UMA PROPOSTA PARA A IDENTIFICAÇÃO DE TENDÊNCIAS DE PESQUISA E DESENVOLVIMENTO EM AGROINFORMÁTICA

MARIA FERNANDA MOURA[1]
BRUNO MALVEIRA PEIXOTO[2]
ROBERTO HIROSHI HIGA[3]
SÍLVIA MARIA FONSECA MASRHUÁ[4]

RESUMO: Neste trabalho experimenta-se a identificação semi-automática das linhas de pesquisa e desenvolvimento em agroinformática refletidas pelos principais fóruns da área. Para isso, tomaram-se as publicações textuais desses fóruns, disponíveis em formato eletrônico, e aplicou-se um processo de mineração de textos. O processo de mineração de textos foi adaptado para esse fim, para permitir o uso de vocabulário controlado da área e o julgamento subjetivo de especialistas do domínio de conhecimento mediante os resultados semi-automaticamente obtidos. Os resultados apresentados, embora parciais, nitidamente mostram que as tendências concentram-se em modelos de simulação, geoprocessamento e análise de dados, em geral, em áreas como análise de mercado e modelagem agroambiental.

PALAVRAS-CHAVE: mineração de textos, análise exploratória de dados, agrupamentos, *thesaurus*, taxonomias, agroinformática

A PROPOSAL FOR TREND IDENTIFYING IN AGRICULTURE INFORMATICS RESEARCH AND DEVELOPMENT

ABSTRACT: This paper presents an experiment on semi-automatic identification of research and development areas in agricultural informatics based on the most expressive forums. To reach this goal, a text mining process was applied on the textual publications available in electronic format. The text mining process was customized to support the use of controlled vocabulary and the subjective judgment by domain specialists in a feedback process. Although the results are not definitive, they present a clear tendency in simulation models, geographic information systems and data analysis, usually concerning agricultural market and environmental agriculture modeling.

KEYWORDS: text mining, exploratory data analysis, clustering, *thesaurus*, taxonomy, agricultural informatics

1. INTRODUÇÃO

A área de informática agropecuária, vista como a tecnologia da informação aplicada a um domínio específico do conhecimento, tem despertado um interesse crescente, que se reflete no aumento ano a ano do número de fóruns científicos e de aplicação em informática agropecuária. Na Europa tem-se o congresso da *European Federation for Information Technologies in Agriculture, Food and the Environment* - EFITA desde 1997. Na Ásia, o congresso da *Asian Federation for Information Technologies in Agriculture* - AFITA, cujas edições podem ser recuperadas eletronicamente. No Brasil, a Sociedade Brasileira de

[1] Doutora em Ciências Mat. e da Computação, Embrapa Informática Agropecuária, fernanda@cnptia.embrapa.br
[2] Mestrando em Ciências da Computação, Instituto de Computação, UNICAMP, brunompeixoto@gmail.com
[3] Doutor em Engenharia Elétrica, Embrapa Informática Agropecuária, roberto@cnptia.embrapa.br
[4] Doutora em Ciências da Computação, Embrapa Informática Agropecuária, silvia@cnptia.embrapa.br

Agroinformática promove o seu congresso bianual, cujos dados de evolução da contribuição em publicações das instituições brasileiras públicas e privadas foram apresentados no congresso de 2009 (Lopes et al, 2009), mostrando o quão significativo este fórum vem se tornando para o agronegócio brasileiro; apesar de ainda ser baixa a participação da iniciativa privada. Além disso, a Embrapa Informática Agropecuária atua desde 1985 especificamente nessa área, bem como alguns outros centros de pesquisa e desenvolvimento da Empresa Brasileira de Pesquisa Agropecuária.

Dessa forma, para realizar uma análise das tendências em pesquisa e desenvolvimento em agroinformática no Brasil e no mundo, considerou-se como representativas as publicações dos congressos da EFITA e AFITA contra as do SBIAgro e a produção científica da Embrapa na área. Com essa análise esperava-se comparar a produção científica nacional e, em particular a da Embrapa, com a produção mundial nesse tema; isto é, quais seriam os pontos que mereceriam maior atenção, fosse pela competitividade ou falha, e quais seriam as principais diferenças.

