Tendências de pesquisa em Agroinformática na Argentina: uma análise histórica

Alfredo Parteli Gomes[1], Sandro da Silva Camargo[2], Yanina Bellini Saibene[3]

1 Campus Santana do Livramento, Instituto Federal Sul-rio-grandense (IFSul),
Santana do Livramento, Rio Grande do Sul, Brasil
alfredogomes@ifsul.edu.br

2 Programa de Pós-Graduação em Computação Aplicada,
Universidade Federal do Pampa & Embrapa Pecuária Sul
Bagé, Rio Grande do Sul, Brasil
sandro.camargo@unipampa.edu.br

3 Estación Experimental Agropecuaria Anguil,
Instituto Nacional de Tecnología Agropecuaria (INTA)
Anguil, La Pampa, Argentina
yanina.bellini@inta.gov.ar

Resumo. Nos dias atuais, dentro do contexto de Big Data, ocorre um crescimento exponencial das fontes de dados, principalmente com dados não estruturados, tais como documentos de texto. Neste cenário, quando se necessita de abordagens que permitam a visualização e análise rápidas de documentos de texto, cresce o uso de nuvens de palavras, as quais permitem apresentar, em uma figura, os termos mais relevantes que compõe os documentos, a partir da frequência de ocorrência destes termos. A fim de analisar as tendências de pesquisa em agroinformática na Argentina, este trabalho faz uma análise visual dos títulos dos trabalhos publicados nos anais das últimas seis edições do Congresso de Agroinformática (CAI), ocorridas entre 2010 e 2017. A abordagem aqui utilizada permitiu identificar os termos de pesquisa mais relevantes em cada uma das edições do CAI, possibilitando a identificação de crescimento ou redução da importância destes termos ao longo dos anos.

Palavras-Chave: nuvens de palavras, tagcloud, visualização de documentos, visualização de dados.

1 Introdução

Os últimos anos foram marcados por inúmeras inovações nas tecnologias de coleta, armazenamento e compartilhamento de dados. Tais dados vêm sendo gerados por pesquisas científicas, internet, smartphones, empresas, governos e outras fontes [6]. No âmbito da pesquisa científica, dados se consolidaram como um dos pilares da ciência moderna, fazendo com que o processo científico orientado a dados dependa da análise rápida e acurada de dados experimentais, sejam eles gerados por observações

ou por simulações [2]. Neste ambiente de constantes desafios, emergiu o conceito de Big Data, que é uma das mais importantes tendências tecnológicas atuais, mostrando potencial para gerar conhecimento a partir de grandes quantidades de dados [1]. Dentre as características comuns ao conceito do Big Data, além da velocidade de geração e do volume, está a variedade dos dados gerados , geralmente marcada pela presença de dados não estruturados, tais como documentos de texto puro.

A fim de avaliar as tendências de pesquisa em agroinformática na Argentina, foi delimitado o contexto do Congresso Argentino de Agroinformática (CAI) que é o principal evento científico do país sobre este tema. No CAI[1] participam pesquisadores, técnicos, desenvolvedores e empresas relacionadas com o setor agroindustrial e são apresentados trabalhos sobre as Tecnologias de Informação e Comunicação (TIC) aplicadas a problemáticas agropecuária, agroindustrial e ambiental, englobando instâncias desde experimentais até comerciais. O CAI é um evento anual que está em sua décima edição e é promovido pela Sociedade Argentina de Informática e Investigação Operativa (SADIO[2]). Neste trabalho, foram analisados os seis últimos anais de edições do CAI, que ocorreram nos anos de 2010, 2011, 2013, 2014, 2016 e 2017, e foram sediados em Buenos Aires, Córdoba, Córdoba, Buenos Aires, Buenos Aires e Córdoba, respectivamente. No ano de 2012 e 2015 não foram gerados anais do CAI, motivo pelo qual estes anos não são considerados neste estudo.

