Mineração de textos biomédicos: uma revisão bibliométrica

Cristiane Raquel Woszezenki

Doutoranda do Programa de Pós Graduação em Engenharia e Gestão do Conhecimento da Universidade Federal de Santa Catarina

Alexandre Leopoldo Gonçalves

Professor Permanente do Programa de Pós Graduação em Engenharia e Gestão do Conhecimento da Universidade Federal de Santa Catarina

A mineração de textos vem sendo, cada vez mais, empregada para automatizar o processo de extração de informações importantes, contidas em textos biomédicos, possibilitando que os pesquisadores fiquem a par do desenvolvimento da biomedicina. Considerando a importância deste campo de pesquisa, este artigo apresenta um mapeamento das publicações científicas sobre mineração de textos biomédicos e discute as principais tarefas desse campo de pesquisa, as quais os pesquisadores têm dedicado maior atenção. Para isso, foi utilizada a bibliometria, uma técnica que permite analisar o desenvolvimento de um campo da ciência, visando identificar suas características. O mapeamento apresentado promove o conhecimento sobre o histórico e o estado atual do campo de pesquisa e disponibiliza insumos, que permitem enriquecer a discussão sobre os possíveis rumos que as pesquisas, na área, têm tomado e as prováveis tendências científicas para os pesquisadores e interessados no tema.

Palavras-chave: *Mineração de textos; Descoberta baseada em literatura; Bioinformática.*

Biomedical text mining: a bibliometrics review

Text mining has being increasingly used to automate the process of extracting the relevant information contained in biomedical texts, enabling researchers to stay abreast of the development of biomedicine. Taking into account the importance of this research field, this paper presents a map of scientific publications on biomedical text mining and discusses the main activities in this field in which the researchers have devoted attention. Thus, we used bibliometrics, a technique that allows us to analyze the development of a field of science to identify its characteristics. The mapping presented promotes knowledge about the history and current state of the research field. Also, it provides inputs aiming to enrich the discussion on possible directions that researches in the area have taken as well as likely scientific trends for researchers and interested in the subject.

Keywords: *Text mining; Literature based discovery; Bioinformatics.*

Recebido em 20.02.2013 Aceito em 14.05.2013

1 Introdução

A pesquisa científica é altamente dinâmica, devido a tecnologias inovadoras, capazes de causar mudanças em campos consolidados e mesmo criar novos campos de investigação. No domínio biomédico, o rápido avanço nas pesquisas tem causado um elevado aumento no número de publicações científicas, com descobertas cada vez mais importantes, impossibilitando que pesquisadores desta área conheçam as últimas informações e tendências em uma quantidade de tempo aceitável. Técnicas de mineração de textos vêm sendo utilizadas desde a segunda metade da década de 90, para automatizar o processo de extração de informações importantes contidas em textos biomédicos, possibilitando que os pesquisadores fiquem a par do desenvolvimento da biomedicina.

A mineração de textos pode ser entendida como o estudo e a prática de extrair padrões, regras e tendências a partir do texto completo de artigos científicos digitais, usando princípios da linguística computacional e métodos analíticos (SULLIVAN, 2001; FAIIAZEE *et al.*, 2012). No domínio biomédico, os trabalhos de mineração de textos buscam, em sua grande maioria, extrair relacionamentos proteína-proteína (ZHANG *et al.*, 2011), gene-gene (TIWARI; ZHANG; CHEN, 2009), gene-proteína (NAEEM *et al.*,

2010), droga-proteína (SHU; HUANG; ZHU, 2012), gene-doença (GONG *et al.*, 2012), entre outros, visando encontrar novos diagnósticos, prevenções e tratamentos (COHEN; HERSH, 2005).

A partir deste cenário, este artigo apresenta duas principais contribuições. A primeira é o mapeamento das publicações científicas sobre mineração de textos biomédicos, visando a identificar as características dos trabalhos, como autores que contribuem para o tema, período das publicações, instituições e países onde o tema é estudado, periódicos nos quais os trabalhos são publicados, entre outros. Para isso, são utilizadas técnicas bibliométricas, que empregam métodos quantitativos na busca por uma avaliação objetiva da produção científica. A segunda contribuição é a identificação e análise das tarefas da mineração de textos biomédicos, as quais os pesquisadores têm dedicado maior atenção. São discutidas as abordagens empregadas nessas tarefas atualmente. Essa análise foi realizada após a seleção e categorização dos estudos identificados na busca sistemática da literatura, de acordo com critérios pré-estabelecidos na seção 3.

O presente estudo é apresentado da seguinte forma: a seguir, são discutidos os aspectos conceituais da mineração de textos biomédicos; logo após, é apresentada a bibliometria como técnica de visualização e mapeamento científico; na sequência, são descritos os procedimentos metodológicos aplicados a este estudo, os resultados observados, as considerações finais e as referências bibliográficas utilizadas.

2 Mineração de textos biomédicos: aspectos conceituais

A Bioinformática é um campo emergente das pesquisas biomédicas e se preocupa com a aplicação de técnicas computacionais para o processamento de dados biológicos ou biomédicos (KIM, 2002). Uma das técnicas amplamente empregadas pela Bioinformática é a mineração de textos, com a finalidade de descobrir novos conhecimentos, a partir do processamento de bases textuais biomédicas.

Atualmente, a principal fonte de informações para a descoberta de conhecimento na área biomédica é a literatura científica armazenada no PubMed/MEDLINE[1]. O MEDLINE disponibiliza mais de 22 milhões de citações e resumos de publicações científicas das áreas da medicina, enfermagem, odontologia, medicina veterinária, biologia, bioquímica, evolução molecular, entre outros.

A mineração da literatura científica biomédica dedica-se a três principais tarefas: classificação de textos, identificação de termos e extração de relacionamentos (COHEN; HERSH, 2005). A classificação de textos objetiva determinar se um documento ou parte de um documento contém características de interesse (COHEN; HERSH, 2005), de forma a reduzir o conjunto de textos a serem processados (SIMPSON, 2012).

[1] PubMed – Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed>. Acesso em: 2 fev. 2013.

A identificação de termos dedica-se a reconhecer automaticamente termos biomédicos no texto, processo este chamado de reconhecimento de entidades. Termos comuns, neste domínio, incluem nomes de genes, proteínas, drogas e suas dosagens, problemas médicos e tratamentos, entre outros (SIMPSON, 2012). Neste sentido, o reconhecimento de entidades é o primeiro passo para a extração de relacionamento entre as mesmas.

A extração de relacionamentos tem por objetivo detectar a ocorrência de relacionamentos entre entidades, como, por exemplo, relacionamentos entre genes, proteínas, genes e doenças, drogas e doenças, entre outros. O relacionamento pode ser explícito ou implícito (GANDRA; PRADHAN; PALAKAL, 2010). Relacionamentos explícitos são conhecidos e estão declarados no texto (por exemplo, "o gene G inibe a proteína P"). Já os relacionamentos implícitos são inferidos a partir da existência de um ou mais relacionamentos explícitos, de forma a gerar hipóteses (que podem guiar experimentos em laboratórios) potenciais para novas descobertas (COHEN; HERSH, 2005). Este tipo de relacionamento é também chamado de associações indiretas (TSURUOKA *et al.*, 2011) ou associações transitivas (JAYADEVAPRAKASH; MUKHOPADHYAY, PALAKAL, 2005).

