Ciencia y Sociedad

A mineração de textos em aplicações de pesquisa e desenvolvimento (P&D)

Alexandre Lucas
Universidade Federal de Santa Catarina
Brasil · alexlucas.al@gmail.com

Angel Freddy Godoy Viera
Universidade Federal de Santa Catarina
Brasil · godoy@cin.ufsc.br

Resumo: A área de Recuperação da Informação (RI) é objeto de pesquisa e estudo na Ciência da Informação e suas técnicas estão presentes dentro dos mais variados processos. A mineração de textos é uma das técnicas do aprendizado de máquina aplicada à Recuperação da Informação (RI) e sua utilização atinge até o mais crítico processo em empresas e organizações, sendo, uma delas, o processo de Pesquisa e Desenvolvimento (P&D). O estudo realiza uma análise de artigos científicos que apresentaram o uso da mineração de textos para aplicação em P&D em três bases de dados referenciais: Web of Science (WoS), SCOPUS, e LISA. Com a utilização da análise bibliográfica, são apresentados aspectos da utilização da mineração de textos, aspectos da aplicação em P&D e detalhes desta aplicação aos artigos que apresentaram pertinência com o objetivo da pesquisa. Os principais resultados são apresentados em tabelas com os agrupamentos dos artigos onde a mineração de textos é utilizada para análise de Patentes, análise de bases especializadas e análise da internet. Foi observado duas grandes vertentes no uso da Mineração de Textos para P&D: na análise de patentes e na análise de bases especializadas onde neste último é predominante o uso na área da saúde.

Palavras-chaves: Recuperação da Informação, Mineração de Textos, Pesquisa e Desenvolvimento, P&D.

Abstract: The area of Information Retrieval (IR) is the subject of research and study in Information Science and its techniques are present within the various processes. The text mining is one of the techniques applied to IR and utilization reaches even the most critical process in companies and organizations one being the process of research and development (R & D) machine learning. The study presents an analysis of scientific articles that showed the use of text mining for use in R & D in three reference databases: Web of Science (WoS), SCOPUS, and LISA. Using content analysis aspects of the use of text mining, aspects of the application in R & D of this application and details of the items that had relevance to the purpose of the research are presented. The main results are presented in tables with groupings of articles Patent analysis, analysis of specialized databases and analysis of internet. Two major strands was observed in the use of Text Mining for R & D: in patent analysis and analysis of specialized databases where the latter is the predominant in the health area.

Keywords: Information Retrieval, Text Mining, Research and Development, R&D.

1 Introdução

O tema recuperação da informação é objeto de pesquisa e estudo na área da Ciência da Informação, mas está presente dentro da vida cotidiana e também em processos de empresas e organizações. A invenção da prensa tipográfica por Gutenberg, em 1445, fez a informação textual crescer de forma exponencial nos anos seguintes. Cinco séculos depois, a discussão sobre a explosão informacional feita por Vannevar Bush, no pós-guerra, fez nascer, ou despontar no mundo ocidental, a Ciência da Informação e um dos seus objetos de estudo: a Recuperação da Informação (RI). Mas foi, com certeza, a explosão informacional provocada pela presença popular da internet, a partir da década de 90, que trouxe o problema de encontrar a informação, relevante e desejada, para posições de importância e destaque. Uma das empresas mais importantes e mais mencionada nestes últimos anos, a Google, tem sua principal atividade comercial baseada na recuperação de informação. Por outro lado, a globalização levou os aspectos econômicos da sociedade para uma escala mundial. A competição é um processo constante e atinge diversos componentes desta nova sociedade. A competição entre empresas é sempre o mais lembrado dos exemplos deste processo, mas ela atinge instituições e nações inteiras. A inovação é considerada um dos elementos mais importantes para aqueles que querem vencer a competição e a Pesquisa e o Desenvolvimento (P&D) é um dos elementos-chave para este intento. Fazer P&D é também um processo de recuperar informação e de gerar novas informações e/ou conhecimento materializado em um novo produto ou processo. Se recuperar a informação já era algo importante, no contexto de P&D, pode ser considerado a diferença entre o sucesso e o fracasso. Como mencionado anteriormente, a *internet* gerou uma explosão informacional e, neste paradigma, já não se trabalha com a informação organizada, em bancos de dados, por exemplo, mas com uma rede quase infinita de dados. Informações que, além de estarem em diferentes formas, por exemplo, as multimídias, podem estar também em estruturas totalmente desorganizadas, como acontece nos portais WEB e nas redes sociais. Umas das técnicas de grande utilização junto com a recuperação da informação é a mineração de textos e seu uso também acontece em processos de P&D, sejam eles empresariais ou governamentais. É nesse entendimento que este artigo se propõe a identificar produções científicas envolvendo a mineração de textos e sua aplicação nos processos da Pesquisa e Desenvolvimento (P&D). Não é uma revisão exaustiva da literatura e nem uma análise bibliométrica, mas uma análise que pretende contribuir para uma visão mais profunda e consolidada da técnica de RI quando aplicada em processos de P&D.

2 Pesquisa e o desenvolvimento (P&D), recuperação da informação e mineração de texto

A atividade ou processo de Pesquisa e Desenvolvimento (P&D) é importante para empresas, instituições e também nações. Com base na bibliografia existente, é possível apresentar elementos dessa importância e alguns aspectos conectados com a Recuperação da Informação e a Mineração de Textos. O termo Pesquisa e Desenvolvimento, abreviado por P&D, é amplamente conhecido e utilizado. Na língua inglesa, fonte das principais referências sobre o assunto, o termo utilizado é *Research and Development* e abreviado por *R&D*. O termo P&D está também frequentemente acompanhado de outros dois termos: Tecnologia e Inovação. Certamente não há empresa, instituição e governo que atue com Tecnologia e Inovação que não realize P&D.