A literatura relata diversas abordagens para avaliação de tendências em pesquisa. Hunt et al. [5] geraram uma análise visual de tendências dos títulos dos artigos publicados nas quatro revistas com ranking mais alto na área de turismo. A pesquisa foi realizada abrangendo quatro décadas, com amostragem nos anos de 1982, 1992, 2002 e 2012. A técnica de visualização utilizada foi a de nuvens de palavras. Haugerud [3] também realizou a análise dos títulos e palavras chave da revista American Ethnologist, analisando quatro décadas com amostragens a cada dez anos. Através do uso de nuvens de palavras, os autores conseguiram identificar as palavras dominantes em cada uma das quatro décadas analisadas. Hearst e Rosner [4] evidenciam o uso de nuvens de palavras para encontrar assuntos pelos quais grupos pessoas estão interessados, e como estes assuntos variam ao longo do tempo. Os autores ainda argumentam sobre a importância da visualização através de nuvens de palavras como sinais ou marcadores de interações entre indivíduos ou grupos sociais e conteúdos de informação.

Com base nas análises da literatura correlata, este trabalho tem como objetivo a identificação e apresentação dos tópicos de pesquisa mais relevantes nas últimas seis edições do Congresso Argentino de Agroinformática, através do uso da técnica de nuvem de palavras para visualização dos títulos dos artigos publicados nos respectivos anais. Como objetivo secundário, serão analisadas as frequências das palavras nas nuvens de palavras de cada ano, a fim de identificar as tendências de pesquisa em cada um dos tópicos identificados como relevantes nas seis edições do Congresso. A partir desta análise de tendência, serão apresentados e discutidos os tópicos com tendência crescente de relevância. Como impactos esperados a partir dos resultados

[1] Site: http://47jaiio.sadio.org.ar/index.php?q=cai
[2] Site: http://www.sadio.org.ar/

apresentados neste trabalho, pretende-se tornar claras as tendências de pesquisa na agroinformática Argentina, a fim de contribuir com a escolha de pesquisadores que tenham interesse em desenvolver trabalhos na área.

O restante do artigo está organizado da seguinte forma: A Seção Material e Métodos apresenta uma breve descrição das características da base de dados utilizada para este estudo e as abordagens utilizadas para Análise de Dados. A Seção Resultados e Discussão apresenta e discute os resultados obtidos a partir da análise dos dados, das nuvens de palavras e análises de tendência. A Seção Conclusões apresenta um resumo das descobertas, as restrições da abordagem utilizada e as perspectivas de trabalhos futuros.

2 Material e Métodos

Este estudo analisou os trabalhos publicados nos seis últimos anais do Congresso Argentino de Agroinformática, ocorridos nos anos de 2010, 2011, 2013, 2014, 2016 e 2017. Estes anais estão disponíveis nas respectivas páginas web dos eventos. Os dados utilizados foram os títulos dos trabalhos publicados em cada uma das edições do evento. A metodologia utilizada foi a seguinte: 1) Obtenção dos anais da edição do evento, 2) Criação de um arquivo em texto puro, para os anais de cada edição, contendo os títulos de todos os trabalhos publicados, 3) Remoção de *stop words*, tais como artigos, preposições, pronomes e conjunções. Tais palavras precisam ser removidas porque tem a tendência de aparecerem muitas vezes e não terem relevância enquanto elementos relevantes no título. 4) Contagem no número de vezes que cada palavra aparece nos títulos de uma determinada edição. Como resultado tem-se uma lista ordenada das palavras e a respectiva contagem de vezes que elas apareceram no texto e, como consequência, tendem a ser mais relevantes enquanto termos de pesquisa [5]. 5) Transformação da contagem absoluta dividindo-a pela quantidade de artigos publicados no evento. Esta atividade precisou ser realizada em função da grande diferença da quantidade de artigos publicados em cada um dos anais. Como exemplos dos valores extremos, o ano de 2013 teve 17 trabalhos publicados. Já no ano de 2016, foram 37 trabalhos. 6) A partir da razão entre a contagem absoluta de menções às palavras e a quantidade de artigos publicados neste ano, foram gerados novos valores representando a frequência relativa das palavras em cada evento. 7) Para cada palavra, foi criado um modelo de regressão linear, a fim de identificar o coeficiente angular do modelo visando inferir a tendência de utilização do termo, com base nas frequências relativas. Coeficientes angulares positivos indicam uma tendência crescente de utilização da palavra nos títulos dos trabalhos. Já coeficientes negativos, indicam uma tendência decrescente.

