foi considerada como um único usuário uma vez que a heurística nos garante que a grande maioria dos usuários de fato não volta ao sistema após um tempo mínimo (BAGLIONI *et al*., 2003).

[13] Um banco de dados relacional é um conceito abstrato que define maneiras de armazenar, manipular e recuperar dados estruturados unicamente na forma de tabelas (também conhecidas como relações).

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.

O resumo dos atributos das tabelas do banco de dados, *clientes* e *tracker,* estão disponíveis, respectivamente, nas Tabelas 5 e 6.

Tabela 5 – Descrição e exemplos dos atributos contidos na tabela *clientes*.

Atributo	Descrição	Exemplo
idsessao	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
pais_cliente	Pais de origem do acesso	Brazil

A tabela *clientes* possui 2.574.763 linhas, que representam o número de usuários distintos que acessaram conteúdos da cultura da cana-de-açúcar, de acordo com a heurística utilizada, 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 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.

Tabela 6 – Descrição e exemplos de atributos contidos na tabela *tracker*.

Atributo	Descrição	Exemplo
idtracker	Identificador com 32 caracteres da sessão	625bf2356fcb803f8e288de486e9210b
idsessão	O mesmo da tabela clientes	----------
local_atual	Página requisitada	Abertura.html
arvore_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
time_stamp	O mesmo da tabela *clientes*	----------

3.3.2 Exploração dos dados

Na fase de exploração dos dados foi realizada a análise descritiva. Com essa análise, buscou-se uma visão geral dos dados que fornecesse características gerais de uso do site: páginas mais visitadas, médias de páginas acessadas por sessão de usuário, árvores de conhecimento mais vistas, tecnologias utilizadas, distribuição espacial dos acessos e estatísticas que fornecessem informações sobre perfis de uso. Foi utilizado o software R para realizar essas análises.

Utilizou-se gráficos de barras, tabelas e porcentagens para mostrar as principais características da Agência Embrapa – cana-de-açúcar e principalmente o perfil e a interação dos usuários com a Agência. Com a informação referente à distribuição de acessos às páginas por sessão, procurou-se mostrar que a escolha da cana-de-açúcar foi a melhor opção para receber o sistema de recomendação. Também foram utilizados gráficos para mostrar a distribuição dos acessos por estado e foi feita a distribuição espacial dos acessos em um mapa do território brasileiro. Todas os resultados da análise exploratória são apresentados no Capítulo 4, seção 4.1.

3.4 Preparação dos dados

Após as fases de compreensão do domínio e entendimento dos dados, prossegui-se para a fase de preparação dos dados. O objetivo principal dessa fase foi a seleção, seguida da preparação dos dados armazenados no sistema de gerenciamento de banco de dados, para um formato adequado à aplicação do algoritmo Apriori (AGRAWAL *et al*., 1993), utilizando a linguagem R.

Os dados disponíveis nas tabelas *clientes* e *tracker* no banco de dados já haviam sido tratados, por isso não houve a necessidade de aplicar procedimentos para limpeza dos dados, identificação de usuários, identificação de sessões e substituição de valores faltantes.

Foi feita a transformação dos dados do banco de dados para uma estrutura de transações, de forma que fosse possível determinar regras por meio do algoritmo Apriori. A forma de armazenamento no banco de dados é ilustrada na Tabela 7, onde cada linha representa um *page view*[14]. Mas é necessária uma estrutura de dados do tipo "cesta de compras" (Tabela 8) como entrada de dados do algoritmo Apriori. Como o sistema de recomendação foi construído como um *script* em R, esta etapa foi realizada como parte da programação do *script*. Para tanto, foi utilizada a função

[14] Um *page view* é o registro de um acesso a uma página específica em um momento do tempo. Uma sessão de usuário é composta de vários *page views.*

read.transactions presente no pacote arules (HAHSLER, 2005).

Como o pacote arules, parte do pacote estatístico R, já possuía o recurso para transformar a estrutura de dados das tabelas do banco para uma estrutura de transações, não foi necessário implementar um *script* para essa tarefa. Assim o motor do sistema de recomendação recebe como entrada os dados da tabela *tracker* e transpõe para a estrutura transações, automaticamente dentro do ambiente R.

Tabela 7 - Exemplo de estrutura de dados presente na tabela *tracker*.

ID	Item (página)	Data do Acesso
001	Praga no Colmo	15/06/2011
001	Praga nas Raízes	15/06/2011
007	Produção	15/06/2011
006	Produção	15/06/2011
004	Praga no Colmo	15/06/2011
004	Praga as Raízes	15/06/2011
004	Produção	15/06/2011

Tabela 8 - Estrutura de dados de acessos separada em sessões de usuário.

ID	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 da composição da base de conhecimento, foi necessário editar a saída do algoritmo Apriori, incluir os títulos das páginas e o link completo das páginas a serem recomendadas. Para tal, foi criada uma tabela no banco de dados, chamada *regras*, onde foram armazenados o título da página antecedente, o link da página antecedente, o título da página consequente, o link da página consequente e o valor da métrica MaxConf (WU *et al*., 2010), definida como segue:

$$MaxConf(A, B) = \text{máximo}\left\{\frac{suporte(A \cup B)}{suporte(A)}, \frac{suporte(A \cup B)}{suporte(B)}\right\} \quad (4)$$

onde o suporte(A *U* B) foi definido da fórmula (2).

Na Tabela 9 é apresentado um exemplo de estrutura dos dados contidos na tabela *regras* do banco de dados.

Para a análise das regras de associação, também foi necessária uma etapa de preparação. A

Agência Embrapa foi projetada de forma que em cada arquivo HTML, as páginas web pudessem ser identificadas univocamente. Por exemplo, o arquivo em HTML referente à página cachaça (CONT**000fiog1ob502wyiv80z4s473agi63ul**.html), onde o "CONT" é um prefixo que indica que se trata de uma página relacionada a algum assunto da Agência Embrapa. O código alfanumérico "**000fiog1ob502wyiv80z4s473agi63ul**" é um identificador único da página. Páginas que oferecem recursos eletrônicos, como arquivos em pdf e vídeos, são nomeadas como "**REC**000fjd7d39l02wyiv809gkz51lf7zyjd.html" onde o prefixo é "REC" ao invés de "CONT".

Tabela 9 - Exemplo de regras armazenadas na tabela *regras* para recomendações.

Título Antecedente	Título Consequente	Antecedente	Consequente	MaxConf
Cana-de-açúcar	Adubação	AG01_453_217200392420.html	AG01_459_217200392421.html	0.86
Pré-produção	Socioeconomia	AG01_9_41020068054.html	AG01_12_41020068054.html	0.85
Extração	Moendas	CONTAG01_103_22122006154 841.html	REC000fxourdm502wyiv8018w i9tn1mhg5i.html	0.83

Foram desenvolvidas algumas funções na linguagem R para esta etapa da preparação de dados, com o propósito de extrair os identificadores de cada página, extrair os links, recuperar os nomes das páginas e calcular algumas métricas, como segue:

- **pegaRegras.R**: recupera as regras de associação armazenadas no banco de dados do sistema de recomendação.
- **extractLink.R**: extrai todos os links de uma página e retorna uma lista com o identificador da página e os links associados.
- **limpaUma.R**: retira informações irrelevantes dos links retornando somente o nome do arquivo de página HTML da Agência Embrapa.
- **limpaRegras.R**: retira informações irrelevantes dos links de várias páginas retornando uma lista com somente os nomes dos arquivos em HTML.
- **checkLink.R:** esta função recebe como entrada um conjunto de regras de associação. Sempre que a página no consequente da regra de associação for uma página para a qual já existe um link, uma flag "TRUE" é adicionada à regra, e "FALSE" caso contrário.
- **pegaSessoes.R:** recebe como entrada os dados da tabela *tracker* e duas datas definidas pelo usuário da função. A partir destas informações a função retorna uma lista com todas as sessões de usuário e as páginas vistas nestas sessões.
- **leTransac.R:** recebe um conjunto de sessões de usuário e retorna uma tabela com o

identificador de sessão, a primeira página vista na sessão e o número de páginas vistas na sessão.

3.5 Modelagem

A partir do conjunto final de dados, já tratados, foram determinadas as regras de associação mais relevantes entre as páginas de conteúdo da Agência de cana-de-açúcar, de forma a oferecer recomendações de conteúdo, baseadas no perfil da comunidade de usuários. Cada regra de associação relacionava somente duas páginas, o antecedente e o consequente, tal que 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.

Após a fase de transformação dos dados, um conjunto de transações foi gerado. Formalmente, seja t_m uma transação associada ao usuário *m*, onde $t_m = \{p_1, p_2 ... p_n\}$ representa o conjunto de *n* páginas visitadas pelo usuário *m*, em uma sessão. O *script* R extrai regras de associação, tendo como entradas o conjunto de transações dos usuários ($t_1 t_2 ... t_m$). Assim, como resultado, as regras geradas são da forma $p_i \rightarrow p_j$, que indicam padrões de acesso dos usuários entre estas páginas *i* e *j*.

Em algumas aplicações, a presença de uma página em uma transação pode ser ponderada pelo tempo que o usuário esteve na página. Assim, páginas que os usuários visualizaram por mais tempo recebem os maiores pesos, uma vez que despertaram maior interesse. Nessa análise optou-se por não atribuir pesos relacionados ao tempo; assim cada página em uma transação teve o mesmo peso.

O algoritmo Apriori inicialmente encontra os grupos de itens ocorrendo frequentemente juntos em transações. Nesse contexto, esses grupos são páginas visitadas, juntas, em uma mesma sessão. Esses conjuntos de itens são denominados conjuntos frequentes de itens. Conjuntos frequentes de itens 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.

No domínio de sistemas web, páginas que estão em um nível mais aprofundado da estrutura de links tendem a ser menos visitadas. Dessa forma, existem abordagens na literatura onde foram implementados algoritmos, em que o usuário pode variar o valor do suporte durante o processo de descoberta das regras (LIU *et al*., 1999).

No contexto desta pesquisa, entretanto, foi utilizado suporte fixo. O suporte foi baixo o suficiente (sup = 0.0005), tal que padrões de páginas com poucos acessos fossem encontrados. Após a geração das regras, estas foram armazenadas em um arquivo csv[15] para posterior consulta e análise. Essas regras foram ordenadas pela confiança e armazenadas no banco de dados da Agência de Informação Embrapa.

