



MINERAÇÃO DE DADOS PARA IDENTIFICAÇÃO DE PRINCIPAIS TERMOS DE BUSCA NO SITE DA APTA REGIONAL

MARCOS VINICIUS PEREIRA **BIGLIAZZI**¹; CRISTINA **FACHINI**²; RICARDO **FIRETTI**³

Nº 12301

RESUMO

A APTA Regional é um departamento que faz parte da Agência Paulista de Tecnologia dos Agronegócios (APTA). Seu site é um importante veículo para a transferência de conhecimentos e, posteriormente, de tecnologias para o seu público alvo. Tendo isso em vista, este trabalho apresenta uma análise, com base na mineração de dados, dos termos de busca utilizados pelos usuários deste site durante os anos de 2009, 2010 e 2011, pretendendo formar, a partir desse trabalho analítico, certas tendências e linhas gerais dos principais temas e das principais cadeias produtivas que obtém destaque nessa busca por informações por parte dos usuários. Utilizando-se a frequência relativa percebeu-se que os termos de maior frequência nestes anos foram “Carne(s)”, “Maturada” e “Piscicultura”.

ABSTRACT

APTA Regional is a department that is part of the Agribusiness Technology Agency of São Paulo State (APTA). Your site is an important vehicle for knowledge transfer and, subsequently, technology transfer for your target audience. This paper presents an analysis based on data mining of the search terms used by users of this site during the years 2009, 2010 and 2011, intending to build, from this analytical work, certain trends and general lines of the main themes and main supply chains that gets highlighted in this search for information by users. Using the relative frequency we realized that the terms of an increasead frequency in recent years were “Meat”, “Matured” and “Pisciculture”.

INTRODUÇÃO

Criada pelo Decreto No. 44.885 de 11 de maio de 2000, a Agência Paulista de Tecnologia dos Agronegócios (APTA) é a instituição de pesquisa da Secretaria de

¹ Bolsista FUNDAP: Graduação em Ciências Econômicas, UNICAMP, Campinas-SP, mbigliazzi@apta.sp.gov.br.

² Orientadora: Pesquisadora, APTA Regional- Sede, Campinas-SP.

³ Colaborador: Pesquisador, APTA Regional- Sede, Campinas-SP.



Agricultura e Abastecimento do Estado de São Paulo. Teve sua organização (reorganização) instituída pelo Decreto 46.448 de 08 de Janeiro de 2002, onde passou a contar em sua estrutura com o Departamento de Descentralização do Desenvolvimento (DDD ou APTA Regional), além de seus Institutos de excelência: Agrônomo de Campinas (IAC), de Tecnologia de Alimentos (ITAL), Instituto de Zootecnia (IZ), Instituto Biológico (IB), Instituto de Economia Agrícola (IEA) e Instituto de Pesca (IP).

A APTA Regional está sediada administrativamente em Campinas e estruturada em 15 Polos Regionais de Desenvolvimento Tecnológico dos Agronegócios e 19 Unidades de Pesquisa e Desenvolvimento. Esse departamento tem a finalidade de articular os Pólos e UPDs sob sua coordenação na geração, adaptação e transferência de conhecimentos científicos e tecnológicos, a partir de uma visão multidisciplinar focada em cada região do estado, contemplando as principais cadeias de produção locais. Desta maneira atua no campo da ciência aplicada onde os resultados de pesquisas fundamentalmente devem atingir os processos de produção da agricultura, pecuária e aquicultura.

É exatamente no objetivo da transferência dos conhecimentos científicos e tecnológicos gerados ou estudados pelo Departamento que encontra-se a grande importância do site da APTA Regional. O site abre um canal direto de comunicação com os agentes que se interessam pelos assuntos tratados pelos Polos, e também o inverso, com o público que os Polos Regionais de Desenvolvimento desejam e necessitam dialogar (produtores rurais principalmente).

É nesse sentido, levando-se em conta a importância do site da APTA Regional nesse trabalho de divulgação e de transferência de conhecimento, que consideramos importante a análise das informações e dados sobre a ferramenta de busca do site. O objetivo deste trabalho é apontar os assuntos e temas mais procurados no site da APTA Regional, captar quais cadeias produtivas tem mais demanda por informações de pesquisa e tecnologia, além de trazer uma indicação mais detalhada, e baseada em dados concretos, dos temas mais relevantes para serem atendidos e divulgados futuramente neste espaço tão importante.