A partir desta lista, e seguindo a abordagem apresentada nos trabalhos correlatos, foi utilizada a técnica de visualização de dados de nuvem de palavras, que possibilita, de maneira rápida, comparar a frequência de uma palavra em relação as demais. Como resultado da nuvem de palavras, é gerada uma figura com uma montagem composta por conjunto de palavras, com diferentes tamanhos de fonte, os quais são proporcio-

nais à quantidade de vezes que a palavra aparece no texto. Fontes maiores representam palavras que aparecem mais vezes em um texto e fontes menores, as que aparecem menos vezes. Para gerar as nuvens de palavras foi utilizada a ferramenta WordArt[3].

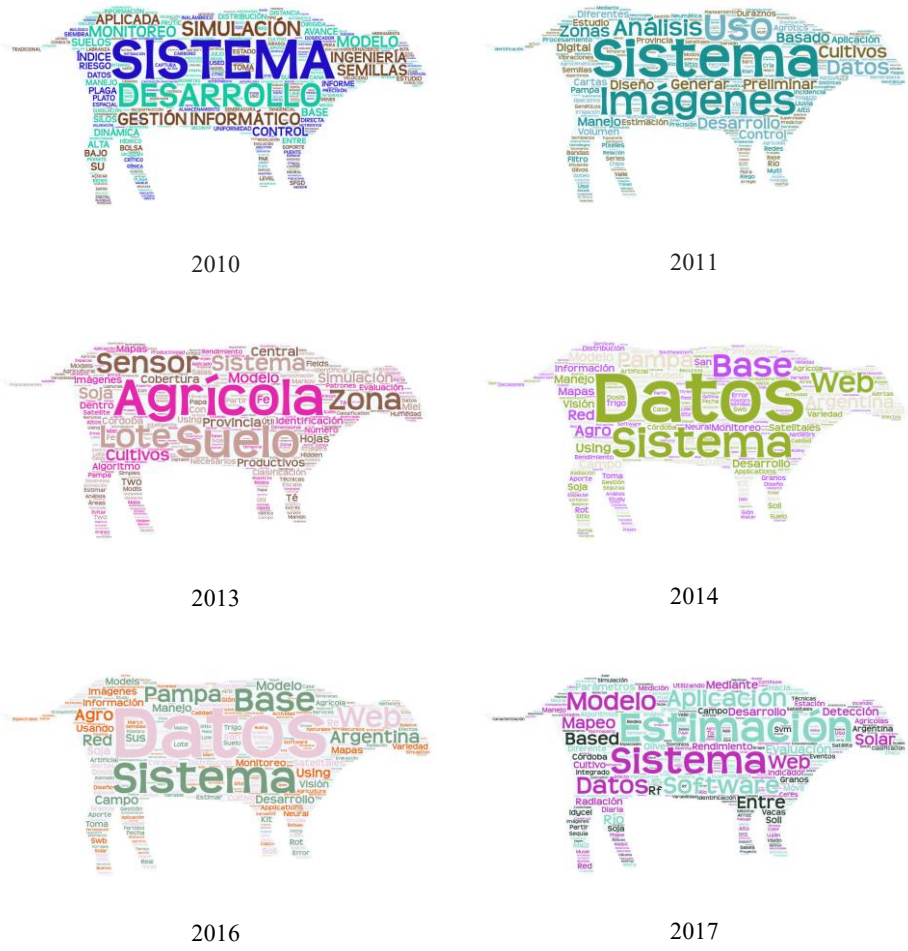
3. Resultados e Discussão

Com a aplicação da abordagem proposta, foram identificados os temas chave que têm sido discutidos no CAI. A abordagem teve início com o somatório da frequência de ocorrência de palavras a partir da análise dos títulos de artigos publicados nos seis últimos anais do CAI. Foram identificadas palavras que ocorreram entre 1 e 34 vezes nos títulos. A Tabela 1 mostra as palavras que ocorreram mais de seis vezes, deixando de serem apresentadas aquelas que ocorreram, em média, uma vez ou menos por edição do evento. As palavras que se destacam com maior frequência na história do CAI são: sistema(34), dados(23), desenvolvimento(20) e modelo(14). Esta situação induz à inferência de que os focos principais do evento têm sido os seguintes: 1) desenvolvimento e divulgação de sistemas, 2) discussões sobre dados, que podem envolver os processos de coleta, armazenamento, processamento e análise, o que poderia conduzir a estudos mais aprofundados sobre esta questão. 3) discussões sobre modelos, que são simplificações de uma realidade ou fenômeno investigado a fim de facilitar a sua compreensão ou permitir a realização de simulações. Também podem ser identificadas as palavras, específicas da área agro, que aparecem mais vezes nos títulos, tais como: manejo(13), agrícola(12), cultivo(11) e campo(10). Tais palavras podem indicar a referência espacial sobre os ambientes onde os problemas estão sendo pesquisados e as províncias onde estão os grupos de pesquisa que mais contribuem com o congresso. Também podem ser identificadas algumas palavras sobre referências espaciais, tais como: Argentina(12), Pampa(11) e Córdoba(9). Tais palavras podem estar indicando a referência espacial sobre os ambientes onde os problemas estão sendo pesquisados.