O algoritmo Apriori tradicional utiliza duas informações para a inferência de regras de associação: suporte e confiança. Contudo, nem sempre essas métricas são suficientes para determinar se padrões encontrados nos dados são significativos. Por essa razão, uma métrica com propriedades de invariância em relação ao número de sessões deve ser utilizada nesse tipo de problema, como por exemplo, Cosine, Kulc, dentre outras (WU *et al*., 2010). Essa propriedade é desejável, pois o acréscimo de visitas não invalida padrões de uso já encontrados.

As regras ranqueadas, utilizando a métrica MaxConf (WU *et al*., 2010), foram então armazenadas em uma tabela no banco de dados, a tabela *regras*. Sempre que um usuário acessa uma página A, tal que exista uma regra no banco de dados (A → B) o sistema recomenda a visualização de B.

Por fim, como muitas páginas da Agência cana-de-açúcar não apresentavam nenhuma regra de recomendação, foi gerada uma lista das top 3 páginas mais vistas a partir de cada página da Agência cana-de-açúcar. Essa lista foi utilizada como forma alternativa de recomendação sempre que não havia regras de associação para algumas páginas da Agência cana-de-açúcar.

3.6 Avaliação

Nesta fase foi realizada uma avaliação do processo, de forma a construir um modelo que tivesse credibilidade. Aqui é importante avaliar detalhadamente o modelo, e rever os passos executados na sua construção para garantir que os objetivos propostos serão alcançados. Com esse fim, a base de conhecimento foi avaliada para verificar se esta trouxe novas ligações entre as páginas por meio das regras. Também foi avaliado o potencial de resumo e indicação dos links principais de uma página por meio da base de conhecimento.

Também foi avaliada a usabilidade do sistema de recomendação. Neste ponto, é importante verificar o impacto das informações para o usuário. Considerando que na última década foram

[15] Formato de arquivo de texto que contém dados, em cada linha, separados por um caractere de separação (em geral, uma vírgula ou um ponto e vírgula).

desenvolvidas várias abordagens para sistemas de recomendação, a escolha da abordagem apropriada pode ser baseada na comparação da performance de diferentes técnicas ou na verificação do impacto do sistema de recomendação para os usuários.

Três abordagens são apresentadas na literatura para avaliação de sistemas de recomendação: experimentos offline, experimentos on-line e estudo dos usuários. Em experimentos offline, por meio de dados simulados, ou mesmos dados anteriores gerados com o uso de um portal, a eficácia do sistema é avaliada. Em estudos on-line, a eficácia do sistema é avaliada em relação aos dados reais dos acessos de usuários e em tempo real. Por fim, em estudos com os usuários, uma amostra é selecionada para utilizar o sistema e reportar em um questionário suas experiências. De todas essas abordagens, somente na primeira não há a necessidade da interação com usuários reais (RICCI *et al*., 2011).

A abordagem offline foi descartada porque não houve comparação de algoritmos. Assim na validação, foi utilizada a avaliação do comportamento dos usuários, por meio de estudos on-line, verificando o impacto das recomendações nos padrões de uso registrados no banco de dados.

Essa avaliação com dados on-line foi feita utilizando uma métrica conhecida na literatura como *bounce rate*, um 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 esta página e abandonaram o site logo em seguida, sem visualizar mais nenhuma outra página. De acordo com Sculley *et al*. (2009), 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[16] 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.

Para o teste Z, sob a hipótese de uma distribuição Binomial, utilizando a aproximação da distribuição Normal, a estatística de teste pode ser definida como:

[16] Um teste paramétrico é um teste estatístico que assume que os dados têm uma certa distribuição de probabilidade. Aqui a distribuição é Normal.

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{pq\left(\frac{1}{m} + \frac{1}{n}\right)}} \sim Normal(0,1) \tag{4}$$

assim, quando $z > z_\alpha$ onde α é o nível de significância do teste, a hipótese nula é rejeitada.

Para o teste qui-quadrado, a estatística do teste é dada por:

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

onde χ^2_{k-1} é uma variável aleatória com distribuição qui-quadrado com *k-1* graus de liberdade, *n* é a contagem de usuários na categoria *i*, p_i a proporção de usuários na categoria *i* e $n \times \hat{p}_i$ são os valores esperados das proporções.

Especificamente no caso de hipótese alternativa ser H_1: $p_1 \neq p_2$, ambos 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.

Também foi feita uma avaliação com 24 participantes, por meio de um questionário de avaliação de sistemas de recomendação (PU *et al.*, 2011). Cada pergunta do questionário foi respondida de acordo com a escala de Likert (BOONE e BOONE, 2012). Em uma questão na escala de Likert, o respondente recebe uma afirmação e deve responder se concorda com ela de acordo com uma de cinco categorias:

1. Discordo totalmente.
2. Discordo.
3. Neutro.
4. Concordo.
5. Concordo totalmente.

Foi feita uma análise descritiva dos dados para os 24 respondentes, mas não foi possível realizar testes estatísticos com relação à diferença entre os grupos de usuários especialistas e não-especialistas, devido ao número reduzido de observações.

3.7 Distribuição

Na última fase do processo, com o modelo já construído, a distribuição ocorreu com a implantação do sistema de recomendação, tal que as recomendações geradas com a modelagem dos dados foram oferecidas aos usuários.

As recomendações foram oferecidas na forma de links. Para a apresentação dos links, foram utilizadas a linguagens PHP para a programação do servidor e a linguagem Javascript para a programação no navegador.

Sempre que um usuário acessa uma das páginas da agência cana-de-açúcar, um *script* desenvolvido em Javascript, rodando no próprio navegador do usuário, dispara um evento de requisição de informação ao servidor, ao final do carregamento da página. Assim, um *script* em PHP do lado do servidor executa as consultas e retorna, para a página, as informações dos links e dos títulos das recomendações daquela página. Assim, os links podem ser visualizados pelo usuário como ilustrado na Figura 6.

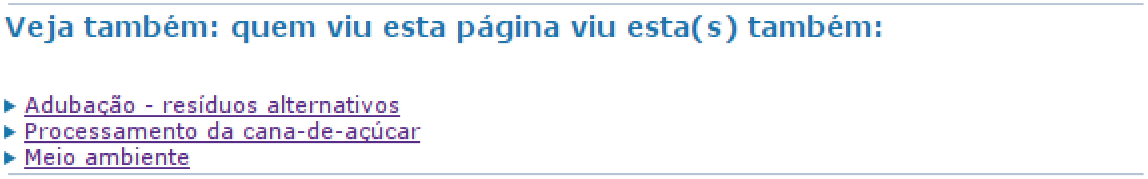


Figura 6: Exemplo da interface para o acesso aos links nas páginas de cana-de-açúcar.

3.8 Softwares utilizados

Os principais softwares utilizados para a criação e implementação do sistema de recomendação foram o pacote estatístico R (versão 2.15.2) e o sistema de gerenciamento de banco de dados PostgresSQL, versão 8.4. A etapa de preparação de dados foi realizada também com recursos de programação do pacote estatístico R e, a etapa de distribuição, foi realizada utilizando-se as tecnologias Javascript e PHP. A plataforma na qual esse trabalho foi desenvolvido e implantado foi o Ubuntu, uma das distribuições do sistema operacional Linux.

O pacote estatístico R é apontado como o principal software utilizado para mineração de

dados em uma pesquisa, mostrada na Figura 7. Por seus recursos e facilidade para programação de *scripts* e funções, foi extensamente utilizado para a etapa de preparação, pois houve a necessidade de manipular strings, dados numéricos, datas e outros formatos; alterar estruturas de dados na etapa de modelagem; armazenamento dos resultados de análises e comunicação com o sistema de gerenciamento de banco de dados.

Na fase de entendimento dos dados, a linguagem R foi utilizada para geração de gráficos e tabelas. Além das funções do pacote base do R, foram necessários outros pacotes como o "plyr" (HADLEY, 2011) para manipulação e preparação dos dados, "reshape" (HADLEY, 2007) para preparação de dados, "rworldmap" (SOUTH, 2011) para criação de mapas, e o pacote "arules" (HAHSLER, 2005) para geração de regras de associação.

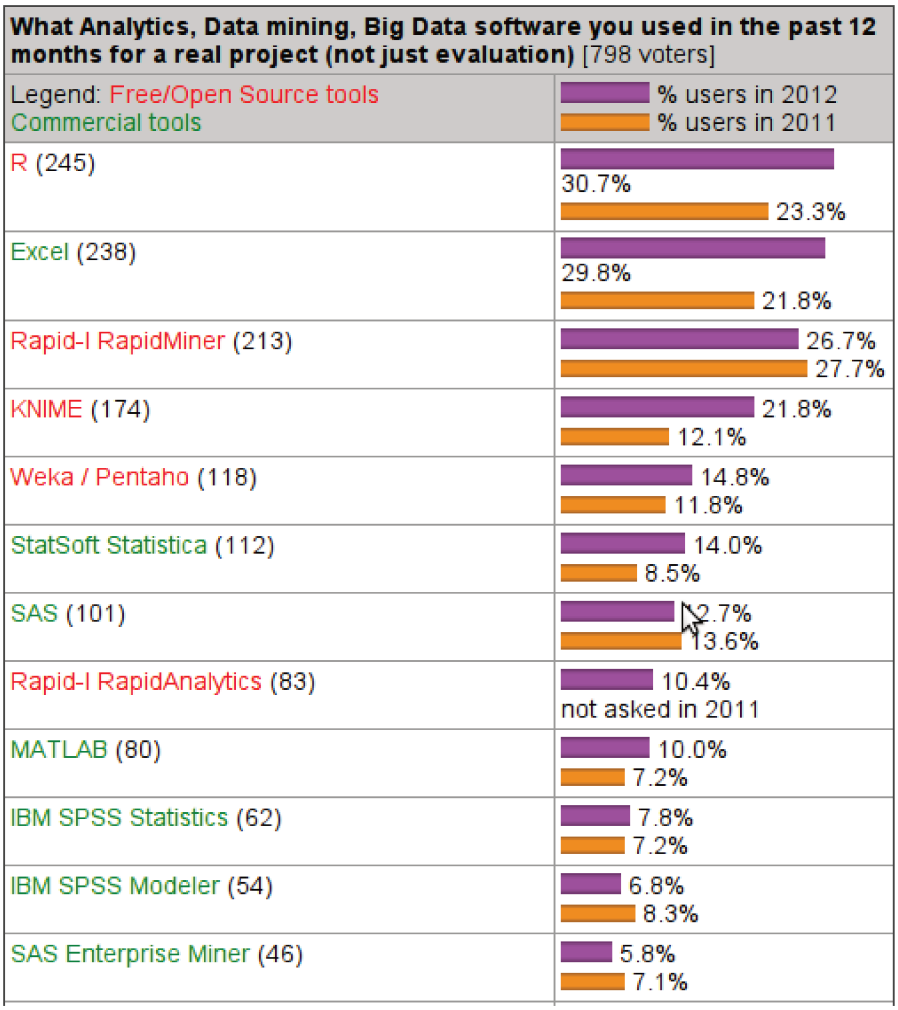


Figura 7: Levantamento publicado em maio de 2012 que apresenta os doze softwares mais utilizados em 2011 e 2012, com 798 participantes.
Fonte: http://www.kdnuggets.com/polls/2012/analytics-data-mining-big-data-software.html

4. RESULTADOS E DISCUSSÃO

O principal resultado desse projeto foi a construção e implantação de um sistema de recomendação para a Agência de Informação Embrapa, especificamente para o portal da cana-de-açúcar, baseado no perfil de acessos dos usuários, tendo como base os dados gerados no acesso às páginas do portal. A base de conhecimento gerada, na forma de regras de associação entre as páginas e o processo de validação também foram resultados importantes.

