Boletim de Pesquisa 28
e Desenvolvimento
ISSN 1677-9266
Dezembro, 2011

Uma estratégia para auxiliar a identificação de portfólios por meio de mineração de textos

Embrapa

ISSN 1677-9266
Dezembro, 2011

Empresa Brasileira de Pesquisa Agropecuária
Embrapa Informática Agropecuária
Ministério da Agricultura, Pecuária e Abastecimento

Boletim de Pesquisa e Desenvolvimento 28

Uma estratégia para auxiliar a identificação de portfólios por meio de mineração de textos

Maria Fernanda Moura
Silvio Roberto Medeiros Evangelista
Silvia Maria Fonseca Silveira Massruhá
Thiago Teixeira Santos

Campinas, SP
2011

Embrapa Informática Agropecuária
Av. André Tosello, 209 - Barão Geraldo
Caixa Postal 6041 - 13083-886 - Campinas, SP
Fone: (19) 3211-5700 - Fax: (19) 3211-5754
www.cnptia.embrapa.br
sac@cnptia.embrapa.br

1ª edição on-line 2011

Dados Internacionais de Catalogação na Publicação (CIP)
Embrapa Informática Agropecuária

Uma estratégia para auxiliar a identificação de portfólios por meio de mineração de textos / Maria Fernanda Moura... [et al.]. - Campinas : Embrapa Informática Agropecuária, 2011.
26 p. : il. - (Boletim de pesquisa e desenvolvimento / Embrapa Informática Agropecuária, ISSN 1677-9266 ; 28).

1. Portifólios. 2. Mineração de textos. 3. Aprendizado de máquina. 4. Seleção de atributos. I. Moura, Maria Fernanda. II. Embrapa Informática Agropecuária. III. Título. IV. Série.

CDD 006.33 (21. ed.)

Sumário

Uma estratégia para auxiliar a identificação de portfólios por meio de mineração de textos

Maria Fernanda Moura[1]
Silvio Roberto Medeiros Evangelista[2]
Silvia Maria Fonseca Silveira Massruhá[3]
Thiago Teixeira Santos[4]

Resumo

Neste trabalho utilizou-se uma estratégia com base em mineração de textos para fornecer uma indicação dos atuais portfólios da Embrapa Informática Agropecuária. Para tanto, utilizaram-se, como dados, os projetos liderados por essa unidade nos últimos anos, de 2004 a 2010. A esses dados aplicou-se um processo de extração semiautomática do conhecimento, utilizando vocabulário controlado, filtragem estatística de atributos, agrupamento hierárquico de documentos, descrição única dos agrupamentos e várias interações usuários do processo. Os resultados obtidos foram satisfatórios, tendo passado apenas por validação subjetiva, mostrando que o processo pode ser aplicado a dados semelhantes.

Termos para indexação: Portfólios, mineração de textos, aprendizado de máquina, seleção de atributos, agrupamento de documentos, descrição de agrupamentos.

---

[1] *Doutora em Ciências de computação e matemática computacional, Pesquisadora da Embrapa Informática Agropecuária, Caixa Postal 6041, Barão Geraldo, 13083-886, Campinas, SP, fernanda@cnptia.embrapa.br*
[2] *Doutor em Engenharia do Software, Analista da Embrapa Informática Agropecuária, silvio@cnptia.embrapa.br*
[4] *Doutora em Computação Aplicada, Pesquisadora da Embrapa Informática Agropecuária, silvia@cnptia.embrapa.br*
[3] *Doutor em Ciências da Computação, Pesquisador da Embrapa Informática Agropecuária, thiago@cnptia.embrapa.br*

A strategy to aid the identification of portfolios through text mining

Abstract

This paper presents a text mining strategy used to estimate some of the Embrapa Agricultural Informatics portfolios. The collected data came from the texts of the projects carried out by that Research Center between 2004 and 2010. A semi-automatic process to extract knowledge was applied to those data, using: controled vocabulary, statistical feature selection, hierarchical clustering, unique cluster labeling and user interactions. The obtained results were considered good through a subjective analysis; and they showed that the process can be applied to similar data.

Indexed terms: Portfolio, text mining, machine learning, feature selection, document cluster, cluster labeling.

Introdução

Na Wikipédia encontramos uma explanação bastante simples e elucidativa sobre portfólios (WIKIPEDIA... 2011). Nesta, eles são classificados como uma relação de trabalhos de um profissional ou uma empresa, ou seja, a coleção de todo o trabalho em andamento na organização, relacionado com o alcance de seus objetivos. E, ainda, é afirmado que toda organização tem um portfólio, mesmo que não o reconheça especificamente. No caso da Embrapa Informática Agropecuária, e de várias outras unidades da Empresa Brasileira de Pesquisa Agropecuária (Embrapa), os portfólios existem, mas nem sempre estão reconhecidos e classificados.

Retornando ao texto da Wikipédia, classes de portfólio poderiam ser um conjunto de aplicações no mercado de ações (portfólio de investimentos), projetos exploratórios de empresas de petróleo (portfólio exploratório), imóveis (portfólio de bens) ou um portfólio de quadros e fotografias, desde que destes espere-se algum tipo de retorno para a empresa ou profissional. Para algumas classes, é fácil observar o retorno, como no caso das três primeiras citadas; já para a última classe, o retorno é medido de forma indireta, tais como no campo social, de divulgação de uma imagem ou estilo pessoal, bem próximo ao caso dos portfólios de pesquisa e desenvolvimento, que acabam compondo bens maiores, no caso específico da Embrapa.

O caso da Embrapa Informática Agropecuária é bem particular, pois a computação por si só não é área fim da Embrapa e sim área meio, e, essa unidade não tem a característica de uma software house. De fato, a Embrapa Informática Agropecuária é uma unidade de pesquisa, desenvolvimento e inovação em informática para a agricultura. Portanto, os portfólios dessa unidade devem compor portfólios da empresa, em pequenas peças do quebra-cabeça total. Logo, identificar e classificar seus portfólios é um trabalho de montagem de parte desse quebra-cabeça.