Além da frequência total, a Tabela 1 também apresenta a frequência individual das palavras em cada edição do evento. Alguns termos têm se destacado em edições específicas, tais como Sistema e Desenvolvimento, na edição de 2010; Sistema, na edição de 2011; Agrícola e Solo, na edição de 2013; Sistema, Base e Dados, nas edições de 2014 e 2016. Na edição de 2017, não houve palavras que se destacassem muito e tivessem sido utilizadas no título de cinco ou mais trabalhos. Outro importante aspecto investigado foi a tendência de crescimento ou diminuição da importância de cada palavra. Neste sentido, foram analisadas as frequências relativas das palavras em cada uma das seis edições do evento, de forma isolada. Esta frequência relativa foi obtida a partir da divisão da frequência da palavra na edição do evento pela quantidade de trabalhos publicados nos respectivos anais , que foram de 28, 26, 17, 18, 37 e 31 trabalhos. Em seguida, foi adaptado um modelo de regressão linear sobre os dados históricos destas frequências relativas. De acordo com a Figura 1, é possível visualizar as

[3] Site: https://wordart.com/

nuvens de palavras geradas com relação a coleta dos dados realizada neste estudo, identificando as principais palavras em destaque e as que menos frequência apresentaram.



2010 2011

2013 2014

2016 2017

Fig. 1. Nuvens de palavras geradas a partir dos títulos dos trabalhos publicados em cada edição do CAI.

Na Figura 2, é apresentado o coeficiente angular da frequência de cada palavra. As palavras com maior tendência negativa são: sistema(-0.019), desenvolvimento(-0.018), solo(-0.08) e imagens(-0.07). As palavras com maior tendência de crescimento são: web(0.017), software(0.014), Argentina(0.013) e dados(0.012). Estas palavras mostram um aumento da investigação com foco em sistemas web e gestão de dados,

com foco na Argentina. Outras palavras tem se mantido com uma tendência constante, tais como: manejo(0.01), cultivo(0.01) e campo(0.01).