MATERIAL E MÉTODOS

A Mineração de Dados ou Prospecção de Dados é basicamente o processo de extração de conhecimento a partir de uma base de dados não numéricos. Essa extração de conhecimentos é feita utilizando-se modelos estatísticos ou técnicas

matemáticas para a percepção de padrões ou associações consistentes entre os dados.

QUONIAN et al. (2001) define a *data mining*, uma das principais técnicas da mineração de dados, como uma técnica para extrair informações, previamente desconhecidas e de grande abrangência, a partir de bases de dados, como subsídio para a tomada de decisões. A aplicação dessa técnica torna possível a transformação de dados em informação e, posteriormente à análise inclusa nesse processo, em conhecimento. A utilização da *data mining* faz com que a difícil análise de dados qualitativos fique mais simplificada, já que há transformação desses em dados quantitativos, expressos por um modelo matemático.

Neste estudo foram utilizados como base de dados os termos de busca utilizados para pesquisa no site da APTA Regional pelos seus visitantes. Esses termos de busca foram capturados através da utilização do AWStats.

O AWStats (Advanced Web Statistics) é uma ferramenta poderosa que analisa arquivos de log, de determinada página da web, ricos em informação. Essa análise é disponibilizada por meio de relatórios e gráficos sobre diversos aspectos desse site, como por exemplo: FTP, servidor de e-mail, tráfego, termos de busca etc.

Uma das informações extraídas a partir do uso dessa ferramenta é exatamente o conjunto de palavras que foram buscadas no site, bem como a frequência mensal em que a mesma palavra foi tema de busca. Para análise feita neste trabalho utilizamos os dados de 2009, 2010 e 2011.

As informações dadas por esta ferramenta, entretanto, precisaram de certo tratamento para que a análise pudesse se dar de forma satisfatória. Antes de aplicarmos sobre as variáveis um instrumental matemático, fizemos uma "limpeza" na base de dados: alguns termos parecidos, que tratavam sobre as mesmas cadeias produtivas, foram agrupados, outros que poderiam se relacionar com mais de uma cadeia de produção foram desconsiderados, objetivando sempre no resultado final termos uma análise mais direta sobre as cadeias produtivas mais importantes e mais procuradas pelos usuários do nosso site. Retiramos da lista de termos as chamadas *stopwords*, palavras apenas acessórias como artigos, preposições, adjuntos nominais e adverbiais, etc. Além disso, ao invés de utilizarmos os dados mensais, como o AWStats nos fornecia, preferimos utilizar o somatório das frequências absolutas de cada ano, e em cada ano de cada mês, chegando ao final do trabalho com frequências relativas dos termos de busca que correspondiam aos três anos de nossos intervalos somados.

As próximas etapas de análise desses dados foram feitas com base nas teorias de ZIPF (1949) e LUHN (1958) sobre a mineração de dados. De ZIPF utilizamos a constatação empírica (exposta na Lei de Zipf) que apresenta uma descrição da distribuição de freqüências de palavras na linguagem humana, que podem ser organizadas em três categorias: poucos termos muito comuns, uma quantidade média de termos de frequência intermediária e, por fim, muitos termos que ocorrem poucas vezes. De LUHN utilizamos a idéia de que a frequência de um termo em um documento, ou em uma base de dados (como é o nosso caso), fornece uma medida útil para a determinação da significância deste mesmo termo (metodologia também utilizada por Firetti, 2012).

Após a determinação da frequência absoluta e relativa dos termos considerados durante todo o intervalo de tempo considerado, a próxima etapa do trabalho se centrou na obtenção da curva de Zipf, utilizando a planilha de cálculos Excel, para o conjunto de termos de busca nos anos considerados. Acrescentou-se a esse gráfico sua curva de tendência utilizando-se uma função de regressão potência com o respectivo valor de R^2 também indicado (quanto mais esse valor aproxima-se de 1,0 melhor o ajuste dos dados obtidos).

Após esse procedimento, separou-se, a partir dos cortes de Lhun, três zonas de frequência de termos, foram elas: Os termos mais comuns (com os termos de busca com freqüências relativas maiores que 6,0%), os termos significativos (com os termos de busca com freqüências relativas entre 5,9% e 3,0%) e os ruídos, ou os termos de menor importância (com os termos de busca com freqüências relativas menores que 2,9%). Para facilitarmos a visualização desses três grupos, foram utilizadas três cores diferentes para cada um deles no gráfico: PRETO para os termos mais comuns, VERMELHO para os termos significativos, e AZUL para os termos de menor importância.

