

Aplicando Técnicas de Bibliometria, Mineração de Texto e Visualização na Identificação de Temas e Tendências de Pesquisa em *e-Learning*

Aline Maria Frascareli^{1,2}, Edson Pinheiro Pimentel¹

¹Universidade Federal do ABC (UFABC)
Santo André – SP – Brasil

²Instituto de Psicologia – Universidade de São Paulo (IPUSP)
São Paulo – SP – Brasil

{aline.frascareli, edson.pimentel}@ufabc.edu.br

Resumo. O tripé formado pelo ensino, pela pesquisa e pela extensão constitui o eixo fundamental da Universidade brasileira e são componentes indissociáveis. No âmbito da pesquisa, a identificação de temas atuais e relevantes em uma determinada área não é uma tarefa trivial, principalmente em um cenário no qual o volume de informações nas bases de dados cresce dia a dia. Objetiva-se mostrar a aplicação de técnicas de bibliometria, mineração de texto e visualização em um processo automático, para identificação de temas e tendências de pesquisa. Para tal, realizou-se um experimento com artigos científicos da área de *e-learning*. Os resultados da análise bibliométrica apontaram um crescimento quantitativo de artigos científicos em *e-learning* no período de 2001 a 2011. A mineração de texto e visualização possibilitou um comparativo da ocorrência e proximidade de termos no período em questão. Espera-se que as técnicas possam ser aplicadas em outros domínios de conhecimento.

Abstract. The tripod formed by education, research and extension is the cornerstone of the Brazilian university and are inseparable components. As part of the research, identifying current and relevant topics in a given area is not a trivial task, especially in a scenario where the volume of information in databases grows everyday. This article aims at showing the application of bibliometric techniques, text mining and visualization in an automatic process to identify themes and trends research. To this end, we carried out an experiment with scientific articles in the field of *e-learning*. The results of the bibliometric analysis indicated a quantitative growth of scientific papers on *e-learning* in the period from 2001 to 2011. The text mining and visualization enabled a comparison of the occurrence and proximity of terms during the concerned period. It is expected that these techniques can also be applied to other fields of knowledge.

1. Introdução

A disponibilidade de informação cresceu e continuará crescendo exponencialmente. A diminuição do custo e a melhoria das tecnologias disponíveis para armazenamento, recuperação e disseminação de informações tem oferecido condições para permitir esse fenômeno de “sobrecarga de informação” [Kielgast and Hubbard 1997]. Da mesma

forma, esse fato tem ocorrido na área científica. O aumento exponencial de informação textual dificulta o entendimento da estrutura e da dinâmica de desenvolvimento de um campo de conhecimento. O volume de informação é tão intenso que torna-se praticamente impossível analisar as informações disponíveis sem o auxílio de tecnologias.

A produção de novos conhecimentos é um importante papel do sistema educacional em todos os níveis e principalmente no Ensino Superior. De acordo com a legislação, o tripé formado pelo ensino, pela pesquisa e pela extensão constitui o eixo fundamental da Universidade brasileira e são componentes indissociáveis. A identificação de temas de pesquisa para projetos de iniciação científica, projetos de graduação ou de pós-graduação geralmente são realizados com base em aspectos diversos, destacando-se principalmente os interesses de pesquisa do orientador e/ou do candidato. Em situações em que não é possível identificar um tema a partir dos interesses de pesquisa, faz-se necessário a prospecção e análise de textos científicos a fim de identificar tendências e lacunas de pesquisa. Entretanto, dado o volume de artigos e textos científicos disponíveis, a análise humana torna-se uma tarefa trabalhosa, ocasionando que, muitas vezes, temas de pesquisa promissores e relevantes sejam desconsiderados. Uma técnica utilizada para apoiar a prospecção de trabalhos científicos é a revisão sistemática que consiste em uma forma de síntese dos resultados de pesquisas relacionados com o problema que o originou [Berwanger et al. 2007]. No entanto, normalmente ela é utilizada em um estágio da pesquisa em que já se tem um problema bem delineado.

