CLUSTER BINÁRIO E MINERAÇÃO DE PATENTES NA INTELIGÊNCIA DE NEGÓCIOS PARA OFERTANTES DE TECNOLOGIA

Tomoe Daniela Hamanaka Gusberti, Dr. Eng.
Secretaria de Desenvolvimento Tecnológico (SEDETEC) e Parque Científico e Tecnológico Zenit, Universidade Federal do Rio Grande do Sul - UFRGS
Praça Argentina, s/n - Prédio 11.102, CEP 90040-020 - Campus Centro, Porto Alegre, RS, Brasil
tomoe.gusberti@pq.cnpq.br

Marco Aurélio Schunke, M.Sc.
Parque Científico e Tecnológico Zenit, Universidade Federal do Rio Grande do Sul - UFRGS
Praça Argentina, s/n - Prédio 11.102, CEP 90040-020 - Campus Centro, Porto Alegre, RS, Brasil
marco.schunke@ufrgs.br

RESUMO

A inovação aberta ocorre desvinculando o fluxo de informação do fluxo de bens e serviços, criando um mercado para conhecimentos, ideias e tecnologias. As ofertantes de tecnologias definem estrategicamente a configuração da oferta tecnológica. Analistas de patentes utilizam-se de informações retirada de um conjunto massivo de dados de patentes, gerando conhecimento útil para a tomada de decisão relacionada ao negócio. *Patinformatics* automatizam a revelação de inteligência contida em um conjunto de patentes, fontes de dados e informações abertas, necessário em ambiente de mercado para ideias e tecnologias. Este artigo apresenta o desenvolvimento do módulo de data mining de um sistema de *Business Intelligence* para empresas e instituições ofertantes em mercados para tecnologias e ideias, baseado em um caso e utilizando-se da premissa de que é possível conduzir descoberta de conhecimento na base de patentes sem se utilizar de mecanismos avançados de inteligência artificial e processamento de linguagem natural.

PALAVRAS CHAVE. Patent informatics, data mining, cluster binário.

ABSTRACT

Open innovation occurs detaching information flow from goods and service flows, creating a market for knowledge, ideas and technologies. In this context, technology suppliers needs to define strategically the configuration of technological offering. Patent analysists use information obtained from a massive amount of patent data, generating knowledge useful to business related decision making. Patinformatics automatize the disclosing of knowledge from patents open data. This paper presents the development of a data-mining module of a Business Intelligence System for companies and institutions offering ideas and technologies, based in a case. The paper also follows a premise that it is possible to reveal knowledge from a patent database without the use of advanced mechanisms of artificial intelligence or natural language processing mechanisms, for example.

KEYWORDS. Patent informatics, data mining, binary cluster.

1. Introdução

Patentes são dados públicos e abertos criados como mecanismos de difusão do conhecimento tecnológico. Ao mesmo tempo que protegem a propriedade intelectual em mercados específicos visando a obtenção de retorno dos investimentos em P&D, a disseminação da informação contida neste tipo de documento a todo o mundo viabiliza a transferência de conhecimento em países subdesenvolvidos e a promoção da evolução da indústria através do incentivo à criação de novos conhecimentos. No meio corporativo, patentes compreendem fonte de dados e informações abertos e relevantes para a tomada de decisão estratégica e competitividade das empresas (Abbas 2014, Ernst 2003, Trippe). A análise de patentes permite antecipar informações referentes a lançamentos de produtos comerciais, tendências tecnológicas, além de prever o ciclo de vida das tecnologias em uma determinada indústria (Trappey et al, 2010).

No entanto, a quantidade de patentes registradas é muito grande e a análise de todo o conjunto de dados é humanamente impossível sem o estabelecimento de mecanismos de descoberta de conhecimento para auxíliar a tomada de decisão.

Existem dois tipos de usuários deste tipo de dados. Identificadores de patentes utilizam o conteúdo de uma patente, verificando oportunidade em cada uma das patentes, seja para utilizá-los em um negócio ou para verificar a patenteabilidade de uma tecnologia. Analistas de patentes utilizam-se de informações retiradas de um conjunto massivo de dados de patentes, gerando conhecimento útil para a tomada de decisão relacionada ao negócio (Moehrle 2010). Este tipo de método de descoberta de conhecimento é necessário em ambiente de mercado para ideias e tecnologias, cerne da economia baseada em conhecimento.