O primeiro problema que se apresentou foram as diferenças entre os tópicos e sub-tópicos desses eventos científicos e tecnológicos. Além disso, em uma pequena amostra aleatória de publicações, verificou-se, subjetivamente, que na descrição das publicações não eram utilizadas palavras-chaves específicas de algum *Thesaurus*. Portanto, não seria possível simplesmente categorizar as publicações e verificar as freqüências dos tópicos ou palavras-chaves ao longo dos anos, para poder compará-los. Para contornar essa dificuldade, fez-se necessário encontrar quais seriam as categorias comuns a toda a coleção de artigos, por meio de um processo de mineração de textos.

No item material e métodos, apresenta-se a metodologia empregada para obter e comparar o desempenho das categorias citadas nas publicações. No item experimentos e resultados, descreve-se como foram selecionadas as bases de artigos comparadas, o processo experimental e os resultados obtidos. A seguir, discorre-se sobre as principais limitações do experimento realizado e possíveis limitações da interpretação dos resultados; e, finalmente apresentam-se algumas linhas de trabalhos futuros.

3. MATERIAL E MÉTODOS

Um processo de mineração de textos foi adaptado (Peixoto e Moura, 2010) para que se pudesse identificar termos em comum entre as diversas publicações, considerando-se que algumas se encontram em inglês e outras em português. A primeira padronização, no processo, foi separar as coleções por línguas e utilizar dois *thesaurus*: o AGROVOC (2010) e o THESAGRO (BINAGRI, 2010). Foram considerados apenas os descritores marcados, em cada *thesaurus*, de modo que não houvesse repetições de vocábulos e utilizaram-se apenas unigramas, bigramas e trigramas; pois nas línguas e *thesaurus* considerados não ocorriam combinações superiores de n-gramas. Para obter-se os termos da área de computação foi utilizada a taxonomia da ACM (IEEE, 2011); e, como esta é disponível apenas em inglês, realizou-se também uma tradução para português (como utilizado no Brasil e em Portugal).

No primeiro passo do processo, separam-se os descritores (dos *thesaurus*), removem-se as *stopwords* de português e inglês, e então aplica-se a ferramenta PreText (Soares et al, 2008) para *stemmizar* os descritores e obter os unigramas, bigramas e trigramas. Esse conjunto de unigramas. bigramas e trigramas, neste trabalho, é considerado o **vocabulário controlado**.

No segundo passo do processo, identificam-se estatisticamente os vocábulos presentes nos artigos; isto é, pela presença ou ausência de algum vocábulo nos textos. Novamente utiliza-se a PreText (Soares et al, 2008) para retirar as mesmas *stopwords,* aplicar o mesmo processo de *stemmização* e obter unigramas, bigramas e trigramas. Esse conjunto de unigramas. bigramas e trigramas, neste trabalho, é considerado o **vocabulário obtido**. Como o número de vocábulos obtidos no segundo processo é bem maior e mais variado que o vocabulário controlado; no terceiro passo, verifica-se qual a intersecção entre esses dois vocabulários e fica-se apenas com a sua intersecção. No quarto passo, verifica-se quantas vezes cada

vocábulo da intersecção aparece em cada documento da coleção e constrói-se uma matriz de incidência, na qual as linhas correspondem aos documentos e as colunas aos vocábulos. No quinto passo, clusteriza-se a matriz de incidência, com o algoritmo *average linkage* e utlizando similaridade de cosseno; então, corta-se o agrupamento usando as recomendações de Marcacini et al (2009) e a seguir rotulam-se os agrupamentos com o método de Moura e Rezende (2010); produzindo uma possível taxonomia de tópicos da coleção de textos em análise. No sexto passo, um especialista do domínio irá analisar subjetivamente os resultados e editar as categorias indicadas em cada nó do agrupamento. É interessante que o especialista tenha uma boa noção de conceitos como vocabulário controlado e uso de *thesaurus*, ou conte com a ajuda de um profissional da área de informação. Nesse passo são, subjetivamente, classificados os termos que podem ser considerados *stopwords* de domínio, categorias de modelos de software ou análise de dados e categorias de áreas de aplicação – como foi considerada, subjetivamente, a divisão da agro-informática neste trabalho. Enquanto o epecialista do domínio não estiver satisfeito com a taxonomia produzida, ele repete todo o processo, desde o terceiro passo. Embora, ao final do processo, as tendências em tópicos fiquem evidentes na taxonomia, para observá-las mais resumidamente e, o seu desempenho ao longo dos anos, decidiu-se por utilizar também gráficos de freqüências. Para isso, calcularam-se os quartis das freqüências dos tópicos por área de aplicação e por modelo computacional, considerando-se como mais significativos apenas os tópicos que aparecem no último quartil - superiores a 75%.

