MINERAÇÃO DE TEXTO E SUAS APLICAÇÕES NA LITERATURA CIENTIFICA – ESTUDO BIBLIOMÉTRICO

Bruno Missi Xavier
Mestrando em Pesquisa Operacional e Inteligência Computacional / UCAM
bmissix@gmail.com

Alcione Dias da Silva
Mestrando em Pesquisa Operacional e Inteligência Computacional / UCAM
diasalcione@gmail.com

Georgia Regina Rodrigues Gomes
Doutora em Informática / UCAM
georgia@ucam-campos.br

Helder Costa
Doutor em Engenharia Mecânica / UFF
helder.uff@gmail.com

Recebido: 03 de abril de 2012. Revisado: 06 de junho de 2012. Aceito: 14 de agosto de 2012. Publicado online: 17 de setembro 2012.

RESUMO

A utilização de técnicas de Mineração de Textos na descoberta de informações relevantes em bases não estruturadas vem crescendo consideravelmente. Este assunto se torna cada vez mais relevante devido ao volume de informações textuais divulgadas na web, como também geradas pelas empresas. O atual trabalho apresenta um estudo bibliométrico seguido de uma revisão bibliográfica da mineração de textos aplicada a indexação de documentos a fim de identificar os artigos mais importantes, os autores de maior relevância, as revistas científicas com maior abertura para o assunto e os artigos mais alinhados com o tema proposto. Quanto a metodologia, foram selecionados 105 artigos publicados entre 1998 e 2011 nas três bases escolhidas para este estudo, Scopus, Isi of knownledge e Scielo. Os resultados apresentados destacam três autores, três revistas científicas e oito artigos de maior influência com o tema proposto. Os resultados gerados contribuem para o melhor entendimento da mineração de textos dando base para novos pesquisadores e auxiliando no referencial teórico.

Palavras chave: Mineração de Textos, Bibliometria, Indexação da Informação

ABSTRACT

The use of Text Mining techniques in the discovery of relevant information in unstructured databases is growing considerably. This issue becomes increasingly relevant due to the volume of textual information disclosed on the web, as well as generated by the companies. The current study presents a bibliometric study followed by a literature review of text mining applied to document indexing to identify the most important articles, the most relevant authors, scientific journals with greater openness to the subject and the articles more closely aligned with the theme. Regarding the methodology, we selected 105 articles published between 1998 and 2011 on three bases chosen for this study, Scopus, SciELO and Isi of knownledge. The results presented highlight three authors, three scientific journals and eight papers with more influence with the theme. The results generated contribute to a better understanding of text mining basis for helping new researchers in the theoretical reference.

Keywords: Text Mining, Bibliometrics, Indexing Information

1. INTRODUÇÃO

O volume de informação textual disponível na internet vem crescendo consideravelmente a cada dia. Segundo Royal Pingdom (2012), no ano de 2008 estavam disponíveis para consulta 186 milhões de websites. Em 2011, foram 555 milhões de sites disponíveis. Assim, técnicas de Mineração de Textos vêm sendo abordadas visando aumentar a qualidade da informação recuperada.

A Mineração de Textos (MT) vem sendo aplicada com sucesso em diversas áreas a fim de descobrir informação não trivial através da definição de padrões (Aranha e Passos, 2006), assumindo desta forma características multidisciplinares abrangendo diversas áreas do conhecimento, entre elas, Ciência Cognitiva, Processamento de Linguagem Natural, Aprendizado de Máquina, Estatística, Recuperação de Informação e, principalmente, Mineração de Dados (Soares, 2008).

O processo de Descoberta de Conhecimento Textual é originado do Knowledge Discovery in Databases (KDD), que tem como objetivos extrair padrões relevantes em bases de dados estruturadas (Fayyad *et al.*, 1996). De forma análoga, KDT representa uma subárea da Inteligência Computacional responsável por extrair informação relevante e ainda oculta de documentos ou bases de dados não estruturadas. O KDT é composto das etapas de pré-processamento, com a finalidade de preparar, transformar, organizar e melhorar a qualidade do texto para a etapa seguinte, o processamento ou Mineração de Textos, que é o objetivo do KDT, onde as técnicas aplicadas variam de acordo com a finalidade, recuperação da informação, indexação, extração da informação, associação de documentos, sumarização, clusterização e classificação/categorização. Por fim a etapa de pós-processamento avalia os resultados da mineração (Barion e Lago, 2008).

O atual trabalho tem o objetivo de analisar a produção científica nacional e internacional relativa a Mineração de Textos aplicada a indexação de documentos, identificando os artigos de maior relevância, os autores mais importantes, as revistas científicas com maior abertura para publicações no tema e os artigos mais alinhados com o estudo realizado. Para isto é apresentado um estudo bibliométrico, seguido de uma revisão bibliográfica dos artigos que apresentaram maior importância. Este trabalho torna-se relevante à medida que apresenta um estudo sobre a evolução do assunto do ponto de vista bibliométrico e orienta novas pesquisas auxiliando no referencial teórico.

2. METODOLOGIA

A bibliometria apresenta-se como um conjunto de métodos matemáticos e estatísticos utilizados para investigar e quantificar os processos de comunicação escrita. Alguns parâmetros passíveis de estudo são autores, palavras-chave, citações, periódicos e publicações, ano de publicação, origem dos trabalhos, áreas do conhecimento entre outros. Os principais fundamentos da bibliometria são: Produtividade de Periódicos (Lei de BradFord), Produtividade Científica (Lei de Lotka) e Freqüência de Palavras (Lei de Zipf) (PAO, 1989).