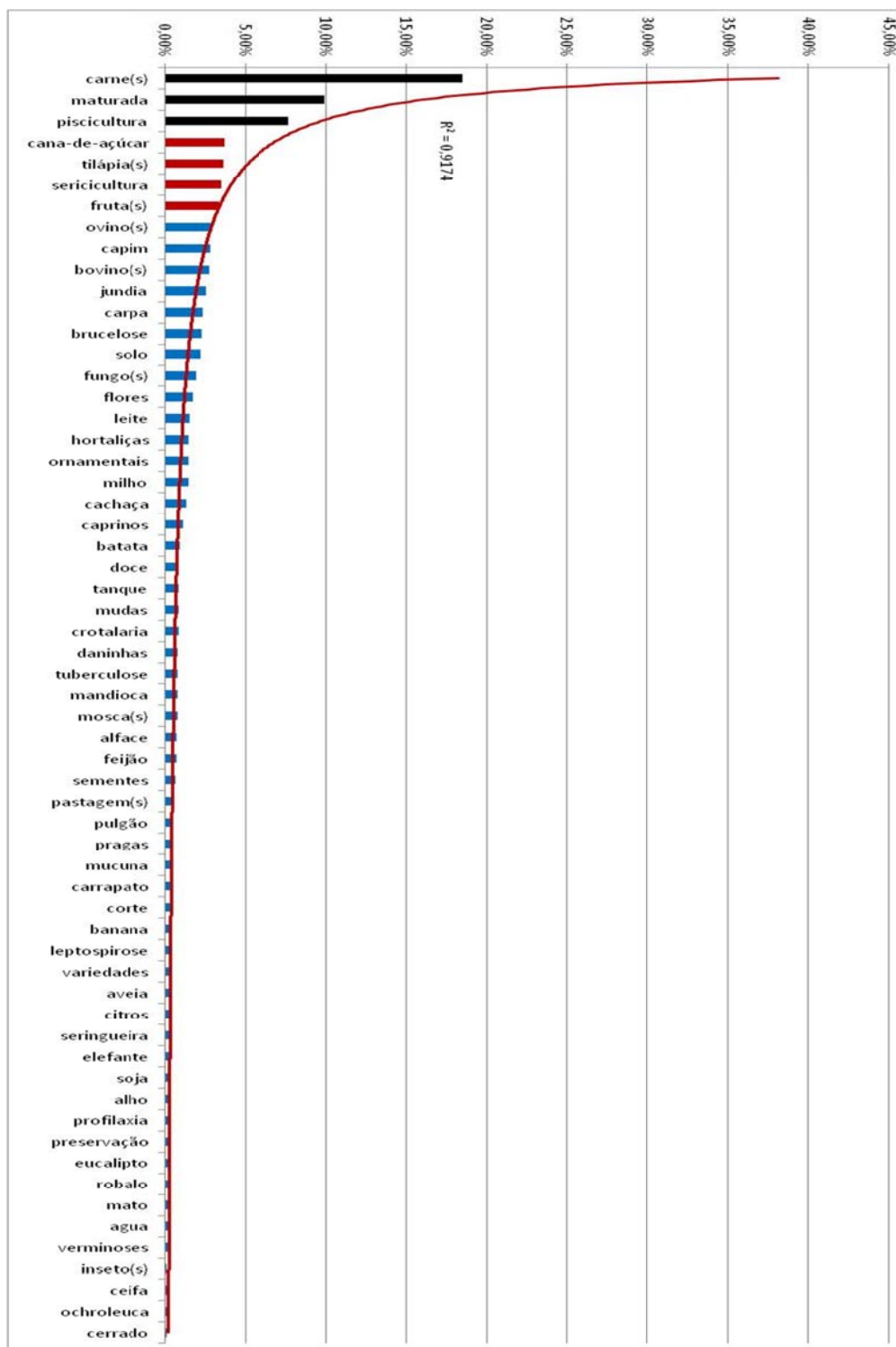
RESULTADOS E DISCUSSÃO

No gráfico resultante da reorganização dos termos, bem como as suas freqüências, consideramos um intervalo que reuniu os anos de 2009, 2010 e 2011. Isso facilitou a nossa análise e permitiu ter uma ideia mais geral das tendências nas buscas por informações no site da APTA Regional durante o intervalo como um todo.