As associações indiretas são objeto de estudo da Descoberta Baseada em Literatura (DBL), uma vertente da mineração de textos biomédicos. As descobertas se dão na forma de conexão implícita entre dois conceitos primários, como, por exemplo, um medicamento para um gene que é a causa de uma doença (HRISTOVSKI *et al.*, 2006). Assim, a DBL pode ser utilizada para fazer a conexão entre conjuntos distintos de literatura, relacionando diferentes disciplinas ou especialidades (GANIZ; POTTENGER; JANNECK, 2006). Suponha que uma comunidade científica sabe que B é uma das características da doença C. Outro grupo de pesquisadores sabe que a substância A afeta B. A descoberta, nesse caso, é o levantamento da associação indireta de A com C, por meio de B (WEEBER *et al.*, 2001).

A DBL foi introduzida por Don Swanson, na década de 1980, quando não existiam ferramentas de mineração automática de textos. Nesta época, as pesquisas manuais realizadas por Don Swanson, na literatura científica, resultaram na associação da "Síndrome de Raynaud" (uma condição que resulta em restrição intermitente do fluxo sanguíneo para os dedos, disparado pelo frio ou estímulos emocionais) com o "óleo de peixe", por meio da "alta viscosidade do sangue" (SWANSON, 1986). Posteriormente a essa importante descoberta, o surgimento de métodos e ferramentas computacionais de mineração de textos biomédicos alavancou a descoberta de conhecimento, a partir da literatura científica.

3 Método

Este trabalho tem natureza exploratória de caráter descritivo (VERGARA, 2003) e faz uso de técnicas bibliométricas. A bibliometria é uma técnica de medição de índices de produção e disseminação do

conhecimento científico (FONSECA, 1986). Seu ponto central é a utilização de métodos quantitativos, na busca por uma avaliação objetiva da produção científica (ARAUJO, 2006).

Os indicadores bibliométricos possibilitam analisar o desenvolvimento de um campo da ciência, de forma a identificar características, como: o crescimento cronológico da produção científica; a produtividade de autores e instituições; a colaboração entre pesquisadores e instituições; o impacto das publicações; e a análise e avaliação de fontes difusoras de trabalhos e a dispersão da produção científica entre as diversas fontes (BUFREM; PRATES, 2005). A observação dessas características, para uma determinada área do conhecimento, revelam sua evolução e as principais tendências das publicações científicas.

3.1 Procedimentos metodológicos

O desenvolvimento deste estudo foi realizado em quatro etapas: 1) coleta de dados; 2) representação e análise dos dados; 3) seleção e categorização dos trabalhos para análise descritiva; e 4) análise descritiva das principais tarefas da mineração de textos biomédicos e suas respectivas técnicas e ferramentas empregadas.

A subseção 3.1.1 explica a etapa da coleta de dados (etapa 1). A seção 4 faz a representação e análise dos dados coletados por meio de tabelas, gráficos e figuras (etapa 2). A seção 5 apresenta os critérios de seleção e categorização de trabalhos para análise (etapa 3) e os resultados da análise descritiva (etapa 4).

3.1.1 Coleta de dados

Para realizar o mapeamento das publicações científicas sobre mineração de textos biomédicos, foi utilizada como fonte de informação a Scopus, uma base de dados que indexa publicações científicas multidisciplinares do mundo todo, sendo reconhecida cientificamente tanto pela quantidade quanto pela qualidade dos periódicos científicos indexados. A escolha pela base Scopus se deu após buscas realizadas em outras bases relevantes para a área de estudo em questão, como a IEEE, a ScienceDirect e a Web of Science (WoS). Os resultados das buscas mostraram que a Scopus indexa um número bem maior de trabalhos do que as outras bases (mais do que o dobro) e contempla pelo menos 80% dos trabalhos indexados em cada uma delas. Assim, constatou-se que os dados coletados da Scopus são suficientes para fornecer um mapa bibliográfico satisfatório da área pesquisada e optou-se por utilizar apenas essa base. Além disso, opção por trabalhar com apenas uma base também levou em conta o fato de que a utilização de mais de uma base de dados dificulta o trabalho da contabilização de certas informações bibliométricas, uma vez que os artigos se repetem em diferentes bases e os dados a serem importados pela ferramenta de apoio à análise devem ser normalizados.

Após a escolha da base de dados, foram estabelecidos os critérios de busca. Com a finalidade de realizar uma ampla cobertura das publicações sobre o tema da mineração de textos biomédicos, foi investigada a área como um todo, de forma a coletar todos os possíveis estudos desenvolvidos. Assim, as buscas foram realizadas, utilizando-se a expressão ***(biomedical AND "text mining") OR ("literature based discovery")***, no campo de busca correspondente a "título, palavras-chave, resumo". A inclusão da expressão *"literature based discovery"* se faz necessária, pois, conforme discutido na seção 2, diversos trabalhos dessa área dedicam-se a levantar associações indiretas entre conceitos biomédicos, fazendo uso da mineração de textos.

Todas as sub-bases disponíveis na Scopus foram utilizadas: *Life Sciences*, *Health Siences*, *Physical Sciences* e *Social Science & Humanities.* O período aplicado à busca foi também o disponível na base até o último ano incompleto: 1960-2012 (setembro, mês em que a busca foi realizada).

A partir dos registros retornados pela base, foi possível gerar um arquivo com as informações bibliométricas dos trabalhos. A obtenção dos resultados (apresentados na Seção 4) foi viabilizada por meio da importação desses arquivos para os *softwares EndNote e HistCite*. Esses *softwares* possibilitam a organização e visualização dos dados bibliográficos provenientes de bases que indexam publicações, permitindo uma análise dos dados mais completa.

4 Representação e análise dos dados

As buscas realizadas retornaram 585 trabalhos científicos. Esses trabalhos foram escritos por 1464 autores, vinculados a 160 instituições, de 46 países diferentes e estão publicados em 230 periódicos. A Tabela 1 sintetiza os resultados gerais da pesquisa.

Tabela 1 – Resultados gerais.

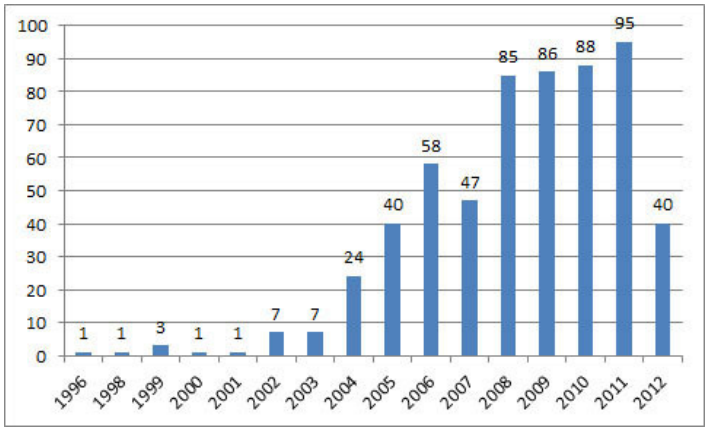
Informações bibliométricas	Quantidade
Artigos	585
Autores	1464
Periódicos	230
Países	46
Instituições	160
Palavras-chave	3134

Fonte: Dados da pesquisa.