Com o objetivo de realizar o levantamento dos dados para análise, foram definidos os seguintes parâmetros a serem aplicados no estudo bibliométrico:

a. Ano de publicação;

b. Citações;

c. Área do conhecimento;

d. Autores;

e. Revista científica.

As bases de artigos científicos que disponibilizam informações sobre os parâmetros selecionados foram: Scopus, Isi of Knowledge e Scielo. Foram aplicados comandos booleanos para unir "text mining" com "indexing", "indexed" e "documents". Desta forma, a chave de busca gerada para a seleção dos artigos foi:

"text mining" AND ("indexing" OR "indexed") AND "documents"

As chaves utilizadas nos mecanismos de busca das páginas web são idênticas. O resultado ainda foi refinado, na intenção de restarem apenas documentos do tipo "artigos", excluindo os demais tipos.

Na base Scopus, a consulta retornou 66 artigos publicados entre 1999 e 2011, em 15 áreas diferentes. Na base Isi of Knowledge foram encontrados 60 artigos publicados entre 1998 e 2011 de 20 áreas. A pesquisa realizada na base Scielo não retornou ocorrências. Para garantir que não existam artigos relacionados ao tema na base Scielo, a mesma pesquisa foi realizada traduzindo os termos para o português, onde, ainda assim, não foram encontradas ocorrências.

Após a coleta dos artigos nas bases, os mesmos foram comparados eliminando registros recorrentes. Nesta tarefa foram encontrados 21 artigos existentes em ambas as bases. Para a contabilização das citações, a referência ao artigo foi sempre ao da base que apresentou o maior número de citações.

Por fim, os artigos passaram por uma breve avaliação com o objetivo de identificar trabalhos que não estejam relacionados com Mineração de Textos e indexação de documentos. A aplicação desta etapa não teve impactos sobre os artigos selecionados.

O conjunto de artigos resultante da aplicação da primeira etapa contou com 105 artigos publicados entre 1998 e 2011.

3. RESULTADOS E DISCUSSÕES

3.1 Estudo bibliométrico

Após a definição do conjunto de artigos a ser analisado, a aplicação da etapa de estudos bibliométricos teve como objetivo responder através de dados quantitativos, três perguntas: Quais são os artigos de maior influência para o estudo proposto? Quem são os pesquisadores com maior número de documentos publicados? Quais são as revistas científicas com maior abertura a publicações no tema?

A Figura 1 demonstra o número de publicações por ano desde 1998 até 2011. O ano de 2008 se destaca pelo maior número de trabalhos publicados, seguido por 2009. As publicações no período de 2007 a 2011, somadas alcançam um percentual de 50,48 % do conjunto de artigos publicados.

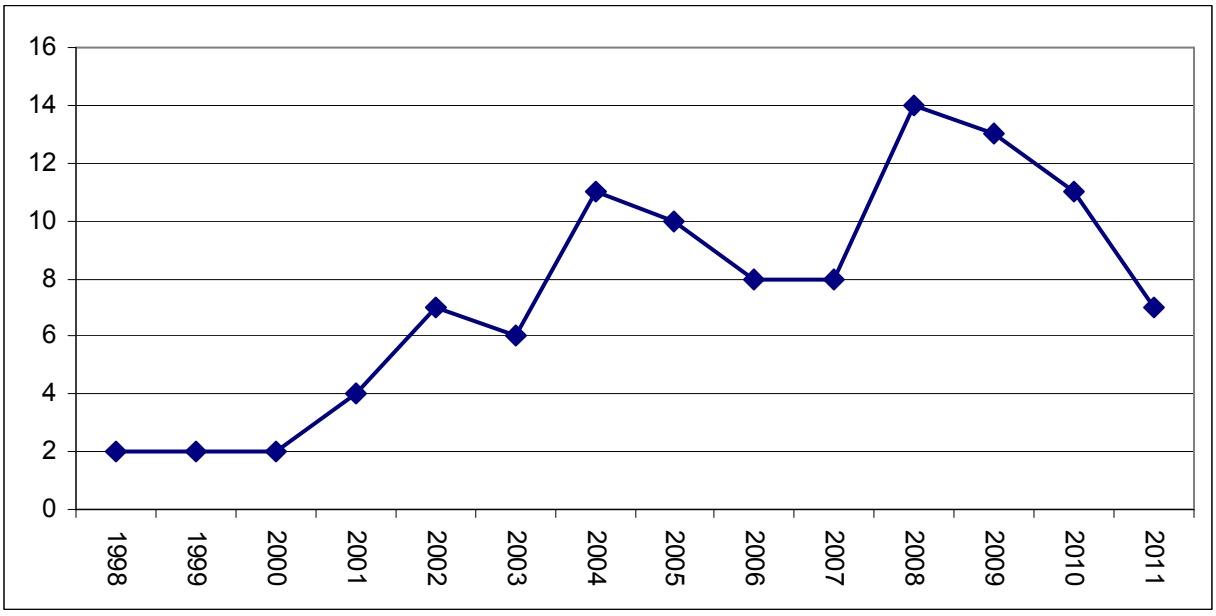

Figura 1 – Número de publicações por ano.

Os artigos selecionados formam fontes de referências para 1147 outros trabalhos desde 1999 até 2011. A Figura 2 apresenta o número de citações de acordo com o ano de publicação dos artigos citados.