1.1. *Open innovation*, mercado para tecnologias

Um dos temas largamente abordados ao se falar em inovação compreende a inovação aberta (*open innovation*). Nesta abordagem, as empresas exploram oportunidades no mercado e/ou conduzem projetos colaborativos, ao invés de apenas investirem em desenvolvimento interno (H. Chesbrough & Rosenbloom, 2002; H. W. Chesbrough, 2003). A inovação aberta ocorre devido a um ambiente no qual é possível desvincular o fluxo de informação do fluxo de bens e serviços, criando um mercado para conhecimentos, ideias e tecnologias. Neste contexto, criam-se ofertantes de tecnologia, compreendidas por empresas e laboratórios de pesquisa, universidades e instituições de pesquisa, inovadores seriais, spin-offs e startups inovadores, focados em desenvolver e licenciar ideias e tecnologias que não se ajustam aos objetivos e contextos da empresa, ou que não possuem a pretensão de fabricar ou comercializar na forma de produtos ou serviços (Arora, Fosfuri, & Gambardella, 2001; Arora & Gambardella, 2010; Gans & Stern, 2003; Natalicchio, Messeni Petruzzelli, & Garavelli, 2014; Teece, 1998).

Neste processo, as ofertantes de tecnologias devem definir estrategicamente se vale a pena ofertar a tecnologia no mercado, ao mesmo tempo em que se define a configuração da oferta tecnológica, e a demandante deve definir qual tecnologia vale a pena adquirir ou desenvolver (interna ou externamente) (Natalicchio et al., 2014; Teece, 1998). Assim, a empresa necessita compreender o valor de tecnologias internas e externas compreendem objetivos da gestão estratégica do negócio. E o sistema de *Business Intelligence* de empresas atuantes no contexto de indústrias dinâmicas, por ser o responsável por auxiliar no processo de decisão estratégica, deve contemplar componentes que viabilizem esta análise.

1.2. Sistemas de Business Intelligence, Patent Informatics

Business Intelligence converte dados em informação útil e, a partir deste, em conhecimento. Através do uso de sistemas de business inteligência formais ou informais, alguns conhecimentos são obtidos, enquanto outros são criados através da análise de dados internos e públicos (Negash, 2004).

Sistema de *Business Intelligence* compreende um sistema complexo que analisa dados de forma a apoiar a tomada de decisão gerencial relacionada ao negócio, utilizando-se de dados

oriundos de diversos sistemas operacionais da corporação (Gangadharan & Swami, 2004; Negash, 2004). Embora sistemas de *business intelligence* baseados em dados internos e estruturados de uma empresa possam ser úteis em um ambiente não muito dinâmico, visando o controle da eficiência dos processos internos, não é suficiente em ambientes dinâmicos. Apesar do custo, dados externos, especialmente as relacionadas à concorrência, são essenciais no contexto de indústrias dinâmicas e inovadoras (Calof & Wright, 2008; Cleland & King, 1975; Gilad & Gilad, 1985; Negash, 2004).

Especificamente, em ambientes dinâmicos tecnologicamente, estes sistemas devem suportar decisões referentes ao portfólio de tecnologias e investimentos. E, portanto, análise e utilização de dados de patentes para tomada de decisão, denominados *patinformatics* ou *patent analytics* são essenciais para viabilizar a análise de inserção de tecnologias em mercados específicos, análise econômica, análise do perfil tecnológico de empresas, determinar tendências tecnológicas, prever desenvolvimentos tecnológicos em algum domínio, planejar estrategicamente a tecnologia, *roadmap* tecnológico, identificação de concorrentes tecnológicos, entre outros. Em termos operacionais, *Patifnormatics* compreende o uso ou desenvolvimento de ferramentas automatizado para revelar a inteligência contida em um conjunto de patentes através de técnicas como visualização, analise de citações e outras técnicas como *text mining* (Abbas, Zhang, & Khan, 2014; Abraham & Moitra, 2001). Em termos de softwares comerciais, há muitos mecanismos de busca e visualização de dados de patentes, como contagem nos países dos inventores/titulares, grandes áreas de conhecimento, CPCs, gráficos de evolução temporal da contagem de patentes preenchidas, mapa de citações, identificação de família de patentes entre outros (Yang, Akers, Klose, & Yang, 2008). A maioria dos softwares parece ser voltada para usuários do tipo identificador de patentes e não muito prático para analistas de patentes.