O Gráfico 1 apresenta a distribuição temporal dos 585 trabalhos encontrados. O primeiro trabalho identificado na área da mineração de textos biomédicos foi realizado por Gordon e Lindsay (1996), no qual os autores replicam a descoberta de Swanson, em 1986, a respeito da conexão entre a Síndrome de Raynaud e o óleo de peixe. Eles fizeram uso de técnicas estatísticas, utilizando as publicações científicas do MEDLINE.

Percebe-se que, a partir deste momento, a mineração de textos biomédicos vem evoluindo, de forma crescente.

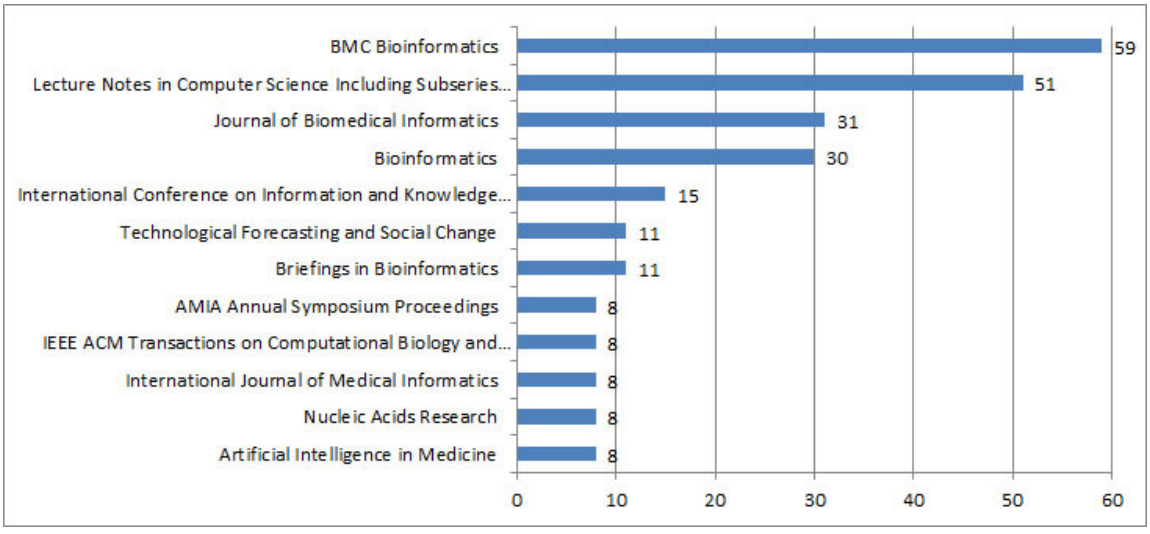
Gráfico 1 – Frequência das publicações por ano no período (1996-2012)



Fonte: Dados da pesquisa.

Na sequência, foram analisados os periódicos com maiores frequências de artigos publicados sobre o tema. O Gráfico 2 apresenta os doze periódicos com maior quantidade de publicações. Quase metade das publicações (248 artigos, representando 42,3% do total de trabalhos) está concentrada em 5% do total de periódicos. Destacam-se os seguintes periódicos: *BMC Informatics*, *Lecture Notes in Computer Science Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics*, *Journal of Biomedical Informatics*, *Bioinformatics*.

Gráfico 2 – Periódicos com maior frequência de publicações

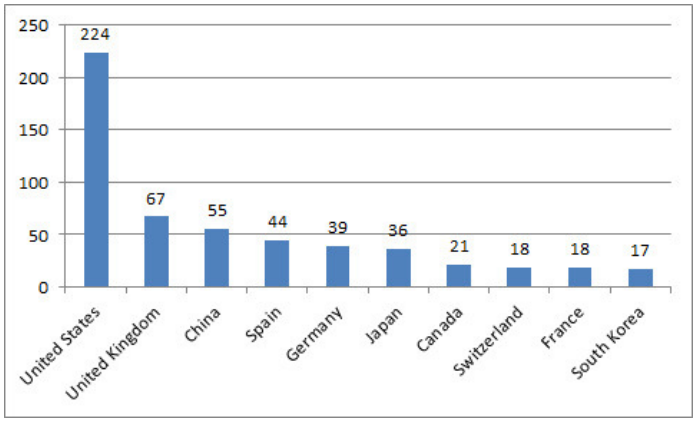


Fonte: Dados da pesquisa.

Quanto aos países de origem das publicações, os Estados Unidos lideram a lista ,com uma frequência de 224 artigos publicados, representando 38% da quantidade total de trabalhos, enquanto que os

62% restantes estão distribuídos entre 45 países. Após os Estados Unidos, destacam-se o Reino Unido, a China, a Espanha, a Alemanha e o Japão. O Gráfico 3 apresenta os dez países com maior quantidade de publicações.

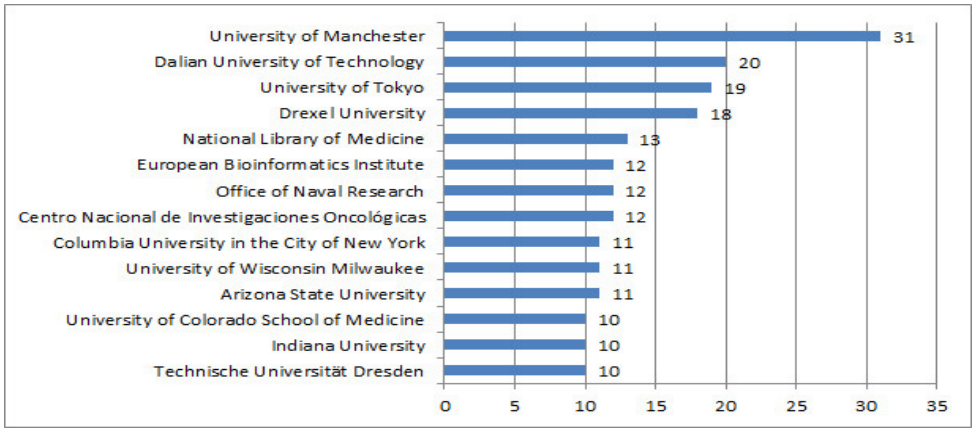
Gráfico 3 – Países com maior frequência de publicações



Fonte: Dados da pesquisa.

Dentre as 160 instituições que estudam sobre a mineração de textos biomédicos, observou-se que o maior número de publicações está distribuído em 14 instituições, conforme o Gráfico 4. As 14 instituições estão localizadas nos cinco primeiros países que mais publicam trabalhos sobre o tema (Gráfico 3). A maior parte dessas instituições (8) é dos Estados Unidos, primeiro colocado na lista dos países. Embora os Estados Unidos sejam o país que mais publicam sobre o tema, a *University of Manchester* do Reino Unido lidera a lista das instituições com 31 trabalhos, seguida pela *Dalian University of Technology,* da China, com 20 trabalhos. Observa-se, também, na lista das instituições que mais publicam, a Biblioteca Nacional de Medicina dos Estados Unidos[2], a maior biblioteca médica do mundo, com 13 publicações. Ela é responsável por criar e manter o PubMed/MEDLINE, a base de estudos científicos mais utilizada para a mineração de textos biomédicos.