4. RESULTADOS E DISCUSSÃO

A base de artigos para ser utilizada no processo precisa estar em formato digital, de modo que possa ser convertida a texto plano não formatado, e, então se possam identificar automaticamente vocábulos e trabalhar os conteúdos dos textos. Dessa forma, foram considerados apenas os anais de congresso que satisfizeram essas condições: AFITA de 1998, 2000, 2002, 2009; EFITA 1997, 1999, 2001, 2003, 2005, 2007 e 2009; e, SBIAGRO 2003, 2005, 2007 e 2009. Considerou-se apenas a produção científica da Embrapa Informática Agropecuária até 1997 e, depois, de dois em dois anos até 2009 - para que os intervalos fossem mais ou menos compatíves com os dos congressos. Para todas as coleções de artigos separaram-se apenas os metadados de título, resumo e palavras-chaves. Nesse experimento disponibilizou-se de apenas dois especialistas de domínio para gerar e editar a taxonomia. Ambos da área de desenvolvimento de software e inteligência artificial, e, um deles também estatístico e com vivência em organização da informação.

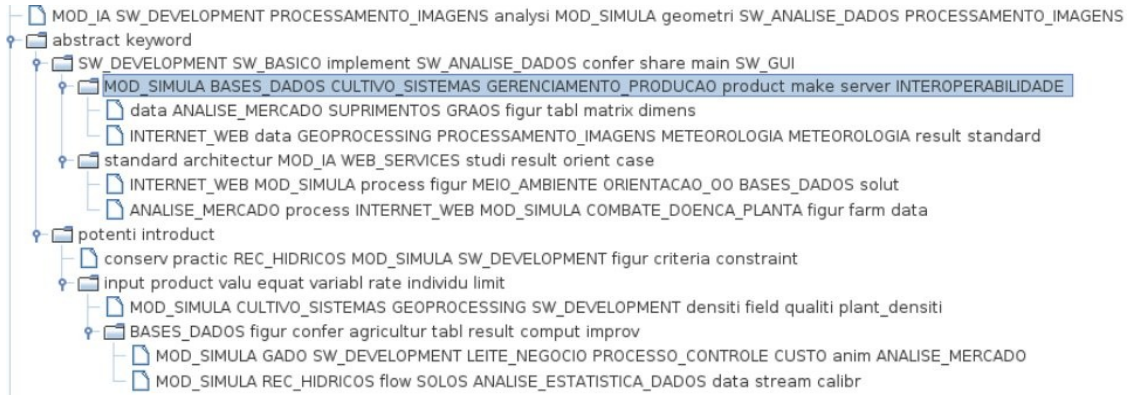

Figure 1: Parte da taxonomia de tópicos do EFITA 2009

Na Figura 1 é ilustrada uma parte da taxonomia obtida para a coleção EFITA 2009. Embora as tendências em tópicos estejam presentes na taxonomia, é difícil visualizá-las como um todo; pois algumas categorias se repetem em ramos diferentes, junto a outras informações. E, vê-