Neste capítulo, são apresentados resultados da análise exploratória dos dados de uso da Agência Embrapa, a arquitetura do sistema de recomendação, bem como a análise da base de conhecimento gerada na forma de regras de associação e, por fim, os resultados referente ao processo de validação do sistema de recomendação.

4.1 Análise exploratória

O conteúdo do portal Agência de Informação Embrapa está estruturado em árvores de conhecimento. Assim, durante o registro dos acessos, a página vista e a árvore a qual a página pertence são armazenadas no banco de dados. A distribuição dos acessos, de acordo com as árvores de conhecimento, é apresentada na Figura 8.

De acordo com a Figura 8, pode-se ver que os acessos à árvore do conhecimento da cana-de-açúcar representam quase metade do total de acessos à Agência Embrapa, seguida pela árvore do agronegócio do leite. Esses dados reforçam a motivação deste trabalho, já que a árvore da cana-de-açúcar é a mais procurada, no âmbito da Agência de Informação Embrapa.

No banco de dados de acessos também são registradas as localizações geográficas das origens dos acessos, tal que é possível determinar o País, Estado e o Município desses acessos. Como pode ser visto na Figura 9, o Estado de São Paulo é a origem do maior volume de acessos, com quase um terço do total.

Existe também um número considerável de sessões onde não foi possível identificar a localidade dos acessos. Mas apesar de mais de um quarto das sessões não terem origem geográfica identificada, o montante de acessos à cana-de-açúcar vindos de São Paulo pode ser ainda maior que o efetivamente identificado.

Por meio da Figura 10, utilizando as coordenadas geográficas de latitude e longitude das origens dos acessos, 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 Zona da Mata no Nordeste são responsáveis pelo grande volume de acessos no país. Essas regiões coincidem com grandes áreas produtoras de cana-de-açúcar.

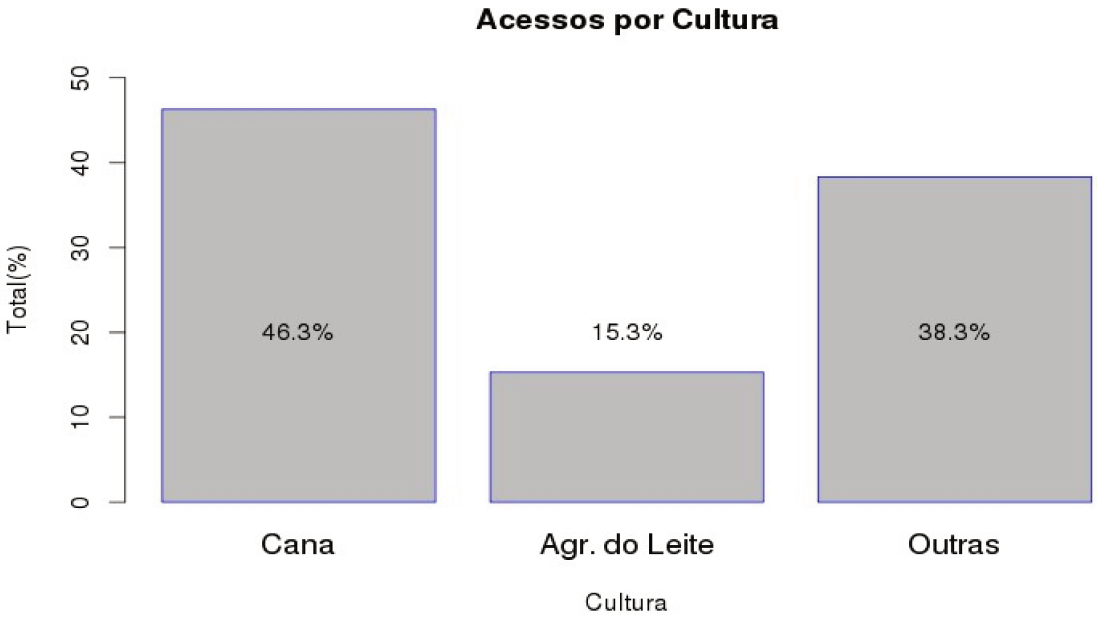


Figura 8: Distribuição do número de sessões de usuário para as árvores de conhecimento das culturas de cana-de-açúcar, agronegócio do leite e todas as outras aglomeradas.

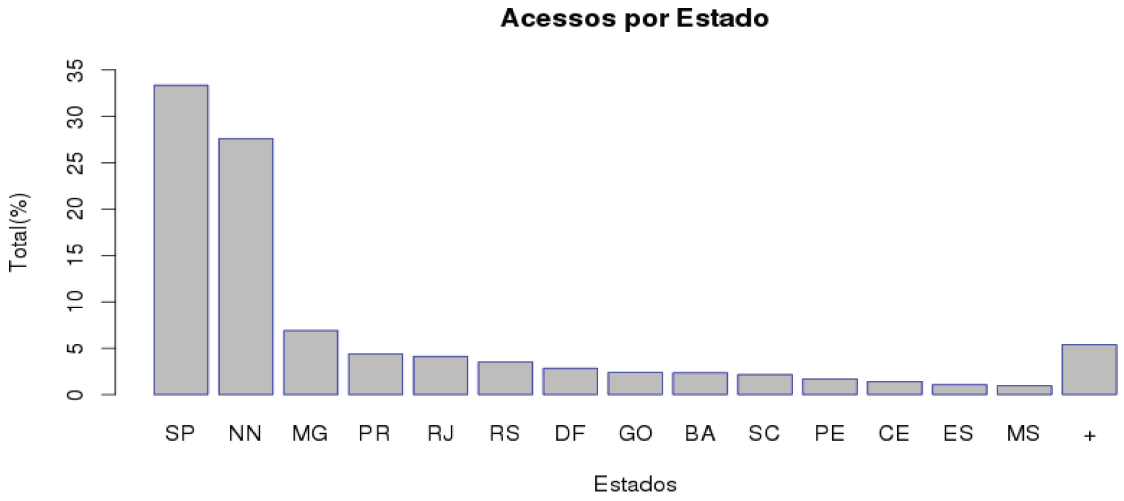


Figura 9: Distribuição do total de acessos à árvore de conhecimento de cana-de-açúcar por estado. A sigla NN indica a porcentagem de sessões de localização não identificada e o + indica o acumulado dos outros estados.

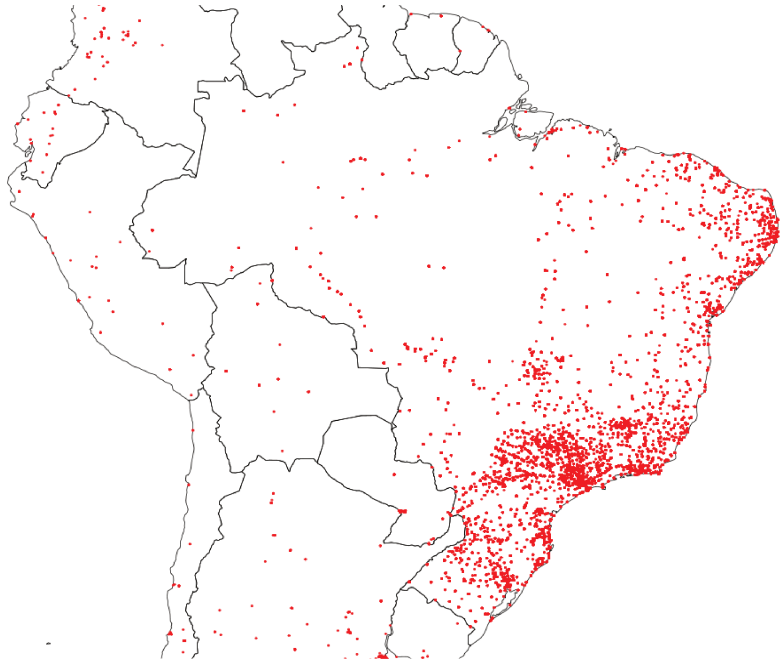


Figura 10: Distribuição espacial dos acessos às páginas da árvore de cana-de-açúcar.

Com relação às tecnologias utilizadas, a grande maioria dos usuários utiliza alguma versão do Windows junto ao navegador Internet Explorer. Nas Figuras 11 e 12, vê-se que o volume de usuários de outros sistemas operacionais é muito menor que os do Windows[17].

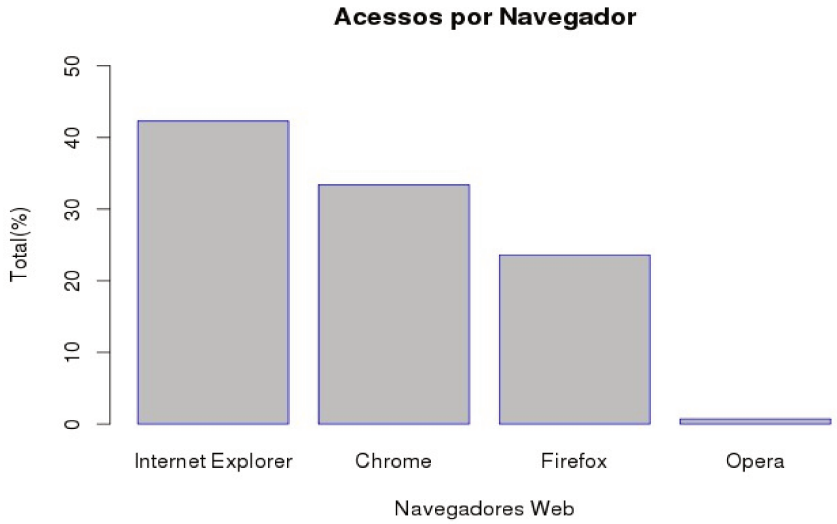


Figura 11: Distribuição dos acessos em relação ao uso dos navegadores.