Gráfico 4 – Instituições com maior frequência de publicação



Fonte: Dados da pesquisa.

[2] National Library of Medicine – Disponível em: <http://www.nlm.nih.gov/>. Acesso em: 2 fev. 2013.

Os autores com maior número de trabalhos publicados sobre a mineração de textos biomédicos, juntamente com seu vínculo institucional (afiliação) e país, são apresentados na Tabela 2. Os dois principais autores são Ananiadou S. e MIYANISHI H., com 17 trabalhos cada um, seguidos de Li Y. e Yang Z., com 16 trabalhos cada um. Os quatro primeiros autores pertencem as duas primeiras instituições que mais publicam (Gráfico 4).

Tabela 2 – Autores com maior número de publicações e sua afiliação

Autor	Número de publicações	Afiliação	País
Ananiadou, S.	17	University of Manchester	United Kingdom
Lin, H.	17	Dalian University of Technology	China
Li, Y.	16	Dalian University of Technology	China
Yang, Z.	16	Dalian University of Technology	China
Hu, X.	14	Henan University	China
Kostoff, R. N.	14	Office of Naval Research	United States
Krallinger, M.	13	National Center of Biotechnology	Spain
Valencia, A.	13	National Center of Biotechnology	Spain
Rebholz-Schuhmann, D.	12	European Bioinformatics Institute	United Kingdom
Tsujii, J.	12	University of Tokyo	Japan

Fonte: Dados da pesquisa.

Por fim, foi analisada a frequência de citações dos estudos sobre a mineração de textos biomédicos. A Tabela 3 apresenta os dez trabalhos mais citados. Observa-se que os dois principais trabalhos são intitulados: *A survey of current work in biomedical text mining*, com 187 citações e *Mining the Biomedical Literature in the Genomic Era: An Overview*, com 127 citações.

Tabela 3 – Artigos mais citados sobre o tema mineração de textos biomédicos

Autor	Título	Ano	Total de citações*
Cohen A. M., Hersh W. R.	*A survey of current work in biomedical text mining*	2005	187
Shatkay H., Feldman R.	*Mining the Biomedical Literature in the Genomic Era: An Overview*	2003	127
Tanabe L. *et al.*.	*MedMiner: An Internet text-mining tool for biomedical information, with application to gene expression profiling*	1999	110
Settles B.	*ABNER: An open source tool for automatically tagging genes, proteins and other entity names in text*	2005	107
Chen H., Sharp B. M.	*Content-rich biological network constructed by mining PubMed abstracts*	2004	107
Tiffin N. *et al.*.	*Integration of text- and data-mining using ontologies successfully selects disease gene candidates*	2005	103
Ananiadou S. *et al.*.	*Text mining and its potential applications in systems biology*	2006	102

Srinivasan P.	*Text Mining: Generating Hypotheses from MEDLINE*	2004	99
Krauthammer M., Nenadic G.	*Term identification in the biomedical literature*	2004	85
Bunescu R. *et al.*.	*Comparative experiments on learning information extractors for proteins and their interactions*	2005	83

Fonte: Dados da pesquisa.

*Mensuradas a partir do *Global Citation Score* – Escore Global de Citações (GCS): quantidade de vezes que os trabalhos foram citados por outros trabalhos, na base Scopus.

5 Identificação das tarefas da mineração de textos biomédicos

Um dos objetivos deste trabalho é a identificação de focos de estudo da área da mineração de textos biomédicos. Para tanto, foi necessário realizar a categorização de cada um dos 585 artigos obtidos na busca sistemática. A categorização foi realizada de acordo com as diversas tarefas executadas dentro da mineração de textos biomédicos (por exemplo, recuperação de documentos, extração de termos, desenvolvimento de ontologias, extração de relacionamentos, entre outros). Isso foi possível mediante leitura e análise (i) do título; (ii) do resumo; e (iii) das palavras-chave de cada artigo. Nos casos em que, após analisar os três campos citados, não fosse possível identificar fidedignamente o foco do estudo, recorreu-se ao texto completo do artigo, para melhor compreensão.

Assim, foi possível identificar as principais tarefas da mineração de textos biomédicos para as quais os pesquisadores têm dedicado maior atenção, sendo elas: (i) recuperação de documentos, (ii) classificação de documentos, (iii) extração de termos, (iv) anotação de documentos, (v) extração de relacionamentos, (vi) mineração de imagens e (vii) desenvolvimento de ontologias. Após a categorização dos estudos, partiu-se para a análise dos textos completos, de forma a identificar as abordagens de soluções exploradas pelos pesquisadores em cada tarefa. A seguir, cada uma dessas tarefas é discutida, juntamente com as respectivas abordagens utilizadas.

5.1 Recuperação de documentos

A interface de busca mais utilizada no domínio biomédico é a interface do PubMed que, embora seja flexível, retorna uma grande quantidade documentos que deve ser analisada pelo usuário, de forma a identificar documentos relevantes (CHOI *et al.*, 2011). Assim, diversos pesquisadores têm trabalhado no desenvolvimento de motores de busca, na tentativa apresentar melhores resultados. Huang e Hsu (2008) criaram o PubMed Smarter que se utiliza de uma árvore de relacionamentos de palavras, de modo que estas possam ser incorporadas na consulta do usuário.

Alguns trabalhos têm se dedicado a incorporar semântica aos motores de busca (DIETZE; SCHROEDER, 2009; CHOI *et al.*, 2011; THOMAS *et al.*, 2012). Choi *et al.* (2011) desenvolveram o BOSS, um motor de busca que indexa segmentos (frases, cláusulas ou sentenças) semanticamente coerentes, em um determinado contexto (por exemplo, objetos biomédicos e suas relações). Para uma busca, BOSS encontra todos os segmentos correspondentes e identifica os objetos de contexto presentes. Como resultado, o usuário recebe uma lista de objetos classificados com seus respectivos segmentos.

Dietze e Schroeder (2009) desenvolveram o GoWeb, um motor de busca para a internet que faz uso da pesquisa tradicional baseada em palavras-chave e filtra os resultados de acordo com categorias fornecidas por ontologias (*GeneOntology* (GO) e *Medical Subject Headings* (MeSH)). Thomaz *et al.* (2012) também desenvolveram um motor de busca semântica para o PubMed, o GeneView, construída sob uma versão semanticamente anotada dos resumos disponibilizados pelo PubMed.

5.2 Classificação de documentos

A classificação de documentos determina se um documento em particular tem características de interesse, ou seja, se possui determinado tipo de informação ou discute algum tópico específico. O usuário tipicamente fornece um *training set* composto pelos documentos, contendo as características de interesse (*training set* positivo) e outro que não tem essas características (*training set* negativo). Esses métodos automaticamente aprendem as características para diferenciar positivos de negativos, usando aprendizado de máquina (GONŹALEZ; IGLESIAS; DIZ, 2012).