se que a taxonomia é incompleta, pois alguns ramos apresentam vocábulos soltos que não foram categorizados, ou que talvez devessem compor alguma categoria. Por exemplo, o ramo hachurado, MOD_SIMULA (model simulation) poderia ser CULTIVO_SISTEMAS (crop systems), que representa uma classe específica de "simulation_models_for_crop_systems“. Esse julgamento subjetivo dos especialistas é importante para evitar o mal uso de algum termo; por exemplo, "warehousing“ como termo agrícola significa silo ou armazém, porém na computação o significado remete ao uso de técnicas e ferramentas inteligentes de modelagem e anáĺise de dados. A categoria DATA_WAREHOUSING, por exemplo, foi formada por "armaz_dad_softw; planej_qualidad_softw; metod_qualidad_softw; armazem_dados_softwar; planejamento_qualidad_softwar" e, GEOPROCESSING por "geoprocessing; geoprocess; sistem_informa_geograf; geoprocess; geoprocessamento; sistema_informaCao_geografica". Dessa forma, ao todo foram encontradas noventa categorias para áreas de aplicação e quarenta e nove para modelos computacionais.

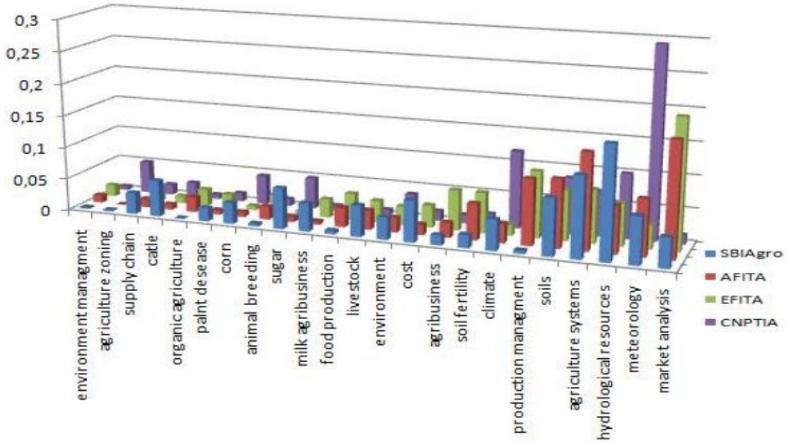

Figure 2: Tendências em áreas de aplicação

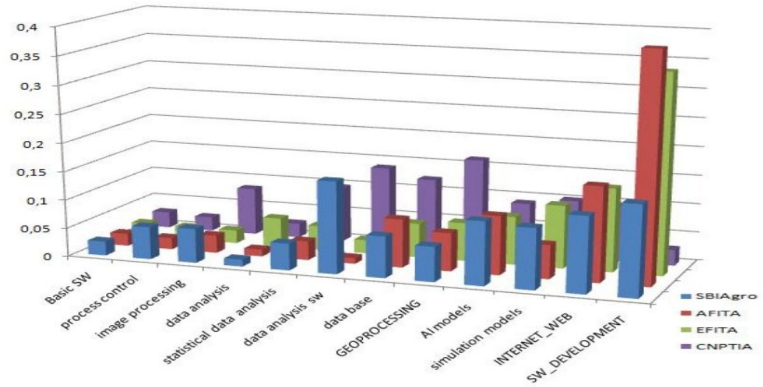

Figure 3: Tendências em modelos computacionais

Ao calcular os dois últimos quartis entre as categorias encontradas, obtiveram-se os gráficos das figuras 2 e 3. Sabendo-se, por experiência própria, que os resultados obtidos mais ou menos correspondem ao que se observa na prática, vemos que para a Embrapa Informática Agropecuária (no gráfico CNPTIA) nos itens *meteorology*, *hydrological resources* e *climate* sobressai-se aos demais. Deve-se lembrar, no entanto, que o EFITA, o AFITA e o SBIAGRO não são os principais fóruns dessas áreas, logo, não poderiam bem representá-las. O tema agricultura de precisão, que está sendo pela primeira vez focado no SBIAGRO, também não está representado nos gráficos acima, dado que também tem fóruns específicos ou aparece dentro do escopo de outros temas como *agriculture systems and soils* que tem grande representatividade de artigos no SBIAGRO. Para as demais áreas observa-se uma certa compatibilidade, exceto por *corn* e *sugar*, que são fortemente presentes em publicações brasileiras. Em relação a modelos computacionais, na análise resumida no gráfico da Figura 3,

é nítido que as publicações da Embrapa (CNPTIA) valorizam mais a análise de dados, o processamento de imagens, o geoprocessamento, a modelagem e simulação e a inteligência artificial, em detrimento ao desenvolvimento de software, embora também desenvolva software. Já nos congressos da área de agroinformática há uma maior concentração no tópico de desenvolvimento de software em detrimento aos demais modelos computacionais e de análise de dados.