Começando a análise da primeira faixa de frequência considerada, a dos termos mais comuns com frequência relativa acima de 6,0%, percebeu-se uma “liderança” dos termos que são relacionados com a Cadeia Produtiva da Bovinocultura de Corte.

Estes termos, “Carne(s)” e “Maturada” apresentam participações bem altas na busca por palavras no site da APTA Regional, representando, respectivamente 18,5% e 9,9% dos termos de busca considerados no trabalho, ficando no grupo PRETO.

Gráfico. Principais termos de busca no site da APTA Regional durante 2009, 2010 e 2011.



Outro termo também pertencente à faixa de frequência PRETO é “Piscicultura”, o que caracteriza uma importância também bastante significativa desta Cadeia Produtiva na demanda por informação dos usuários do site. Temos então, a partir desta análise, que as Cadeias Produtivas mais importantes nas buscas no site da APTA Regional são: a Bovinocultura de Corte e a Piscicultura.

A próxima faixa de frequência é também muito importante, pois nela são considerados termos significativos, com frequências relativas entre 5,9% e 3,0%. Nesta faixa podemos captar temas que, apesar de não representarem os principais termos pesquisados, demonstram um grande potencial até mesmo para um futuro crescimento na demanda por informações em anos posteriores.

Nesta faixa de frequência encontramos quatro termos: o primeiro deles corresponde a uma das principais commodities agrícolas do Estado de São Paulo, a cana-de-açúcar (3,72%). É notada a importância desta cultura tanto em termos históricos no nosso país quanto, posteriormente, no encadeamento que este tipo de produto gera, podendo ser convertido em açúcar refinado posteriormente a partir da atividade industrial, ou até mesmo em produtos como a cachaça (termo também buscado no site com frequência de 1,31%) que tanto podem ser industriais quanto artesanais.

O próximo termo de busca em destaque neste intervalo de frequência é o termo “Tilápia(s)” que corrobora plenamente com a constatação da importância da Cadeia Produtiva da Piscicultura. A Tilápia é um tipo de peixe que corresponde no Brasil a uma das espécies mais resistentes e de boa reprodução, o que faz com que sua criação corresponda a uma atividade bastante rentável. Isso explica em grande parte a grande procura no site da APTA Regional de informações sobre esta produção, essas informações são bastante úteis aos produtores que através de novas tecnologias ou métodos de criação podem fazer com que esta atividade aumente sua eficiência consideravelmente. Também é conveniente dizer que existe um corpo importante de pesquisadores em piscicultura do departamento, distribuídos em oito dos quinze Polos Regionais, que pesquisam principalmente a tilápia.

Posteriormente, encontramos ainda no grupo dos termos significativos a palavra “Sericultura” com cerca de 3,6% de participação em todas as buscas feitas nestes anos. A Sericultura é uma parte da Zootecnia que trata do estudo e da criação do Bicho-da-seda. Esta constatação é muito interessante, pois percebemos que essa cadeia produtiva, muitas vezes pouco comentada, ocupa a importância nas demandas dos usuários do site por informações. Durante todo o intervalo de tempo considerado esta demanda se fez presente, fazendo com que esta palavra estivesse no grupo dos

termos significativos, em sexta posição em relação a 112 termos de busca analisados, mostrando a importância da divulgação de informações científicas e novas tecnologias que fazem parte desta atividade. Ressalta-se que do departamento possui uma unidade de pesquisa e desenvolvimento (UPD) em Gália que gera pesquisas referentes a essa cadeia.