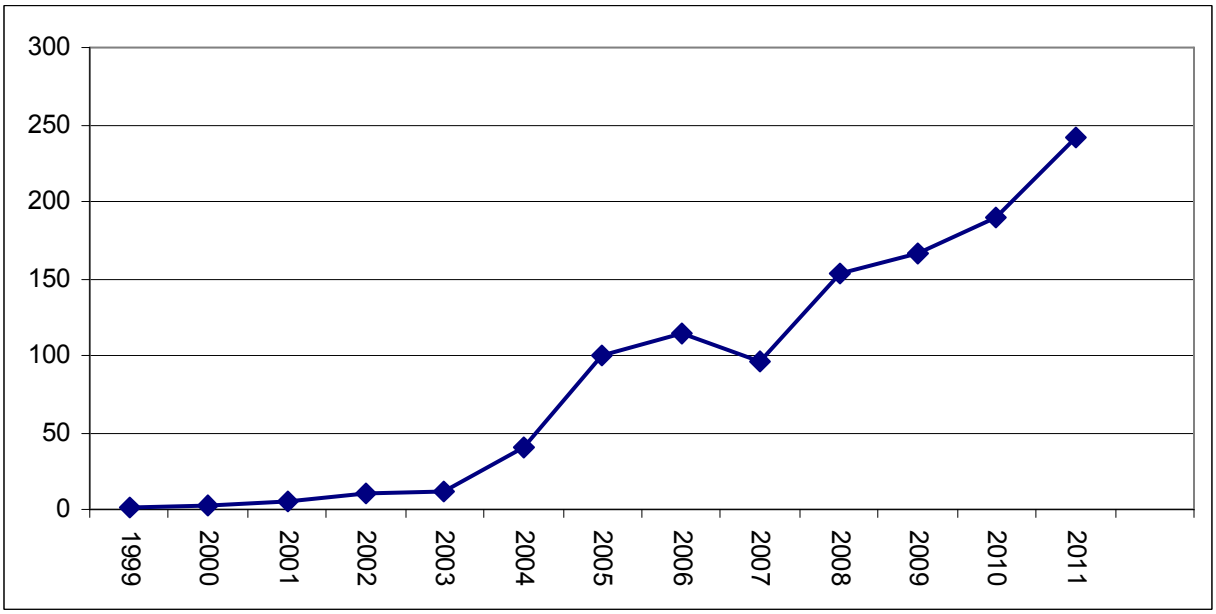

Figura 2 – Gráfico de citações por ano.

Através dos gráficos da Figura 1 e da Figura 2, é possível perceber um crescente interesse pelo tema. A cada ano, mais trabalhos são publicados com referências à Mineração de Textos. Isto devido ao aspecto multidisciplinar e pela possibilidade de aplicação em diversas áreas do conhecimento

A Figura 3 apresenta o gráfico das áreas que mais publicam estudos relacionados a Mineração de Textos para indexação de documentos. Destacam-se Computer Science (Ciência da Informação), logo após, Engineering (Engenharia), seguida por Mathematics (Matemática), Biochemistry, Genetics and Molecular Biology (Bioquímica, Genética e Biologia Molecular), e Medicine (Medicina).