5. CONCLUSÕES E TRABALHOS FUTUROS

A análise realizada possui alguns limitantes que precisariam ser minimizados, como a questão do número de especialistas de domínio, que julgaram as taxonomias obtidas, e a questão de polissemia no vocabulário obtido. Futuramente devem-se incluir especialistas de agricultura e profissionais de informação ou terminologia do domínio entre os especialistas da área, para obter taxonomias mais precisas. Ainda, para identificar e estudar melhor a tendências, aumentar tanto a amostra de publicação da Embrapa, considerando outros centros que também atuam em agroinformática (embora não seja missão dos mesmos), e aumentar a amostra de congressos, para cobrir os de geoprocessamento, meteorologia e mudanças climáticas. E, finalmente, cruzar as informações de área de aplicação e modelos computacionais, a fim de verificar se as tendências se conservam ou diferem significativamente. Além disso, pretende-se realizar uma análise cruzada da participações de instituições públicas e privadas nessa publicações, similarmente ao trabalho realizado por Lopes et al (2009), porém para todos os fóruns considerados neste trabalho. Assim, novas análises bi-anuais serão realizadas, a fim de se aprimorar e atualizar os resultados obtidos até aqui, bem como o ferramental utilizado.

6. REFERÊNCIAS

AGROVOC, FAO. Agriculture Information Managment Standards – AIMS. AGROVOC – Agriculture Vocabulary. In: http://aims.fao.org/website/AGROVOC-Thesaurus/sub ou http://www.icpa.ro/AgroWeb/AIC/RACC/Agrovoc.htm. December, 2010.

BINAGRI, Biblioteca Nacional de Agricultura . Thesaurus Agrícola Nacional – Thesagro. In: http://www.agricultura.gov.br/portal/, December, 2010.

IEEE Computer Society **ACM Taxonomy. In:** IEEE Computer Society – Keywords. Disponível em: http://www.computer.org/portal/web/publications/acmtaxonomy, 09/05/2011.

LOPES, M.A. GARREFA, F.H.R. PAULA, J.A. Um Estudo da Evolução da Quantidade de Artigos Publicados nos Congressos da Associação Brasileira de Agroinformática. In: Congresso da Associação Brasileira de Agroinformática (SBIAGRO), 2009, Viçosa-MG. Anais do Congresso da SBIAGRO, 2009.

MARCACINI, R.M. ; MOURA, M. F. ; REZENDE, S. O. Uma Abordagem para Seleção de Grupos Significativos em Agrupamento Hierárquico de Documentos. In: Encontro Nacional de Inteligência Artificial, 2009, Bento Gonçalves - RS. Anais do VII Encontro Nacional de Inteligência Artificial (ENIA), 2009.

MOURA, M. F. ; REZENDE, S. O. A Simple Method for Labeling Hierarchical Document Clusters. In: International Conference on Artificial Intelligence and Applications, 2010, Innsbruk - Austria. Proceedings of the 10th IASTED International Conference on Artificial Intelligence and Applications. Anaheim, Calgary, Zurich: Acta Press, 2010. v. 1. p. 336-371.

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.

SOARES, M. V. B., PRATI, R. C. MONARD, M. C. PreTexT II: Descrição da Reestruturação da Ferramenta de Pré-Processamento de Textos. Relatório Técnico, Instituto de Ciências Matemáticas e da Computação, ICMC/USP, São Carlos, n.333, 2008. In:http://www.icmc.usp.br/~biblio/BIBLIOTECA/rel_tec/RT_333.pdf.