Análise das Tendências de pesquisa no Congresso de Agroinformática da Argentina: um estudo regional

Alfredo Parteli Gomes[1], Bruno Techera da Motta[2], Sandro da Silva Camargo[3], Yanina Bellini Saibene[4]

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

[2] Campus Santana do Livramento, Instituto Federal Sul-rio-grandense (IFSul), Santana do Livramento, Rio Grande do Sul, Brasil
brunomotta@ifsul.edu.br

[3] 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

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

Abstract. Dado o aumento do volume de dados de diferentes tipos e origem, e a viabilidade de análise e tendências através de técnicas de Mineração de Dados, é possível se obter maior conhecimento sobre os dados armazenados e com isso visualizar um cenário atual do contexto que representam estes dados estruturados ou não estruturados. Com este conceito, desenvolve-se em paralelo o uso de tecnologia representada por nuvens de palavras, cuja finalidade é representar em uma figura, as palavras mais frequentes dentro de um documento. Com o objetivo de identificar quais as tendências das pesquisas do Congresso Argentino de Agroinformática (CAI), este trabalho faz uma abordagem nos dados referentes aos títulos dos artigos publicados neste evento no período entre 2008 a 2018, apontando as quais as Províncias da Argentina que mais publicaram e suas as respectivas tendências de pesquisa dos autores dentro deste Congresso.

Palavras-chaves: Congresso Argentino de Agroinformática, Mineração de Dados, Nuvem de Palavras, visualização de dados.

1 Introdução

A revolução tecnológica possibilitou que a informação digitalizada seja fácil de obter, processar, armazenar, distribuir e transmitir. Dado o avanço da informática e suas tecnologias associadas e sua expansão nos mais diversos aspectos da vida, continua-se coletando e armazenando-se grandes volumes de dados. Transformar em conhecimento este grande volume de dados é desafiante. A Mineração de Dados (MD) objetiva dar forma e sentido à explosão de informação que atualmente pode ser armazenada [5,7].

As novas tecnologias de gerenciamento de dados são fruto do avanço nas tecnologias de hardware, armazenamento, rede e modelos de computação, como a virtualização e computação em nuvem. Como resultado deste processo, surge o termo Big Data como uma nova tendência que permite as organizações coletar, armazenar, gerenciar e manipular volume de dados em menor tempo e com maior acurácia [1].

Big data é usado para avaliar um volume massivo de dados estruturados e não estruturados. Atualmente, os dados estão presentes em diferentes tipos de dados: Estruturada, dados numéricos em bancos de dados tradicionais e Texto não estruturado, documentos, e-mail, vídeo, áudio e transações financeiras.

Na literatura, existem vários trabalhos associados que utilizam metodologias aplicadas à ciência de dados e a bibliotecnologia, nas quais fazem referência ao avanço permanente da disciplina e a necessidade de integrar estratégias inovadoras de acordo ao campo de conhecimento [9].

Ao mesmo tempo, é evidente que na literatura científica, aborda-se em repetidas ocasiões o conceito de tendências para destacar as principais correntes que emergem em cada área. Teodorescu et al. [8] propõe uma análise de produção científica por países, colaboradores e citações. Já Kawalec [6], foca nas publicações académicas espanholas e suas correlações entre diretórios e bases de dados documentais de acesso aberto. Han et al.[4] baseiam-se na pesquisa de colaboração internacional e a formação de redes académicas, a partir da identificação de países e instituições que produzem conhecimento na ciência da informação. Camargo et al.[2] faz uma abordagem com ênfase na colaboração científica no Congresso Argentino de Agroinformática a partir da identificação de redes de colaboração científica entre as instituições que publicaram em determinados períodos no CAI.

Com o objetivo de dar sequencia nas sugestões levantadas por Gomes et al.[3] no último Congresso Argentino de Agroinformática (CAI) realizado em Buenos Aires, este trabalho apresenta os resultados de uma pesquisa sistemática dos títulos dos artigos publicados no CAI1, um dos principais eventos científicos do país sobre este tema. O evento tem a participação de pesquisadores, técnicos, desenvolvedores e empresas relacionadas com o setor agroindustrial onde são apresentados trabalhos sobre as Tecnologias de Informação e Comunicação (TIC) dedicados a tratar problemas do agronegócio. O CAI iniciou em 2008 e já está em sua décima primeira edição, e faz parte do evento denominado Jornadas Argentinas de Informática (JAIIOs) organizadas pela Sociedade Argentina de Informática e Investigação Operativa (SADIO2). A presente pesquisa, analisou os títulos dos artigos oriundos das quatro Províncias que mais publicaram no CAI: Buenos Aires, Santa Fé, Córdoba e La Pampa. O período avaliado é entre os anos 2008 e 2018, agrupados em três períodos de três anos. Nos anos de 2012 e 2015 não foram gerados anais do CAI, motivo pelo qual estes anos não são considerados neste estudo.