Diversos trabalhos realizam a classificação de documentos, explorando o conhecimento representado pelas ontologias (KASTRIN; HRISTOVSKI, 2008; YOO; HU; SONG, 2007). A classificação de documentos também faz uso de técnicas de agrupamento (*clustering*), de forma a agrupar documentos similares (que discutem o mesmo tópico de interesse) em grupos (KANG *et al*., 2010).

5.3 Extração de termos

No domínio biomédico, termos de interesse são genes (TCL1A), proteínas (PKB), drogas (*warfarin*), doenças (*Raynaud's Syndrom*), entre outros. Essas entidades específicas são também chamadas de "entidades nomeadas" e a tarefa de identifica-las é chamada de *Named Entity Recognition* (NER).

O reconhecimento de entidades é um pré-requisito para a extração de relacionamentos. Três principais abordagens são utilizadas: abordagem léxica, baseada em regras e baseada em aprendizado de máquina. A abordagem léxica considera um dicionário, contendo nomenclaturas das entidades (TOHIDI; IBRAHIM; AZMI, 2011). Já a abordagem baseada em

regras constrói regras ou padrões que podem ser encontrados na literatura (GONG *et al.*, 2009).

Por outro lado, a abordagem baseada em aprendizado de máquina (MUNKHDALAI *et al.*, 2011) utiliza *training sets*: um positivo, contendo características que identificam as entidades (por exemplo, a presença de hifens, dígitos ou colchetes como em COS-X ou BRCA1) e um negativo, que não contém essas características. Essa utiliza modelos estatísticos, sendo os dois principais: *Conditional Random Fields* (CRF) (CAMPOS; MATOS; OLIVEIRA, 2010) e o *Support Vector Machine* (SVM) (JU; WANG; ZHU, 2011).

Os métodos baseados em regras dependem de regras elaboradas manualmente para descrever a composição das entidades e seu contexto, o que acarreta um consumo de tempo considerável. Métodos baseados em dicionários trabalham com coleções de nomes extensas. Métodos baseados em aprendizagem de máquina dependem da qualidade, da quantidade e da seleção de conjunto de características adequados para os *training sets* (WEI, 2009).

Percebe-se que os métodos mais amplamente utilizados para a extração de termos são o baseado em dicionário e o baseado em aprendizagem de máquina. Além disso, diversos trabalhos exploram a combinação de diferentes abordagens para a tarefa de extração de termos, como abordagens baseada em regras e estatísticas (HUA *et al.*, 2011), baseada em dicionários e estatísticas (LIN; LI; YANG, 2007), baseada em regras e dicionários (GONG *et al.*, 2009).

5.4 Anotação de documentos

A anotação de dados textuais não estruturados com metadados estruturados, que podem ser lidos por máquina, é um importante passo para outras tarefas da mineração de textos, como recuperação e classificação de documentos e a extração de informações. A anotação dos documentos nas bases de dados pode ser realizada manualmente por especialistas que fazem a leitura e interpretação de cada artigo e inferem conceitos existentes a partir de vocabulários controlados (ontologias, terminologias, taxonomias) a serem utilizados para anotação dos artigos. Artigos selecionados para inclusão no PubMed, por exemplo, são anotados com termos MeSH, um tesauro utilizado para indexar documentos (MORCHEN *et al.*, 2008).

Embora o processo manual confira qualidade e confiabilidade para as anotações, ele demanda um grande consumo de tempo. Assim, diversos trabalhos tem se dedicado a desenvolver técnicas para extrair anotações automáticas do texto (KRALLINGER *et al.*, 2012). Morchen *et al.* (2008), por exemplo, propõem um método automatizado para anotar documentos, utilizando uma abordagem probabilística.

Os vocabulários controlados têm sido fundamentais no processo de anotação, tanto manual quanto automático. Existem diversos tipos de

vocabulários controlados que são utilizados pelos pesquisadores para anotar as publicações:

a)ontologias: como a GeneOntology (KRALLINGER *et al.*, 2012);

b)taxonomias: como a NCBI *Taxonomy*, contendo a classificação e nomenclatura de 10% de todas as espécies do planeta (BADA *et al.*, 2012); e

c)tesauros: como o UMLS, MeSH (PLAKE *et al.*, 2009).

5.5 Extração de relacionamentos

A extração de relacionamentos é a tarefa primordial da mineração de textos biomédicos, uma vez que é a principal responsável pela extração de informações relevantes que podem gerar hipóteses potenciais para novas descobertas. Quase metade dos estudos identificados na busca sistemática, da literatura realizada neste trabalho, possui como foco a extração de relacionamentos.

O objetivo da extração de relacionamentos é identificar associações entre entidades previamente extraídas na tarefa de extração de termos. Diversos tipos de relacionamentos têm sido identificados: genes com doenças, genes com genes, genes com drogas, doenças com drogas, proteínas com proteínas, entre outros. Os relacionamentos extraídos podem gerar novas hipóteses para guiar pesquisas em laboratório.

Grande parte dos trabalhos utilizam métodos baseados na coocorrência para identificar relacionamentos entre entidades (JAYADEVAPRAKASH; MUKHOPADHYAY, PALAKAL, 2005; TSURUOKA *et al.*, 2011). Esses métodos extraem relacionamentos por analisar a frequência conjunta em que duas entidades coocorrem em um *corpus*. A ideia por trás dessa abordagem é que entidades que aparecem frequentemente juntas em pedaços de textos provavelmente estão relacionadas (GARTEN; COULET; ALTMAN, 2010). Contudo, a coocorrência produz um grande número de falsos relacionamentos (HRISTOVSKI, 2006). Para tratar desse problema, outras abordagens são aplicadas para melhorar os resultados, tais como: classificações probabilísticas utilizando *Naive Bayes* (LI *et al.*, 2010) e *Misture Aspect Model* (ZHU *et al.*, 2006), Teoria de Conjuntos Difusos e Redes Livres de Escala (WREN *et al.*, 2006).

Além disso, a coocorrência não revela a natureza dos relacionamentos (HRISTOVSKI, 2006). Assim, diversos trabalhos fazem uso do Processamento de Linguagem Natural (PLN), explorando a sintaxe das sentenças para obter informações sobre os relacionamentos explicitando-os na forma "A afeta/vincula-se a/regula/interage com B" (HRISTOVSKI, 2006; SHARMA; SWAMINATHAN; YANG, 2010). Métodos baseados na sintaxe criam regras para extrair relacionamentos de interesse. Por exemplo, o padrão <droga> <ação> <gene>, onde

<droga> e <gene> são entidades reconhecidas e <ação> pode ser qualquer verbo de uma lista como "inibe", "induz", "regula", "é metabolizado por" (GARTEN; COULET; ALTMAN, 2010).

Vários estudos baseiam-se em técnicas semânticas para extrair relacionamentos, como *Latent Semantic Indexing* (KOSTOFF *et al.*, 2008), *Semantic Role Labeling* (BARNICKEL *et al.*, 2009), *Conditional Random Fields* (TIWARI; ZHANG; CHEN, 2009), *Natural Language Processing* (HRISTOVSKI *et al.*, 2006), *Predication-based Semantic Indexing* (COHEN *et al.*, 2011), *Latent Semantic Analysis* (COHEN; SCHVANEVELDT; WIDDOWS, 2010). O conhecimento semântico contido em ontologias ou terminologias biomédicas também tem sido utilizado para a extração de relacionamentos (KANG *et al.*, 2011).