Por fim, o último termo que faz parte desse intervalo de frequência é “Fruta(s)”. Este termo, apesar de não fazer alusão a uma cultura agrícola específica, denota uma importância considerável Fruticultura em geral na demanda por informações pesquisadas no site. Fazem parte desta cadeia diferentes tipos de atividades produtivas ligadas às frutas, tanto a produções extensivas, quanto aquelas que demandam uma mão-de-obra e instrumentos mais especializados e que se fazem em propriedades mais diminutas. As informações sobre esse tipo de produção são também muito importantes para os diferentes tipos de produtores rurais na consideração dos melhores produtos e também das melhores formas de produção considerando suas especificidades. Possivelmente o interesse do público por esse assunto é ainda geral, uma vez que os demais termos do grupo azul apresentam poucas cadeias produtivas frutícolas específicas (citricultura e bananicultura) (até a frequência de 0,14%). O contrário, por exemplo, da Horticultura, que não aparece como termo único, porém dentro do grupo AZUL, encontram-se termos relacionados como flores, hortaliças, batata (doce), mandioca, alface e alho, medicinais e alcachofra, que juntos somam 3,66% da frequência.

No terceiro grupo concentraram-se o restante dos termos, ou seja, aqueles que apresentaram frequências relativas abaixo de 2,9% até 0,01%, demonstradas no gráfico até 0,14%⁴.

Percebemos prontamente que os primeiros termos deste grupo são, em grande parte, relacionados aos temas que já foram apresentados. Temos os termos “Ovino(s)”, “Capim”, “Bovino(s)”, “Brucelose”, “Leite” e etc, que se relacionam fortemente, não apenas com a Bovinocultura de Corte, mas também com a produção de Leite e também outras formas de pecuária, denotando à essa área produtiva uma importância ainda maior do que a que já tinha sido caracterizada na parte inicial da discussão sobre nossos resultados.

⁴ 53 termos representando juntos 0,274% dos termos buscados não foram apresentados no gráfico para melhor visualização do mesmo, e por representarem baixa significância dentro do conjunto estudado.

Outros termos se relacionam da mesma forma, mas com a Cadeia Produtiva da Piscicultura, são eles: “Jundiá”, “Carpa”, “Tanque” e etc. Estes termos relacionam-se fortemente com esta atividade e representam palavras que foram também amplamente buscadas durante os anos considerados pelos usuários do site, mostrando a importância da disponibilização de informações sobre este assunto por parte do Departamento nesta ferramenta de divulgação.

Apesar da grande quantidade de termos nesta faixa de frequência podemos indicar alguns temas e Cadeias Produtivas que, segundo nossos resultados, se fizeram importantes nesta demanda por informações captadas. O termo “Fungo(s)”, com cerca de 1,9% de frequência relativa, mostra uma considerada demanda por informações sobre este reino animal no site. Essas informações podem se concentrar tanto nos temas de sanidade animal e vegetal, onde os fungos podem ser causadores de doença, como também na produção de alguns tipos de fungos como cogumelos (shitake, shimeji, etc) e outros produtos vastamente utilizados na culinária.

Outro tipo de Cadeia Produtiva que pode ser indicada é a de produção de flores. Os termos “Flore(s)” e “Ornamentais” representam juntas mais de 3,0% dos termos buscados durante esses anos. A produção de flores é uma atividade agrícola bastante importante, principalmente para alguns municípios paulistas como Holambra e etc. A colocação de informações sobre este tipo de atividade é muito importante para a divulgação de técnicas e novas tecnologias aos produtores rurais que, como vimos pelos nossos resultados, buscam informações sobre este tipo em nosso site.

Alguns outros tipos de termos podem ser destacados, como: “Solo”, “Mudas”, “Sementes” e etc, termos estes que não correspondem alguma cultura agrícola específica, mas que estão bastante presentes nas atividades agrícolas e por isso representam aspectos importantes na busca e na divulgação de informações por parte do site da APTA Regional.

Alguns outros termos pertencentes aos temas de sanidade vegetal também podem ser destacados, como daninhas, mosca(s), pulgão, pragas, inseto(s), fitorremediação, virose(s) e fitopatogênico, que juntos somam 2,79% da frequência relativa. Além os referentes a sanidade animal como carrapatos, brucelose, tuberculose, leptospirose entre outros. Esta importância na consideração de problemas nos cultivos agrícolas ou criações de animais dos produtores rurais é também um indicativo do tipo de demanda por informações trazidas pelo público que vai até o site da APTA Regional.

CONCLUSÃO

O objetivo desse estudo estava centrado na captação dos principais termos e, conseqüentemente, dos principais temas pesquisados pelos usuários do site da APTA Regional nos anos de 2009, 2010 e 2011. Com essa constatação pretendemos conseguir delimitar, ou pelo menos indicar, os principais temas buscados nesse site, e – tendo em vista a importância dessa ferramenta na consolidação da transferência de conhecimentos ao público – tentar indicar as cadeias produtivas mais importantes nessa demanda por informações.