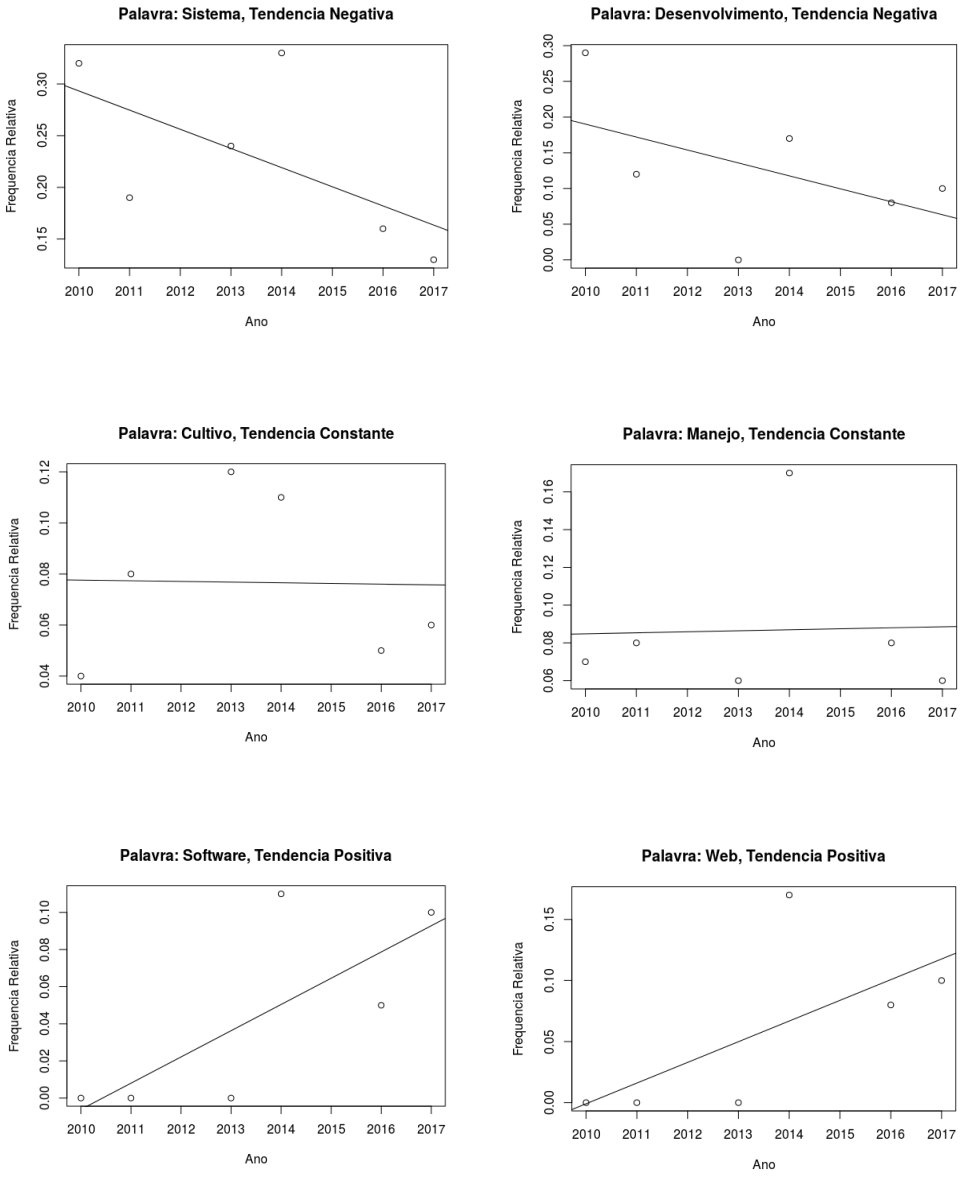


Fig. 2. Gráfico das palavras com maior tendência negativa, constante e positiva.

Tabela 1. Frequência absoluta das palavras em cada uma das edições do CAI, somatório da frequência absoluta para todas as edições, e tendência apresentada pelo coeficiente angular do modelo de regressão linear.

Palavra	2010	2011	2013	2014	2016	2017	Todos	Tendência
Sistema	9	5	4	6	6	4	34	-0,019
Datos	2	3	1	7	7	3	23	0,012
Desarrollo	8	3	0	3	3	3	20	-0,018
Modelo	2	1	2	3	3	3	14	0,006
Base	2	1	0	5	5	0	13	0,003
Manejo	2	2	1	3	3	2	13	0,001
Aplicación	1	2	1	2	2	4	12	0,008
Simulación	3	1	2	2	2	2	12	-0,003
Agrícola	0	1	5	2	2	2	12	0,005
Argentina	0	1	1	4	4	2	12	0,013
Gestión	3	1	2	2	2	2	12	-0,003
Cultivo	1	2	2	2	2	2	11	0,001
Monitoreo	3	1	0	3	3	1	11	-0,002
Pampa	1	1	1	4	4	0	11	0,003
Imágenes	1	3	2	3	0	1	10	-0,007
Estimación	1	1	1	1	2	4	10	0,01
Evaluación	1	0	2	2	2	3	10	0,009
Información	2	1	0	3	3	1	10	0,001
Campo	0	1	1	3	3	2	10	0,01
Suelo	2	0	5	1	1	0	9	-0,008
Córdoba	1	0	2	2	2	2	9	0,006
Web	0	0	0	3	3	3	9	0,017
Satelitales	0	1	1	3	3	1	9	0,007
Soja	1	0	2	2	2	1	8	0,003
Rendimiento	0	1	1	2	2	2	8	0,008
Clasificación	0	0	3	1	1	3	8	0,009
Uso	1	3	2	0	0	1	7	-0,01
Algoritmo	0	1	4	1	1	0	7	-0,003
Identificación	0	1	2	1	1	2	7	0,004
Distribuición	2	1	0	2	2	0	7	-0,004
Radiación	1	0	0	2	2	2	7	0,008
Granos	0	0	1	2	2	2	7	0,01
Software	0	0	0	2	2	3	7	0,014

Na Figura 3, é apresentada a nuvem de palavras dos títulos dos CAI de 2010 a 2017, mostrando um panorama histórico das discussões realizadas ao longo deste evento.