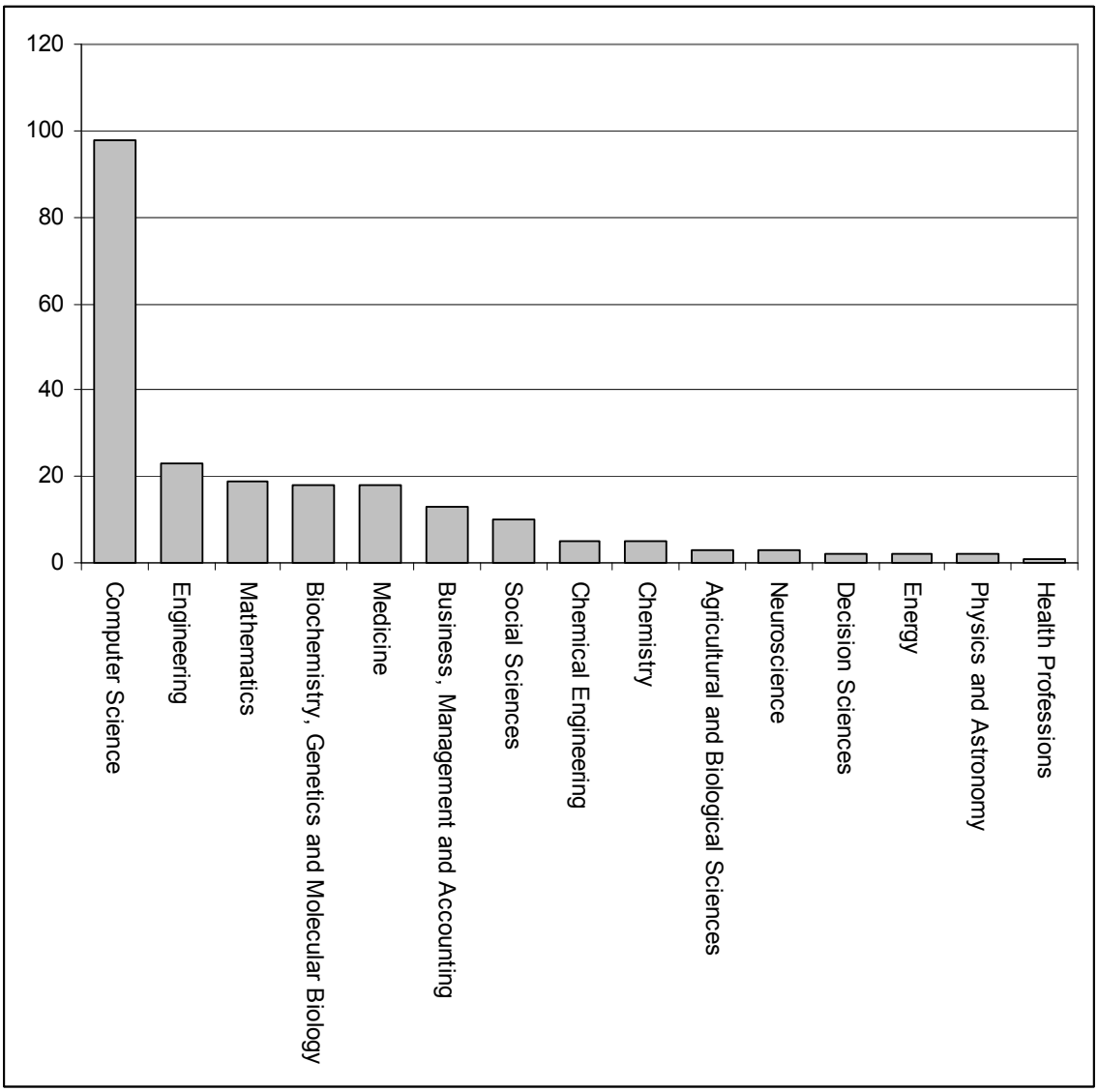

Figura 3 – Gráfico de publicações por áreas de conhecimento.

Para identificar os trabalhos de maior relevância no tema, a Figura 4 apresenta o ranking dos 20 artigos mais citados nas diversas áreas do conhecimento.

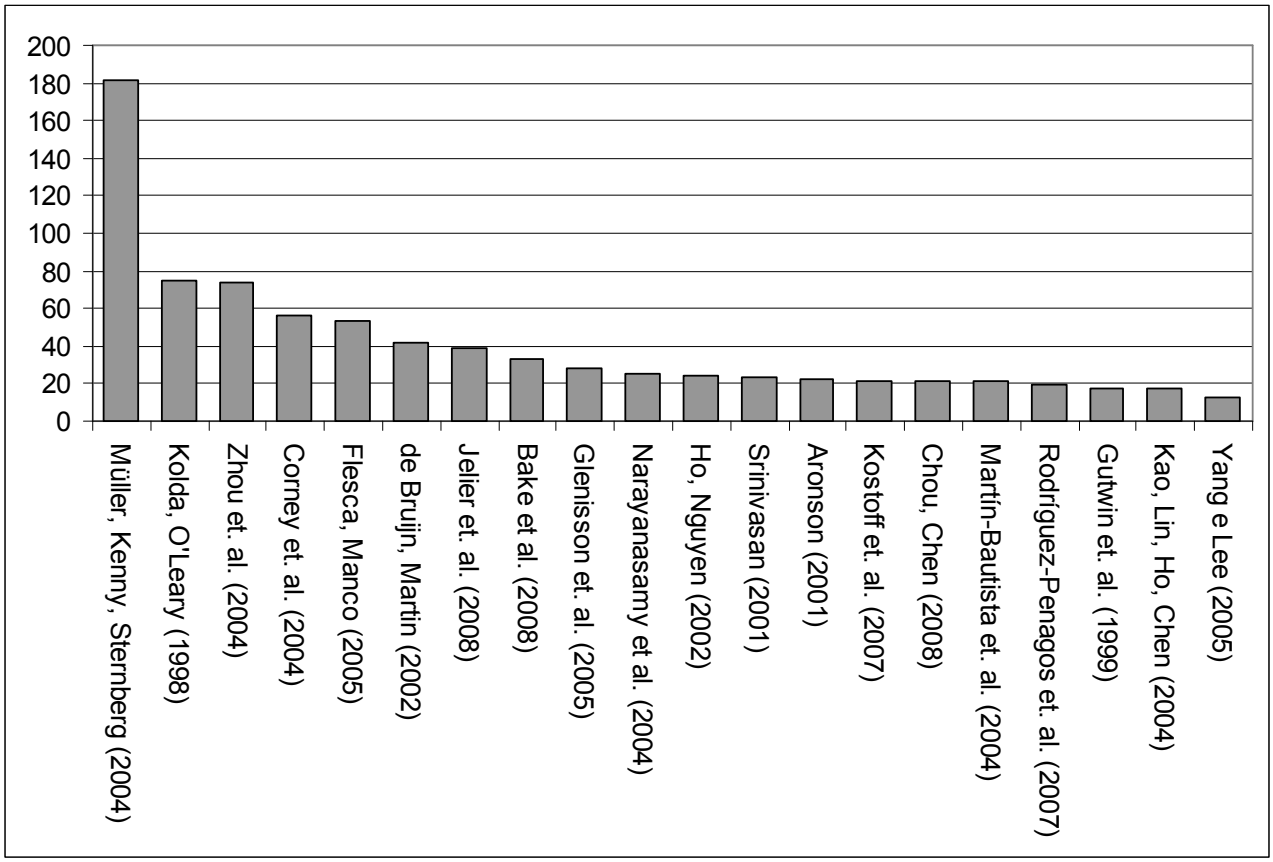

Figura 4 – Ranking dos artigos mais citados.

Os autores dos 20 trabalhos mais citados foram Müller *et al.* (2004), Kolda e O'leary (1998), Zhou *et al.* (2004), Corney *et al.* (2004), Flesca *et al.* (2005), De Bruijn e Martin (2002), Jelier *et al.* (2008), Baker *et al.* (2008), Narayanasamy *et al.* (2004); Glenisson *et al.* (2005); (Baker *et al.*, 2008), Ho e Nguyen (2002), Srinivasan (2001), Aronson (2001), Kostoff *et al.* (2007), Chou e Chen (2008), Martín-Bautista *et al.* (2004), Rodríguez-Penagos *et al.* (2007), Gutwin *et al.* (1999), Kao *et al.* (2004) e Yang e Lee (2005). Destes, observa-se que o artigo mais citado é (Müller *et al.*, 2004), com 182 referências ao seu trabalho, seguido por (Kolda e O'leary, 1998) e (Zhou *et al.*, 2004), com 75 e 74 referências respectivamente. Nota-se também que 2004 foi o ano onde três dos quatro trabalhos mais citados foram publicados.