As áreas de bibliometria, mineração de texto e visualização têm despertado interesse da comunidade científica por apresentar técnicas e ferramentas capazes de sintetizar esse aglomerado de dados e apoiar a identificação de tópicos relevantes e tendências de pesquisas. A bibliometria é o estudo dos aspectos quantitativos da produção, disseminação e uso da informação registrada com a utilização de métodos matemáticos e estatísticos [Macias-Chapula 1998]. Já a mineração de texto busca analisar e processar dados semi-estruturados e não estruturados de um conjunto de textos [Miner et al. 2012]. Nesse contexto, técnicas de visualização de informação evidenciam-se como um componente adicional no processo de mineração de um conjunto de documentos.

No âmbito deste trabalho, pretende-se abordar meios para a recuperação de artigos científicos que ofereçam suporte para identificação de tendências e tópicos para pesquisa, principalmente ao desenvolvimento da iniciação científica, por envolver um perfil de pesquisador iniciante. A partir dos resultados obtidos nos mapas, há a possibilidade de se elaborar novas estratégias de pesquisa com os termos que emergiram do processamento. Este artigo tem por objetivo mostrar a aplicação de técnicas de bibliometria, mineração de texto e visualização em um processo automático para identificação de temas e tendências de pesquisa. Para tanto, utilizou-se artigos científicos recuperados de duas bases de dados da *Web of Science*, de forma a apresentar um panorama da área de *e-learning* no período de 2001 a 2011.

Na Seção 2 são apresentados trabalhos que utilizaram técnicas bibliométricas, de mineração de texto e visualização; a Seção 3 descreve os procedimentos metodológicos; na Seção 4 está a apresentação e discussão dos resultados da pesquisa, seguidos da Seção 5, com as considerações e trabalhos futuros.

2. Bibliometria, Mineração de Texto e Visualização

As técnicas bibliométricas, de mineração de texto e visualização são aplicadas na ciência, visando identificar e organizar áreas de pesquisa conforme especialistas, instituições, publicações, periódicos, citações ou textos; descobrir relações entre eles; examinar mudanças, tais como a velocidade de crescimento; localizar e mapear redes de colaboração científica; e identificar o impacto da estratégia e financiamento da pesquisa aplicado pelo governo ou outras agências. As técnicas dão suporte e complementam o julgamento humano [Shiffrin and Börner 2004].

Alguns dos estudos identificados na literatura apresentam como foco de análise os termos de indexação, cabeçalhos de assunto ou palavras-chave extraídas dos títulos ou resumos [Callon et al. 1991, Ding et al. 2001, Ronda-Pupo and Ángel Guerras-Martin 2012]. Outros métodos sugeriram combinar a análise de co-citação com a análise de palavras [Braam et al. 1991, Liu et al. 2010].

Há trabalhos que trataram da integração de técnicas de mineração de textos e bibliometria aplicadas em texto completo. Glenisson et al. (2005) combinaram a análise de texto completo e métodos tradicionais de bibliometria [Glenisson et al. 2005]. Em um outro estudo que se utilizou do texto completo, a técnica de mineração de texto foi integrada com dados bibliométricos a fim de apresentar características do periódico Cortex¹ e traçar o perfil de pesquisa da comunidade usuária [Kostoff et al. 2005].

O trabalho de Janssens (2007) verificou a hipótese de que a precisão de agrupamento e classificação de domínios de conhecimento é melhorada pela incorporação de algoritmos e técnicas de mineração de texto e bibliometria. O estudo contribui significativamente ao demonstrar em detalhes a complementaridade dos métodos de mineração de texto e bibliometria e ao propor esquemas para a integração. Ressalta-se que a metodologia foi aplicada à área de bioinformática [Janssens 2007].

Aplicações das técnicas de mineração de texto e bibliometria também foram identificados nas áreas de *e-marketing* [He et al. 2012], *mobile learning* [Hung and Zhang 2011] e *e-learning* [Hung 2012].

Artigos científicos são o maior mecanismo dos pesquisadores para apresentar os resultados de suas pesquisas. Uma coleção de artigos em uma determinada área pode revelar importantes aspectos de sua evolução, como a necessidade de estudos de novos tópicos. Diante desse fato, o trabalho de Alencar et al. (2012) buscou identificar a evolução de conteúdos a partir de mapas baseados em similaridade de conteúdo de artigos [Alencar et al. 2012].