Neste sentido, esta pesquisa busca analisar as tendências dos títulos de artigos publicados no CAI avaliando as Províncias que mais apresentaram trabalhos neste Congresso, a fim de contribuir com pesquisadores e com a região.

Desde esta perspectiva, o artigo está organizado da seguinte forma: A Seção Material e Métodos faz uma breve abordagem na descrição das características da base de dados utilizada para esta pesquisa e as técnicas 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 trabalho avaliou os artigos publicados no período entre 2008 e 2018 do Congresso Argentino de Agroinformática. A base de dados foi obtida

[1] Site: http://48jaiio.sadio.org.ar/simposios/CAI

[2] Site: http://www.sadio.org.ar/

através da organização do Congresso. Desta base de dados foram utilizados os títulos dos trabalhos publicados em cada uma das edições do evento, o ano de publicação e a procedência do trabalho, ou seja, o nome da Província da Argentina do autor. Foram identificados 294 trabalhos em 9 eventos do CAI, totalizando 1113 pesquisadores e colaboradores. Esta pesquisa utilizou a seguinte metodologia: 1) Obtenção da base de dados do evento, 2) Criação de um arquivo em texto puro, para os anais de cada período, contendo o ano, os títulos, e as Províncias de todos os trabalhos publicados, 3) Classificação dos artigos por Província agrupados em períodos de três anos, 4) Remoção de stop words, tais como artigos, preposições, pronomes e conjunções. É necessário remover estas palavras, pois têm a tendência de repetirem-se muitas vezes e não apresentarem relevância em relação aos elementos relevantes no título. 5) Classificar numericamente a frequência que cada palavra aparece nos títulos de um determinado período. Com esta classificação, desenvolver uma lista ordenada das palavras e sua respectiva contagem de vezes que elas apareceram no texto e, como consequência, tendem a ser mais relevantes em termos de pesquisa [3]. 6) Transformação da contagem absoluta dividindo-a pela quantidade de artigos publicados em cada período. Esta atividade precisou ser realizada em função da diferença da quantidade de artigos publicados em cada um dos períodos. Por exemplo, no caso da Província de Buenos Aires, ela apresentou 35 (2008-2009-2010), 16 (2011-2013-2014), e 30 (2016-2017-2018) trabalhos publicados respectivamente nos três períodos avaliados. 7) A partir da razão entre a contagem absoluta de menções às palavras e a quantidade de artigos publicados neste período, foram gerados novos valores. 8) Para cada palavra, foi criado um modelo de regressão linear, a fim de identificar o coeficiente angular do modelo a fim de inferir a tendência de utilização do termo. Coeficientes angulares positivos indicam uma tendência crescente de utilização da palavra nos títulos dos trabalhos. Já coeficientes negativos, indicam uma tendência decrescente.

Para complementar o estudo desta pesquisa, foi utilizada a técnica de visualização de dados de nuvem de palavras através da ferramenta WordArt3, que possibilitou, de maneira rápida, comparar a frequência de uma palavra em relação as demais. Como resultado da nuvem de palavras, é gerada uma figura

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

com uma montagem composta por conjunto de palavras, com diferentes tamanhos de fonte, os quais são proporcionais à quantidade de vezes que a palavra aparece no texto. A maiores fontes representam palavras que aparecem mais vezes em um documento e fontes menores, as que aparecem menos vezes.

3 Resultados e Discussão

A primeira ação da proposta desta pesquisa após coletar os dados, foi realizada a preparação dos dados para eliminar os títulos repetidos, pois a base de dados original se apresenta indexada por autor e colaborador, com isto, é comum encontrar títulos duplicados. Após a fase de preparação dos dados, foi possível ordenar os títulos por Províncias, e observar a quantidade de trabalhos por Províncias (Tabela 1). Com está visualização prévia dos dados, decidiu-se estudar as quatro Províncias que mais publicaram no CAI, dado que o somatório dos trabalhos deste grupo totalizou 224 artigos, ou seja, superou os 76% dos trabalhos publicados em todos os eventos do CAI (2008-2018).