A Figura 5 apresenta o gráfico dos autores que publicaram mais de um artigo no conjunto de documentos selecionado.

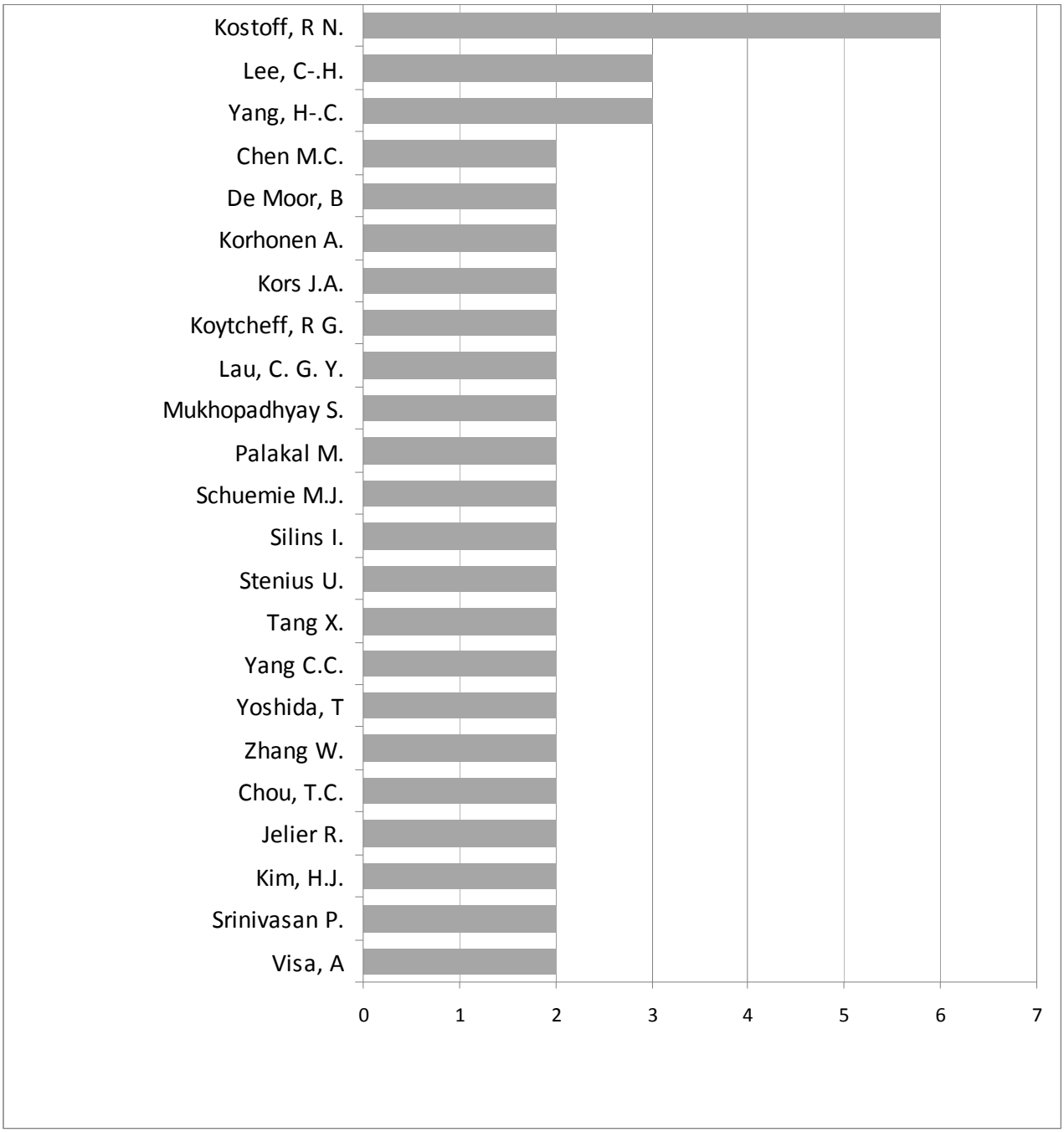

Figura 5 – Gráfico dos autores.

O gráfico da Figura 5 apresentou o número de publicações por autores. Ronald N. Kostoff publicou seis trabalhos na base selecionada, seguido por Chung-Hong Lee e Hsin-Chang Yang, os dois com três trabalhos publicados.

Um dos objetivos desta etapa é identificar as revistas científicas com grande abertura na área de Mineração de Textos. A Figura 6 apresenta o gráfico das revistas com mais de duas publicações no tema.