[17] O Windows é um sistema operacional para computadores pessoais, desenvolvido pela empresa Microsoft.

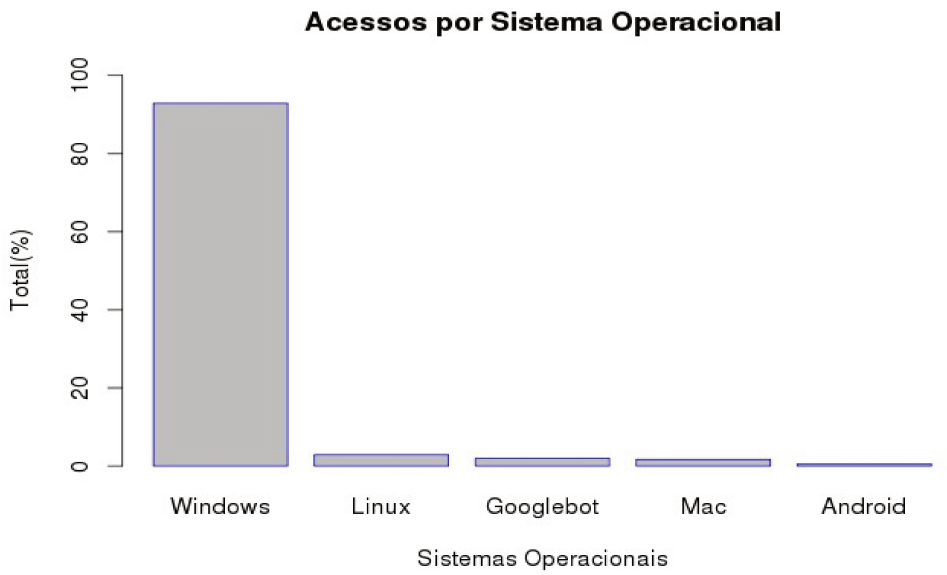


Figura 12: Distribuição dos acessos com relação ao sistema operacional utilizado.

De acordo com as estatísticas do padrão mundial de uso da internet, a maioria dos usuários de computador realmente utiliza alguma versão do Windows, mas o navegador mais utilizado é o Google Chrome (NET APPLICATIONS, 2012). Na Agência Embrapa, essa estatística, com maior relevância do navegador Internet Explorer, que é o navegador padrão do sistema operacional Windows, pode indicar que a maioria dos usuários da Agência Embrapa não esteja alinhada com o uso de tecnologias quanto usuários de outros tipos de portais, como portais de tecnologia. Em outras palavras, boa parte dos acessos é feita por profissionais que não são especialistas em tecnologia da informação. Por esta razão, o sistema de recomendação torna-se essencial para auxiliar esses profissionais no processo de navegação e busca de conteúdo.

Como as páginas de cana-de-açúcar representam mais da metade dos acessos da Agência Embrapa, procurou-se determinar quais as páginas mais acessadas da agência cana-de-açúcar. A Tabela 10 apresentada as seis páginas mais acessadas da agência cana-de-açúcar.

Tabela 10 – Contagem de visitas nas top 6 mais visitadas.

Páginas	Ocorrências	%
Processamento da cana-de-açúcar	117383	8,17
Plantio	58902	4,10
Cachaça	58303	4,06
Fabricação do açúcar	52837	3,68
Fermentação	48408	3,37
Geração de energia elétrica	41210	2,87

Das páginas mais vistas, pode-se perceber que a maioria está relacionada aos produtos feitos a partir da cana-de-açúcar. Somente uma página é relacionada ao plantio. Só a página relativa ao processamento da cana-de-açúcar apresenta mais que o dobro dos acessos da segunda página mais acessada da Agência cana-de-açúcar.

Na Tabela 11, pode-se verificar que em mais de 83% das sessões de usuário, os visitantes entram em uma página e abandonam o portal. Este é um indicativo que a maioria dos usuários pode não estar encontrando a informação desejada. De acordo com Sculley *et al*. (2009), 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.

Tabela 11 – Contagem de sessões de usuário com uma ou mais páginas vistas.

Páginas vistas por sessão	Número de sessões	%
1	2122441	83,27
2	237456	9,32
3	72240	2,83
4	35139	1,38
5	21313	0,84
6	14823	0,58

No caso da Agência Embrapa, devido ao alto número de sessões de usuários que abandonaram a Agência logo na primeira visualização de página, a métrica é um indicativo que o portal necessitava de alguma meio de suporte aos usuários, além dos recursos de procura já implementados.

4.2 Arquitetura do sistema de recomendação

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 13 ilustra a arquitetura geral do sistema de recomendação da agência cana-de-açúcar.

Na Figura 13, a esquerda, tem-se as tarefas que são executadas no servidor e, na direita, as tarefas que são executadas no navegador. Sempre que um usuário requisita uma das páginas no seu navegador, o navegador envia uma requisição de página para o servidor Apache. Este por sua vez, ao

devolver a página ao navegador, envia também *scripts* desenvolvidos em javascript, que são responsáveis por enviar ao servidor as informações do usuário, como IP, página acessada, horário do acesso e etc. Essas informações são gravadas no servidor por um *script* em PHP e são a fonte de dados para o sistema de recomendação.

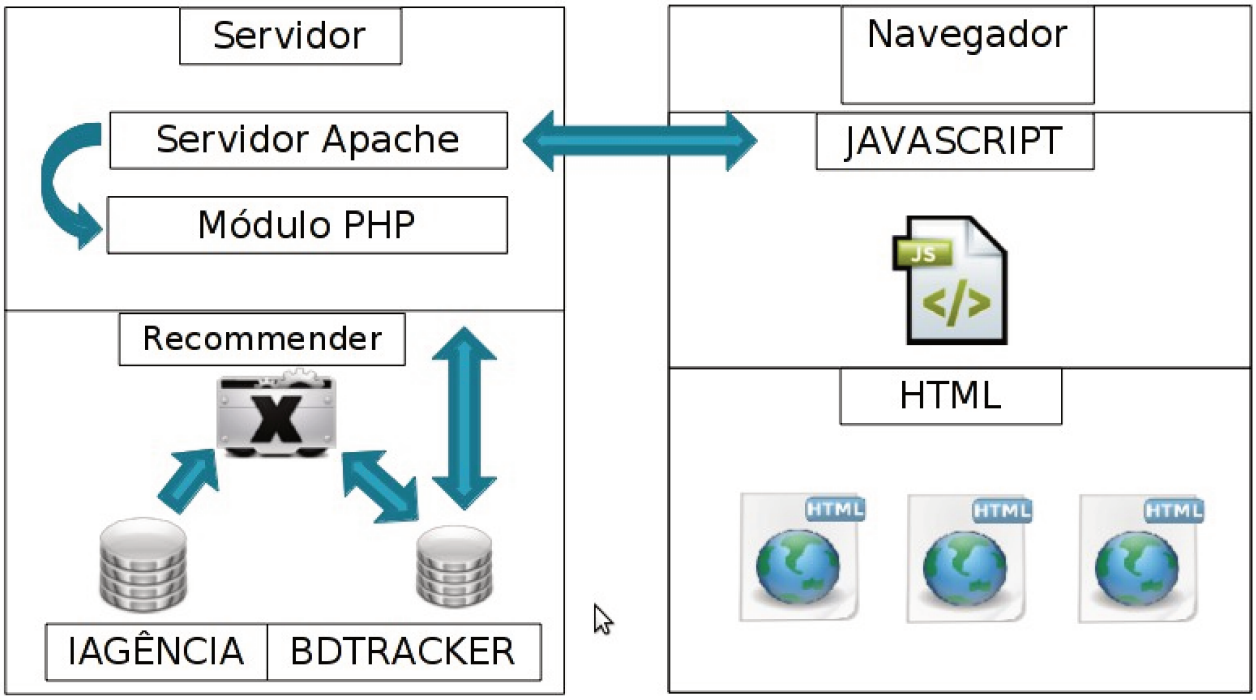


Figura 13: Arquitetura do sistema de recomendação da Agência Embrapa de Informação.

Com o sistema de recomendação em operação, o mesmo *script* responsável por enviar informações dos clientes (usuários), também requisita ao servidor recomendações para aquelas páginas ou para as páginas vistas. Se houver recomendações no banco de dados *bdtracker*, estas são devolvidas (enviadas aos usuários) e a página é atualizada, por meio dos *scripts*, mostrando as atualizações.

O *engine*[18] de recomendação foi escrito em R. Ele é responsável pela aquisição dos dados no banco de dados *bdtracker*, tratamento e transformação dos dados, geração das regras e gravação das regras no banco. Informações contidas no banco de dados iAgência, 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 no *bdtracker* para consultas posteriores.

[18] Um *engine,* que pode ser traduzido por motor, é um termo muito utilizado em TI para designar a parte mais importante de um sistema de software.

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.

4.3 Base de conhecimento

Um dos resultados mais importantes desse trabalho foi a base de conhecimento gerada. Essa base de conhecimento, 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).

A base de conhecimento, gerada para a árvore de cana-de-açúcar, contém vinte e oito regras de associação entre as páginas. Essas regras relacionam páginas de conteúdo textual e páginas com recursos eletrônicos, como arquivos em pdf e vídeos. Na Tabela 12 é apresentada a base de conhecimento de regras que está em operação no sistema de recomendação.

Além das regras geradas especificamente para a árvore de cana-de-açúcar, o sistema de recomendação foi construído de tal forma que durante o processo de geração das recomendações, são geradas regras de associação para as páginas de todas as árvores de conhecimento (agronegócio do leite, milho, feijão, manga, gado de corte, entre outras). 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 confiança mínima de 0,51.

Na análise dos links de todas as páginas antecedentes, nas 28 regras, um resultado importante é que todas 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.

Como se pode observar na Tabela 13, somente seis páginas possuem mais de uma recomendação. Além de as páginas apresentarem muitos links (não são recomendações, mas hiperlinks para outras páginas), estas apresentam somente 1 ou no máximo 4 recomendações. Esse resultado é particularmente importante pois mostra que a base de conhecimento tem um potencial de resumo e de direcionar o usuário aos links mais importantes.