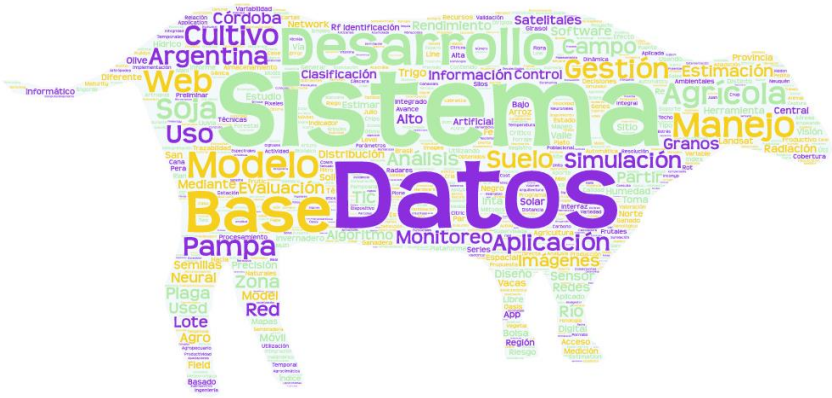


Fig. 3. Nuvem de palavras geradas a partir dos títulos dos trabalhos publicados nas seis edições do CAI.

4. Conclusão

Este trabalho investigou as tendências de pesquisa em agroinformática a partir das análises dos títulos dos trabalhos publicados nos seis últimos anais do Congresso Argentino de Agroinformática. Foram identificadas as palavras com maior frequência nos títulos dos seis eventos e, posteriormente, as palavras com maior frequência em cada uma das edições do evento. A partir da lista de palavras mais frequentes em cada uma das edições, foi feita uma ponderação pela quantidade de trabalhos publicados em cada ano do evento, obtendo-se a frequência relativa das palavras. Foram então utilizados modelos de regressão linear para identificar as tendências de crescimento e redução de uso das palavras ao longo das edições.

Como trabalho futuros, podem ser incluídos o uso das palavras chave, juntamente com os títulos, para analisar as tendências. Outra alternativa seria o uso das árvores de palavras para contribuir na compreensão do contexto que cada palavra foi utilizada dentro dos títulos. Neste contexto, para cada palavra *x* qualquer, usada com grande frequência, seriam gerados novos ramos com as palavras mais utilizadas em conjunto com a palavra *x*. Referências geográficas da filiação dos autores também poderiam ser utilizadas para gerar as nuvens de palavras contextualizadas nas províncias em que os estudos são realizados. Desta forma, poderiam ser compreendidos as peculiaridades de pesquisa de cada região.

Referencias

1. Ahmed, Zaheeruddin. Data Management and Big Data Text Analytics. Special Issue - National Conference on "Novel Trends in Computer Science" (TECHSA-17) p.140-144. 2017.
2. Ailamaki, Anastasia; Kantere, Verena; Dash, Debabrata. Managing Scientific Data. Communications of the ACM, V.53 N.6. p. 68-78, 2010.
3. Haugerud, Angelique. Editor's foreword: AE's keywords by decade. American Ethnologist. V. 40, N. 1. p. 1-5. 2013.
4. Hearst, Marti A.; Rosner, Daniela. Tag Clouds: Data Analysis Tool or Social Signaller? In: Proceedings of 41st Annual Hawaii International Conference on System Sciences. HICSS '08, Hawaii. 2008.
5. Hunt, Carter A. et al. A visual analysis of trends in the titles and keywords of topranked tourism journals. Current Issues in Tourism, V.17, N.10. p. 849-855. 2014.
6. Rossell, David. Big Data and Statistics: A Statistician's Perspective. Metode Sci Stud J., V.5. p. 143–149. 2015.