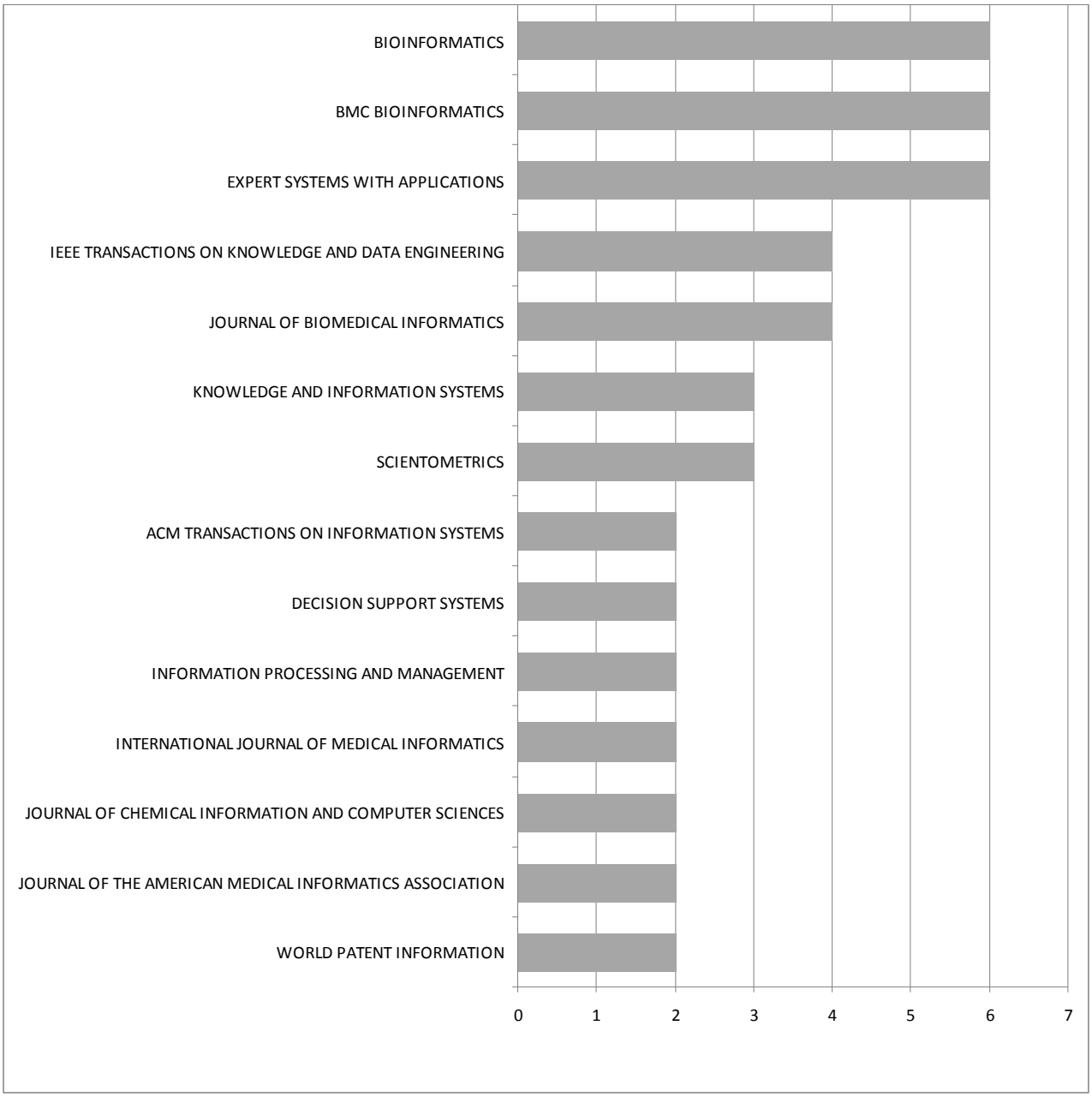

Figura 6 – Gráfico das revistas que mais publicaram no tema.

Analisando o gráfico, nota-se que grande parte das revistas que tem publicado artigos sobre Mineração de Textos está ligada a área das biomédicas. Bioinformatics, Bmc Bioinformatics e Expert Systems With Applications, publicaram cada uma, seis artigos.

3.2 Revisão Bibliográfica

A etapa de revisão bibliográfica busca fazer um apanhado geral das publicações consideradas de maior relevância para o tema proposto, considerando para isto, os resultados da etapa anterior. Os artigos selecionados para a revisão bibliográfica foram os cinco primeiros na ordem de maior influência para o tema. Além disto, são apresentados cinco artigos que foram considerados interessantes.

Müller *et al.* (2004) e Corney *et al.* (2004) descrevem a criação de um sistema de Mineração de Textos para auxiliar a tarefa de recuperação de informação em bases de artigos científicos na área biológica. Segundo (Müller *et al.*, 2004), para refinamento nas buscas são utilizadas ontologias que compreendem 14.500 termos, divididos em 33 categorias. Buscas feitas em bases de 19.800 documentos demonstram um resultado de 45% a 95% na recuperação de informações biológicas. Müller et al. (2004) ainda cita que a ontologia aumenta em aproximadamente três vezes a eficiência das pesquisas. (Corney *et al.*, 2004) demonstram resultados na ordem de 20,31% para recuperação de documentos a partir dos resumos, com precisão de 55,07%, enquanto que com textos completos atingiu recuperação de 43,6% e precisão de 51,25%.

Zhou *et al.* (2004) apresentam um sistema chamado "PowerBioNE", de reconhecimento de entidades nomeadas na área biomédica. Através do Modelo Oculto de Markov (Hidden Markov Model – HMM) são integrados diversos recursos como, termos morfológicos, padrões de formação de palavras, part-of-speech entre outros. Além disto, o algoritmo do K-Vizinho mais Próximo (K-Nearest Neighbor – K-NN) é utilizado para resolver problemas de escassez de dados. Os resultados apresentados demonstram superioridade do modelo algorítmico HMM e K-NN em relação aos modelos de Back-off HMM, Linear Interpolate HMM e Suport Vector Machine.

Kolda e O'leary (1998) propõe a substituição do método Simple-Value Decomposition (SVD) pelo SemiDiscrete Decomposition (SDD). Os resultados demonstram que a metodologia proposta requer vinte vezes menos capacidade de armazenamento e a metade do tempo de processamento.