Em primeiro lugar, deve-se classificar as subáreas da computação, grande área sob a qual classificamos o trabalho desenvolvido pela Embrapa Informática Agropecuária, de acordo com a taxonomia da Association for Computing Machinery (ACM) (ASSOCIATION FOR COMPUTING MACHINERY, 2011). Nessa taxonomia, em linhas gerais, tem-se as se-

guintes subáreas de interesse da Embrapa Informática Agropecuária: organização de sistemas de computadores, software, dados, computação matemática, sistemas de informação, metodologias da computação e aplicações. As três principais subáreas dessa taxonomia, nas quais a Embrapa Informática Agropecuária atua mais fortemente, são: metodologias da computação, matemática da computação e sistemas de informação. A subárea de metodologias da computação envolve manipulação algébrica, inteligência artificial, computação gráfica, processamento de imagens e visão computacional, reconhecimento de padrões, modelagem e simulação, e processamento de textos. A subárea de matemática da computação envolve análise numérica, matemática discreta, probabilidade e estatística e software matemático. A área de sistemas de informação envolve modelos e princípios, gerenciamento de bases de dados, armazenamento e recuperação da informação, sistemas de aplicações e interfaces, e apresentação da informação. Logo, identificar os possíveis portfólios dessa unidade é um trabalho de validação dessa classificação de subáreas e da aplicação dessas técnicas aos domínios de conhecimento e portfólios da Embrapa.

Para elencar os portfólios atuais e validá-los, poder-se-ia fazer uma leitura meticulosa de todas as últimas publicações da unidade e projetos submetidos ou em andamento. Decidiu-se por tentar encontrar padrões de maneira semiautomática entre os projetos finalizados, em andamento e submetidos nos últimos seis anos (entre 2004 e 2010). O ideal seria que todos esses projetos estivessem armazenados no mesmo sistema e no mesmo padrão de descrição, porém, devido a recentes mudanças nesses sistemas de controle dos projetos, procurou-se recuperar, da forma mais completa possível, os projetos. Sobre os projetos já recuperados, utilizou-se a metodologia TopTax (MOURA, 2009) com seu último conjunto de ferramentas implementadas – denominado TaxEdit (MOURA et al., 2010). Os resultados obtidos mostram que essa estratégia foi útil, pois apontou as principais áreas de atuação da Embrapa Informática Agropecuária, tanto no contexto geral da área de agricultura como de soluções tecnológicas de computação aplicadas à agricultura.

Assim, na seção material e métodos explica-se a montagem do processo como uma adaptação da metodologia TopTax, requisitos e materiais necessários. Na seção de resultados e discussão, além do experimento com os projetos da unidade, observou-se como a estratégia utilizada poderia

ser expandida. E, finalmente, nas conclusões também é elucidado como deveria ser o caso ideal.

Material e métodos

Nesta seção, explicações são fornecidas desde a ideia geral até a concepção final da metodologia utilizada. Assim, primeiramente explica-se a ideia básica, que vem da metodologia TopTax, e, então, como foram organizados os dados para serem trabalhados, como foi realizada a escolha do vocabulário a ser utilizado, como foi realizada a escolha de filtros de atributos para o modelo de aprendizado de máquina escolhido, como utilizar o modelo e editar seus resultados na TaxEdit, e, finalmente, a validação dos tópicos encontrados.

A metodologia TopTax

A TopTax é uma proposta metodológica para a construção de taxonomias de tópicos. A ideia de taxonomias de tópicos, nessa metodologia, é definida como uma hierarquia de assuntos de um domínio de conhecimento, obtida a partir de uma amostra de documentos desse domínio considerada satisfatória por um especialista ou grupo de especialistas desse domínio. À essa amostra de documentos, aplicam-se critérios de escolha de atributos e modelo de aprendizado para construir a taxonomia, obter seus descritores e trabalhá-los, e, por fim, escolhe-se um modelo de validação do processo. Algumas dessas escolhas, julgadas oportunas à época da publicação da metodologia, são elucidadas em Moura (2009). O esquema geral da metodologia é ilustrado na Figura 1

De acordo com a Figura 1, a TopTax é uma instanciação do processo de mineração de textos. Essa instanciação permite acoplar novos projetos individuais de pesquisa e desenvolvimento, obtendo-se um ambiente de testes de hipóteses e validação, que pode ser configurado de acordo com

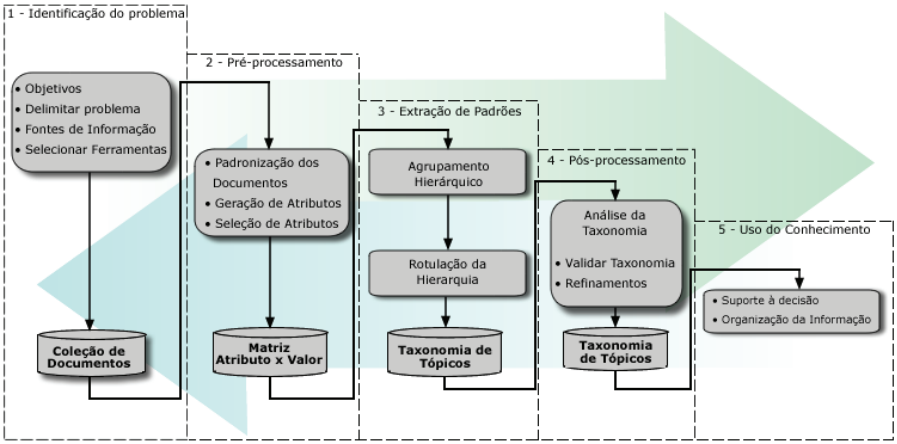

Figura 1. Metodologia TopTax.
Fonte: Moura (2009).

cada aplicação ou teste de métodos. As tarefas de cada etapa fornecem medidas de qualidade, permitindo que se decida pelo prosseguimento ou retorno a etapas do processo. As etapas são basicamente:

- **Pré-processamento**: após a escolha da amostra de documentos, aplicam-se procedimentos de geração ou extração de atributos, que podem ser: modelos puramente estatísticos de verificação de presença/ausência de palavras ou conjuntos de palavras nos textos; de construção de n-gramas, de acordo com algum critério de especificação deles, ou de remoção de inflexões das palavras; construção de n-gramas a partir de algum critério linguístico como uso de radicalização ou lematização; algum outro critério de especificação de expressões regulares; ou métodos híbridos de construção ou geração de atributos; ou, ainda, do uso de algum vocabulário controlado e de seu reconhecimento na amostra de documentos – ou seja, utilizando-se a intersecção das palavras encontradas no texto com o vocabulário controlado (PEIXOTO; MOURA, 2010). Obtidos os candidatos a atributos (termos, n-gramas, etc) nos textos é necessário aplicar-lhes algum filtro de seleção, pois sempre o número de atributos é alto e sua presença na amostra de textos é esparsa. Por fim, escolhe-se uma forma de representação da amostra de textos que possa ser lida por uma ferramenta de aprendizado de máquina. Na represen-

tação de matriz atributo-valor, utilizada na TopTax, cada documento da amostra corresponde a uma linha da matriz, cada atributo corresponde a uma coluna e, cada célula pode representar presença/ausência, frequência ou alguma outra medida de interesse.