Tabela 1. contagem de artigos publicados por Província no período 2008-2018 no CAI.

Província	Artigos Publicados
Buenos Aires	81
Córdoba	56
Santa Fé	45
La Pampa	43
San Juan	12
Entre Ríos	9
Rio Negro	9
Chaco	8
Misiones	5
San Luis	5
Corrientes	4
Mendoza	4
Chubut	3
Catamarca	3
Santiago del Estero	2
Jujuy	2

Neuquén	2
Tucumán	1
Total de Artigos	294

Após análise dos dados obtidos das quatro Províncias citadas anteriormente, constatou-se que algumas palavras estivram presentes em pelo menos três Províncias, como por exemplo “Desarrollo”, “Sistema”, “Datos”. Estas palavras tendem a indicar o tipo de investigação que está se desenvolvendo nestas regiões. Em relação ao estudo por Província, pode-se observar: 1) “Sistema” é o termo que está presente nas quatro Províncias, 2) tanto na Província de Buenos Aires(81), como em Córdoba(57) e La Pampa(43), a palavra “Desarrollo” apesar de apresentar uma frequência alta, destaca-se uma tendência negativa, observa-se também que o termo foi muito utilizado no primeiro período (2008-2010), porém nos períodos seguintes teve uma frequência baixa nos títulos dos trabalhos. 3) as Províncias de Córdoba e Santa Fé apresentam em comum a tendência positiva do termo “Modelo” e “Datos”. 4) Em termos de regionalização, foi possível identificar que pesquisadores da Província de La Pampa, foram os que mais utilizaram o termo “La Pampa”, isto deve-se a investigações relacionadas diretamente com a Província de La Pampa, característica presente nos trabalhos desenvolvidos pelos autores. 5) outra observação é o uso do termo “Soja”, que apareceu apenas na Província de Santa Fé, e encontra-se em tendência negativa.

Conforme a Figura 1, pode-se observar os gráficos com a regressão linear da frequência relativa de cada palavra agrupado por Província.