Flesca *et al.* (2005) discutem uma abordagem para detecção de semelhanças em arquivos no formato XML. A proposta consiste em linearizar a estrutura do documento, transformando-o em uma cadeia de seqüências numéricas. Para identificar similaridades utiliza-se a comparação através de análise de freqüência. Os resultados experimentais demonstram eficácia do método comparado à forma padrão.

Guo *et al.* (2011) aplica a Mineração de textos para estruturar informações contidas em artigos científicos na área de Biomedicina a fim de otimizar o processo de pesquisas. Foram selecionadas três bases de diferentes tipos e granularidades. A primeira baseada em seleção de nomes, a segunda em Zonas Argumentativas (Argumentative Zones ou AZ) e por fim Núcleos de Conceitos Científicos (Core Scientific Concepts ou CoreSC). A proposta tinha foco em resumir textos sobre o tema "Avaliação de Risco de Câncer". Os resultados apresentados demonstram boa confiabilidade para os resumos onde foi aplicado aprendizado de máquina.

Zhang *et al.* (2011) afirma que a representação de texto inclui duas tarefas: Indexação e definição de pesos. Neste sentido, é apresentada uma comparação entre TF-IDF, LSI e Multi-Words para representação de texto, utilizando coleções de documentos nos idiomas Inglês e Chinês para avaliar a recuperação da informação e categorização de textos. Os resultados em fase experimental indicam que para a categorização da informação, LSI é melhor que os outros dois métodos nas duas coleções de documentos. Ainda, LSI produziu o melhor desempenho para recuperação de informação em textos de língua inglesa. Os resultados demonstraram que LSI é mais eficiente semântica e estatisticamente.

Chau e Yeh (2003) discutem uma nova abordagem para descoberta da informação em documentos multilíngües com destaques para documentos textuais na Web. Através da aplicação de um conceito multilíngüe que codifica a relação conceito-termo utilizando um mapa para auto-organização.

Urbain *et al.* (2009) apresentam um modelo para recuperação de informação bidimensional combinando conceitos de semântica e estatística com vários níveis de aprofundamento nos contextos do documento. A pesquisa foi aplicada a coleção TREC 2005 Genomics. Os resultados demonstram promissoras melhorias ultrapassando o estado da arte em 15,28%.

Jahiruddin *et al.* (2010) apresenta uma framework chamada Biomedical Knowledge Extraction and Visualization (BioKEVis) construída com o intuito de identificar chaves para estruturação de textos não-estruturados ou semi-estruturados. A aplicação do BioKEVis é específica para documentos de texto na área da biomedicina e aplica análise lingüística e LSA para a identificação de chaves conceituais. BioKEVis se apresentou uma boa opção para o desenvolvimento de aplicações no intuito de extrair informação para a área biomédica.

4. CONSIDERAÇÕES FINAIS

Através das informações quantitativas apresentadas no estudo bibliométrico, é possível responder às seguintes perguntas: Quais são os artigos de maior influência para o tema proposto? Quem são os pesquisadores com maior número de documentos publicados? Quais são as revistas científicas com maior abertura a publicações no tema?

Os artigos científicos com maior influência no tema, que são referência para diversos outros trabalhos, foram apresentados na Figura 4. Dentre eles, destacam-se em particular Müller *et al.* (2004), Kolda e O'leary (1998) e Zhou *et al.* (2004), com 182, 75 e 74 referências respectivamente.

Os autores considerados mais importantes neste trabalho são apresentados na Figura 5, destacando-se Ronald N. Kostoff com seis artigos publicados, Chung-Hong Lee e Hsin-Chang Yang com três artigos cada um.

Por fim, as revistas que se destacaram neste trabalho foram Bioinformatics, Bmc Bioinformatics e Expert Systems With Applications, cada uma com seis publicações no tema.

A etapa de revisão da literatura buscou definir quais os artigos mais alinhados com o tema proposto. Para isto, foi selecionado um conjunto de dez artigos, sendo os cinco mais citados e cinco considerados os mais interessantes.

O artigo considerado mais alinhado com o tema foi Zhang *et al.* (2011), apresentando estudos diretamente ligados a indexação de documentos. Logo após Urbain *et al.* (2009), Kolda and O'Leary (1998) e Zhou *et al.* (2004), propondo métodos para otimizar o processo de recuperação da informação. Por fim, Müller *et al.* (2004), Corney *et al.* (2004), Zhou *et al.* (2004) e Jahiruddin *et al.* (2010) apresentando novas ferramentas para recuperação da informação.

5. REFERÊNCIAS

ARONSON, A. R. **Effective mapping of biomedical text to the UMLS metathesaurus: The MetaMap Program.** Journal of the American Medical Informatics Association, p. 17-21, 2001 2001. ISSN 1067-5027.

BAKER, C. J. O.; KANAGASABAI, R.; ANG, W. T.; VEERAMANI, A.; LOW, H. S.; WENK, M. R. **Towards ontology-driven navigation of the lipid bibliosphere.** BMC Bioinformatics, v. 9, n. SUPPL. 1, 2008.

CHAU, R.; YEH, C. H. **A concept-based inter-lingua and its applications to multilingual text and Web mining.** In: LIU, J. C. Y. M. Y. H. J. (Ed.). Intelligent Data Engineering and Automated Learning, v.2690, 2003. p.756-760. (Lecture Notes in Computer Science). ISBN 0302-9743 3-540-40550-X.

CHOU, T. C.; CHEN, M. C. **Using incremental PLSI for threshold-resilient online event analysis.** IEEE

Transactions on Knowledge and Data Engineering, v. 20, n. 3, p. 289-299, 2008.