Tabela 12 – Base de conhecimento com 28 regras de associação entre páginas da cultura da 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 ambiente	Diagnóstico agroambiental	0.00009567	0,51
Meio ambiente	Impactos	0.00017338	0,51

Tabela 13 – 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

Nos testes com o restante das páginas da Agência, o resultado é bem distinto: 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 na reestruturação do portal.

4.4 Validação

4.4.1 Taxa de rejeição

A validação do sistema de recomendação ocorreu de duas formas: por meio de um questionário feito a um grupo de 24 usuários, dentre especialistas e não especialistas, e de uma métrica denominada taxa de rejeição (SCULLEY *et al*., 2009).

Dado um conjunto de sessões de usuário, isto é, um conjunto de páginas vistas por um usuário da Agência, a fração de sessões com uma única visualização em relação a todas as sessões da Agência, de acordo com Sculley *et al*. (2009), é um indicativo de que os usuários podem não encontrar a informação desejada.

Quando a taxa de rejeição de um site, de um 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 toda a Agência Embrapa, a taxa de rejeição só para as páginas da cana-de-açúcar e a taxa somente das páginas de cana-de-açúcar 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.

Como se pode ver na Tabela 14, nas linhas destacadas tem-se as páginas que apresentaram variação na taxa de rejeição e que, em pelo menos um dos testes estatísticos, essa variação foi significativa. As diferenças entre as proporções de rejeição foram testadas utilizando um teste não-paramétrico (teste qui-quadrado) e um teste paramétrico (teste Z).

Observando-se os dados da Tabela 14, 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. Assim, pela falta de observações, os valores dos testes estatísticos não são confiáveis, necessitando-se de mais observações para tirar mais conclusões.

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 de recomendação 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.

Tabela 14 - Estatísticas sobre a resposta ao sistema de recomendação para as recomendações.

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

Algumas páginas se destacaram 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 esta página. Em particular, no caso da página de “Custos e Rentabilidade”, houve uma queda muito acentuada. Essa queda pode estar também associada à visibilidade das recomendações, pois de todas as 20 páginas presentes, esta é a página em que as recomendações estão mais visíveis.

No geral, as taxas de rejeição apresentaram uma 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 ser devido à baixa exposição dos links de recomendação; b) o conteúdo de algumas delas já solucionaram as necessidades de muitos usuários e; c) falta de observações em algumas páginas que são menos visitadas.

4.4.2 Questionário

O questionário utilizado neste trabalho foi elaborado a partir do modelo descrito em (PU *et al*., 2011). No questionário, foram avaliadas algumas características do sistema de recomendação:

1. **Qualidade das recomendações:** Acurácia (Q2 e Q3), Familiaridade (Q4), Atratividade (Q5), Novidade (Q6, Q7 e Q8), Diversidade (Q9).
2. **Interface:** Q10 e Q11.
3. **Facilidade de uso:** Q12.
4. **Utilidade:** Q13.
5. **Atitude:** Q14.
6. **Intenções:** Q15 e Q16.

No item 5, a “Atitude” mede subjetivamente a aceitação do sistema de recomendação, enquanto o item 6 tenta captar as intenções do usuário com relação ao uso do sistema no presente e no futuro. Uma cópia do questionário utilizado se encontra no Anexo.

O questionário foi elaborado com a utilização da escala de Likert (BOONE e BOONE, 2012). Assim, os usuários responderam questões com uma numeração que variava de 1 até 5, onde 1 indicava forte discordância e 5 forte concordância. Para análise desses dados, considerou-se que as respostas são ordenadas, mas que não indicam numericamente um intervalo, isto é, que a diferença entre 1 e 2 seria a mesma diferença entre 4 e 5. Os dados brutos estão apresentados na Tabela 15.

O questionário foi aplicado a 24 usuários, sendo que destes, 3 eram especialistas em cana-de-açúcar e 21 não-especialistas. Devido ao volume pequeno de respostas, os resultados deste questionário são usados nesse trabalho como um primeiro indicativo da resposta dos usuários ao sistema. Seria necessário um estudo com mais usuários para permitir inferências e comparações entre o grupo de especialistas e não-especialista, uma vez que neste estudo somente três especialistas responderam ao questionário.

Tabela 15 – Dados brutos coletados em 16 questões para 24 usuários, dentre especialistas e não-especialistas.

Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15	Q16
Especialista	5	5	5	5	3	5	5	5	5	4	4	5	5	5	5
Especialista	5	5	4	4	5	3	4	5	4	4	4	3	5	5	3
Especialista	4	2	4	3	3	4	3	4	4	2	2	3	2	3	3
Não Especialista	5	5	4	5	5	4	5	5	5	5	5	5	5	5	5
Não Especialista	4	5	5	5	5	5	3	2	4	5	4	4	4	5	3
Não Especialista	2	4	1	5	5	5	4	5	5	4	5	4	4	5	4
Não Especialista	5	5	4	4	4	5	5	3	4	4	5	5	5	5	5
Não Especialista	2	3	2	3	4	4	4	3	4	4	1	3	3	3	3
Não Especialista	2	3	2	3	3	3	3	3	3	4	4	3	4	3	3
Não Especialista	5	5	4	5	5	5	5	5	5	5	5	5	5	5	5
Não Especialista	5	4	4	4	3	3	3	4	4	3	3	5	5	5	5
Não Especialista	3	5	4	5	4	5	5	4	4	3	5	4	4	5	5
Não Especialista	5	5	4	5	4	5	5	5	5	4	5	5	5	5	5
Não Especialista	3	4	4	3	4	3	5	5	5	4	3	5	4	4	5
Não Especialista	3	5	3	3	4	3	3	4	5	5	5	3	3	3	3
Não Especialista	2	3	1	2	4	3	4	4	4	4	4	3	3	4	3
Não Especialista	4	4	4	4	3	3	3	3	3	2	2	3	3	5	3
Não Especialista	3	4	4	4	4	5	5	5	5	5	4	4	4	5	4
Não Especialista	3	5	3	4	2	3	4	3	4	4	4	5	3	3	2
Não Especialista	4	5	5	5	3	3	5	3	5	5	4	4	4	5	5
Não Especialista	5	5	3	5	5	5	5	5	5	5	5	5	5	5	5
Não Especialista	5	5	4	4	4	5	4	3	4	5	5	5	4	5	5
Não Especialista	2	5	1	5	4	5	5	4	5	3	5	5	5	5	5
Não Especialista	4	5	5	4	3	4	5	3	5	5	4	5	5	5	5

Da Tabela 16, pode-se verificar que em todas as questões houve respostas muito desfavoráveis e também respostas muito favoráveis (máximo e mínimo). A média e a mediana indicam que na maioria das questões houve concordância com as afirmações. Os desvios também não foram grandes, ficando a maioria próxima de um. Pelas estatísticas descritivas houve alta concordância dos usuários em quase todas as questões.

Tabela 16 – Estatística descritivas do resultado das questões.

	Média	**Desvio Padrão**	**Mediana**	**Mínimo**	**Máximo**
Q2	3,75	1,19	4	2	5
Q3	4,42	0,88	5	2	5
Q4	3,50	1,25	4	1	5
Q5	4,13	0,90	4	2	5
Q6	3,88	0,85	4	2	5
Q7	4,08	0,93	4	3	5
Q8	4,25	0,85	4.5	3	5
Q9	4,00	0,96	4	2	5
Q10	4,42	0,65	4.5	3	5
Q11	4,04	0,91	4	2	5
Q12	4,08	1,14	4	1	5
Q13	4,21	0,88	4.5	3	5
Q14	4,13	0,90	4	2	5
Q15	4,50	0,83	5	3	5
Q16	4,13	1,04	5	2	5

Na Figura 14 é apresentada a distribuição dos níveis em cada uma das questões, para os 24 respondentes. Para a codificação da numeração do questionário para as classes, utilizou-se a correspondência: 1 – Discordo totalmente, 2 – Discordo, 3 – Neutro, 4 – Concordo e 5 – Concordo totalmente. Como se pode verificar, em todas as questões há concordância (Concordo e Concordo Totalmente) na maioria das observações. As questões 2, 7, 9 e 16 são as que apresentam menor taxa de concordância.

A questão 10, sobre a interface, apresenta a maior taxa de concordância. Isto indica que a forma de apresentação e o layout são adequados para a maioria dos usuários. As maiores discordâncias ocorrem nas questões 2, 4 e 12. Como a questão 2 trata da acurácia das recomendações, e a maioria

dos usuários não são especialistas, um certo nível de rejeição era esperado. No caso dos especialistas, as respostas foram 100% positivas.

As questões 4 e 12 tratam da familiaridade e da facilidade de uso, respectivamente. Ambas as questões foram as únicas que apresentaram algum nível de total discordância. Na questão 4, considerando que a maioria dos usuários não é especialista, esperava-se que estes não estivessem familiarizados com os itens. Para os especialistas, a questão 4 apresentou familiaridade de 100% novamente.

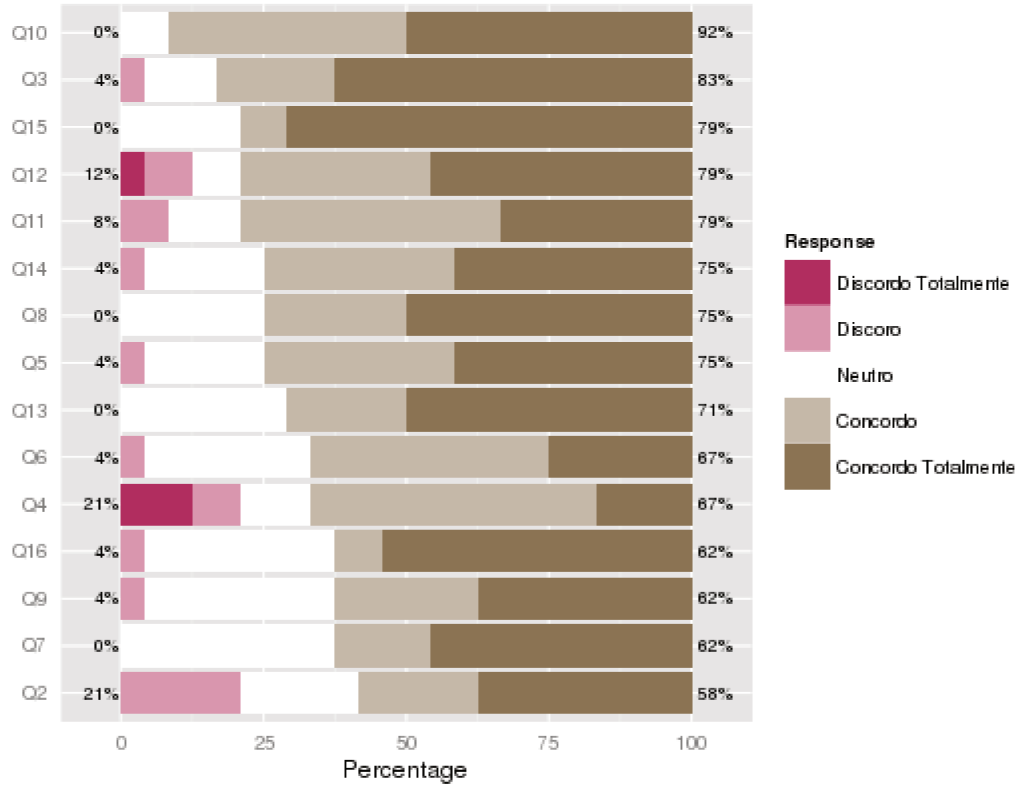


Figura 14: Distribuição das classes por questão. No eixo à esquerda a porcentagem acumulada das classes "Discordo" e "Discordo Totalmente" e, à direita, a porcentagem acumulada das classes "Concordo" e "Concordo Totalmente".