No trabalho de van Eck e Waltman (2011) é apresentado o *software VOSviewer*, que inclui a funcionalidade de mineração de texto. *VOSviewer* é capaz de criar e visualizar mapas bibliométricos. Uma aplicação apresentada no estudo buscou identificar o impacto de citação de diferentes tópicos cobertos pelo *Journal of the American Society for Information Science and Technology* (JASIST). Utilizando a funcionalidade de mineração de texto do *VOSviewer*, um mapa de termos foi elaborado, baseado nos títulos e resumos de todas as publicações recuperadas no JASIST no período de 2005 a 2009 [Eck and Waltman 2011].

¹<http://www.cortexjournal.net/>

3. Procedimentos Metodológicos

Esta seção mostrará a aplicação de técnicas de bibliometria, mineração de texto e visualização automatizando o processo de identificação de temas e tendências de pesquisa. Mais especificamente, as técnicas são aplicadas na área de *e-learning*, escolhida por ser uma área emergente e que conecta a Educação e a Informática.

Odakura e Barvinski (2012) apresentaram o panorama das publicações de Informática na Educação no ano de 2011, basicamente no SBIE (Simpósio Brasileiro de Informática na Educação), WIE (Workshop de Informática na Escola) e RBIE (Revista Brasileira de Informática na Educação), utilizando um levantamento manual [Odakura and Barvinski 2012]. No entanto, não apresentam técnicas para estudos de temas ou tendências.

Descreve-se, a seguir, as etapas utilizadas no presente estudo.

3.1. Coleta de dados

Para a identificação dos artigos científicos relacionados à área de *e-learning*, foram consultadas duas bases de dados bibliográficas abrangidas pela *Web of Science: Science Citation Index Expanded (SCI)* e *Social Sciences Citation Index (SSCI)*.

No recurso de busca avançada, dois termos foram empregados para a elaboração da estratégia de pesquisa, limitando os campos título, resumo e/ou palavras-chave (representada na *Web of Science* por “TS”): TS=(e-learning OR elearning).

Além disso, delimitou-se o tipo de documento a *ARTICLE*, no período de 1º de janeiro de 2001 a 31 de dezembro de 2011. A busca foi realizada no dia 15 de agosto de 2012 e retornou 2695 artigos.

Dez títulos de revistas foram escolhidos para compor o grupo de dados de textos para a mineração de texto e visualização. Os dados foram exportados em grupos de 500 registros (limite máximo permitido pela base) no formato *bib*², contendo informações de título do artigo, autores, título da revista, ano e resumo.

3.2. Análise dos dados: bibliometria

A partir dos recursos disponíveis na *Web of Science*, um conjunto de indicadores bibliométricos foram gerados para apresentar um panorama das pesquisas na área de *e-learning*. Estes indicadores foram os seguintes: o crescimento quantitativo de artigos por ano, distribuição geográfica dos artigos e revistas com maior número de artigos publicados. Parte das variáveis definidas para análise foram apresentadas no trabalho de Hung (2012).

3.3. Mineração de texto e visualização

Após a etapa de coleta de dados, elaborou-se um único arquivo *bib*. O aplicativo *Mendeley*³ foi utilizado para exportação do arquivo *bib* para o formato *EndNote XML*. Dois grupos foram definidos, de acordo com o ano de publicação dos artigos: 2006 a 2008 e 2009 a 2011. Convém destacar que o *Mendeley* é um gerenciador de referências e, além

²<http://www.bibtex.org/>

³<http://www.mendeley.com/>

de possibilitar a exportação das referências, é capaz de remover possíveis duplicidades de registros, principalmente para estudos que utilizarem dados de diversas bases.

Em paralelo, desenvolveu-se uma função *script* em MATLAB⁴ para preparar os dados para a utilização na ferramenta VOSviewer⁵. A função recebe como entrada um arquivo EndNote XML e retorna três vetores: os vetores *titles* e *abstracts*, que possuem respectivamente os títulos e resumos dos artigos; e o vetor *onlytitles*, são aqueles registros que não possuem resumos. Nessa função utilizou-se o comando MATLAB *xmlread* para criar um objeto que navega sobre a árvore XML. Como a estrutura XML do EndNote é conhecida, ficou fácil identificar de quais nós seriam recuperadas as informações. Em seguida, gerou-se um arquivo texto com os dados das revistas selecionadas. Os arquivos contém os dados de títulos e resumos dos artigos em seus respectivos períodos.