Michael Porter (1999) inicia o capítulo 6 de seu livro, intitulado *Competição: Estratégias Competitivas Essenciais*, da seguinte forma:

> A prosperidade nacional não é algo herdado, mas sim o produto do esforço criativo humano. Não é algo que emana dos dotes naturais de um país, de sua força de trabalho, das taxas de juros ou do valor da moeda, como insistem os economistas clássicos. A competitividade de um país depende da capacidade da sua indústria de inovar e melhorar. As empresas conquistam uma posição de vantagem em relação aos melhores competidores do mundo em razão das pressões e dos desafios. [...]

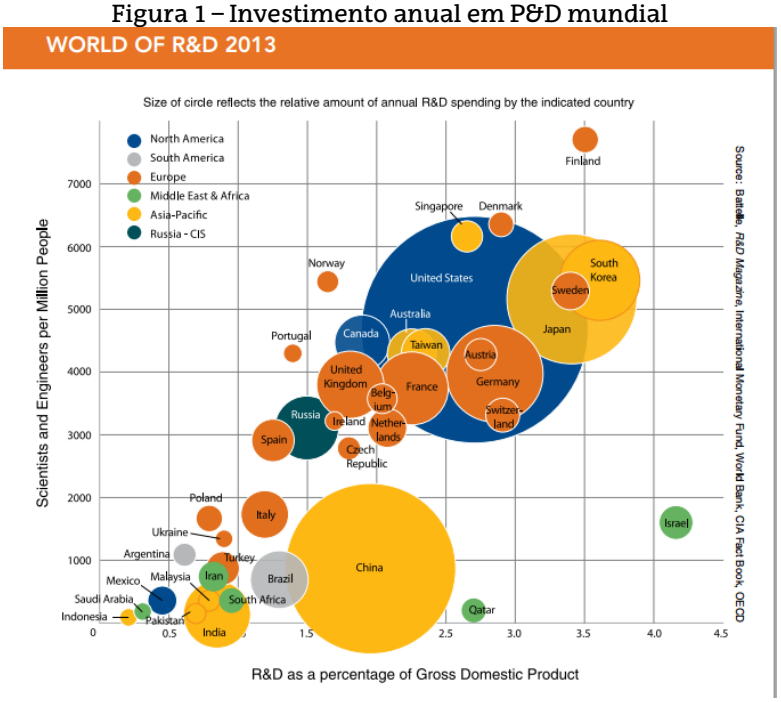
> Num mundo de competição global crescente, os países se tornaram mais, e não menos, importantes. À medida que os fundamentos da competição se deslocam cada vez mais para a criação e assimilação do conhecimento, aumenta a importância dos países (p.167).

Para Schumpeter, importante economista tcheco da metade do século XX, o crescimento de uma empresa ou nação depende da inovação (Schumpeter, Opie, 1934), incluindo a inovação em tecnologia, *marketing* e organização. Historicamente, além disso, as inovações tecnológicas têm tido a influência mais significativa nas mudanças. Desde 1960, devido ao impacto e importância da inovação tecnológica, a previsão da tecnologia do futuro e a antecipação das necessidades do mercado de uma nova tecnologia têm recebido importantes considerações (Choi, Kim, Yoon, Kim, Lee, 2013).

Segundo Lin e Chen (2005) a inovação tecnológica é a mais poderosa fonte de vantagem competitiva para as empresas modernas. Lin e Chen (2005) referenciam Sadowski e Roth (1999) que constataram que as empresas líderes de tecnologia se destacaram em quatro áreas de gestão de tecnologia, incluindo: estratégia; gerenciamento de portfólio; planejamento; e, por último, desenvolvimento/processos de transferência. Sadowski e Roth (1999) ainda mostram que estratégia corporativa para desenvolver e explorar os recursos tecnológicos de uma empresa tem um impacto profundo sobre o seu desempenho a longo prazo. Ressaltam que um pressuposto fundamental para a estratégia de portfólio de tecnologia, originada da teoria moderna de portfólio, é que uma empresa pode reduzir os riscos e aproveitar as oportunidades de negócios derivados da economia de escopo através da realização de um portfólio de diferentes empresas, mercados ou recursos.

Retirado do relatório 2014 Global R&D Funding Forecast, da revista RD Magazine (Wadsworth, 2013), temos o infográfico mostrado na Figura 1, que apresenta os investimentos em P&D realizados pelos mais diversos países do mundo, mostrando, em conjunto com uma relação entre valores investidos, a relação com o Produto Interno Bruto – PIB e correlacionando com a razão entre o número de cientistas e engenheiros para cada 1 milhão de habitantes do mesmo país.

Figura 1 – Investimento anual em P&D mundial



Fonte: 2014 Global R&D Funding Forecast, R&D MAGAZINE. (Wadsworth, 2013)

No infográfico original, é possível identificar os países pelos continentes através das cores. O investimento em P&D está calculado com base no percentual do Produto Interno Bruto – PIB ou do inglês *Gross Domestic Product* (*GDT*). Uma análise inicial do gráfico mostra que os três países que mais investem em P&D, em valor absoluto, são

os Estados Unidos, China e Japão, percebido pelo tamanho dos círculos. Porém, considerando o investimento em P&D na relação com o PIB, temos Israel, Coreia do Sul e Finlândia com os maiores percentuais. É possível também perceber que os países com valores mais altos de P&D em percentual do PIB apresentam uma relação mais alta de cientistas e engenheiros por milhão de habitantes, círculos mais no alto e mais à direita no gráfico. É importante lembrar que esta relação faz sentido, pois os principais atores ou profissionais que atuam em P&D são os cientistas e engenheiros. Cabe ainda ressaltar a posição do Brasil, que apresenta valores de investimento em P&D menores que 1,5% do PIB e abaixo de 1.000 cientistas e engenheiros por milhão de habitantes, porém, com valor absoluto maior do que a Espanha e do que a Itália, além de ser o país mais representativo da América do Sul.