Desta forma, este artigo apresenta o desenvolvimento do módulo de *data mining* de um sistema de B*usiness Intelligence* para empresas e instituições ofertantes em mercados para tecnologias e ideias. Especificamente, explicita se a definição de algoritmos para análise de dados para viabilizar o módulo de *data mining* de patentes do sistema, baseado em um caso.

2. Metodologia

Este artigo apresenta uma pesquisa exploratória e aplicada, visando o desenvolvimento conceitual do módulo de *patent informatics* como uma sequência de procedimentos algorítmicos para o data mining. O desenvolvimento ocorreu com base na análise e compreensão do problema, definição de fonte de dados e variáveis pertinentes, e análise de métodos e medidas de similaridade apropriados. A validação conceitual foi realizando utilizando parcialmente softwares comercialmente disponíveis, sem integrar a partes desenvolvidas pela equipe para viabilizar a flexibilidade necessária em um experimento para gerar insights de melhoria. Linguagem PHP foi utilizada para desenvolvimento do *Crawler*, o tratamento de dados e a análise foram realizados utilizando MS Excel e PASW Statistics 18.0. O método foi aplicado em um caso genérico de discussão ampla e aberta pela comunidade acadêmica, com significativa discussão de políticas públicas e de investimento em pesquisa na área, que permite amplo domínio com quase ausência de gap cognitivo quanto a resultados esperados para fins de comparação da adequação dos resultados por especialistas. Os resultados foram discutidos com grupo multidisciplinar e confrontados com a literatura para promover discussões referentes à adequação ao uso pretendido.

3. Resultados

3.1. Análise e compreensão do problema

Dentre as diversas aplicações e objetivos de patinformatics, este projeto focou no módulo de patinformatics de um sistema de BI que permite analisar o panorama tecnológico de determinado segmento da indústria, sendo o caso específico, a indústria de próteses.

A indústria modifica-se através da evolução de paradigmas tecnológicas, compreendida por fases que demandam tecnologias e projetos de inovação distintos. No início, até a definição do design dominante, processos não coordenados são conduzidos para obtenção de novos produtos (designs alternativos) candidatos. Após a definição do design dominante, o foco é em inovação de processo e a indústria necessitará tecnologias de processos novos para sua viabilização em escala. Posteriormente, passa-se a fase de busca de conhecimentos para otimizar a integração de tecnologias e minimização de custos, e inovações ocorrem geralmente relacionados a fatores relacionados a produção (sistêmica) (Abernathy & Clark, 1985; Abernathy & Utterback, 1978; Bodas Freitas, Marques, & Silva, 2013; Utterback & Abernathy, 1975).

Considerou se que a evolução temporal de grupos significativos de patentes registradas pode auxiliar na identificação dos denominados designs alternativos, compreendidos por apostas tecnológicas dos centros de pesquisa e desenvolvimentos de todo o mundo. A quantidade de patentes, no entanto é massiva e análise individual é humanamente impossível, sendo requerido mineração de dados para auxiliar na busca e descoberta de conhecimento a partir de grandes quantidades de dados. Usualmente este processo de descoberta do conhecimento envolve atividades divididas nas etapas seguintes etapas: a seleção, limpeza e integração dos dados, a transformação dos dados, a mineração dos dados e a avaliação e apresentação dos resultados (Fayyad, Piatetsky-Shapiro, & Smyth, 1996).

3.2. Definição de fonte de dados e variáveis pertinentes

A definição do modelo de decisão, quanto a fonte de dados e variáveis pertinentes está descrito a seguir.

3.2.1. Definição de variáveis

Estrutura do conteúdo da patente é bem definido, compreendo conteúdos básicos do tipo texto (titulo, resumo, texto completo e reivindicações, nome dos inventores e titulares), datas (preenchimento, publicação entre outros), e códigos de classificação.

Os códigos de classificação são frequentemente citados como úteis para busca ou até para definição de grupos. Foram desenvolvidos para permitir a padronização, a ponto de alguns autores (como KANG et al., 2007) os denominarem de “cluster manualmente atribuído”, havendo quem indique inclusive a busca utilizando os mesmos, em substituição a busca por palavras chaves em campos de texto, uma vez que estes podem ser variáveis e de difícil uso, especialmente devido à complexidade, multiplicidade de idiomas possíveis devido a diversidade de nacionalidade dos órgãos de proteção da propriedade intelectual, e também de estilos e vocabulários devido ao domínio das áreas de conhecimento utilizados pelos inventores.