- **Extração de padrões**: como o interesse é gerar uma estrutura hierárquica de relacionamento entre os documentos, nesta etapa pode-se utilizar agrupamento (escolhendo-se entre os diversos métodos disponíveis na literatura), análise de correspondência hierárquico, decomposição Latent Semantic Analysis e aplicar-lhe uma estruturação hierárquica, etc. A estrutura hierárquica produzida não tem uma leitura imediata e simples, assim, foi proposto o Robust Labeling Up Method (RLUM) (MOURA; REZENDE, 2010), para após obtida uma estrutura hierárquica, encontrar os descritores de cada grupo de forma única e concisa, bem como, aplicar podas à árvore original. Finalmente, precisa-se de formas de visualização da solução.

- **Validação**: como esta tarefa é uma análise exploratória de dados, cujo processo é cíclico, o ideal é que um especialista do domínio, junto a um especialista em métodos quantitativos avançados, a execute. Nesta etapa, o ambiente que implementa uma adaptação da TopTax, deverá fornecer ao especialista formas de interagir com os métodos, podendo selecionar partes do resultado como palavras-chaves, categorias pré-definidas ou palavras indesejáveis na análise (que serão retiradas desta), como na TaxEdit. O ciclo se repete até que o especialista esteja satisfeito com o resultado.

- **Uso do conhecimento**: o uso do conhecimento, desde o início do processo, define as adaptações necessárias e onde se almeja utilizar os resultados. Logo, no caso específico deste trabalho, quer-se chegar a indicações de portfólios.

Adaptações para a identificação dos portfólios

Neste item mostra-se como foi configurada a aplicação da metodologia TopTax à extração de tópicos que reflitam a área de atuação da Embrapa

Informática Agropecuária, individualmente. Ou seja, tópicos que indiquem as suas áreas de pesquisa e tecnologias desenvolvidas ou em desenvolvimento, bem como sua inserção aos portfólios de toda a Embrapa. Para isso, precisou-se definir a amostra de documentos, obter e selecionar atributos adequados, encontrar uma configuração satisfatória para a matriz atributo-valor, gerar e editar a taxonomia de tópicos, bem como interpretá-la e avaliá-la.

Amostrando os documentos

Primeiramente, separaram-se os documentos de interesse, escolhendo-se apenas os projetos liderados pela Embrapa Informática Agropecuária. Desta forma, desconsideraram-se aqueles outros projetos em que a unidade possui apenas planos de ação ou atividades. Outra questão foi considerar-se apenas os projetos em andamento ou não. O limitante dessa questão foi o sistema gerenciador de projetos, como houve algumas mudanças nos últimos anos, consideraram-se apenas os projetos que constam do Ideare (sistema de gerenciamento de projetos da Embrapa à época deste trabalho); e poucos outros, já encerrados ou submetidos, que puderam ser recuperados, contendo as seguintes partes:

- Título completo
- Palavras-chaves
- Resumo
- Objetivo geral
- Objetivos específicos
- Caracterização: podendo constar nesta a natureza da pesquisa, ecossistema, grandes temas, área e subárea e cadeia produtiva
- Detalhamento: contextualização, justificativa, hipóteses ou questões técnico-científicas, estratégia de ação, estratégia de gestão, metodologia e impactos.
- Resultados
- Metas

As informações listadas foram subjetivamente julgadas como suficientes para caracterizar projetos, metodologias e seus resultados. Desta forma, quarenta e oito projetos satisfizeram essas condições.

Escolha dos atributos

Como já mencionado, as áreas de atuação próprias da Embrapa Informática Agropecuária podem ser classificadas pela taxonomia da ACM. Porém, todo o domínio de aplicação é a agricultura e todos os projetos obedecem rigorosamente às classificações listadas pela Embrapa. Assim, decidiu-se, após algumas tentativas de diferentes soluções de extração e geração de atributos, trabalhar com um vocabulário controlado que refletisse a junção dessas áreas e subáreas.

Os termos para representar essa junção vieram: da taxonomia da ACM (traduzida para o português); do Thesagro, que é o thesaurus nacional de agricultura mantido pela Binagri (BINAGRI, 2011); e, da classificação utilizada pela Embrapa para as áreas e subáreas de conhecimento dos empregados e projetos da empresa utilizada no sistema Sistema de Informações sobre Recursos Humanos (SIRH). A junção dos termos correspondeu simplesmente à união dos n-gramas obtidos após seu pré-processamento pela ferramenta TaxEdit.

Na TaxEdit está implementado um processo de stemmização (remoção de inflexões) para construir n-gramas. N-gramas são palavras simples ou compostas, por exemplo: agricultura ou seu stem agricult é um unigrama; agricultura familiar ou a forma stemmizada agricult-famili é um bigrama, etc. A importância do uso desse processo de remoção de inflexões está na diminuição da dimensionalidade da matriz atributo-valor, pois cada stem corresponde a várias palavras que lhe são semanticamente próximas, evitando que se trabalhe com dimensões ainda mais altas. Após realizado esse processo tem-se o vocabulário controlado a ser utilizado para representar os documentos. Na Figura 2 mostra-se como selecionar os arquivos com os termos e gerar o vocabulário controlado na ferramenta TaxEdit.

A seguir, no processo, geram-se os n-gramas da amostra de documentos, da mesma forma que foi gerado o vocabulário controlado, ou seja, utilizando-se stemmização e combinando os stems até o número considerado no vocabulário controlado (clicando o botão Criar Vocabulário Controlado, ilustrado na Figura 2). Então, os atributos a serem considerados serão os n-gramas presentes na intersecção dos n-gramas da amostra de docu-

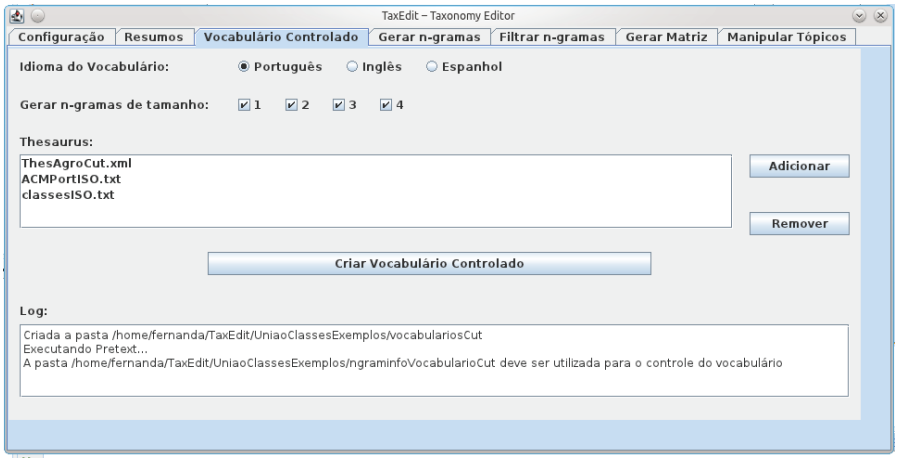

Figura 2. Gerando vocabulário controlado.

mentos e do vocabulário controlado. Para extrair tais n-gramas dos textos, especificam-se as opções da pasta Gerar N-gramas na TaqxEdit como ilustrado na Figura 3 .