Para identificar como é compreendido o P&D e a Inovação Tecnológica no Brasil, pode-se utilizar o entendimento empregado pelo governo federal brasileiro através do DECRETO FEDERAL Nº 5.798, DE 7 DE JUNHO DE 2006, que regulamenta os incentivos fiscais às atividades de pesquisa tecnológica e desenvolvimento de inovação tecnológica, de que tratam os arts. 17 a 26 da Lei no 11.196, de 21 de novembro de 2005.

O decreto, em seu artigo 2º, vai definir o que é inovação tecnológica e P&D tecnológico em inovação tecnológica para determinar que atividades podem receber o incentivo fiscal, (Brasil, 2006).

Art. 2o Para efeitos deste Decreto, considera-se:

I. inovação tecnológica: a concepção de novo produto ou processo de fabricação, bem a agregação de novas funcionalidades ou características ao produto ou processo que implique melhorias incrementais e efetivo ganho de qualidade ou produtividade, resultando maior competitividade no mercado;
II. pesquisa tecnológica e desenvolvimento de inovação tecnológica, as atividades de:
    a. pesquisa básica dirigida
    b. pesquisa aplicada
    c. desenvolvimento experimental
    d. tecnologia industrial básica
    e. serviços de apoio técnico.

Para apresentar uma relação entre P&D e a Recuperação da Informação (RI), Porter e Newman (2011) mostram que, durante décadas, P&D era sinônimo de pesquisa e desenvolvimento interno. Ou seja, somente desenvolvido dentro das empresas ou instituições, seja do ponto de vista operacional quanto do informacional. No entanto, nos últimos anos, a atenção à P&D externo tornou-se significativa, em grande parte, em conjunto com Inteligência Técnica Competitiva (CTI). Grandes empresas, órgãos de defesa e outros têm a necessidade de saber quais tecnologias seus principais concorrentes estão perseguindo (Inteligência Competitiva). Complementam ainda que, dado o aumento implicações dos resultados da gestão da propriedade intelectual, também é essencial saber o que está acontecendo em uma determinada área tecnológica antes de se investir pesadamente em P&D (Inteligência Técnica).

Sobre informação e P&D também encontramos em De Abreu, Da Costa Vianna Franca e Pereira Sinzato (1999):

> ...a informação, sob o impacto da utilização de tecnologia de informação tem uma influência cada vez maior na organização do futuro. A introdução de novas tecnologias de informação nas organizações ampliou as potencialidades da informação recurso estratégico, a velocidade com que a interação entre gestão e informação ocorre e a qualidade desta ligação. Estes avanços tecnológicos modificam as relações entre tempo e espaço. Enfatiza Giddens (1991), as distâncias temporais e espaciais cobertas pelas novas tecnologias tornam o passo de vida cada vez mais rápido. É como se o mundo encolhesse ou fosse uma "vila global". O fenômeno da globalização cria um mundo sem fronteiras, onde a vantagem competitiva é concedida aos centros detentores de tecnologias ajustadas ao mercado de interesse da empresa e que apresentam investimentos consistentes em pesquisa e desenvolvimento (P&D) (p. 322).

Segundo Borges (1995), a competitividade de uma empresa é diretamente proporcional à sua capacidade de obter informação, processá-la e disponibilizá-la, de forma rápida e segura. Assim sendo, a informação adquire caráter estratégico de apoio à tomada de decisão nas organizações. Para tal, a cultura organizacional deve se adaptar às constantes mudanças que ocorrem no ambiente externo das organizações (p. 13).

Segundo Matellart (2002), vivemos na Sociedade da Informação e nesta sociedade, a produção e venda de informações contribuem de maneira considerável para as economias mais desenvolvidas. Um dos fenômenos desta sociedade é o aparecimento da propriedade intelectual, do *Copyright,* em 1709, e também das patentes. Infelizmente, aparece também a espionagem industrial como curso natural deste processo. Mas vale lembrar o alerta de Burke (2003), "a aquisição do conhecimento depende não só da possibilidade do acesso a acervos de informação, mas também da inteligência, pressupostos e práticas individuais".

Com a informação assumindo o papel principal desta nova fase de nossa sociedade, mas trazendo ainda muitos aspectos e características da sociedade anterior, a industrial, temos o uso massivo de sistemas informacionais nos quatros cantos do globo terrestre, principalmente no ambiente empresarial. E, neste cenário de uma infinidade de bancos de dados de todos os tipos e portes, é possível afirmar que (Todesco, Carretero, Duran, 2007):

A informação é considerada chave para alcançar a vantagem competitiva;

Ela é considerada vital para tomada de decisões e se encontram nas bases de dados corporativas;

Estas bases são montanhas de dados disseminados por todas as partes;

A chave para ganhar vantagem competitiva reside na obtenção de inteligência desses dados: converter dados em conhecimento.

Está claro que a informação é importante, mas conseguir recuperá-la conforme o momento e a necessidade do usuário é fundamental. Ingwersen (1992) no prefácio do seu livro coloca:

> Recuperação da informação abrange os problemas relacionados com o armazenamento eficaz, acesso e busca de informações necessárias pelos indivíduos. Atualmente, a informação continua a crescer exponencialmente, diversificando em muitas formas e meios de comunicação. Neste labirinto de recuperação complexo existe uma clara necessidade de maior esforço que visa adaptar o desempenho de RI para as demandas dos usuários. Umberto Eco fez aprender seu irmão William chamando a atenção, em um momento de reflexão ao visitar a biblioteca, o problema fundamental na recuperação de informação é a forma de conectar o texto e sua potencialidade para fornecer informações para o leitor individual. Contribuição a estes esforços contínuos de harmonia entre a informação e o usuário, o objetivo desta publicação é apresentar e aumentar as exigências teóricas e operativas necessárias para um desempenho eficaz, em particular de intermediários, na interação de recuperação de informação.

Segundo Manning, Raghavan e Schütze (2009), a Recuperação da Informação (RI) pode ser definida como: " ...encontrar materiais (geralmente documentos) de natureza não estruturada (geralmente texto) que satisfaça uma necessidade de informação de dentro grandes coleções (geralmente armazenados em computadores)"(p. 1).