Finalmente, com relação à questão 12, apesar de quase 80% dos usuários concordarem que o sistema é fácil de usar, como houve um nível moderado de rejeição, este é um ponto a ser observado e corrigido. Possivelmente a alteração do local dos links de recomendação na página podem melhorar a experiência dos usuários. Esse resultado pode também estar relacionado à queda na taxa de rejeição das páginas, onde os links estavam mais evidentes.

Na Figura 15, com a distribuição detalhada das respostas, pode-se verificar que as questões 3 e 15 apresentam a maior parte das respostas na categoria “Concordo Totalmente”. A questão 3 trata da qualidade das recomendações e a questão 15 trata das intenções do usuário em usar ou indicar o sistema para outras pessoas. Esse resultado é importante, pois indica que para a maioria dos usuários analisados o sistema oferece boas recomendações e que estes voltariam a usá-lo.

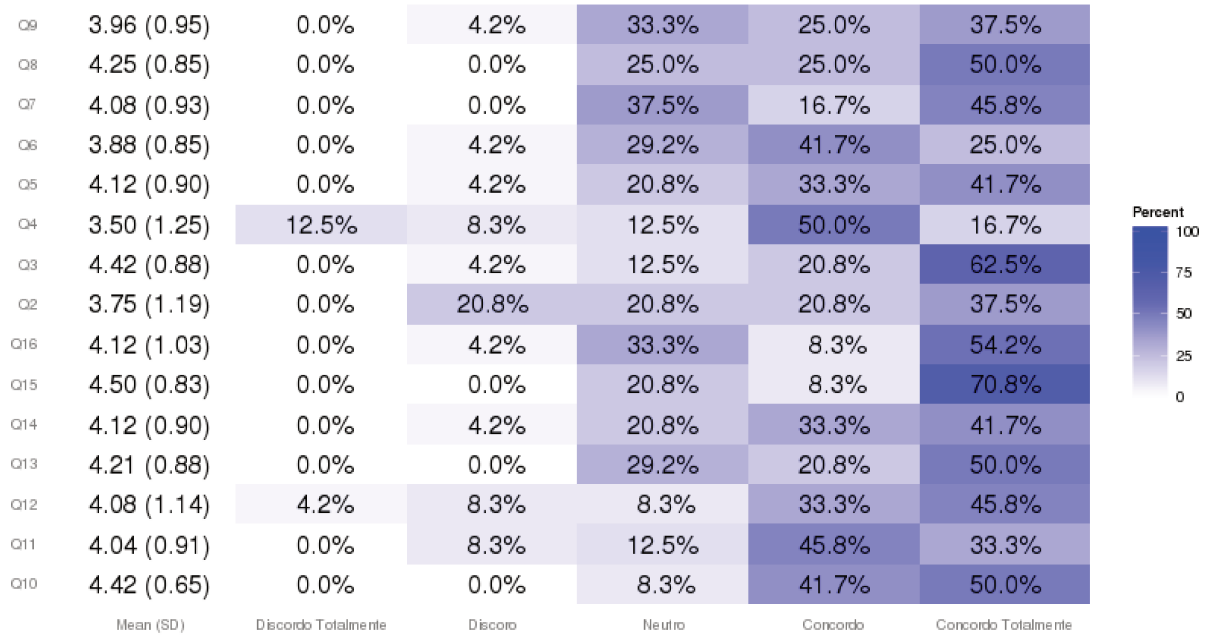


Figura 15: Distribuição detalhada das respostas por classe.

Baseado na diminuição da taxa de rejeição de quase metade das páginas da Agência cana-de-açúcar, e também dado o nível elevado de aceitação mostrado na análise do questionário, verifica-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 encontrados em Jorge *et al*., (2002) e Putman (2010). Em Jorge *et al*. (2002), o sistema de recomendação foi avaliado em relação à recomendação aleatória, obtendo um resultado satisfatório. Já em Putman (2010), um sistema de recomendação baseado em regras de associação e em conteúdo foi capaz de obter uma queda da taxa de rejeição global de um portal de 86% para aproximadamente 40%. Comparando com o resultado desse trabalho, algumas páginas também obtiveram redução significativa da taxa de rejeição e também houve uma boa aceitação por parte dos usuários.

Além dos resultados da interação com o usuário, também foi gerada uma importante base de conhecimento na forma de regras de associação, que no caso da Agência cana-de-açúcar, pode ser usada para determinar os links mais importantes presentes nas páginas. Essas regras estendidas para todas as árvores do conhecimento da Agência Embrapa, como foi mostrado, também obtiveram relação entre páginas não interligadas, fornecendo indícios de que o sistema proposto pode atuar também oferendo novos caminhos de navegação, se estendido a toda Agência Embrapa.

5. CONCLUSÕES E TRABALHOS FUTUROS

Conclui-se ser possível diminuir a taxa de rejeição de um sistema de informações técnicas agrícolas sobre cana-de-açúcar, por meio de um sistema de recomendação baseado em regras de associação, inferidas a partir do uso do portal pela comunidade.

Pela revisão de literatura, este é um dos primeiros trabalhos sobre a aplicação de sistemas de recomendação a sistemas de informações agrícolas. Como foi mostrado no trabalho, existe muita oferta de informações on-line na agricultura, mas não se encontrou outras aplicações de sistemas de recomendação relacionadas a esses repositórios de informação.

A partir dos históricos de acessos 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. Essa base de conhecimento atuou como uma forma de resumo, indicando quais links nas páginas poderiam ser mais importantes.

Do total de páginas na árvore de cana-de-açúcar, 20 delas receberam recomendações, e destas, mais da metade tiveram a taxa de rejeição diminuída com significância estatística. Ainda assim, para algumas páginas, o volume de dados coletado no período de exposição do sistema de recomendação não foi grande o suficiente para tirar conclusões, necessitando-se assim de um prazo maior de coleta de dados.

O questionário com 24 respondentes mostrou resposta positiva com relação aos usuários: as médias para as respostas foram próximas de 4 (nível de concordância) e, para as características analisadas do sistema de recomendação, os usuários não apresentaram nenhuma rejeição significativa. Ainda assim, verificou-se que o sistema pode melhorar a exposição dos links nas páginas: a) a questão que avaliou a facilidade foi a que recebeu mais avaliações negativas e; b) as páginas com links mais evidentes tiveram as taxas de rejeição diminuídas de forma mais significativas.

Como foi mostrado nesse trabalho, um sistema de recomendação para informações agrícolas em cana-de-açúcar pode ser útil para facilitar o acesso aos conteúdos de um sistema como a Agência Embrapa. Foi gerada uma base de conhecimento de regras que pode ser utilizada futuramente para reestruturação do portal e a taxa de rejeição de quase metade das páginas da Agência cana-de-açúcar foi diminuída. Também, pelo questionário, o sistema parece satisfatório do ponto de vista dos usuários.

Ainda assim, metade das páginas não apresentaram diminuição das taxas de rejeição, alguns itens do questionário apresentaram discordância e nas páginas de cana-de-açúcar não houve recomendações ligando páginas que já não tivessem links. Para determinar as causas desses resultados, é necessário mais tempo de exposição do sistema de recomendação na Agência e a coleta de mais avaliações de usuário.

A seguir são listadas algumas possibilidades para trabalhos futuros:

- Implantar um módulo de recomendação por conteúdo, de forma que usuários que buscam conteúdos que destoam da média, ou mesmo tem um perfil muito diferente da comunidade, recebam recomendações de materiais de seu interesse.
- Expandir o sistema de recomendação para outras culturas agrícolas, páginas sobre criações e árvores de conhecimento temáticas, de forma a validar a técnica de recomendação para uma ampla variedade de temas em sistemas de informações tecnológicas agrícolas.
- Implantar um módulo para extração de acessos sequenciais à Agência de Informação Embrapa, pois existem sequências de páginas que podem indicar caminhos importantes de navegação, que além de poderem ser utilizados como recomendação, podem ser utilizados para modificar a estrutura do portal.

6. REFERÊNCIAS BIBLIOGRÁFICAS

ADOMAVICIUS, G.; TUZHILIN A. Using data mining methods to build customer profiles. **Computer**, Washington, v. 34, n. 3, p. 74-82, 2001.

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.

AGRAWAL, R.; SRIKANT, R. Fast Algorithms for Mining Association Rules. **Proceedings Of The 20th International Conference On Very Large Data Bases**, Santiago, 1994.

ANJOS, I. A. dos.; FIGUEIREDO, P. A. M. de. **Aspectos fitotécnicos do plantio.** In: Cana-de-açúcar. - DINARDO-MIRANDA, L. L.; VASCONCELOS, A. C. M. de; LANDELL, M. G. de A. (Eds.) Campinas: Instituto Agronômico, 1 edição, 882p., 2010.

ANSARI, A.; ESSEGAIER, S.; KOHLI, R. Internet Recommendation Systems. **Journal Of Marketing Research,** New York, p. 363-375. ago. 2000.

BALABANOVIC, M.; SHOHAM, Y. "Fab: Content-Based, Collaborative Recommendation. **Communications Of The Acm**, v. 40, n. 3, p.66-72, 1997.