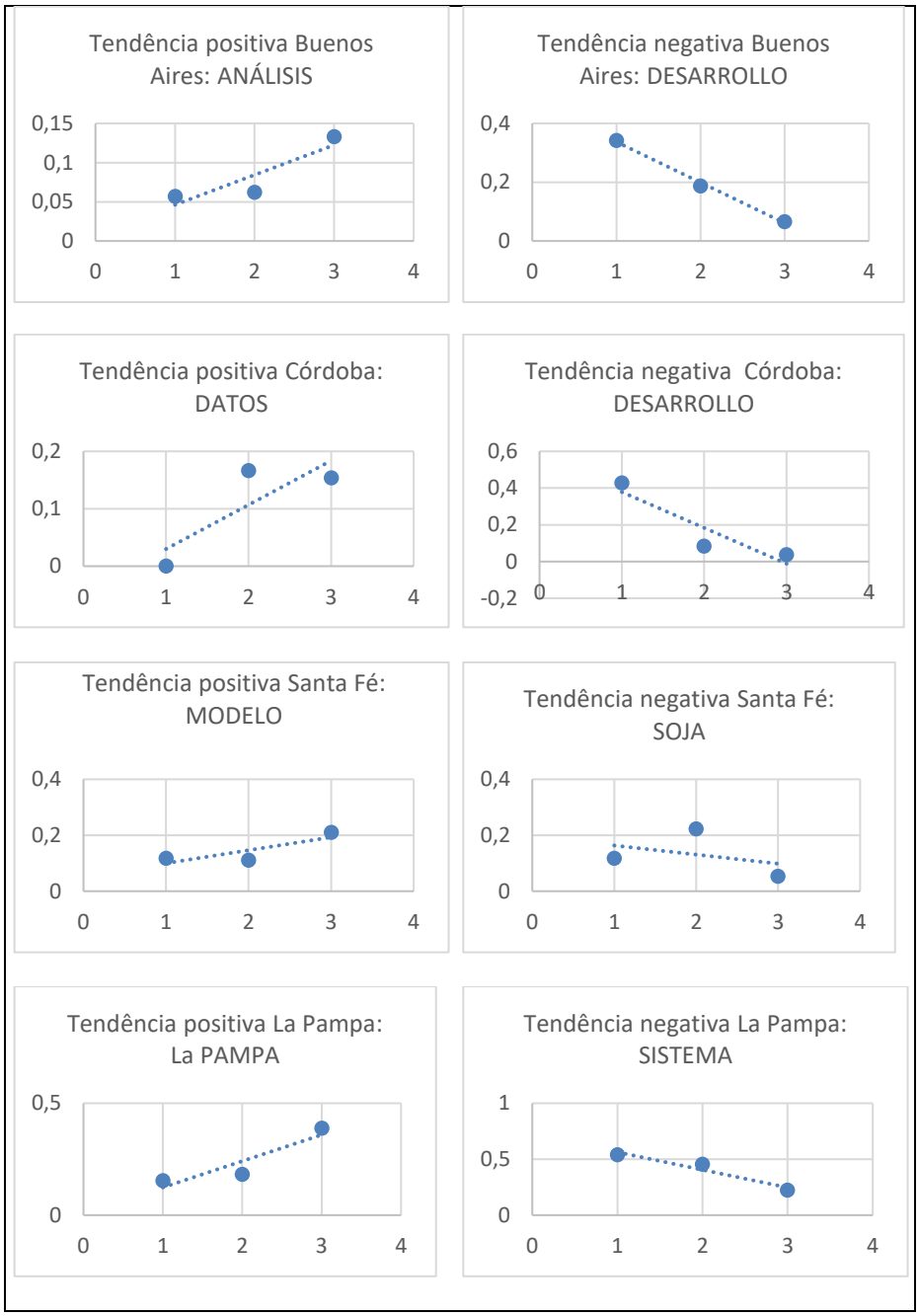

Fig. 1. Gráfico das palavras com maior tendência positiva e negativa das Províncias de Buenos Aires, Córdoba, Santa Fé e La Pampa.

Das quatro Províncias analisadas (Buenos Aires, Córdoba, Santa Fé e La Pampa), as palavras com maior tendência de crescimento por Província respectivamente, são: “Análisis”(0,012), “Datos” (0,026), “Modelo”(0,022) e “La Pampa”(0,060). Estas palavras mostram uma tendência de investigação com foco em desenvolvimento de sistemas e gestão de dados e ou uso de ciência de dados para o tratamento da informação, além do destaque para a palavra regionalizada “La Pampa” que representa um estudo específico nesta região da Argentina. Com relação à tendências negativas, aparecem respectivamente as palavras: tanto em Buenos Aires como em Córdoba, a palavra “Desarrollo” (-0,062 e -0,018) aparece em destaque, visto que sua frequência é maior apenas nos primeiros anos (2008-2009-2010), já em Santa Fé a palavra com menor tendência é “Soja” (-0,011) e por último em La Pampa, “Sistema” (-0,036) (Tabela 2). Os períodos representados na Tabela 2, correspondem respectivamente: 1- 2008, 2009 e 2010. 2- 2011, 2013 e 2014. 3- 2016, 2017 e 2018.

Tabela 2. tabela com as frequências absoluta de cada Província e com suas respectivas tendências.

BUENOS AIRES					
Palavra	**Período**			**Total**	**Tendência**
	1	**2**	**3**		
DESARROLLO	12	3	2	17	-0,062
SISTEMA	10	2	4	16	-0,037
DATOS	7	2	4	13	-0,019
INTA	3	1	1	5	-0,012
MANEJO	3	3	2	8	-0,006
AGROPECUARIA	4	1	2	7	-0,012
HERRAMIENTA	2	4	2	8	0,000
DISEÑO	2	2	3	7	0,006
ANÁLISIS	2	1	4	7	0,012
INFORMACIÓN	2	1	1	4	-0,006
INFORMÁTICO	4	1	0	5	-0,025
Total Trabalhos					**81**
CÓRDOBA					
Palavra	**Período**			**Total**	**Tendência**
	1	**2**	**3**		
DESARROLLO	3	2	1	6	-0,018
SISTEMA	2	4	4	10	0,018