É importante ressaltar que os relacionamentos podem ser diretos, como no caso de relacionamentos extraídos por meio de regras baseadas em sintaxe (por exemplo, "A inibe B") ou podem ser indiretos (implícitos ou transitivos), quando não estão explicitamente declarados no texto e podem ser inferidos a partir de outros relacionamentos. Os relacionamentos implícitos são interesse de uma grande parte dos trabalhos que focam na extração de relacionamentos (TSURUOKA *et al.*, 2011; MIYANISHI; SEKI; UEHARA, 2010; SWANSON; SMALHEISER; TORVIK, 2006). A ideia, por trás disso, é formular hipóteses potenciais para novas descobertas que ainda não foram comprovadas e publicadas.

A DBL é a principal responsável por estudar técnicas para extrair relacionamentos indiretos. O modelo de descoberta mais conhecido é o modelo ABC de Swanson (SWANSON, 1986), pioneiro na formulação de associações indiretas. Por exemplo, considerando a afirmação em um determinado artigo de que "A afeta B", e a informação de que "B afeta C" apresentada por outro artigo, pode-se derivar a afirmação implícita de que "A afeta C" (SMALHEISER, 2011). Ou seja, a premissa dessa abordagem é que existem duas disciplinas ou estruturas de conhecimento científico que não se comunicam diretamente. Contudo, parte do conhecimento de um domínio pode complementar o conhecimento do outro (WEEBER, 2003).

5.6 Mineração de Imagens

A mineração de imagens é uma tarefa que tem despertado atenção de pesquisadores do domínio biomédico, uma vez que as imagens contêm informações importantes sobre os resultados das pesquisas. Ela procura extrair informações relevantes das imagens e legendas contidas nas publicações científicas, que podem ser usadas para diversos propósitos.

Ishii *et al.* (2010) extraem informações de imagens a partir de suas legendas e de sentenças no texto que fazem referência a estas figuras. Chen, Shatkay e Blostein (2006) utilizam a mineração de imagens para fazer a triagem de documentos, ou seja, classificar os documentos relevantes a serem anotados. Para cada documento, é criada uma

descrição de acordo com as imagens e, posteriormente, é aplicado o algoritmo de classificação *Naive Bayes* sob essas descrições. Eles utilizam imagens de microscopia de fluorescência, que contêm informações importantes sobre a localização e o comportamento de proteínas.

Coelho *et al.* (2010) desenvolveram um sistema que analisa imagens de documentos biomédicos e suas respectivas legendas, possibilitando que o usuário possa consultar por informações contidas nessas imagens. As figuras são divididas em diferentes painéis tratados separadamente. Cada painel pode ser classificado em seis diferentes classes, cada uma representando diferentes informações. A classificação é realizada por meio de uma abordagem baseada em aprendizagem de máquina, que faz uso de um algoritmo de aprendizagem chamado redução de risco empírica e de um classificador baseado em LIBSVM.

5.7 Desenvolvimento de Ontologias

Vocabulários controlados como taxonomias, tesauros e ontologias fornecem uma representação semanticamente estruturada do conhecimento no domínio biomédico. Esses recursos são utilizados para realizar várias tarefas da mineração de textos biomédicos, sobretudo a anotação de documentos (veja seção 5.6). Assim, diversos pesquisadores, nesta área, preocupam-se com a criação ou manutenção de vocabulários controlados.

Existem diversas terminologias que contemplam diferentes focos de estudo do domínio biomédico (*Gene Ontology*, *Cancer Ontology*, *Cell Ontology*, MeSH, entre outros). Entretanto, novas ontologias vêm sendo desenvolvidas na medida em que ontologias existentes não possibilitam análises específicas (POLPINIJ, 2011).

Uma vez que o desenvolvimento de ontologias é um processo trabalhoso e custoso, existem esforços para automatizar este processo. Inniss *et al.* (2006) utilizam técnicas de mineração de textos e PLN para extrair, a partir de descrições verbais transcritas de especialistas em retina, pares características-atributos de degeneração macular relacionada à idade. Esses pares são utilizados para produzir colaborativamente uma ontologia nesse domínio. Wächter, Fabian e Schroeder (2011) apresentam um processo semiautomático para a construção de ontologias. Antezana *et al.* (2008) propõem uma API, chamada ONTO-PERL, para o desenvolvimento e análise de ontologias no domínio biomédico.

Gacitua *et al.*. (2009) propõem um método colaborativo para o desenvolvimento de ontologias, envolvendo engenheiros de ontologias e especialistas do domínio. Esse método tem por objetivo proporcionar um maior grau de automação na elaboração da ontologia, bem como tornar mais eficiente a comunicação e a troca de informações entre engenheiros e especialistas de domínio que estão geograficamente distribuídos.

6 Considerações Finais

Considerando a relevância da área da mineração de textos biomédicos, para facilitar e agilizar a extração de informações da literatura biomédica, de forma que os pesquisadores possam explorá-las na condução de novas pesquisas e, tendo em vista a ausência de estudos que identifiquem e apresentem o desenvolvimento e o *mainstream* desse campo de pesquisa, este artigo foi desenvolvido de forma a ajudar no preenchimento dessa lacuna, disponibilizando para pesquisadores e interessados no tema, um mapa com o perfil das publicações sobre mineração de textos biomédicos. Esse mapeamento promove o conhecimento sobre o histórico e o estado atual desse campo de pesquisa.

Percebeu-se a evolução crescente do interesse dos pesquisadores pela mineração de textos biomédicos desde 1996 até os dias atuais. Os EUA são o país que mais contribuem para esse campo de pesquisa, possuindo a maior parte das instituições que mais publicam sobre o tema.

Diversas tarefas fazem parte da mineração de textos biomédicos, recebendo destaque neste estudo: recuperação de documentos, classificação de documentos, extração de termos, anotação de documentos, extração de relacionamentos, mineração de imagens e desenvolvimento de ontologias. Constatou-se que a grande maioria das publicações é dedicada a extração de relacionamentos explícitos ou implícitos entre termos biomédicos. A extração de termos é a segunda atividade que mais recebe atenção dos pesquisadores.

Diferentes métodos e técnicas são empregados para a realização de cada atividade na tentativa de obter resultados mais eficazes e/ou tornar o processo mais eficiente. Percebe-se também que os pesquisadores têm combinado duas ou mais abordagens para a solução de um determinado problema.

O estudo realizado, neste trabalho, gera várias oportunidades de pesquisas e contribui para a compreensão do quadro bibliográfico da mineração de textos biomédicos no nível internacional e disponibiliza insumos que permitem enriquecer a discussão sobre os possíveis rumos que as pesquisas na área têm tomado e, ainda, as prováveis tendências científicas para os pesquisadores e/ou interessados no tema.

Referências

AHLERS, C. B. *et al.* Using the literature-based discovery paradigm to investigate drug mechanisms. *In:* AMIA ANNUAL SYMPOSIUM, 2007, Chicago. *Proceedings…* Sheraton Chigago: AMIA Symposium, 2007. p. 6-10.