BAGLIONI, M., FERRARA, U., ROMEI, A., RUGGIERI, S., & TURINI, F. Preprocessing and mining web log data for web personalization. **Lecture Notes in Artificial Intelligence** , v. 2829, pp. 237-249, 2003

BASU, C.; HIRSH, H.; COHEN, W. Recommendation as Classification: Using Social and Content-Based Information in Recommendation. **Recommender Systems. Papers From 1998 Workshop**: Technical Report WS-98-08, AAAI Press, 1998.

BERTIN, P. R. B.; LEITE, F. C. L.; PEREIRA, F. do A. Embrapa technological information: a bridge between research and society. A**gricultural Information orldwide**, v.2, n.1, p. 10-18, 2009.

BILLSUS, D.; CLIFFORD, A. B.; CRAIG, E.; Brian, G.; MICHAEL, P. Adaptive Interfaces for Ubiquitous Web Access. **Communications Of The Acm**: The Adaptive Web, New York, v. 45, n. 5, p.34-38, 2002.

BOONE, H. Jr.; BOONE, D. Analyzing Likert Data. **Journal of Extension,** Morgantown , v. 50, n. 2, p.1-5, 2012

BOSCHI, R. S. **Analise da precipitação pluvial e de veranicos no estado do Rio Grande do Sul por meio de tecnicas de mineração de dados.** 2010. 105 f. Dissertação (Mestrado) - Faculdade de Engenharia Agrícola, Universidade Estadual de Campinas, Campinas, 2010.

BURKE, R. Knowledge-based recommender systems. **Encyclopaedia of Library and Information Systems**, vol. 69, Supplement 32. A. Kent, Ed. 2000.

CABENA, P., HADJINIAN, P., STADLER, R., VERHEES, J. & ZANASI, A. Discovery Data Mining. From Concept to Implementation. Prentice Hall. 1997.
CENTRO DE ESTUDOS AVANÇADOS EM ECONOMIA APLICADA – CEPEA/USP. Cadeia Agroindustrial da Cana-de-açúcar. Piracicaba, 2010. Disponível em:<http://cepea.esalq.usp.br/pibpec/>. Acesso em 11 de mar. de 2013.

CHAPMAN, P.; CLINTON, J.; KERBER, R.; KHABAZA, T.; REINARTZ, T.; SHEARER, C.; WIRTH, R. **CRISP-DM 1.0:** step-by-step data mining guide. Illinois: SPSS, 78p. 2000.

CLAYPOOL, M.; GOKHALE, A.; MIRANDA, T.; MURNIKOV, D.; NETES, D. COMBINING Content-Based and Collaborative Filters in an Online Newspaper. **Proc. Acm Sigir '99 Workshop Recommender Systems: Algorithms And Evaluation**, ago. 1999.

CHINCHULUUN, A.; XANTHOPOULOS, P. Data Mining techniques in agricultural and environmental sciences. **Agricultural and Environmental Sciences**, v. 1, n. June, p. 26-40, jan. 2010.

DENOBILE, T. **Modelo de gestão estratégica com foco no cliente para comercialização de produtos orgânicos.** 2005. 152 f. Dissertação (Mestrado) - Faculdade de Engenharia Agrícola, Universidade Estadual de Campinas, Campinas, 2005.

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

DILLON, S. L.; SHAPTER, F. M.; HENRY, R. J. Domestication to Crop Improvement: Genetic Resources for Sorghum and Saccharum. **Annals Of Botany**, London, v. 100, n. 5, p.975-989, 2007.

DUHAN, N.; SHARMA, A. K.; BHATIA, K. K. Page Ranking Algorithms: A Survey. **Ieee International Advance Computing Conference**, Patiala, p.1530-1537, 2009.

EMBRAPA. Agência de Informação Embrapa. Disponível em www.agencia.cnptia.embrapa.br. Acessado em Janeiro de 2012

FAYYAD, U.; PIATETSKI-SHAPIRO, G.; SMYTH, P. From Data Mining to Knowledge Discovery: an overview. In: **Advances In Knowledge Discovery & Data Mining.** Menlo Park: American Association for Artificial Intelligence, p. 1-34, 1996.

FRANCISCO, V. L. F. dos S. Acesso do setor rural à internet no Estado de São Paulo. **Informações econômicas**, SP, São Paulo, v. 33, n. 5, p. 53-56, maio, 2003.

FRATERNALI, P. Tools and Approaches for Developing Data-Intensive Web Applications: A Survey. **Acm Computing Surveys**, Milan, v. 31, n. 3, p.227-263, 1999.

GAUDER, M.; GRAEFF-HOENNINGER, S.; CLAUPEIN, W. The impact of a growing bioethanol industry on food production in Brazil. **Applied Energy,** Stuttgart, v. 88, n. 5, p.672-679, 2011.

GOLDBERG, D.; NICHOLS, D.; OKI, B. M.; TERRY, D. Using collaborative filtering to weave an information tapestry. **Communications Of The Acm**: Special issue on information filtering, New York, v. 35, n. 12, p.61-70, 1992.

HADLEY, W. The Split-Apply-Combine Strategy for Data Analysis. **Journal of Statistical Software** , v. 40, n. 15, p. 1-29, 2011.

HADLEY, W. Reshaping Data with the reshape Package. **Journal of Statistical Software** , v. 21, n. 12, p. 1-20, 2007.

HAHSLER, M., GRUEN, B., HORNIK, K. arules - A Computational Environment for Mining Association Rules and Frequent Item Sets. **Journal of Statistical Software** , v. 14, n. 15, p. 1-25, 2005.

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

HERLOCKER, J. **Understanding and Improving Automated Collaborative Filtering Systems.** 2000. 136 f. Tese (Doutorado) - University Of Minnesota, Minnesota, 2000.

HILL, C. M.; MALONE, L. C.; TROCINE, L. Data Mining and Traditional Regression. In: BOZDOGAN, Hamparsum. **Statistical Data Mining and Knowledge Discovery.** Knoxville: Chapman & Hall/crc, 2003. p. 17.

ITTOO, A. R.; YIYANG, Z.; JIANXIN, J. A text mining-based recommendation system for customer decision making in online product customization. **Icmit 2006 Proceedings - 2006 Ieee International Conference On Management Of Innovation And Technology**, Singapore, v. 1, p.473-477, 2006.

JANNACH, D.; ZANKER, M.; FELFERNIG, A.; FRIEDRICH, G. Recommender systems: an introduction. Cambridge: Cambridge University Press, 2011. 335 p

JAN, Z.; ABRAR, M.; BASHIR, S.; MIRZA, A. M. **Seasonal to Inter-annual Climate Prediction Using Data Mining KNN Technique** Wireless Networks, Information **Anais** 2009 Disponível em: <http://www.springerlink.com/index/v1822u8p6647187h.pdf>. Acesso em: 10 dez. 2012

JORGE, A.; ALVES, M.; AZEVEDO, Recommendation with association rules: a web mining application **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. **International Journal of Applied Mathematics and Computer Science,** Wrocław, v. 19, n. 1, p.165-186, 2009.

KOHLHEPP, G. Análise da situação da produção de etanol e biodiesel no Brasil. **Estudos Avançados**, São Paulo, v. 24, n. 68, p.223-253, 2010.

KRIEGEL, H. P., BORGWARDT, K. M., KRÖGER, P., PRYAKHIN, A., SCHUBERT, M., ZIMEK, A. Future Trends in data mining. **Data Mining and Knowledge Discovery,**

Hingham, v.15, n.1, p.87-97, aug., 2007.

KUI, F.; JUAN, W.; WEIQIONG, B. Research of Optimized Agricultural Information Collaborative Filtering Recommendation Systems.**Communications In Computer And Information Science**, Changsh, v. 134, p.692-697, 2011.

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.

KUMAR, A.; KANNATHASAN, N. A Survey on Data Mining and Pattern Recognition Techniques for Soil Data Mining. **International Journal of Computer Science Issues**, Chennai, v. 8, n. 3, p.422-428, 2010.

KURGAN, L. A.; MUSILEK, P. A survey of Knowledge Discovery and Data Mining process models. **The Knowledge Engineering Review**, Alberta, v. 21, n. 1, p.1-24, 2006.

LANG, K. NewsWeeder: Learning to. **Proceedings Of The Twelfth International Conference On Machine Learning**, San Francisco, p.331-339, 1995.

LINDEN, G.; SMITH, B.; YORK, J. Amazon.com Recommendations:Item-to-Item Collaborative Filtering. **IEEE Internet Computing**, New York, v. 4, n. 1, p.76-80, Jan. 2003.

LIU, B.; HSU, W.; MA, Y. Association rules with multiple minimum supports. **Proceedings Of The Acm Sigkdd International Conference On Knowledge Discovery & Data Mining**, San Diego, ago. 1999.

MAES, P.; SHARDANAND, U. Social information filtering: Algorithms for automating "word of mouth". **Human Factors in Computing Systems**. Proceedings…, p. 210-217, 1995.

MARISCAL, G.; MARBÁN, O.; FERNÁNDEZ, C. A survey of data mining and knowledge discovery process models and methodologies. **The Knowledge Engineering Review**, Cambridge, v. 25, n. 2, p.137-166, 2010.

MEIRA, Carlos Alberto Alves. **Processo de descoberta de conhecimento em bases de dados para a analise e o alerta de doenças de culturas agrícolas e sua aplicação na ferrugem do cafeeiro.** 2008. 105 f. Tese (Doutorado) - Departamento de Faculdade de Engenharia Agrícola, Universidade Estadual de Campinas, Campinas, 2008.

MIDDLETON, S. E.; SHADBOLT, N. R.; ROURE, D. C. Ontological user profiling in recommender systems. **ACM Transactions On Information Systems**, New York, v. 22, n. 1, p.54-88, 2004.

MILLER, B. N.; ALBERT, I. S.; LAM, K.; KONSTAN, J.A.; RIEDL, J. MovieLens Unplugged: Experiences with an Occasionally Connected Recommender System. **Proc. Int'l Conf. Intelligent User Interfaces**, 2003.

MOBASHER, B.; COOLEY J.; SRIVASTAVA, J. Automatic personalization based on Web usage mining. **Communications of the ACM**, v. 43, n. 8, p. 142-151, 2000.

MUCHERINO, A.; PAPAJORGJI, P.; PARDALOS, P. M. A survey of data mining techniques applied to agriculture. **Operational Research**, v. 9, n. 2, p. 121-140, 2009.

MOONEY, R. J.; ROY, L. Content-based book recommending using learning for text categorization. **Proceedings Of The Acm International Conference On Digital Libraries**, Austin, p.195-204, 2000.