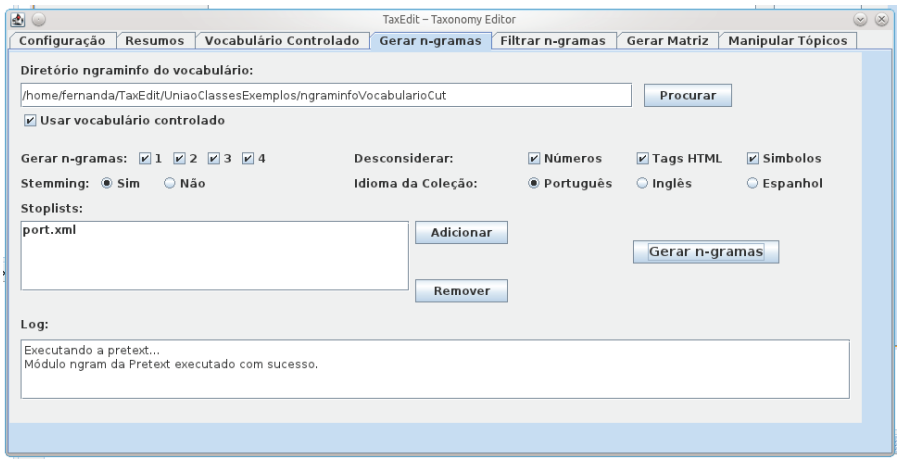

Figura 3. Interseccionando o vocabulário controlado com os n-gramas da amostra de documentos.

Filtros e representação da matriz atributo-valor

Apenas interseccionar os n-gramas pode não ser um filtro suficiente; o número de atributos, em geral, ainda é muito grande e a matriz atributo-valor muito esparsa. Para realizar uma redução simples e eficiente, pode-se calcular *stopwords* para a amostra de documentos, isto é, quais unigramas são estatisticamente insignificantes para discriminar esses documentos. Para isso, desenha-se um gráfico da frequência dos unigramas, e eliminam-se os unigramas com frequências muito altas e muito baixas. Sobre frequências muito baixas, não há acordo na literatura, então procura-se não utilizar cortes muito radicais, isto é, procura-se cortar apenas as frequências muito próximas a 1. Esses cortes são escolhidos subjetivamente. Na Figura 4, é ilustrado o gráfico com as frequências de todos os unigramas na parte a e, depois, na parte b, um corte aproximadamente executado nos pontos de inflexão da curva ilustrada na parte a. Se após isso, na TaxEdit for escolhido gerar *stopwords da coleção* (palavras a serem retiradas da análise dessa amostra de documentos), e elas forem usadas junto ao modelo de geração de n-gramas (conforme opções ilustradas na Figura 3, adicionar um novo arquivo de *stopwords* além do "port.xml"), então ter-se-ão cortes que interferem na formação dos n-gramas de modo geral[5].

Gerados os n-gramas, isto é, o conjunto de atributos de interesse, é necessário construir a matriz atributo-valor. Essa matriz é construída de acordo com o método que vai ser utilizado na extração de padrões, e, muitas vezes também se volta a esse passo para modificar a sua forma de construção, de acordo com os resultados que vão sendo observados na extração de padrões. No problema de identificação de portfólios, após várias explorações de resultados, optou-se por utilizar a medida tf-idf para relacionar os documentos e seus atributos e, então, gerar os agrupamentos. Essa medida provoca uma inversão dos pesos dos atributos mais frequentes e menos frequentes, pois multiplica-se a frequência no documento por uma razão inversa à frequência do atributo na coleção; isto é, o objetivo é atingir um grau de irrelevância para o atributo (denominado termo na área de recuperação de informação, por isso a letra t na medida), isto é, se o atributo aparecer em muitos documentos deve ser considerado menos im-

[5] *Para maiores detalhes em Moura et al. (2008).*

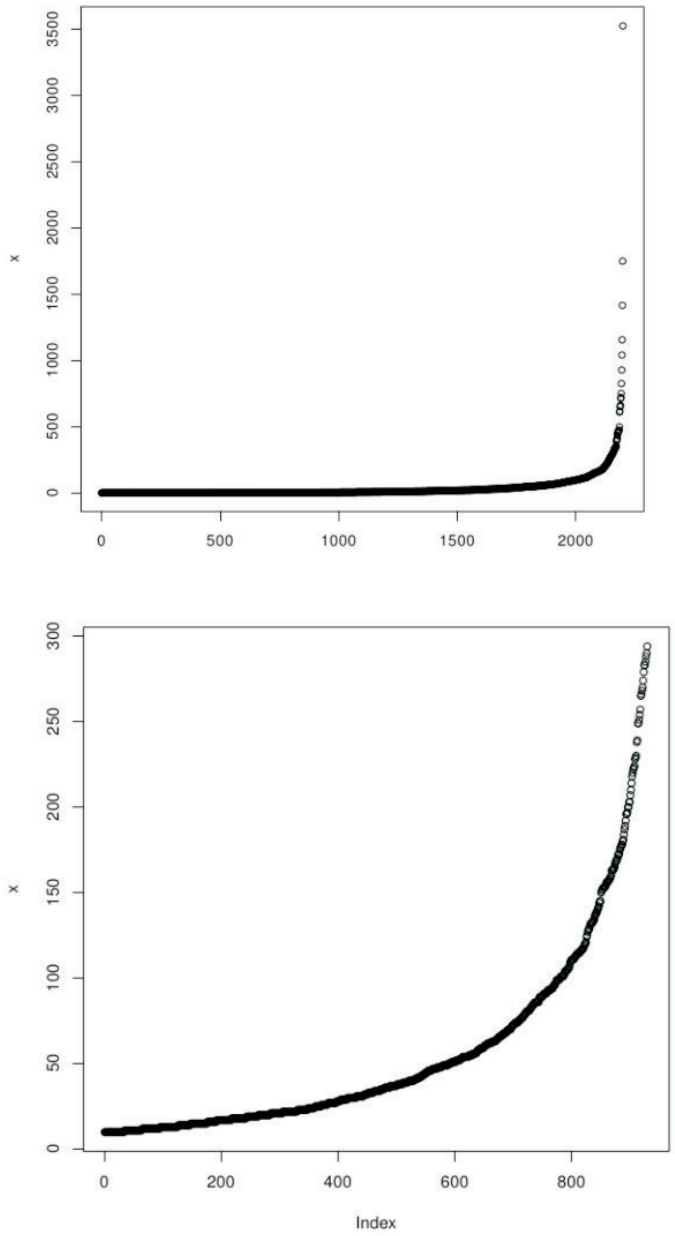

Figura 4. Parte a e parte b.