ANTEZANA, E. *et al*. ONTO-PERL: An API for supporting the development and analysis of bio-ontologies. *Bioinformatics*, v. 24, n. 6, p. 885-887, 2008.

ARAUJO, C. A. Bibliometria: evolução, história e questões atuais. *Em Questão*, Porto Alegre, v. 12, n. 1, p. 11-32, 2006.

BADA, M. *et al.* Concept Annotation in the CRAFT corpus. *BMC Bioinformatics*, v. 13, p. 161, 2012.

BARNICKEL, T. *et al.* Large scale application of neural network based semantic role labeling for automated relation extraction from biomedical texts. *PLoS ONE*, v. 4, n. 7, 2009.

BUFREM, L.; PRATES, Y. O saber científico registrado e as práticas de mensuração da informação. *Ciência da Informação*, Brasília, v. 34, n. 2, p. 9-25, 2005.

CAMPOS, D.; MATOS, S.; OLIVEIRA, J. L. Recognition of gene/protein names using conditional random fields. *In:* JOINT CONFERENCE ON KNOWLEDGE DISCOVERY AND INFORMATION RETRIEVAL, 2010, Valencia. *Proceedings…* 2010. p. 275-280.

CHEN, N.; SHATKAY, H.; BLOSTEIN, D. Use of figures in literature mining for biomedical digital libraries. *In:* SECOND INTERNATIONAL CONFERENCE ON DOCUMENT IMAGE ANALYSIS FOR LIBRARIES, 2006, Lyon. *Proceedings…* 2006. p. 180-197.

CHOI, J. *et al.* BOSS: A biomedical object search system. *In:* ACM FIFTH INTERNATIONAL WORKSHOP ON DATA AND TEXT MINING IN BIOMEDICAL INFORMATICS, 2011, New York. *Proceedings…* 2011. p. 19-26.

COELHO, L. P. *et al.* Structured literature image finder: Extracting information from text and images in biomedical literature. *Lecture Notes in Bioinformatics*, v. 6004, p. 23-32, 2010.

COHEN, A. M.; HERSH, W. R. A Survey of Current Work in Biomedical Text Mining. *Briefings in Bioinformatics*, v. 6, n. 1, p. 57-71, 2005.

COHEN, T. *et al.* Finding schizophrenia's prozac emergent relational similarity in predication space. Aberdeen. *In*: FIFTH INTERNATIONAL SYMPOSIUM ON QUANTUM INTERACTION, 2011, Aberdeen. *Proceedings…* 2011. p. 48-59.

COHEN, T.; SCHVANEVELDT, R.; WIDDOWS, D. Reflective Random Indexing and indirect inference: A scalable method for discovery of implicit connections. *Journal of Biomedical Informatics*, v. 43, n. 2, p. 240-256, 2010.

DIETZE, H.; SCHROEDER, M. GoWeb: A semantic search engine for the life science web. BMC Bioinformatics, v. 10, n. 10, 2009.

FAIIAZEE, H. *et al.* Text mining in bioinformatics: past, present and future. *In:* INTERNATIONAL CONFERENCE ON INFORMATION RETRIEVAL AND KNOWLEDGE MANAGEMENT, 2012, Kuala Lumpur. Proceedings… 2012. p. 327-330.

FONSECA, E. N. *Bibliometria*: teoria e prática. São Paulo: Cultrix; Ed. da USP, 1986.

GACITUA, R. *et al.* A collaborative workflow for building ontologies: A case study in the biomedical field. *In:* THIRD IEEE INTERNATIONAL CONFERENCE ON RESEARCH CHALLENGES IN INFORMATION SCIENCE, 2009, Fès. *Proceedings…* 2009. p. 121-128.

GANDRA, P.; PRADHAN, M.; PALAKAL, M. J. Biomedical association mining and validation. *In*: INTERNATIONAL SYMPOSIUM ON BIOCOMPUTING, Calicut, 2010. *Proceedings…* 2010.

GANIZ, M. C.; POTTENGER, W. M.; JANNECK, C. D. Recent advances in literature based discovery. *Journal of the American Society for Information Science and Technology*, 2006.

GARTEN, Y.; COULET, A.; ALTMAN, R. B. Recent progress in automatically extracting information from the pharmacogenomic literature. *Pharmacogenomics*, v. 11, n. 10, p. 1467-1489, 2010.

GONG, L. *et al.* Prediction of autism susceptibility genes based on association rules. *Journal of Neuroscience Research*, v. 90, n. 6, p. 1119-1125, 2012.

GONG, L. J. *et al.* A hybrid approach for biomedical entity name recognition. *In*: INTERNATIONAL CONFERENCE ON BIOMEDICAL ENGINEERING AND INFORMATICS, 2., 2009, Tianjin. *Proceedings…* 2009. p. 1-5.

GONŹALEZ, R. R.; IGLESIAS, E. L.; DIZ, L. B. Applying balancing techniques to classify biomedical documents: An Empirical study. *International Journal of Artificial Intelligence*, v. 8, p. 186-201, 2012.

GORDON M. D.; LINDSAY R. K . Toward discovery support systems: a replication, re-examination, and extension of Swanson's work on literature-based discovery of a connection between Raynaud's and fish oil. *Journal of the American Society for Information Science*, v. 47, n. 2, p. 116-128, 1996.

HRISTOVSKI, D. *et al.* Exploiting Semantic Relations for Literature-Based Discovery. *In:* AMIA ANNUAL SYMPOSIUM, 2006, Washington. *Proceedings…* 2006. p. 349-353.

HUA, Y. *et al.* Combination method of rules and statistics for abbreviation and its full name recognition. *Advances in Intelligent and Soft Computing*, v. 110, p. 707-714, 2011.

HUANG, Y. F; HSU, C. H. PubMed smarter: query expansion with implicit words based on gene ontology. *Knowledge-Based Systems*, v. 21, n. 8, p. 927-933, 2008.

INNISS, T. R. *et al*. Towards applying text mining and natural language processing for biomedical ontology acquisition. *In*: INTERNATIONAL WORKSHOP ON TEXT MINING IN BIOINFORMATICS, 1., 2006, Arlington. *Proceedings…* 2006. p. 7-14.

ISHII, N. *et al.* Figure classification in biomedical literature to elucidate disease mechanisms, based on pathways. *Artificial Intelligence in Medicine*, v. 49, n. 3, p. 135-143, 2010.

JAYADEVAPRAKASH, N.; MUKHOPADHYAY, S.; PALAKAL, M. Generating association graphs of non-co-occurring text objects using transitive methods. *In:* 2005 ACM SYMPOSIUM ON APPLIED COMPUTING, 2005, Santa Fe. *Proceedings…* 2005. p. 141-145.

JU, Z.; WANG, J.; ZHU, F. Named entity recognition from biomedical text using SVM, *In*: INTERNATIONAL CONFERENCE ON BIOINFORMATICS AND BIOMEDICAL ENGINEERING, 5., 2011, Beijin. *Proceedings…* 2011. p. 1-4.

KANG, B. C. *et al*. Document clustering of MEDLINE abstracts based on non-negative matrix factorization using local confidence assessment. *Biochip Journal*, v. 4, n. 4, p. 336-349, 2010.