Seguem ainda complementando que, desta forma, a RI costumava ser uma atividade para apenas alguns profissionais: bibliotecários de referência, estagiários e pesquisadores profissionais e similares. Agora, o mundo mudou e centenas de milhões de pessoas se envolvem na RI a cada dia quando usam um *site* de busca na web ou procuram seu *e-mail*. Recuperação da informação está se tornando a forma dominante de acesso à informação, ultrapassando a busca no estilo banco de dados tradicional.

Rijsbergen (1995), em seu livro sobre recuperação da informação, considera que o termo é amplamente utilizado e, muitas vezes, vagamente definido. Infelizmente, a palavra informação pode ser muito enganadora. No contexto da Recuperação de Informação (RI), a informação, no sentido técnico dado na teoria da comunicação de Shannon, não é facilmente medido. De fato, em muitos casos, pode-se descrever

adequadamente o tipo de recuperação simplesmente substituindo o termo 'documento' pelo termo 'informação'. A definição perfeita e simples é dada por Lancaster (1968) citado por Rijsbergen (1995):

> Recuperação da informação é o termo convencionalmente, embora um pouco impreciso, aplicada ao tipo de atividade. Um sistema de recuperação de informação não 'informa' (ou seja, não altera o conhecimento de) o usuário sobre o assunto de sua investigação. Limita-se a informar sobre a existência (ou não existência) e o paradeiro dos documentos relativos ao seu pedido.

Sobre a importância na extração de textos na RI, encontramos em Nanba, Ishino e Takezawa (2012) que trata-se da extração de relações que referem-se ao método de detecção e identificação de relações semânticas predefinidas dentro de um conjunto de entidades, em documentos de texto eficientes (Zelenco, AOne & Richardella, 2003; Zhang, Zhou, e AITI, 2008). A importância de tal método foi reconhecida pela primeira vez na *Message Understanding Conference* (MUC, 2001), realizada entre 1987-1997, sob a supervisão da DARPA (*Defense Advanced Research Projects Agency*). Depois disso, a *Automatic Content Extraction* (ACE, 2009) *workshop* facilitou inúmeras pesquisas entre 1999-2008 e foi promovido pelo NIST um novo projeto. Atualmente, a oficina é realizada todos os anos, sendo o maior fórum mundial para comparação e avaliação de novas tecnologias na área de extração de informações, a extração de relação, a extração de evento, e extração de informação temporal. Este *workshop* é conduzido um subcampo de *Text Analytics Conference* (TAC, 2012), que está atualmente sob a supervisão do *National Institute of Standards and Technology* (NIST).

Para Feldman e Sanger (2007), a Mineração de Textos é uma nova e excitante área de pesquisa de ciência da computação que tenta resolver a crise da sobrecarga de informação através da combinação de técnicas de mineração de dados, aprendizagem de máquina, processamento de linguagem natural, recuperação de informação e gestão do conhecimento. Da mesma forma, a detecção de *links* - uma abordagem rápida para a análise do texto em que são compartilhados e construídos muitos dos elementos-chave de Mineração de Texto - também fornece novas ferramentas para as pessoas aproveitarem melhor os recursos de dados textuais. Segundo ainda Feldman e Sanger (2007), detecção de *links* depende de um processo de construção de redes de objetos interconectados através de várias relações, a fim de descobrir padrões e tendências. As principais tarefas de detecção de *links* são extrair, descobrir e unir evidências esparsas de grandes quantidades de fontes de dados, para representar e avaliar o significado das evidências relacionadas, e de aprender os padrões para orientar a extração, descoberta, e conexão de entidades.

Com o referencial teórico colocado, a Mineração de texto, a Recuperação da Informação, o Processo de Pesquisa e Desenvolvimento, a importância da informação e seu uso como estratégia competitiva compõem o objeto de interesse da pesquisa. Cabe agora, nos procedimentos e nas opções metodológicas, delimitar as condições de recuperação do material para análise e entendimento das aplicações da mineração de texto nos processos de P&D.

4 Procedimentos e opções metodológicas

Uma forma de identificar a produção científica de um determinado tema é a consulta em bases especializadas. Desta forma, com o objetivo de identificar a produção científica que tratasse de Mineração de Textos e sua utilização em aspectos de P&D, realizou-se uma pesquisa conforme algumas definições e opções estabelecidas para o propósito deste trabalho.

É importante notar que o conteúdo apresentado neste artigo é parte dos estudos realizados no contexto do grupo de pesquisa Recuperação de Informação e Tecnologias Avançadas (RITA) e integrante do processo de desenvolvimento de uma dissertação para o mestrado sobre Inteligência de Negócios em Institutos de Ciência Tecnologia e Inovação, do programa de pós-graduação em Ciência da Informação da Universidade Federal de Santa Catarina (PGCin/UFSC).

Quanto à natureza da pesquisa, ela foi exploratória e delineada por um estudo bibliográfico, na concepção de Gil (2002, p.52-53), e também qualitativa no que concerne ao tipo de análise.

Testes preliminares nas bases BRAPCI e Scielo com termos em português mostraram não haver resultado relevante para pesquisa. Assim, definiu-se realizar a pesquisa em bases internacionais e utilizar termos para recuperação da informação em inglês.

Com relação às bases de dados foram selecionadas três que indexam os periódicos nas áreas da Ciência da Informação, Ciências Sociais Aplicadas em geral, Computação e Engenharia do Conhecimento. São elas: *Web of Science* (WOS), SCOPUS e *Library and Information Science Abstracts* (LISA).

O *corpus* da pesquisa foi constituído por todos os artigos publicados, nas bases selecionadas, até maio de 2014.

Os termos definidos para pesquisa foram:

Mineração de Dados: utilizado o termo em inglês *Text Mining;*

Pesquisa e Desenvolvimento - P&D: utilizado o termo abreviado em inglês de Research *and Development* R&D;

A equação de busca ficou constituída da seguinte forma: ("Text Mining" *AND* R&D).