portante do que se ele aparecer em poucos documentos. O fator de escala faz com que o atributo que é pouco frequente na coleção seja mais notado pela sua medida final. Assim, atributos que aparecem frequentemente em um documento, mas não aparecem em vários documentos, são mais raros e recebem maior peso, pois podem ser mais interessantes à análise dos documentos que os muito comuns. Além disso, como a TaxEdit vem trabalhando com algoritmos de agrupamento hierárquico *bottom-up*, o ideal é que o atributo apareça em pelo menos dois documentos, para conseguir discriminar um grupo. Para especificar isso, para cada n-grama a ser colocado na matriz atributo-valor, especifica-se a "df ≥ 2", isto é, "*document frequence*" pelo menos dois. Na Figura 5, estão ilustradas as opções de especificação da construção da matriz atributo-valor.

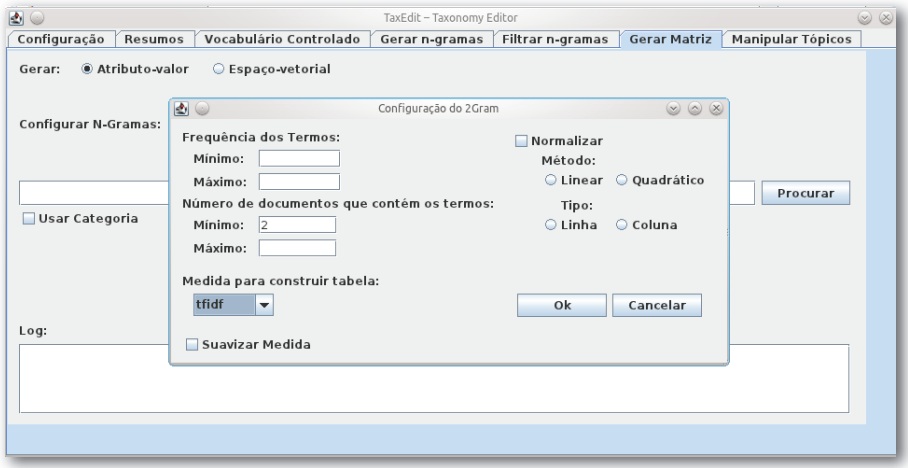

Figura 5. Especificando a configuração da matriz atributo-valor.

Extração do conhecimento

Para agrupar os documentos, utilizou-se um algoritmo de agrupamento hierárquico *bottom up*, com medida de similaridade dada por cosseno e utilizando a estratégia average, implementado pela biblioteca "*proxy*" do Software R (http://www.r-project.org/). A seguir, aplicou-se o algoritmo do RLUM, implementado na TaxEdit, a fim de se obter descritores únicos, concisos e sem repetição ao longo dos ramos da hierarquia. Além disso, o

RLUM promove podas da hierarquia original, de acordo com a independência estatística dos grupos.

Na Figura 6, observa-se o resultado do agrupamento logo após a aplicação do software R; e na Figura 7, o resultado após a aplicação do RLUM. Nos dois casos, os descritores foram ordenados pelos maiores valores de

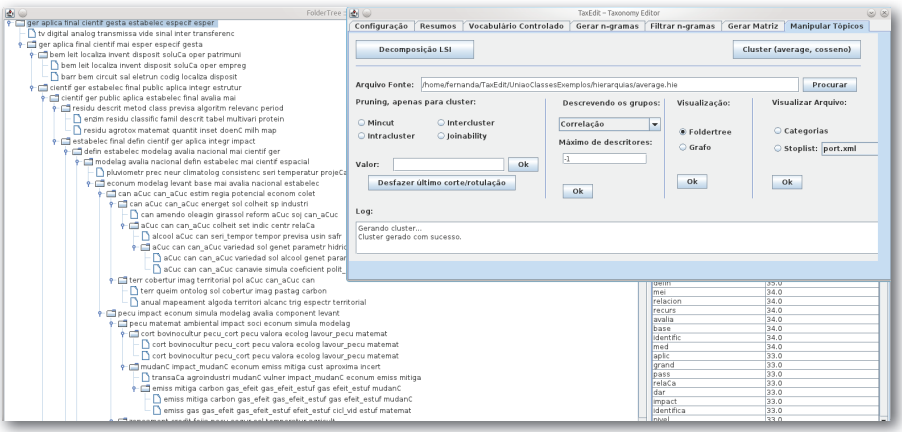

Figura 6. Hierarquia após a agrupamento e com seus descritores ordenados por correlação.

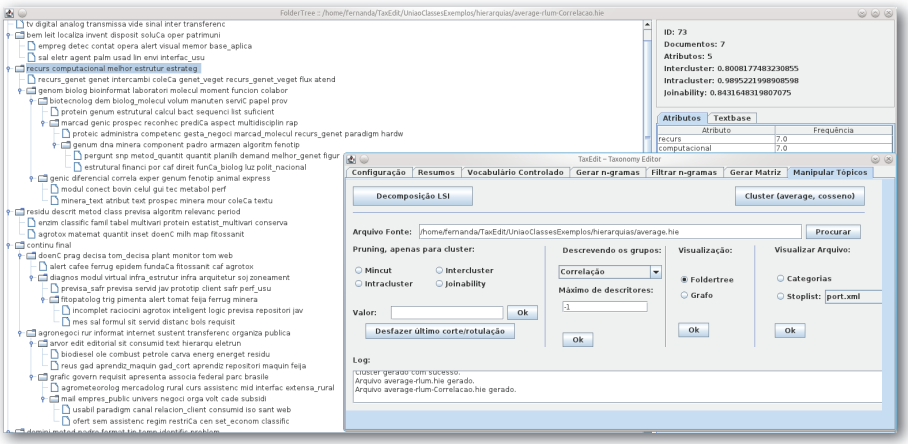

Figura 7. Hierarquia após aplicação do RLUM e com descritores ordenados pela correlação.

correlação entre os descritores encontrados e o agrupamento. Deve-se notar que a leitura e interpretação dos grupos na Figura 7 é um pouco mais clara que na Figura 6, pois, na primeira, a repetição de termos e o excesso de ramos não dão dicas claras sobre cada grupo, embora os ramos mais próximos à raiz forneçam uma boa ideia dos assuntos gerais tratados pela amostra de documentos.