NAKAGAWA, M.; MOBASHER, B. Impact of site characteristics on recommendation models based on association rules and sequential patterns. **Proceedings of the IJCAI'03 Workshop on Intelligent Techniques for Web Personalization,** Acapulco**,** p.1-10, 2003.

NET APPLICATIONS**. Desktop Top Browser Share Trend.** Disponível em: <http://www.netmarketshare.com/>. Acesso em: 31. dez. 2012.

NEVES, M. F.; CONEJERO, M. A.. Sistema agroindustrial da cana: cenários e agenda estratégica. **Economia Aplicada**, Ribeirão Preto, v. 11, n. 4, Dec. 2007 .

NONATO, R. T. **Aplicação de mineração de dados na identificação de áreas cultivadas com cana-de-açúcar em imagens de sensoriamento remoto no Estado de São Paulo.** 2011. 128 f. Dissertação (Mestrado) - Faculdade de Engenharia Agrícola, Universidade Estadual de Campinas, Campinas, 2010.

OLIVEIRA, D. R. M. S. ; OLIVEIRA, S. R. M ; SOUZA, M. I. F. Agencia de Informação Embrapa - uma ferramenta para gestão do conhecimento. In: Oitava Conferencia Iberoamericana em Sistemas, Cibernética e Informática - CISCI 2009, 2009, Orlando, FL. Oitava Conferencia Iberoamericana en Sistemas, Cibernética e Informática - CISCI 2009, p. 113-119, 2009.

OVERMYER, S P. What's Different about Requirements Engineering for Web Sites? **Requirements Engineering**, Philadelphia, v. 5, n. 1, p.62-65, 2000.

PARIK, T. S.; PATE, Neil; SCHWARTZMA, Yael. A Survey of Information Systems Reaching Small Producers in Global Agricultural Value Chains. **2007 International Conference On Information And Communication Technologies And Development**, Bangalore, 2007.

PAZZANI, M.; BILLSUS, D. "Learning and Revising User Profiles: The Identification of Interesting Web Sites. **Machine Learning**, New York, v. 27, n. 3, p.313-331, 1997.

PANNAGIO, R. **Arcabouço genérico baseado em técnicas de agrupamento para sistemas de recomendação.** 2010. 61 f. Dissertação (Mestrado) - Universidade Estadual de Campinas, Campinas, 2010.

PEDRONETTE, D. **Uma plataforma de serviços de recomendação para bibliotecas digitais.** 2008. 95 f. Dissertação (Mestrado) - Universidade Estadual de Campinas, Campinas, 2008.

PERKOWITZ, M.; ETZIONI, O. Towards adaptive Web sites: Conceptual framework and case study. **Artificial Intelligence**, v. 118, n. 1-2, p. 245-275, abr. 2000.

PIATETSKY-SHAPHIRO, G. & FRAWLEY, W. Knowledge Discovery in Databases. AAAI/MIT Press. 1991.

PIERRAKOS, D.; PALIOURAS, G.; PAPATHEODOROU, C.; SPYROPOULOS, C. D. Web Usage Mining as a Tool for Personalization: A Survey. **Journal User Modeling And User-adapted Interaction**, Hingham, v. 13, n. 4, p.311-372, nov. 2003.

PU, P.; CHEN, L.; HU, R. A user-centric evaluation framework for recommender system **5º ACM conference on Recommender systems. Anais** 2011 Disponível em: <http://dl.acm.org/citation.cfm?id=2043962>. Acesso em: 21 jan. 2013.

PUTMAN, T. **Towards adaptative retrieval and recommendation of higher education programmes.** 2010. 158 f. Dissertação (Doutorado) – Departament of Matemathics and Computer Science, Eindhoven University of Technology, Eidhoven, 2010.

REATEGUI, E. B.; CAZELLA, S. C. Sistemas de Recomendação. In ENIA 2005: Encontro Nacional de Inteligência Artificial. **Anais**, p. 306–349, 2004.

RESNICK, P; VARIAN, H R. Recommender Systems. **Communications Of The ACM**, New York, v. 40, n. 3, p.55-58, 1997.

REZENDE, S. O.; PUGLIESI, J. B.; MELANDA, E. A.; DE PAULA, M. F. Mineração de dados. In: REZENDE, S. O. **Sistemas inteligentes**: fundamentos e aplicações. São Paulo: Manole, p. 307-336, 2005.

RICCI, F.; ROKACH, L.; SHAPIRA, B.; KANTOR, P. B. (Ed.). Recommender systems handbook. New York: Springer, 2011. 842 p.

ROMANI, L. A. S.; AVILA, A. M. H.; ZULLO, J.; CHBEIR, R.; TRAINA, C.; JR.; TRAINA, A. J. M. CLEARMiner: a new algorithm for mining association patterns on heterogeneous time series from climate data. **In Proceedings of the 2010 ACM Symposium on Applied Computing**. ACM, New York, NY, USA, 900-905.

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. **Anais** New York, New York, USA: ACM Press, 2009. Disponível em: <http://portal.acm.org /citation.cfm?doid=1557019.1557161> .

SHEN, Y.; ZENGFENG, Z.; HONGSHENG, W. Agricultural information technology development and innovation path. **Internation Conference on Electronics, Communications and Control (ICECC)**, Zhejiang, p.2512-2515, 2011.

SINGH, B.; SINGH, H. Web data mining research: a survey. **Ieee International Conference On Computational Intelligence And Computing Research**, Lucknow, p.661-670, 28 dez. 2010.

SOBOROFF, I; NICHOLAS, C. Combining Content and Collaboration in Text Filtering. **Proc. Int'l Joint Conf. Artificial Intelligence Workshop: Machine Learning For Information Filtering**, ago. 1999.

SOUTH, A. rworldmap: A New R Package for Mapping Global Data. **The R Journal**, v. 3, n. 1, p. 35- 43, 2011.

SOUZA, M. I. F.; VIAN, C. E. de F.; MARIN, F. R. ; SAKAI, R. H.; LOPES, P. C. Informação tecnológica sobre cana-de-açúcar na internet. Análises e Indicadores do Agronegócio, v. 4, p. 1-5, 2009.

THANGAMANI, C.; THANGARAJ, P. Survey on Web Usage Mining: Pattern Discovery and Applications. **International Journal Of Computer Science And Information Security**, Sathy, v. 9, n. 10, p.78-83, out. 2011.

TUZHILIN, A. Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions. **IEEE Transactions On Knowledge And Data Engineering**, New York, v. 17, n. 6, p.734-749, jun. 2005.

UNIÃO DA AGROINDÚSTRIA CANAVIEIRA DE SÃO PAULO (UNICA). **Dados e Cotações - Estatísticas.** Disponível em: <http://www.unica.com.br/dadosCotacao/estatistica>. Acesso em: 2. ago. 2011.

UNIÃO DA AGROINDÚSTRIA CANAVIEIRA DE SÃO PAULO (UNICA). **Final Report of 2011/2012 Harvest Season.** Disponível em: <http://www.unicadata.com.br/listagem.php?idMn=72> Acesso em: 5. dez. 2012.

VIBHA, L.; HARSHAVARDHAN, G. M.; PRASHANTH, S. J.; SHENOY, P. DEEPA; VENUGOPAL, K. R.; PATNAIK, L. M. A hybrid clustering and classification technique for soil data mining, **Information and Communication Technology in Electrical Sciences, 2007. ICTES. IET-UK International Conference**, pp.1090-1095, 2007.

WU, T.; CHEN, Y.; HAN, J. Re-examination of interestingness measures in pattern mining: A unified framework. **Data Mining and Knowledge Discovery,** Illinois, v. 21, n. 3, p. 371-397, 2010.

WU, X.; KUMAR, J.; QUINLAN J.; GOSH, J.; YANG, Q.; MOTODA, Y.; MCLACHLAN G. J.; NG, A.; LIU, B.; YU, P. S.; ZHOU, Z.; STEINBACH, M.; HAND, D. J.; STEINBERG, D. Top 10 algorithms in data mining. **Knowledge and Information Systems**, v. 14, n. 1, p. 1-37, 2008.

YANG, H.; PARTHASARATHY, S. On the use of constrained associations for web log mining. **Proceedings of the 4-rd International Workshop on Mining Web Data for Discovering Usage Patterns and Profiles, WEBKDD 2002, MiningWeb Data for Discovering Usage Patterns and Profiles,** Edmonton, v. 2703, p.100-118, 2003.

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.

YU, K.; SCHWAIGHOFER, A.; TRESP, V.; XIAOWEI, X.; KRIEGEL, H. P. Probabilistic Memory-Based Collaborative Filtering. **IEEE Transactions On Knowledge and Data Engineering**, New York, v. 16, n. 1, p. 56-69, jan. 2004.

ANEXO

Questionário de Validação

* Identificação

1. Selecione seu perfil com relação à cana-de-açúcar. (Especialista ou Não Especialista)

* Qualidade das recomendações

* Acurácia

2. Os itens recomendados são do meu interesse.

1 2 3 4 5

3. O sistema oferece boas recomendações.

1 2 3 4 5

* Familiaridade

4. Alguns itens recomendados são familiares a mim.

1 2 3 4 5

* Atratividade

5. Os itens recomendados são atrativos.

1 2 3 4 5

* Novidade

6. Os itens recomendados a mim foram novos e interessantes.

1 2 3 4 5

7. O sistema de recomendação é educacional.

1 2 3 4 5

8. O sistema de recomendação ajuda a encontrar novos materiais de leitura.

1 2 3 4 5

*** Diversidade**

9. Os itens recomendados são diversos (sugestões sobre temas diferentes).

1 2 3 4 5

*** Avaliação da interface**

10. Os títulos das recomendações são claros.

1 2 3 4 5

11. O layout das recomendações é atrativo.

1 2 3 4 5

*** Percepção de facilidade de uso**

12. Eu encontrei as recomendações facilmente.

1 2 3 4 5

*** Percepção de utilidade**

13. Os itens recomendados influenciaram minha navegação.

1 2 3 4 5

*** Atitude**

14. As recomendações me deixaram mais confiante sobre o que ler e ver.

1 2 3 4 5

☐ ☐ ☐ ☐ ☐

*** Intenções de comportamento**

15. Eu usaria o sistema de recomendação novamente.

1 2 3 4 5

☐ ☐ ☐ ☐ ☐

16. Eu vou falar sobre o sistema de recomendação para interessados em cana-de-açúcar.

1 2 3 4 5

☐ ☐ ☐ ☐ ☐