Os campos utilizados para as buscas com a equação de busca foram: título, resumo e palavras-chave em forma simultânea e foram selecionados somente os resultados apresentados como artigo, na classificação para o tipo de documento.

Ao total foram recuperados 36 registros, alguns dos quais apresentavam textos completos, e outros, somente resumos para seleção dos artigos a serem estudados. Para os registros que apresentavam somente o resumo, a insuficiência de informação para identificação dos aspectos pertinentes foi considerada para uma decisão de descarte.

Por último, os artigos que que não apresentavam pertinência com o objeto da pesquisa, mesmo contendo os termos utilizados para recuperação, também sofreram o descarte.

Com o total de 20 artigos selecionados, utilizou-se a análise qualitativa e buscou-se identificar dois aspectos:

Aspecto da Mineração de Texto: caracterização do tipo de mineração do texto utilizado;

Aspecto da Aplicação dividido em duas partes:

- Caraterização geral do objetivo da utilização da mineração de texto visando uma categorização;
- Caracterização mais específica da utilização da mineração de texto, visando esclarecer melhor a aplicação.

Também foi utilizado o infográfico formado pela 'nuvem de palavras' para demonstrar a frequência de aparecimento dos principais termos constantes tanto nos títulos dos artigos quanto os constantes nas palavras-chave.

4 Tratamento e análise dos dados

O tratamento e análise dos dados consistiu inicialmente na identificação de artigos idênticos que estavam indexados por bases diferentes. Procedeu-se então a leitura dos resumos visando selecionar somente artigos pertinentes ao objetivo do trabalho. Dos 36 artigos recuperados, tabela 1, foram selecionados somente 20 artigos. A etapa subsequente retirava dos textos os aspectos relacionadas a técnica de Mineração de Texto e da Aplicação no processo de P&D. Por último, com uma categorização optou-se por agrupar os artigos segundo estas categorias e estruturadas em tabelas conforme estão apresentados abaixo.

A Tabela 1 apresenta a quantidade de artigos recuperados nas bases de dados pesquisadas.

Tabela 1 – Artigos recuperados por base de dados

Base de Dados	Artigos recuperados (qtde.)
WoS	16
SCOPUS	19
LISA	1

No infográfico em forma de nuvem de palavras, apresentado na Figura 2, é mostrada a análise da ocorrência dos termos traduzidos em português com a retirada dos termos 'texto', 'mineração' e 'P&D'.

Figura 1 – Nuvem de Palavras para Títulos Traduzidos

Utilizado o site http://www.jasondavies.com/wordcloud,
para desenvolver a 'nuvem de palavras'.

A análise da nuvem de palavras apresenta o destaque para alguns termos que podem caracterizar elementos da aplicação da mineração de textos nos documentos identificados. Observa-se que os termos informações, tecnologia, patentes, pesquisa, previsão, tendências são termos com alta frequência nos documentos recuperados. Uma das leituras juntando estes termos poderia ser: Uma FERRAMENTA para SUMARIZAÇÃO de PATENTES visando a PREVISÃO/TENDÊNCIAS.

A principal análise do trabalho foi direcionada para identificar usos ou aplicações que, organizados em categoria, mostrassem como a mineração de texto apoia os processos de P&D. Este resultado é apresentado em tabelas, separado por categorias. Estas categorias foram estabelecidas com base nos próprios artigos, mas arbitradas durante a construção da pesquisa.

Análise de patentes

Na categoria Análise de Patentes estão agrupados os artigos recuperados que utilizaram a Mineração de Textos para analisar bases especializadas de patentes.

Tabela 2 – Categoria Análise de Patentes

Título	Aspecto da Mineração de Texto	Aspecto da Aplicação	Artigo
Automatic patent document summarization for collaborative knowledge systems and services	Uso de uma ontologia combinada baseada em TF-IDF e uma abordagem com conceito de agrupamento.	Sistema automático para extrair informações de patentes. Metodologia proposta extrai, conjuntos, e integra o conteúdo de uma patente para obter um resumo e um diagrama de árvore conjunto de termos-chave. Busca melhorar a partilha de conhecimento e colaboração entre os membros da equipe de P & D.	Trappey; Trappey; Wu, 2009.

An R&D knowledge management method for patent document summarization	Reconhecimento da palavra chave e uso do processo de stopping, stemming and splitting e calculo de relevância pela frequência dos termos.	Sumarização de Patentes para gestão de P&D. Método automático de sumarização patente para abstração conhecimento exato e gestão do conhecimento eficaz de P&D.	Trappey; Trappey, 2008.
An SAO-based text-mining approach for technology roadmapping using patent information	A abordagem baseada em palavra-chave, utiliza técnicas de mineração de texto em conjunto com o processamento de linguagem natural (NLP).	Reduzir custo no desenvolvimento de Road Maps de Tecnologia Extrai palavras-chave importantes que representem o conteúdo de documentos importantes e descobre padrões que contenham implicações tecnológicas.	Choi, et al, 2013.
Identifying technology trends for R&D planning using TRIZ and text mining	Utilização da técnica de KeyGraph na mineração de Texto.	Identificação sistemática de tendências das tecnologias em patentes. Utilizando o método de mineração de textos sobre patentes sobre MRAM os princípios fundamentais e tendências evolutivas da TRIZ (Theory of Inventive Problem Solving, or Teoriya Reshniya Izobretatelskikh Zadatch).	Wang; Chang; Kao, 2010.
Nanopatenting patterns in relation to product life cycle	Utilização de Software VantagePoint e Thomson Data Analyzer na análise com uso de termos definidos por especialistas.	Análise de patentes sobre nanotecnologia como um indicador de inovação. Análise de patentes e classificação do estágio de P&D de nano desenvolvimentos por meio da identificação de 3 estágios do ciclo de vida.	Alencar; Porter; Antunes, 2007.
Text mining as a valuable tool in foresight exercises: A study on nanotechnology	Utilização de palavras- chave com termos de especialistas e revistas especializadas.	Apoio ao processo de tomada de decisão relacionada com o estabelecimento de políticas de CT&I e atividades no Brasil. Utilizado a mineração de texto para análise de: · Artigos científicos análise bibliométrica em nível internacional; · Patentes análise bibliométrica em nível internacional; · Mapeamento das capacidades de recursos humanos no Brasil.	De Miranda Santos et al, 2006.
A systematic approach for identifying technology opportunities: Keyword-based morphology analysis	Mineração de Palavras-chave com base em análise morfológica.	Apoio a análise Morfológica (MA), uma técnica qualitativa de representação da tecnologia de previsão (TF), utilizada para identificar oportunidades tecnológicas. Análise de patentes. Desenvolvimento de um dicionário de tecnologia pela análise fatorial por palavras-chave que são extraídos de documentos de patentes através da mineração de texto.	Yoon; Park, 2005.
A technology forecasting method using text mining and visual apriori algorithm	Uso da mineração com regras de associação (ARM) e visualização.	Análise de Patentes com a Tecnologia de Previsão (TF) para planejamento de politicas de P&D Método para a análise de dados de patentes, usando uma combinação de mineração de texto e o algoritmo Apriori (VA). Experimento utilizando documentos de patentes relativas à tecnologia de banco de dados recuperados do Patent and Trademark Office dos Estados Unidos.	Jun, 2014.