Validação dos resultados obtidos

O processo de validação dos resultados é cíclico, ou seja, depende de várias interações do usuário com a ferramenta TaxEdit e de várias iterações do processo. O usuário precisa ter uma boa noção do domínio de conhecimento de onde a amostra é obtida e contar com algum conhecimento de análise exploratória de dados ou contar com o auxílio de outro usuário junto a ele com esse perfil. Isso permite identificar atributos pertinentes, ou não, à análise de dados. No caso dos portfólios, gestores da Embrapa Informática Agropecuária e pesquisadores da área de mineração de dados e textos executaram e validaram o processo conjunta e subjetivamente.

Resultados e discussão

O estudo de caso deste trabalho corresponde aos projetos da Embrapa Informática Agropecuária, considerando a premissa de que, se o processo é validado com esses projetos, o mesmo poderia ocorrer para todas as demais unidades da Embrapa, dado que a forma de proposição de projetos é padronizada em toda a empresa. Assim, o processo foi aplicado como mostrado na Figura 8, ou seja, iniciou-se com quarenta e oito projetos e foram indicados oito grandes grupos entre eles.

Para entender cada grupo, é necessário abrir cada ramo, individualmente, e, então, cada lista de atributos do ramo (verificar seus significado semântico, cada stem é associado à sua origem) e cada lista de documentos associada a cada ramo (observando os conteúdos dos documentos quando desejado); essa observação é detalhadamente ilustrada na Figura 9.

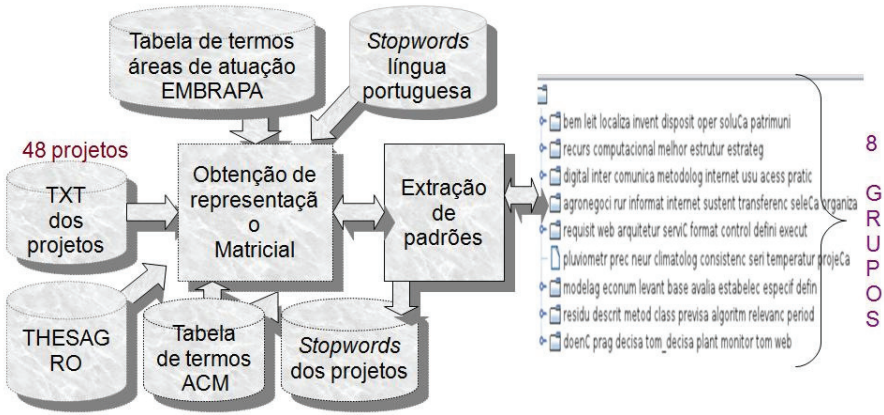

Figura 8. Processo para a Embrapa Informática Agropecuária.

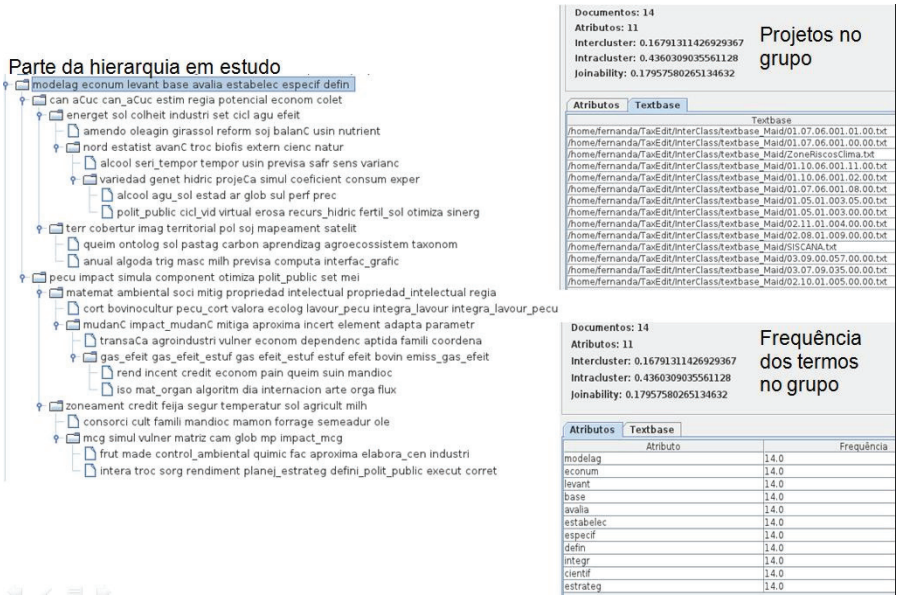

Figura 9. Observação detalhada de cada grupo obtido.

Os oito grupos identificados são melhor observados na Figura 10 e seus detalhamentos são desdobrados nos próximos itens.

Deve-se observar que, como mostrado na Figura 10, os grupos encontrados não se enquadram exatamente em categorias pré-definidas. Eles emergem naturalmente, mediante alguma junção por similaridade, dos dados analisados. Assim, os grupos encontrados correspondem à amostra de projetos coletados. Eles não necessariamente serão mantidos caso a amostra seja aumentada ou diminuída, pois dependem das relações de similaridade estabelecidas entre os dados amostrados.

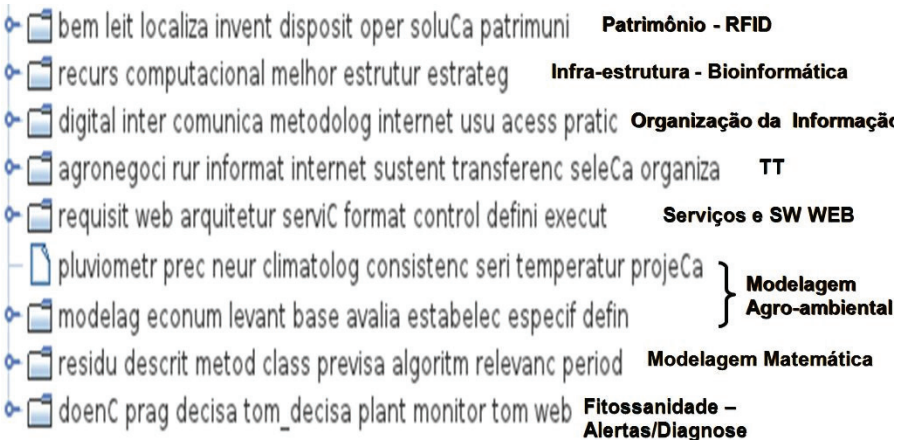

Figura 10. Focando os oito grupos encontrados.