No entanto, a literatura cita algumas das limitações da utilização de códigos de classificação para busca ou definição de grupo se devem a (Kang et al., 2007; Montecchi, Russo, & Liu, 2013; Simmons, 2005): (i) grande número de classes (em torno de 250.000 para CPC, 70.000 para IPC, 245.000 para ECLA/ICO, 160.000 para USPC) que dificulta a seleção de todos os códigos passíveis de se ajustarem ao objetivo da análise; (ii) a complexidade e heterogeneidade das classificações; (iii) o fato da quantidade de códigos nas patentes registrares seguirem a tendência óbvia de seres humanos em se conformarem ao encontrarem uma certa quantidade de códigos descritores ao invés de tentarem explorar todas as opções; (iv) a existência de diversos códigos de classificação conforme a base de dados/país de registro); (v) a evolução do sistema conforme a evolução da área de conhecimento, que ocorre naturalmente de forma posterior a evolução propriamente dita.

Por esta razão, muitas pesquisas recentes na área de patinformatics vem investindo em métodos mais modernos de inteligência artificial, tais como processamento de linguagem natural (PLN), além de *textmining* para posterior agrupamento para visualização de dados. Esta estratégia, embora receba bastante atenção na academia devido ao seu potencial (Abbas et al., 2014), possui limitações. Além de demandar significativo esforço de processamento, tecnologias podem ter sido desenvolvidos sob influência de áreas de conhecimento diferentes e, portanto,

suas patentes podem ter sido redigidas utilizando vocabulário e perspectivas diferentes. Os dicionários usualmente disponíveis não são técnicos, nem abrangem todas as línguas possíveis, além de precisarem ser constantemente atualizados, considerando a evolução natural da linguagem científica e técnica.

Este projeto utiliza-se da premissa de que é possível conduzir data mining e descoberta de conhecimento na base de patentes sem se utilizar de mecanismos avançados de inteligência artificial e processamento de linguagem natural, utilizando-se de classificação CPC (*Cooperative Patent Classification*). A CPC é atribuída manualmente, não apenas por quem conduz a solicitação de registro, mas também envolve uma análise meticulosa por analistas na agência de proteção da propriedade intelectual (Montecchi et al., 2013). A literatura a indica como algo relevante para caracterização do conteúdo, sendo utilizada inclusive como critério para avaliação da abrangência ou escopo Tecnológico de uma patente (Ernst, 2003).

3.2.2. Definição de fonte de dados

Quanto à fonte de dados, as patentes possuem validade nacional e devem ser registradas em diversos países de forma a garantir um mercado global. No entanto, obter dados de diversos países, tratar conteúdos codificados em formatos, língua e linguagens distintas e, principalmente estabelecer e incorporar a correspondência entre eles requer esforços de processamento (e, portanto, de desenvolvimento) consideráveis. Por esta razão, considerou se que a base de patentes americana seria suficiente para fins propostos, uma vez que compreende o sistema de propriedade intelectual mais forte, além de mercado forte e significativo para tecnologias. Desta forma, a base de dados utilizada foi a da USPTO.

Esta escolha não seria apropriada caso o sistema se propusesse a analisar a patenteabilidade de determinada patente, como requerida por identificadores de patentes, porém considerou-se que não já grandes riscos para analistas de tecnologia que necessitam compreender o mercado de forma geral.

As patentes para comporem a base local de analise foram obtidas através de um *crawler* atuando sobre site do USPTO. Além do número da patente, buscou se o CPC, utilizando o documento *html* disponibilizado pelo mesmo site. O *crawler* utilizou se do sistema de busca rápida por palavras chaves da USPTO em texto completo. A busca por CPC, embora discutida como vantajosa na literatura, não foi empregada pois a base de CPC e muito grande e é humanamente inviável a classificação fidedigna e exaustiva, embora possua uma estrutura hierárquica e seja relativamente confiável por ser atribuído por pessoal especializado e treinado (agentes do órgão de proteção de propriedade intelectual).