Emerging technology forecasting using new patent information analysis	Construção de uma matriz códigos de patentes-IPC (PICM) recuperados utilizando técnicas de mineração de texto. Essa matriz é utilizada para a modelagem ETF (Previsão de Tecnologia Emergente)	Previsão de tecnologias emergentes e Planejamento de P&D através de patentes. Proposto uso de Códigos internacionais de classificação de Patentes de uma tecnologia alvo, através de um modelo emergente da previsão tecnológica, combinando inferência estatística e redes neurais para a construção de modelo de análise de informações de patentes. Realizado estudo em nanotecnologia como a tecnologia alvo.	Jun; Lee, 2012.
A text-mining-based patent network: Analytical tool for high-technology trend	Mineração de texto é executado para transformar documentos em dados vetoriais de palavras chave e matriz de incidência.	Proposição de uma análise baseada em rede e um método alternativo para a análise de citação de patentes. Usando um conjunto de dados ilustrativos, o processo global de desenvolvimento da rede de patente é descrita. Novos índices como o índice de tecnologia centralidade, o índice de ciclo de tecnologia e grupos de palavras-chave de tecnologia são sugeridas para análise quantitativa em profundidade.	Yoon; Park, 2004.

Análise de bases especializadas

Na categoria Análise de Bases Especializadas estão agrupados os artigos recuperados que utilizaram a Mineração de Textos para analisar bases especializadas, com exceção daquelas que contenham patentes. Em sua maioria são bases da área da saúde.

Tabela 3 – Categoria Análise de Bases Especializadas

Título	Aspecto da Mineração de Texto	Aspecto da Aplicação	Artigo
Mining external R&D	Processo sistemático de mineração de bases de referência com proposição de 5 estágios: Literature review, Research profiling, Tech mining, Structured knowledge discovery, Literature-based discovery.	Obtenção Tecnológica externa à empresa pela mineração de texto de bases de dados. Utilização de perguntas com a Gestão de Tecnologia (MOT). Utilização de quadro técnico para mineração com base em questões recorrentes e que pode ser abordada através de 200 ou mais potenciais indicadores de inovação empíricos.	Porter; Newman, 2011.
Development and application of a keyword-based knowledge map for effective R&D planning	Utilização de vetores das palavras chave.	Nova abordagem para a geração de mapas de conhecimento de bancos de dados grande dificuldade de análise. Geração de cinco tipos de mapas de conhecimento (mapa do P&D core, mapa tendência de P&D, mapa da concentração do P&D, mapa de relação do P&D e mapa de cluster de P&D).	Yoon; Lee; Lee, 2010.
Co-authorship Network Analysis: A Powerful Tool for Strategic Planning of Research, Development and Capacity Building Programs on Neglected Diseases	Uso de software de mineração de texto VantagePoint com uso de filtros apropriados do Institute for Scientific Information - ISI.	Identificar Publicações de autores brasileiros sobre as doenças tropicais negligenciadas e indexadas pela WoS. Apoiar o planejamento estratégico brasileiro para financiar a pesquisa, desenvolvimento e capacitação sobre doenças tropicais negligenciadas.	Morel et al, 2009.
Literature-related discovery (LRD): Potential treatments for Raynaud's phenomenon	Busca de frases especificas em artigos científicos e clusterização.	Busca de artigos com informações sobre a síndrome de Raynaud. Busca de artigos com informações sobre a síndrome de Raynaud na base de dados MEDLINE.	Kostoff, et al, 2008.