Os dez títulos de revistas selecionados foram:

1. *Lectures Notes in Computer Science*,
2. *Computers & Education*,
3. *Educational Technology & Society*,
4. *British Journal of Educational Technology*,
5. *Expert Systems with Applications*,
6. *Journal of Universal Computer Science*,
7. *Journal of Computer Assisted Learning*,
8. *IEEE Transactions on Education*,
9. *International Journal of Engineering Education*,
10. *IEEE Transactions on Learning Technologies*.

Os artigos publicados no período de 2006 a 2011 foram agrupados em dois períodos para as análises:

- Grupo 1 (2006 a 2008): 301 artigos foram recuperados, mas 3 registros não apresentaram resumos e foram excluídos, restando para análise 298 artigos.
- Grupo 2 (2009 a 2011): 268 artigos foram recuperados, mas 2 registros não apresentaram resumos e foram excluídos, restando para análise 266 artigos.

4. Apresentação e Discussão dos Resultados

Tendências de publicações no tempo

A Figura 1 indica o número de artigos publicados em *e-learning* por ano, no período de 2001 a 2011. É possível notar um crescimento significativo de artigos publicados na área a partir de 2008.

Áreas de pesquisa

As pesquisas em *e-learning* concentraram-se em duas áreas: Ciência da Computação (39.147%) e Educação (34.471%). Estudos também foram identificados nas áreas de Engenharia (13.506%), Psicologia (5.158%) e Ciência da Informação (4.861%). Outras áreas corresponderam a 2.857% dos artigos publicados.

Países

Os artigos recuperados apontaram os Estados Unidos como o país mais produtivo na

⁴<http://www.mathworks.com/products/matlab/>

⁵<http://www.vosviewer.com/>

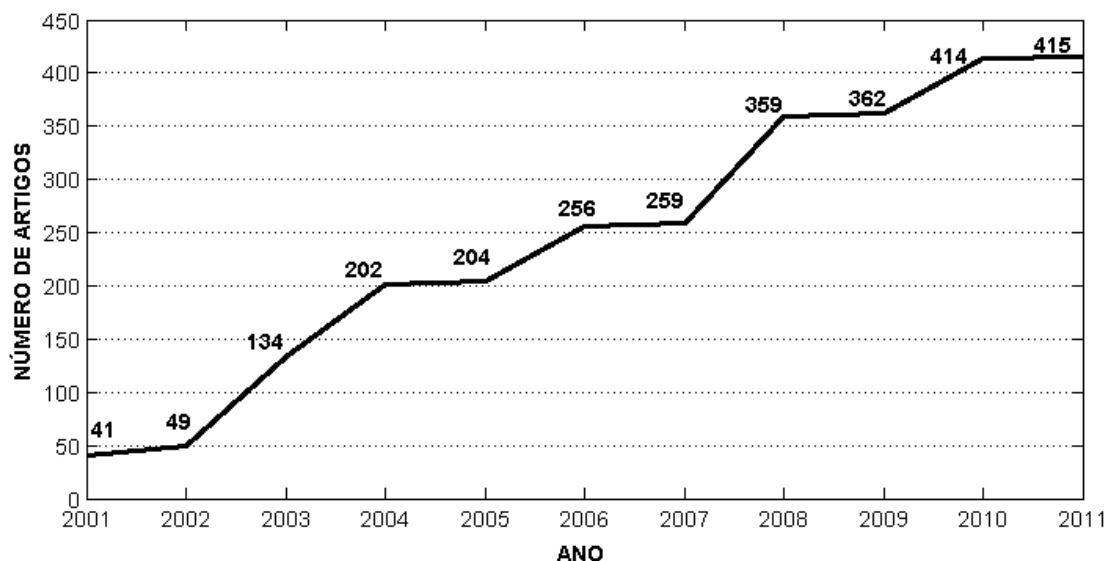


Figura 1. Número de artigos publicados por ano na área de *e-learning* no período de 2001 a 2011

área de *e-learning* com 360 registros (13.358%). Em seguida, a Inglaterra (11.577%), Taiwan (10.019%), Espanha (9.091%) e Alemanha (7.087%). O Brasil apresenta 26 artigos (0.965%) no período em estudo.