Validação dos resultados esperados

Alguns resultados apresentaram-se entre os esperados, ou seja, conseguiu-se observar assuntos, áreas e subáreas de atuação na agricultura e as soluções das áreas e subáreas da computação associadas a elas. Por exemplo, na Figura 11, ilustra-se o assunto de modelagem agroambiental, dividido entre as áreas e subáreas presentes nos projetos liderados pela unidade. E na Figura 12, ilustram-se algumas indicações de áreas e subáreas da computação sob o tema de monitoramento, dentro de modelagem agro-ambiental.

A visualização dos resultados, embora apresente algumas facilidades na ferramenta, não se mostrou adequada para um relatório final. Assim,

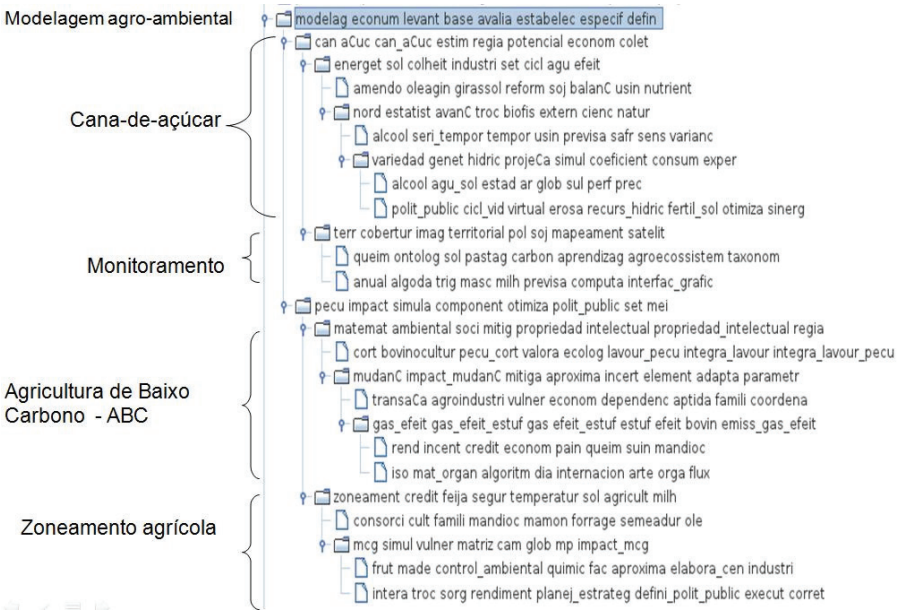

Figura 11. Divisão de grupos da modelagem agroambiental.

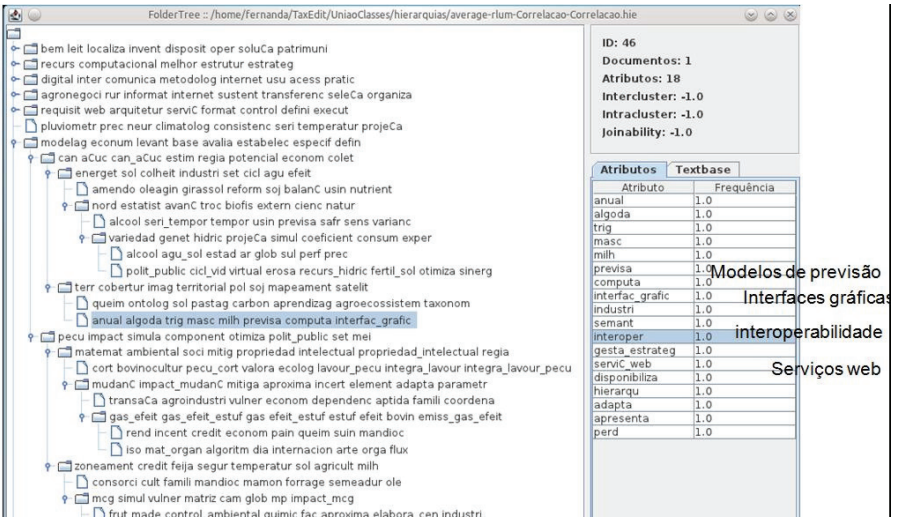

Figura 12. Algumas áreas/subáreas da computação na modelagem agroambiental.

após validar os resultados, optou-se por tabelá-los. As tabelas construídas refletem as áreas de atuação em agricultura na horizontal e, na vertical, as áreas e subáreas da computação relacionadas a cada qual; como ilustrado na Figura 13. Ainda, para melhor compreender como se relacionam os projetos propostos, também sua sigla é colocada na última coluna.

Portifólios EMBRAPA		Tecnologias específicas	Computação científica – análise de dados			Projetos
Modelagem Agro-ambiental	Cana-de-açúcar Culturas (soja, milho)	Potencial econômico Nutrientes Variedades Previsão de safra Políticas públicas	Bancos de dados Desenvolvimento de software	Séries temporais Modelos lineares generalizados Redes neurais	Modelagem e simulação	Reforma de Canavial Monitoramento Safra Parametrização CANEGRO SISCANA Simulação Cenários (Assad)
	Monitoramento	Cobertura territorial Imagens satélite	Bancos de dados Desenvolvimento de software	Representação do conhecimento - ontologias	Geoprocessamento	INTAGRO MAPAGRI
	ABC	Pecuária – emissão de gases Efeito estufa Mitigação de efeitos Políticas públicas	Bancos de dados Desenvolvimento de algoritmos	Análise de incertezas Modelos probabilísticos	Econometria	AVISAR SCAF BD-PECUS
	Zoneamento	Políticas de crédito/consórcios Agricultura familiar Mudanças climáticas globais (MCG) Simulação de cenários Controle ambiental	Bancos de dados Serviços web	Seleção de atributos	Modelagem e simulação Geoprocessamento Imagens satélite	Zoneamento riscos climáticos SCAF-Gestão SCAF-Cenários
Matemática Computacional		Análise relações proteínas/enzimas Resíduos de agrotóxicos Fitossanidade	Algoritmos Análise multivariada	Classificação Previsão Estatística descritiva	Modelagem e simulação Mineração de dados	Avaliação biomatemática STING

Figura 13. Exemplo da tabulação de resultados.