Mapping of spices research in Asian countries	Lista de palavras-chave classificados na ordem da freqüência e ocorrência no ano de abrangência. Uso de Software Data e Text Mining (DTM) dedicada à análise exploratória de dados numéricos e textuais multivariados.	Análise de Bancos de dados nos campos de resumo e indexação para identificar os movimentos das pesquisas dos países asiáticos em especiarias. Análise do Banco de dados HORT-CD, publicado pela CABI (Centro de Biociência Agrícola International), Reino Unido, Londres. Os Serviços de Resumos e Indexação sobre este tema foi escolhido para ser a fonte base de dados do estudo.	Senthilkumaran; Amudhavalli, 2007.
The SINAMED and ISIS projects: Applying text mining techniques to improve access to a medical digital library	Integração da categorização de textos e técnicas de sumarização nos processo de busca e browsing.	Acesso a informação sobre um domínio médico específico. O acesso a informação sobre registros clínicos de pacientes e documentação científica relacionada, no âmbito de dois projetos de pesquisa diferentes: SINAMED e ISIS.	De Buenaga, et al, 2006.
Automated extraction and visualization of information for technological intelligence and forecasting	Análise em grandes bancos de dados, de fácil acesso, utilizando software de mineração de texto. Aumento da recuperação através de macros (scripts de programação) que sequenciam automaticamente as medidas necessárias para gerar determinados produtos de informação desejados.	Previsão tecnológica empírica (TF) na gestão de tecnologia e inovação. Aumento da utilização gerencial de grandes volumes de informações disponíveis. Desenvolvimento de processos semi automatizados para gerar conhecimento útil. Geração de uma família de mapas tecnologia que ajudam o desenvolvimento de uma tecnologia alvo. Geração de indicadores de inovação de atividade de P&D.	Zhu; Porter, 2002.
Unblocking blockbusters: Using boolean text-mining to optimise clinical trial design and timeline for novel anticancer drugs	Pesquisa de uma série termos que representam conjunto principal de interesse. Correção aritmética para diferentes freqüências. A análise estatística não paramétrica foi realizada pelo cálculo do qui-quadrado.	Usando cenários relativos à oncologia através de mineração de dados da literatura científica. Pesquisas foram realizadas usando a última versão de busca e recuperação baseada em texto do PubMed, um serviço da National Library of Medicine desenvolvido pelo Centro Nacional de Informações sobre Biotecnologia usado para integrar as principais bases de dados (incluindo a PubMed Central, Revistas, Livros, OMIM).	Epstein, 2009.

Análise da internet

Na categoria Análise da *Internet* estão agrupados os artigos recuperados que utilizaram a Mineração de Textos para analisar a *internet* de forma geral, ou seja, análise dos *sites* de uma maneira ampla.

Tabela 3 – Categoria Análise da Internet

Título	Aspecto da Mineração de Texto	Aspecto da Aplicação	Artigo
Web mining based extraction of problem solution ideas	*Text Mining* na WEB, considera dependências de domínio e aspectos linguísticos	Identificação automática de novas ideias. O estudo de caso identifica novas ideias tecnológicas do programa de pesquisa de defesa alemão, com base em projetos de P&D existentes	Thorleuchter; Van Den Poel, 2013.

Collective SME approach to technology watch and competitive intelligence: The role of intermediate centers	Ferramentas de mineração de texto nas atividades de análise de informações.	Atividades de detectação de oportunidades e ameaças em um estágio inicial e facilitar as informações para decidir e executar as estratégias adequadas. Análise das possíveis maneiras de introduzir soluções de mineração de texto para as Pequenas e médias empresas, descrevendo soluções metodológicas e operacionais de mineração de textos de baixo custo.	Izquierdo; Larreina, 2005.

O agrupamento nas categorias escolhidas surge na perspectiva de que as pesquisas do uso da Mineração de Textos em processos de P&D possam reunir informações similares e abordagens complementares quando Recuperando Informação de Patentes, em bases de especializadas e ou na internet de forma geral. Porém, não fez parte deste trabalho apresentar estas caracterizações de forma mais profunda ou ainda encontrar elementos não aplicáveis em cada um dos propósitos.

5 Considerações finais

A análise dos aspectos identificados nos artigos demonstra a aplicabilidade da técnica de mineração de textos nos processos de P&D. Alguns artigos enfocam a técnica da Mineração de Textos, outros a aplicação, e há aqueles objetivam a sistematização, o método de recuperar informação, utilizando a Mineração de textos para um determinado fim específico dentro do processo de P&D.

Pela categorização observa-se 2 grandes vertentes no uso da Mineração de Textos para P&D: na análise de patentes e na análise de bases especializadas onde neste último é predominante o uso na área da saúde.

A categoria escolhida para o agrupamento dos artigos resultou em um grupo por tipo de bases de dados, porém, os termos da nuvem de palavras formadas com os títulos poderiam sugerir outros agrupamentos como por exemplo: Tecnologia, Tipos de Previsão/ Tendências, elaboração de Mapas e ainda na Sumarização.

Por último, foi percebido, ao final da pesquisa, que alguns artigos não foram recuperados por não apresentarem a forma abreviada do termo *Research and Development, R&D,* mas somente a forma por extenso sem a sigla. Deste modo, uma complementação da equação de busca e uma nova pesquisa deveria ser realizada.

Referências

Brasil, Decreto federal nº 5.798, de 7 de junho de 2006. Recuperado em 10 jun.2014 de ‹http://www.receita.fazenda.gov.br/legislacao/decretos/2006/dec5798.htm.›.

Borges, M. E. N. (1995). A informação recurso gerencial das organizações na sociedade do conhecimento. *Ciência da informação*, Brasília, v. 24, n. 2. Recuperado em 18 mar. 2009 em ‹http://revista.ibict.br/index.php/ciinf/article/viewpdfinterstitial/551/500›.

Choi, S., Kim, H., Yoon, J., Kim, K., & Lee, J. Y. (2013). An SAO-based text-mining approach for technology roadmapping using patent information.R&D Management, 43(1), 52-74.

de Abreu, A. F., & Sinzato, C. I. P. (1999). Acesso à informação–promovendo competitividade em P&D com o uso de tecnologia de informação. Ci. Inf, 28(3), 322-332.

Feldman, R., & Sanger, J. (2007). The text mining handbook: advanced approaches in analyzing unstructured data. Cambridge University Press.

Giddens, A. (2013). Modernity and self-identity: Self and society in the late modern age. Stanford University Press.

Gil, A. C. (2002). Como elaborar projetos de pesquisa. São Paulo, 5, 61.

Ingwersen, P. E. R. (1992). Information Retrieval Interaction. Taylor Graham. Recuperado em 10 jun. 2014 de http://pure.iva.dk/ws/files/31047349/Ingwersen_IRI.pdf

Nanba, H., Ishino, A., & Takezawa, T. (2012). Automatic Compilation of Travel Information from Texts: A Survey. INTECH Open Access Publisher. Recuperado em 10 jun. 2014 de http://www.intechopen.com/books/theory-and-applications-for-advanced-text-mining/automatic-compilation-of-travel-information-from-texts-a-survey

Lancaster, F. W. (1968). Information retrieval systems; characteristics, testing, and evaluation.