CORNEY, D. P. A.; BUXTON, B. F.; LANGDON, W. B.; JONES, D. T. **BioRAT: Extracting biological information from full-length papers.** Bioinformatics, v. 20, n. 17, p. 3206-3213, 2004.

DE BRUIJN, B.; MARTIN, J. **Getting to the (c)ore of knowledge: mining biomedical literature.** International Journal of Medical Informatics, v. 67, n. 1-3, p. 7-18, Dec 4 2002. ISSN 1386-5056.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. **Knowledge discovery and data mining: Towards a unifying framework.** Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, 1996.

FLESCA, S.; MANCO, G.; MASCIARI, E.; PONTIERI, L.; PUGLIESE, A. **Fast detection of XML structural similarity.** IEEE Transactions on Knowledge and Data Engineering, v. 17, n. 2, p. 160-175, 2005.

GLENISSON, P.; GLANZEL, W.; JANSSENS, F.; DE MOOR, B. **Combining full text and bibliometric information in mapping scientific disciplines.** Information Processing & Management, v. 41, n. 6, p. 1548-1572, Dec 2005. ISSN 0306-4573.

GUO, Y.; KORHONEN, A.; LIAKATA, M.; SILINS, I.; HOGBERG, J.; STENIUS, U. **A comparison and user-based evaluation of models of textual information structure in the context of cancer risk assessment.** BMC Bioinformatics, v. 12, 2011.

GUTWIN, C.; PAYNTER, G.; WITTEN, I.; NEVILL-MANNING, C.; FRANK, E. **Improving browsing in digital libraries with keyphrase indexes.** Decision Support Systems, v. 27, n. 1-2, p. 81-104, Nov 1999. ISSN 0167-9236.

HO, T. B.; NGUYEN, N. B. **Nonhierarchical document clustering based on a tolerance rough set model.** International Journal of Intelligent Systems, v. 17, n. 2, p. 199-212, Feb 2002. ISSN 0884-8173.

JAHIRUDDIN; ABULAISH, M.; DEY, L. **A concept-driven biomedical knowledge extraction and visualization framework for conceptualization of text corpora.** Journal of Biomedical Informatics, v. 43, n. 6, p. 1020-1035, 2010.

JELIER, R.; SCHUEMIE, M. J.; VELDHOVEN, A.; DORSSERS, L. C. J.; JENSTER, G.; KORS, J. **A. Anni 2.0: A multipurpose text-mining tool for the life sciences.** Genome Biology, v. 9, n. 6, 2008.

KAO, H. Y.; LIN, S. H.; HO, J. M.; CHEN, M. S. **Mining Web informative structures and contents based on entropy analysis.** IEEE Transactions on Knowledge and Data Engineering, v. 16, n. 1, p. 41-55, Jan 2004. ISSN 1041-4347.

KOLDA, T. G.; O'LEARY, D. P. **A semidiscrete matrix decomposition for latent semantic indexing in information retrieval.** Acm Transactions on Information Systems, v. 16, n. 4, p. 322-346, Oct 1998. ISSN 1046-8188.

KOSTOFF, R. N.; KOYTCHEFF, R. G.; LAU, C. G. Y. **Global nanotechnology research metrics.** Scientometrics, v. 70, n. 3, p. 565-601, Mar 2007. ISSN 0138-9130.

MARTÍN-BAUTISTA, M. J.; SÁNCHEZ, D.; CHAMORRO-MARTÍNEZ, J.; SERRANO, J. M.; VILA, M. **A. Mining web documents to find additional query terms using fuzzy association rules.** Fuzzy Sets and Systems, v. 148, n. 1, p. 85-104, 2004.

MÜLLER, H. M.; KENNY, E. E.; STERNBERG, P. W. **Textpresso: An ontology-based information**

retrieval and extraction system for biological literature. PLoS Biology, v. 2, n. 11, 2004.

NARAYANASAMY, V.; MUKHOPADHYAY, S.; PALAKAL, M.; POTTER, D. A. **TransMiner: Mining transitive associations among biological objects from text.** Journal of Biomedical Science, v. 11, n. 6, p. 864-873, 2004.

RODRÍGUEZ-PENAGOS, C.; SALGADO, H.; MARTÍNEZ-FLORES, I.; COLLADO-VIDES, J. **Automatic reconstruction of a bacterial regulatory network using Natural Language Processing.** BMC Bioinformatics, v. 8, 2007.

SRINIVASAN, P. **MeSHmap: a text mining tool for MEDLINE.** Proceedings / AMIA . Annual Symposium. AMIA Symposium, p. 642-646, 2001.

URBAIN, J.; GOHARIAN, N.; FRIEDER, O. **A dimensional retrieval model for integrating semantics and statistical evidence in context for genomics literature search.** Computers in Biology and Medicine, v. 39, n. 1, p. 61-68, Jan 2009. ISSN 0010-4825.

YANG, H. C.; LEE, C. H. **A text mining approach for automatic construction of hypertexts.** Expert Systems with Applications, v. 29, n. 4, p. 723-734, Nov 2005. ISSN 0957-4174.

ZHANG, W.; YOSHIDA, T.; TANG, X. **A comparative study of TF*IDF, LSI and multi-words for text classification.** Expert Systems with Applications, v. 38, n. 3, p. 2758-2765, 2011.

ZHOU, G. D.; ZHANG, J.; SU, J.; SHEN, D.; TAN, C. L. **Recognizing names in biomedical texts: A machine learning approach.** Bioinformatics, v. 20, n. 7, p. 1178-1190, 2004.