Para este fim foi então aplicada uma metodologia de mineração de dados. Houve a transformação de nossa base de dados em um conhecimento mais sistematizado, visualizado através de um gráfico, expressando as freqüências relativas dos termos de busca dos usuários do site em três grupos, ou melhor, três faixas de freqüência: Uma primeira faixa com poucos termos muito significativos, a segunda com uma quantidade média de termos de significância intermediária e, por fim, uma terceira, com muitos termos de menor significado, ou seja, apresentando menores freqüências relativas.

Analisando-se os dados, a primeira conclusão que se mostra evidente em nossos resultados é a grande importância, nas buscas por informações dos usuários do site, da Cadeia de Bovinocultura de Corte. Esse fato analisado no trabalho é importante do ponto de vista da constatação dos principais temas que são buscados pelos usuários do site no seu conteúdo. Mostra que esse tema ainda deve ser muito trabalhado e divulgado, para que se tenham sempre novas informações importantes e materiais que auxiliarão no desenvolvimento dessa cadeia produtiva. O mesmo se aplica para a Cadeia Produtiva de Piscicultura, que tem a própria palavra “Piscicultura” neste grupo de termos mais comuns e é ainda mais destacada com algumas palavras que estão presentes na segunda faixa de freqüências considerada.

A análise da segunda zona de abrangência é muito importante. É nesse intervalo que percebemos, durante todos os anos, os termos também bastante significativos (em termos de freqüência relativa), mas que não apresentam freqüência suficiente para fazerem parte do primeiro grupo. Esses termos podem ser considerados, de certa forma, como os termos inovadores, ou melhor explicando, termos que apresentam potencial de significância e que podem vir a fazer parte, no futuro, das palavras mais buscadas em nosso site.

Termos ligados a Cadeia Produtiva da Piscicultura se mostraram presentes nessa faixa como já havíamos dito, o que corrobora ainda mais para a importância dessa cadeia na demanda por informações por parte dos usuários. Além destes, o termo



“Cana-de-açúcar” e o termo “Sericicultura” (ligado a criação do bicho-da-seda) também se mostraram presentes nesta faixa, o que indica a importância destas atividades bem específicas na busca por informação pelos usuários do site.

Sobre os termos pertencentes da terceira faixa de frequência, vimos que é muito difícil apresentarmos destaques ou tendências, já que essa faixa tem nesta análise uma quantidade de 105 palavras. Apesar disso, alguns temas podem ser destacados, como: Sanidade Vegetal, Produção de Flores, Fungos, Pecuária, diversos tipos de produções específicas, por exemplo, Milho, Batata, Mandioca, Feijão, e etc.

É com base nesses resultados, nesses aspectos que foram possíveis serem destacados, que podemos visualizar com mais clareza os temas procurados pelos usuários do site da APTA Regional durante esse período. Essas informações são muito preciosas e, levando-se sempre em conta a função do site como veículo de informação e de transferência de conhecimentos e tecnologia, importantes para a tomada de decisões sobre o conteúdo que irá ser disponibilizado daqui pra frente.

REFERÊNCIAS

APTA Regional. Quem somos. <http://www.aptaregional.sp.gov.br/index.php/quem-somos>. Acesso em 10 de maio de 2012.

FIRETTI, R. Utilização da técnica de mineração de dados para determinação de padrões distintivos identificadores de sobreposições nos projetos de pesquisa em andamento na APTA. 2012. Disponível em: http://www.aptaregional.sp.gov.br/index.php/component/docman/doc_download/1006-pesquisas-institutos-apta-x-apta-regional-sobreposicao-ou-complementacao-?Itemid=269 . Acesso em 15/5/2012.

LUHN, H.P. **The automatic creation of literature abstracts**. IBM Journal, 159-165, 1958.

QUONIAN, L.; TARAPANOFF, K.; ARAÚJO JÚNIOR, R.H.; ALVARES, L. Inteligência obtida pela aplicação de data mining em bases de teses francesas sobre o Brasil. **Ciência e Informação**, v.30, n.2, p.20-28, 2001.

ZIPF, G.K. **Human behavior and the principle of the least effort**. Cambridge, MA:AddisonWesley, 1949.