3.3. Definição da estratégia

Considerando os fatores descritos acima, definiu-se como estratégia de análise o agrupamento de patentes oriundas da USPTO quanto a seu conteúdo Tecnológico. Como passos principais, seguem-se os mesmos apresentados na literatura para agrupamento de patentes, porém com algumas alterações. Embora a literatura aposte mais em tecnologias de informação emergentes como *natural language processing* (NLP) e *semantics* (Bonino, Ciaramella, & Corno, 2010), o presente artigo defende que há ainda bastante potencial a ser explorado em *data mining* com soluções mais simples, utilizando-se de agrupamento baseado na codificação CPC disponível. Especificamente, observa-se na literatura a utilização de medida de similaridade distância euclidiana ou Manhattan, e *k-means* e hierárquico como métodos de agrupamento (Abbas et al., 2014; Chang, 2012; Trappey, Trappey, & Wu, 2010). Neste artigo, optou-se por utilizar medida de similaridade *Dice*, apropriada aos dados binários obtidos, com Análise hierárquica utilizando método de agrupamento *within groups*. Uma outra dificuldade considerada foi a dificuldade de processamento da análise por limitações dos recursos computacionais disponíveis em determinadas situações. Considerou-se que o uso de amostragem poderia ser viável, contanto que garantida a aleatoriedade na amostragem, dado o objetivo de se analisar o panorama tecnológico.

Embora as definições CPC no nível de subgrupo sejam mais detalhadas e interessantes para descrever a tecnologia, a grande quantidade de categorias disponíveis (mais de 70.000) não só pode dificultar a análise (obtenção de cluster), mas também indica que a exaustão das possibilidades de classificação por seres humanos não é viável. Considerando que as definições CPC seguem uma estrutura hierárquica de desdobramento das áreas e conteúdos de conhecimento, optou-se por utilizar uma estratégia de agrupamento em duas etapas, compreendidas por agrupamento no nível de grupo, seguida por agrupamento no nível de subgrupo.

3.4. O experimento para validação conceitual

Para a aplicação experimental, foi selecionado o caso de diagnóstico do panorama de pesquisa e desenvolvimento (e registro de patentes) na área de órteses e próteses. O caso foi selecionado por ser uma área tecnológica: (i) ainda não consagrada, cujos usuários finais e sociedade carecem e anseiam por soluções; (ii) no qual ainda não se estabeleceu um design dominante e o paradigma ainda está para ser definido; (iii) havendo uma grande quantidade de investimentos em pesquisa e desenvolvimento conduzidos no mundo; (iv) um tema de comoção social devido a relação com problemas de saúde, existindo discussão ampla e aberta pela comunidade acadêmica, com significativa discussão de políticas públicas e de investimento em pesquisa na área, que permite identificação fácil do resultado esperado para fins de comparação da adequação dos resultados.

3.4.1. Busca e formação da base local de análise

Para operacionalizar a estruturação da base necessária para a condução da análise, desenvolveu-se um *crawler*, um robô da web, desenvolvido para varrer e extrair dados (número da patente, CPC, ano) de uma quantidade predefinida de patentes do website da USPTO (*United States Patent and Trademark Office*) através da pesquisa de palavras chaves, em linguagem PHP.

As palavras chaves utilizadas foram "*prothesis*" e "*prosthesis*", para contemplar variantes de digitação. E a busca foi efetuada entre os dias 5 e 4 de abril de 2016. A Tabela 1 apresenta o retorno obtido com o Crawler.

Tabela 1: quantidade de retornos obtidos pelo *crawler* em comparação com dados existentes na base original

	Qqtde no USPTO	Obtidos	
		Patentes	CPC
Prosthesis	28330	28330 (100%)	26053 (92%)
Prothesis	2010	2010 (100%)	2010 (100%)

O retorno menor na fase de obtenção do CPC deve-se a diversidade de codificações no caso de patentes do tipo design. A base compreendia de algumas patentes repetidas, o qual, após a retirada de registros redundantes, resultou em 26.598 patentes. Destes ainda estavam incluídas patentes do tipo RE (*Reissue* ou *Re-examination*), aos quais se considerou melhor omitir da base para evitar duplicidade de registros, totalizando uma base de 26.490 patentes.

3.4.2. Tratamento de dados

Os dados obtidos foram armazenados em base SQL e exportados para formato CSV para Excel para tratamento de dados em MS Excel®. A base era compreendida por 26.490 patentes com 17.188 códigos CPC no nível de subgrupo e, transformado para nível de grupo, 367 grupos.