DATOS	2	1	5	8	0,026
Total Trabalhos					**57**

SANTA FÉ					
Palavra	**Período**			**Total**	**Tendência**
	1	**2**	**3**		
SISTEMA	4	1	4	9	0,000
MODELO	2	1	4	7	0,022
SOJA	2	2	1	5	-0,011
DATOS	1	2	1	4	0,000
SIMULACIÓN	1	2	1	4	0,000
Total Trabalhos					**45**

LA PAMPA					
Palavra	**Período**			**Total**	**Tendência**
	1	**2**	**3**		
SISTEMA	7	5	4	16	-0,036
INFORMACIÓN	4	1	3	8	-0,012
DESARROLLO	3	2	2	7	-0,012
GESTIÓN	3	1	1	5	-0,024
DATOS	2	1	3	6	0,012
LA PAMPA	2	2	7	11	0,060
Total Trabalhos					**43**

Fig. 2. Nuvens de palavras geradas a partir dos títulos dos trabalhos publicados por cada Província no período 2008-2018: Buenos Aires, Córdoba, Santa Fé e La Pampa.

Na Figura 2, pode-se observar as nuvens de palavras de cada Província.

Nuvem de Palavras: Buenos Aires

Nuvem de Palavras: Córdoba

Nuvem de Palavras: Santa Fé

Nuvem de Palavras: La Pampa

Esta visualização de dados é indicada para uma interpretação clara e rápida, na qual é possível observar as palavras mais frequentes em cada caso durante o período 2008-2018.

4 Conclusão

Este trabalho é uma ampliação de abordagem uma abordagem realizada no CAI 2018 [3], porém foi possível comprovar as tendências de pesquisa em agroinformática de autores argentinos, regionalizadas por Províncias e avaliados em três períodos de três anos.

Como resultado, após a metodologia aplicada, foram identificadas palavras com maior frequência nos títulos agrupadas por Províncias e, posteriormente, aplicado o modelo de regressão linear para identificação das tendências positivas e negativas que cada região está pesquisando.

Estudos posteriores, podem adicionar o uso das palavras chave, juntamente com os resumos, para analisar as tendências e relacionar com trabalhos anteriores para comprovar se a metodologia está adequada. Além disso, será possível incluir os trabalhos oriundos de instituições estrangeiras e assim colaborar no processo de internacionalização do CAI.

Concluindo, espera-se que este trabalho possa auxiliar aos pesquisadores de toda a Argentina que participam do CAI identificar o que está relevante e qual é a área que ainda necessita estudo em termos de pesquisa científica em agroinformática.

5 Referências

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. Camargo, Sandro et al. Congreso argentino de agroinformática: Un análisis bibliometrico. In: X Congreso de AgroInformática (CAI)-JAIIO 47 (CABA, 2018).
3. Gomes, Alfredo Parteli; Camargo, Sandro; Bellini Saibene, Yanina. Tendências de pesquisa em Agroinformática na Argentina: uma análise histórica. In: X Congreso de AgroInformática (CAI)-JAIIO 47 (CABA, 2018).
4. Han, Pu; Shi, Jin; Li, Xiaoyan; Wang, Dongbo; Shen, Si; Su, Xinning (2014). International collaboration in LIS: global trends and networks at the country and institution level. Scientometrics, vol. 98, no. 1 (January), p. 53–72.
5. Hernandéz Orallo, José; Ferri Ramirez, C; Ramirez Quintana, M. J. Introducción a la Minería de Datos. 2004.
6. Kawalec, Anna (2013). Research trends in library and information science based on Spanish scientific publication 2000 to 2010. Malaysian journal of library & information science, vol. 18, no. 2, p. 1–13.
7. Mitra, Sushmita; Acharya, Tinku. Data mining: multimedia, soft computing, and bioinformatics. John Wiley & Sons, 2005.

8. Teodorescu, Daniel; Andrei, Tudorel (2011). The growth of international collaboration in East European scholarly communities: a bibliometric analysis of journal articles published between 1989 and 2009. Scientometrics, vol. 89, no. 2, p. 711–722.

9. Tsay, M. A bibliometric analysis and comparison on three information science journals: JASIST, IPM, JOD, 1998-2008. Scientometrics, v. 89, n. 2, p. 591- 606, Nov. 2011. Disponível em: <http://link.springer.com/article/10.1007%2Fs11192-011-0460-4>. Acesso em: 01 mai. 2019.