Assim, de modo geral, os resultados foram muito positivos e dentro do esperado; isto é, permitiram identificar, com bastante precisão, as competências técnicas das equipes da Embrapa Informática Agropecuária e suas relações com os portfólios da empresa. É nítido que a Embrapa Informática Agropecuária tem uma equipe multidisciplinar, o que viabiliza a condução de projetos voltados à área de agricultura, bem como viabiliza projetos com cunho mais computacional ou de matemática computacional; vê-se também que métodos quantitativos avançados fazem parte da maioria dos projetos. Os métodos quantitativos avançados, segundo a classificação da Embrapa correspondem à formulação de modelos de tratamento de informação, suporte à decisão e sistemas econômicos, sociais e naturais, todos com base em: modelagem e simulação, previsão, classificação, modelagem e simulação de cenários agrícolas, métodos quantitativos experimentais, geoestatística, estatística, mineração de dados, mineração de textos, economia e econometria.

Resultados não esperados

Embora os resultados tenham sido muito bons, eles também mostraram algumas falhas nas descrições de projetos da unidade. Essa observação, de forma alguma, denigre a análise realizada, apenas reforça pontos para futuras proposições de projetos. Há uma tendência muito evidente na descrição dos projetos de elucidar o que será feito em termos de pesquisa e desenvolvimento para a agricultura e, muito pouco, das inovações em computação, que ocorrem em muitos dos projetos. Essas inovações, ou pesquisas adaptativas, deveriam estar mais presentes e melhor descritas na metodologia e estratégia de ação dos projetos. De fato, o reflexo dessas pesquisas só é observado analisando-se a produção científica da unidade e, mesmo assim, timidamente, como apresentado no trabalho de Moura et al. (2011).

Talvez com o enfoque atual dado pela empresa em transferência de tecnologia e resultados de projetos voltados a ela, esse resultados inovativos passem a ter uma presença mais marcante nos textos dos projetos, além de serem timidamente contabilizados nos resultados.

Limitações observadas

Alguns problemas com a ferramenta implementada e formatação dos dados foram observados durante o processo, obrigando os analisadores dos dados a tomarem providências emergenciais para concluir o trabalho, como a construção manual da tabela de resultados.

A seleção de atributos na TaxEdit carece de uma melhoria de processo. É necessário integrar-lhe um processo melhorado de seleção de n-gramas e do uso de uma ferramenta mais robusta para extrair e gerar os n-gramas. Também seria útil a integração de uma visualização hiperbólica dos resultados, dado que, por vezes, a hierarquia é bastante extensa, e esse tipo de visualização é uma boa solução para esses casos. E, futuramente, a geração de tabelas cruzadas, dado que se conheçam, a priori, as possíveis classes a aparecerem em cada linha ou coluna.

Os dados precisam estar limpos, sem títulos como “Caracterização”, “Hipóteses”, etc, pois a frequência dessas descrições nos textos atrapalha a análise. Eles devem vir diretamente de uma consulta à base de dados do sistema Ideare, ou do sistema que for à época da análise. Ainda os textos devem estar em arquivos TXT codificados no padrão ISO e nenhum outro.

Conclusões

A estratégia adotada de extrair padrões de forma semiautomática dos projetos da Embrapa Informática Agropecuária para identificar seus portfólios mostrou-se bastante útil e promissora. Os portfólios puderam ser identificados, especialmente refletindo as competências técnicas das equipes da unidade.

Os resultados também evidenciam que a Embrapa Informática tem uma equipe multidisciplinar, o que tem viabilizado a condução de projetos voltados à área de agricultura, bem como de projetos com cunho mais computacional ou de matemática computacional e híbridos deles todos. Também foi evidenciado que que métodos quantitativos avançados fazem parte da maioria dos projetos e estão fortemente atrelados às competências das equipes.

Esses resultados indicam que se pode repetir esse tipo de análise para amostras de dados semelhantes com boas chances de sucesso. A experiência confirmou que os analisadores de dados devem ter um pouco de experiência em análise exploratória de dados e uma boa experiência com o domínio de conhecimento da amostra de dados.

Referências

ASSOCIAÇÃO FOR COMPUTING MACHINERY. **The ACM Computing Classification**: [1998 Version] valid through 2011. Disponível em: em:<http://dl.acm.org/ccs.cfm?part=author&coll=portal&dl=GUIDE>. Acesso em: 13 nov. 2011.

BINAGRI (Brasil). **Thesagro**. Disponível em: <http://snida.agricultura.gov.br:81/binagri/html/Cen_Thes1.html>. Acesso em: 10 de ago. 2011.

MOURA, M. F. **Contribuições para a construção de taxonomias de tópicos em domínios restritos utilizando aprendizado estatístico**. 2009. 137 f. Tese (Doutorado em Ciências de Computação e Matemática Computacional) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, SP.

MOURA, M. F.; PEIXOTO, B. M.; HIGA, R. H.; MASSRUHÁ, S. M. F. S. Uma proposta para a identificação de tendências de pesquisa e desenvolvimento em agroinformática. In: CONGRESSO BRASILEIRO DE AGROINFORMÁTICA, 8., 2011, Bento Gonçalves. **Anais**... Florianópolis: UFSC; Pelotas: UFPel, 2011. Não paginado. SBIAgro 2011.

MOURA, M. F.; NOGUEIRA, B. M.; CONRADO, M. S.; SANTOS, F. F.; REZENDE, S. O. Making good choices of Non-Redundant N-gramwords. In: INTERNATIONAL WORKSHOP ON DATA MINING AND ARTIFICIAL INTELIGENCE, 1.; INTERNATIONAL CONFERENCE ON COMPUTER AND INFORMATION TECHNOLOGY, 11., 2008, Khulna. **Proceedings**... Los Alamitos: IEEE Computer Society, 2008. v. 1. p. 64-71. ICCIT 2008.

MOURA, M. F.; MERCANTI, E.; PEIXOTO, B. M.; MARCACINI, R. M.; TAMADA, T.; LIMA, A. F.; SANTOS, F. F. dos. **TaxEdit - Taxonomy Editor V 2.0. Versão 1.0**. Campinas: Embrapa Informática Agropecuária, 2011. 1 CD-ROM.

PEIXOTO, B. M.; MOURA, M. F. Análise histórica de tópicos de publicações em agroinformática. In: MOSTRA DE ESTAGIÁRIOS E BOLSISTAS DA EMBRAPA INFORMÁTICA AGROPECUÁRIA, 6., 2010, Campinas. **Resumos**... Campinas : Embrapa Informática Agropecuária, 2010. p. 23-26.

WIKIPEDIA, Potfólios, Wikipedia a Enciclopédia Livre. Disponível em: <http://pt.wikipedia.org/wiki/Portf%C3%B3lio>. Acesso em: 3 nov. 2011.

Embrapa

Informática Agropecuária

Ministério da
Agricultura, Pecuária
e Abastecimento

GOVERNO FEDERAL
BRASIL
PAÍS RICO É PAÍS SEM POBREZA

CGPE 9777