KANG, B. C. *et al.* Semantic data integration to biological relationship among chemicals, diseases, and differential expressed genes. *Biochip Journal*, n. 5, v. 1, p. 63-71, 2011.

KASTRIN, A.; HRISTOVSKI, D. A fast document classification algorithm for gene symbol disambiguation in the BITOLA literature-based discovery support system. *In*: AMIA ANNUAL SYMPOSIUM, 2008, Whashington. *Proceedings…* 2008. p. 358-362.

KIM, J. H. Bioinformatics and genomic medicine. *Genetics in Medicine*, v. 4, p. 62S-65S, 2002.

KOSTOFF, R. N. *et al.* Literature-related discovery (LRD): water purification. *Technological Forecasting and Social Change*, v. 75, n. 2, p. 256-275, 2008.

KRALLINGER, M. *et al.* How to link ontologies and protein-protein interactions to literature: Text-mining approaches and the BioCreative experience. *Database Journal of biological databases and curation*, 2012.

LI, X. *et al*. A mouse protein interactome through combined literature mining with multiple sources of interaction evidence. *Amino Acids*, v. 38, n. 4, p. 1237-1252, 2010.

LIN, H.; LI, Y.; YANG, Z. Incorporating dictionary features into conditional random fields for gene/protein named entity recognition. *Lecture Notes in Computer Science*, v. 4819, p. 162-173, 2007.

MIYANISHI, T.; SEKI, K.; UEHARA, K. Hypothesis generation and ranking based on event similarities. *In*: ACM SYMPOSIUM ON APPLIED COMPUTING, 25., 2010, Sierre. Proceedings… 2010. p. 1552-1558.

MÖRCHEN, F. *et al.* Anticipating annotations and emerging trends in biomedical literature. *In*: ACM INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 14., 2008, Las Vegas. *Proceedings…* 2008. p. 954-962.

MUNKHDALAI, T. *et al.* BFSM: Finite state machine learned as name boundary definer for bio named entity recognition. *In*: INTERNATIONAL CONFERENCE ON AWARENESS SCIENCE AND TECHNOLOGY, 2011, Dalian. *Proceedings…* 2011. p. 344-349.

NAEEM, H. *et al.* MiRSel: Automated extraction of associations between microRNAs and genes from the biomedical literature. *BMC Bioinformatics*, v. 11, p. 135, 2010.

PLAKE, C. *et al.* GoGene: gene annotation in the fast lane. *Nucleic Acids Research*, v. 37, p. W300-W304, 2009.

POLPINIJ, J. The cancerology ontology: Designed to support the search of evidence-based oncology from biomedical literatures, *In*: INTERNATIONAL SYMPOSIUM ON COMPUTER-BASED MEDICAL SYSTEMS, 2011, Bristol. *Proceedings…* 2011. p. 1-6.

SHARMA, A.; SWAMINATHAN, R.; YANG, H. A verb-centric approach for relationship extraction in biomedical text, *In*: IEEE FOURTH INTERNATIONAL CONFERENCE ON SEMANTIC COMPUTING, 2010, Pittsburgh. *Proceedings…* 2010. p. 377-385.

SHU, G.; HUANG, X.; ZHU, S. A consensus method for prioritising drug-associated target proteins. *International Journal of Data Mining and Bioinformatics*, v. 6, n. 2, p. 178-195, 2012.

SIMPSON, M. S.; DEMNER-FUSHMAN, D. Biomedical text mining: a survey of recent progress. *In:* AGGARWAL, C. C.; ZHAI, C. *Mining Text Data*. Springer US, 2012. cap. 14 , p. 465-517.

SMALHEISER, N. R. Literature-based discovery: Beyond the ABCs. *Journal of the American Society for Information Science and Technology*, v. 63, n. 2, p. 218-224, 2011.

SULLIVAN, D. *Document warehousing and text mining*. New York: John Wiley & Sons, 2011.

SWANSON, D. R. Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspectives in Biology and Medicine*, v. 30, n. 1, p. 7-18, 1986.

SWANSON, D. R.; SMALHEISER, N. R.; TORVIK, V. I. Ranking indirect connections in literature-based discovery: The role of medical subject headings. *Journal of the American Society for Information Science and Technology*, v. 57, n. 11, p. 1427-1439, 2006.

THOMAS, P. *et al.* GeneView: a comprehensive semantic search engine for PubMed. *Nucleic Acids Research*, v. 40, n. W1, p. W585-W591, 2012.

TIWARI, R.; ZHANG, C.; CHEN, W. B. Extraction of coexpression relationship among genes from biomedical text using dynamic conditional random fields. *In*: IEEE INTERNATIONAL SYMPOSIUM ON COMPUTER-BASED MEDICAL SYSTEMS, 2009, Albuquerque, *Proceedings…* 2009. p. 1-4.

TOHIDI, H.; IBRAHIM, H.; AZMI, M. A. Statistical character-based syntax similarity measurement for detecting biomedical syntax variations through named entity recognition. *Communications in Computer and Information Science*, v. 136, p. 164-178, 2011.

TSURUOKA, Y. *et al.* Discovering and visualizing indirect associations between biomedical concepts. *Bioinformatics*, v. 27, n. 13, p. i111-i119, 2011.

VERGARA, S. C. *Projetos e relatórios de pesquisa em administração*. São Paulo: Atlas, 2003.

WÄCHTER, T.; FABIAN, G.; SCHROEDER, M. DOG4DAG: Semi-automated ontology generation in OBO-Edit and Protégé, *In*: INTERNATIONAL WORKSHOP ON SEMANTIC WEB APPLICATIONS AND TOOLS FOR THE LIFE SCIENCES, 4., 2011, London. *Proceedings…* 2011. p. 119-120.

WEEBER, M. Advances in Literature-Based Discovery. *Journal of the American Society for Information Science and Technology*, v. 54, n. 10, p. 913-925, 2003.

WEEBER, M. *et al.* Using concepts in literature-based discovery: simulating Swanson's Raynaud-fish oil and migraine-magnesium discoveries. *Journal of the American Society for Information Science and Technology*, v. 52, n. 7, p. 548-557, 2001.

WEI, C. H. *et al.* Normalizing biomedical name entities by similarity-based inference network and de-ambiguity mining. *In*: IEEE INTERNATIONAL CONFERENCE ON BIOINFORMATICS AND BIOENGINEERING, 9., 2009, Taichung. *Proceedings…* 2009. p. 461- 466.

WREN, J. D. Using fuzzy set theory and scale-free network properties to relate MEDLINE terms. *Soft Computing*, v. 10, n. 4, p. 374-381, 2006.

YOO, I.; HU, X.; SONG, I. Y. Biomedical ontology improves biomedical literature clustering performance: A comparison study. *International Journal of Bioinformatics Research and Applications*, v. 3, n. 3, p. 414-428, 2007.

ZHANG, Y. *et al.* Protein-protein interaction extraction based on improved all-paths kernel. *Journal of Computational and Theoretical Nanoscience*, v. 8, n. 10, p. 1925-1932, 2011.

ZHU, S. *et al.* Application of a new probabilistic model for mining implicit associated cancer genes from OMIM and Medline. *Cancer Informatics*, v. 2, p. 361-371, 2006.