Revistas

Revistas que mais apresentaram artigos na área de *e-learning* no período de 2001 a 2011 estão indicados na Tabela 1:

Tabela 1. Revistas com artigos em *e-learning* no período de 2001 a 2011

Revista	% de 2695 artigos
<i>Lectures Notes in Computer Science</i>	10.798%
<i>Computers & Education</i>	5.826%
<i>Educational Technology & Society</i>	4.861%
<i>British Journal of Educational Technology</i>	2.746%
<i>Expert Systems with Applications</i>	2.189%

Mineração de texto e visualização

A funcionalidade de mineração de texto do *VOSviewer* permite a criação de mapas de termos baseados em um conjunto de documentos. Um mapa de termo é um mapa bidimensional no qual os termos são localizados de tal forma que a distância entre dois termos pode ser interpretada como uma indicação do relacionamento entre eles. Quanto menor a distância entre dois termos, maior é a relação entre eles [Eck and Waltman 2011].

A Figura 2 mostra o resultado da aplicação do *VOSviewer* no primeiro período (2006–2008). As cores indicam a densidade dos termos, variando do azul (menor densidade) ao vermelho (maior densidade). Os termos que se destacaram nesse período e que apresentam forte relação com a área foram: *domain*, *ontology*, *collaboration*, *communication*, *tutor*, *recommendation*, *individual learner*, *learning process*, *instruction*, *semantic*

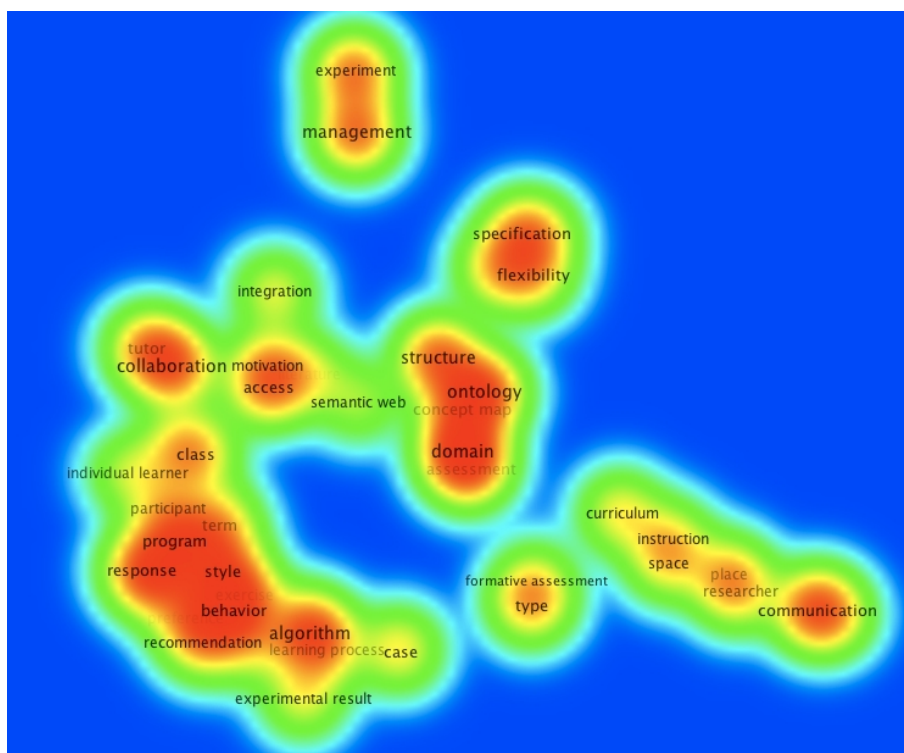


Figura 2. Grupo 1: Mapa de termos dos artigos publicados no período de 2006 a 2008 nas revistas selecionadas

web, *formative assessment*. No primeiro mapa, os termos *ontology* e *semantic web* estão próximos, assim como *tutor* e *collaboration*.