Lin, B. W., & Chen, J. S. (2005). Corporate technology portfolios and R&D performance measures: a study of technology intensive firms. R&D Management, 35(2), 157-170.

Manning, C. D., Raghavan, P., & Schütze, H. (2009). Introduction to information retrieval (Vol. 1, p. 496). Cambridge: Cambridge university press.

Mattelart, A. (2002). História da sociedade da informação. Loyola.

Porter, A. L., & Newman, N. C. (2011). Mining external R&D. Technovation, 31(4), 171-176.

Porter, M. (2004). Estrategia competitiva. Elsevier Brasil.

Rijsbergen, C. J. (1995) One introduction. In: Information Retrieval. University of Glasgow. Recuperado em 10 jun. 2014 de http://www.dcs.gla.ac.uk/Keith/Preface.html

Schumpeter, J. A. (1934). The theory of economic development: An inquiry into profits, capital, credit, interest, and the business cycle (Vol. 55). Transaction publishers.

Todesco, Jose Leomar; Carreteiro Díez, Luis Eugenio; Duran, Alfonso (2007) Business intelligence (business intelligence). [slides]. In.: Curso de business intelligence. Escuela complutense latinoamericana, Florianópolis.

Wadsworth, J. (2013) Gráfico R&D as a percentage of gross domestic product. 2014 global r&d funding forecast. R&D Magazine, v. 55, n. 6, p. 6.

Corpus da pesquisa

Alencar, M. S. M., Porter, A. L., & Antunes, A. M. S. (2007). Nanopatenting patterns in relation to product life cycle. Technological Forecasting and Social Change, 74(9), 1661-1680.

Choi, S., Kim, H., Yoon, J., Kim, K., & Lee, J. Y. (2013). An SAO-based text-mining approach for technology roadmapping using patent information.R&D Management, 43(1), 52-74.

de Buenaga, M., Maña, M., Gachet, D., & Mata, J. (2006). The SINAMED and ISIS Projects: Applying Text Mining Techniques to Improve Access to a Medical Digital Library. In Research and Advanced Technology for Digital Libraries (pp. 548-551). Springer Berlin Heidelberg.

de Miranda Santo, M., Coelho, G. M., dos Santos, D. M., & Fellows Filho, L. (2006). Text mining as a valuable tool in foresight exercises: A study on nanotechnology. Technological Forecasting and Social Change, 73(8), 1013-1027.

Epstein, R. J. (2009). Unblocking blockbusters: using boolean text-mining to optimise clinical trial design and timeline for novel anticancer drugs. Cancer informatics, 7, 231.

Izquierdo, J., & Larreina, S. (2005). Collective SME Approach to Technology Watch and Competitive Intelligence: The Role of Intermediate Centers. In Knowledge Mining (pp. 181-189). Springer Berlin Heidelberg.

Jun, S. (2014). A Technology Forecasting Method using Text Mining and Visual Apriori Algorithm. Appl. Math, 8(1L), 35-40.

Jun, S., & Lee, S. J. (2012). Emerging Technology Forecasting Using New Patent Information Analysis. International Journal of Software Engineering & Its Applications, 6(3).

de Miranda Santo, M., Coelho, G. M., dos Santos, D. M., & Fellows Filho, L. (2006). Text mining as a valuable tool in foresight exercises: A study on nanotechnology. Technological Forecasting and Social Change, 73(8), 1013-1027.

Kostoff, R. N., Briggs, M. B., & Lyons, T. J. (2008). Literature-related discovery (LRD): Potential treatments for multiple sclerosis. Technological Forecasting and Social Change, 75(2), 239-255.

Morel, C. M., Serruya, S. J., Penna, G. O., & Guimarães, R. (2009). Co-authorship network analysis: a powerful tool for strategic planning of research, development and capacity building programs on neglected diseases. PLoS Negl Trop Dis, 3(8), e501.

Porter, A. L., & Newman, N. C. (2011). Mining external R&D. Technovation, 31(4), 171-176.

Schoeneck, D. J., Porter, A. L., Kostoff, R. N., & Berger, E. M. (2011). Assessment of Brazil's research literature. Technology Analysis & Strategic Management, 23(6), 601-621.

Senthilkumaran, P., & Amudhavalli, A. (2007). Mapping of spices research in Asian countries. Scientometrics, 73(2), 149-159.

Thorleuchter, D., & Van den Poel, D. (2013). Web mining based extraction of problem solution ideas. *Expert Systems with Applications*, *40*(10), 3961-3969.

Trappey, A. J., Trappey, C. V., & Wu, C. Y. (2009). Automatic patent document summarization for collaborative knowledge systems and services. Journal of Systems Science and Systems Engineering, 18(1), 71-94.

Trappey, A. J., & Trappey, C. V. (2008). An R&D knowledge management method for patent document summarization. Industrial Management & Data Systems, 108(2), 245-257.

Zhu, D., & Porter, A. L. (2002). Automated extraction and visualization of information for technological intelligence and forecasting. Technological forecasting and social change, 69(5), 495-506.

Yoon, B., & Park, Y. (2004). A text-mining-based patent network: Analytical tool for high-technology trend. *The Journal of High Technology Management Research*, *15*(1), 37-50.

Yoon, B., & Park, Y. (2005). A systematic approach for identifying technology opportunities: Keyword-based morphology analysis. Technological Forecasting and Social Change, 72(2), 145-160.

Yoon, B., Lee, S., & Lee, G. (2010). Development and application of a keyword-based knowledge map for effective R&D planning. Scientometrics, 85(3), 803-820.

Wang, M. Y., Chang, D. S., & Kao, C. H. (2010). Identifying technology trends for R&D planning using TRIZ and text mining. R&d Management, 40(5), 491-509.