A Figura 3 apresenta o resultado gerado com os artigos do segundo período (2009–2011). Os termos *communication*, *tutor*, *recommendation*, *learning process* e *experimental result* mantiveram-se presentes no segundo período da análise. Contudo, é possível notar que novos termos surgiram, tais como: *community*, *internet*, *mobile learning*, *game* e *simulation*. Uma razão para o surgimento desses termos no segundo período delimitado, pode ser justificada pela popularização de equipamentos como *smartphones* e *tablets* e o amplo uso da internet, o que pode ter gerado como consequência o aumento de estudos nessa perspectiva.

É importante destacar que, para a validação dos termos obtidos, a análise dos resultados foi auxiliada por um pesquisador de Informática na Educação, que atua na área de *e-learning* há dez anos. Durante a validação, o especialista destacou a ocorrência de termos nos mapas não comuns à área, o que pode ser decorrência da técnica utilizada pela ferramenta. Alguns desses termos foram excluídos durante o processo de análise. Para continuidade da pesquisa, essa questão deverá ser estudada com maior profundidade.

5. Considerações e Trabalhos Futuros

O estudo buscou aplicar técnicas de bibliometria, mineração de texto e visualização como um processo automático para identificação de temas e tendências de pesquisa. A aplicação foi realizada na área de *e-learning* considerando dois períodos distintos: 2006 a 2008 e 2009 a 2011.

- Ding, Y., Chowdhury, G. G., and Foo, S. (2001). Bibliometric cartography of information retrieval research by using co-word analysis. *Information Processing & Management*, 37(6):817–842.
- Eck, N. J. V. and Waltman, L. (2011). Text mining and visualization using VOSviewer. *ISSI Newsletter*, 7(3):50–54.
- Glenisson, P., Glänzel, W., Janssens, F., and Moor, B. D. (2005). Combining full text and bibliometric information in mapping scientific disciplines. *Information Processing and Management: an International Journal*, 41(6):1548–1572.
- He, W., Chee, T., Chong, D., and Rasnick, E. (2012). Using bibliometrics and text mining to explore the trends of e-marketing literature from 2001 to 2010. *International Journal of Online Marketing*, 2(1):16–24.
- Hung, J. L. (2012). Trends of e-learning research from 2000 to 2008: use of text mining and bibliometrics. *British Journal of Educational Technology*, 43(1):5–16.
- Hung, J.-L. and Zhang, K. (2011). Examining mobile learning trends 2003-2008: a categorical meta-trend analysis using text mining techniques. *Journal of Computing in Higher Education*, pages 1–17.
- Janssens, F. (2007). Clustering of scientific fields by integrating text mining and bibliometrics. Master's thesis, Katholieke Universiteit Leuven.
- Kielgast, S. and Hubbard, B. A. (1997). Valor agregado à informação: da teoria à prática. *Ciência da Informação*, 26(3).
- Kostoff, R. N., Buchtel, H. A., Andrews, J., and Pfeil, K. M. (2005). The hidden structure of neuropsychology: text mining of the journal cortex: 1991-2001. *Cortex*, 41(2):103–115.
- Liu, X., Yu, S., Janssens, F., Glänzel, W., Moreau, Y., and Moor, B. D. (2010). Weighted hybrid clustering by combining text mining and bibliometrics on a large-scale journal database. *Journal of the American Society for Information Science and Technology*, 61(6):1105–1119.
- Macias-Chapula, C. A. (1998). O papel da informetria e da cienciometria e sua perspectiva nacional e internacional. *Ciência da Informação*, 27(2):134–140.
- Miner, G., Elder, J., Hill, T., Nisbet, R., Delen, D., and Fast, A. (2012). *Practical text mining and statistical analysis for non-structured text data applications*. Academic Press, Oxford.
- Odakura, V. V. V. A. and Barvinski, C. A. (2012). SBIE, WIE e RBIE: publicações da comunidade brasileira de Informática na Educação. In *Workshop de Desafios da Computação Aplicada à Educação*, Curitiba. SBC.
- Ronda-Pupo, G. A. and Ángel Guerras-Martin, L. (2012). Dynamics of the evolution of the strategy concept 1962–2008: a co-word analysis. *Strategic Management Journal*, 33(2):162–188.
- Shiffrin, R. M. and Börner, K. (2004). Mapping knowledge domains. *PNAS*, 101(1):5183–5185.