A Figura 1 apresenta a distribuição de quantidade de códigos CPC atribuídos por patente. As patentes da base apresentavam em média 10,65 códigos CPC em nível de subgrupo (ou grupo principal), sendo a moda 2 e o máximo de 129 códigos. Após transformação, obtêm-se, para CPC de nível de grupo, média de 3,99 códigos, moda de 2 e máximo de 29 códigos por patente.

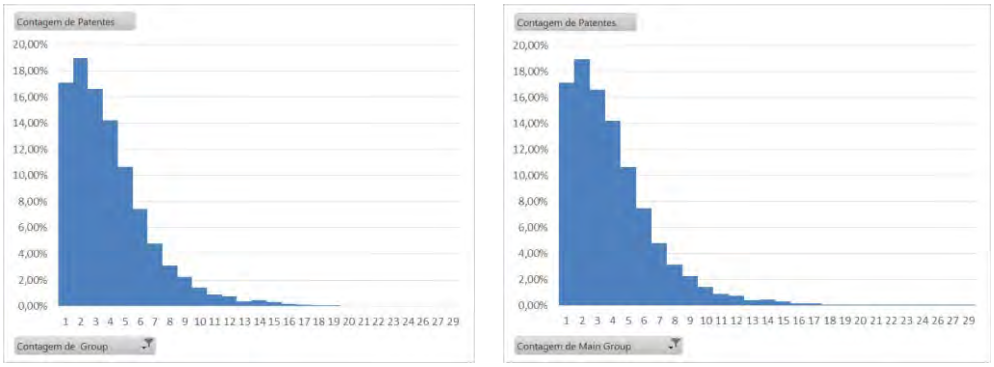

Figura 1: distribuição de quantidade de códigos CPC atribuídos por patente na base obtida

A partir dos dados extraídos, conduziu-se tratamento de dados para obtenção de matriz binária do tipo patentes x CPC, seja no nível grupo ou subgrupo. Em um primeiro momento, a matriz de 367 grupos de CPC por 26.490 patentes, utilizando planilha dinâmica em MS EXCEL® 32 bits. No entanto, ao tentar conduzir a obtenção de cluster em PASW Statistics 18 64bits, o mesmo indicou falta de memória para processar a análise.

A obtenção, tratamento e análise de dados foi conduzido em computador pessoal cuja configuração está disposta na Figura 2:

Windows 7 professional 64 bits
HP Compaq elite 8300 SFF
Intel® Core™ i5-3570 CPU 3.40 GHz
4,00 GB Memória RAM

Figura 2: configuração do computador pessoal utilizador

Para viabilizar a análise, conduziu-se amostragem, utilizando gerador de números aleatórios do MS Excel® (distribuição binomial, p= 0,2). A análise foi conduzida com amostragem de 16% o qual foi viável de processamento no equipamento utilizado (p=0,6; 0,5; 0,4 e 0,3 também foram testados).

3.4.3. Análise de agrupamento, Resultados e Interpretação dos resultados

A partir da amostra, gerou-se matriz binária patentes x CPC, conduziu-se cluster hierárquico, método *within groups*, utilizando medida Dice and Sorensen, em PASW SPSS 18. O método visou obter agrupamentos cuja distâncias médias de seus membros fossem menores e, os grupos, mais homogêneos.

Tabela 2: Segmento da tabela contendo descrição dos agrupamentos obtidos utilizando CPCs nível de grupo

CPC Definição	N Cluster	101 1	540 2	1196 3	672 4	340 5	258 6
Filters; Devices providing patency to tubular structures; Prostheses; Accessories	A61F2	17%	62%	79%	76%	39%	69%
Joints	A61F2002	12%	60%	70%	55%	8%	60%
Fixations or connections for prostheses classified in groups A61F 2/00 to A61F 2/26or A61F 2/82or A61F 9/00or A61F 11/00or subgroups thereof	A61F2220	0%	1%	39%	27%	0%	62%
Special features of prostheses classified in groups A61F 2/00to A61F 2/26 to A61F 2/82or A61F 9/00or A61F 11/00or subgroups thereof	A61F2250	1%	0%	24%	67%	4%	1%
Orthopaedic methods or devices for non-surgical treatment of bones or joints […]; Nursing devices	A61F5	77%	0%	0%	0%	0%	0%

A medida de similaridade Dice and Sorensen foi selecionada por omitir a concordância 0-0 e atribuir o dobro de peso à concordância 1-1 (Choi, Cha, & Tappert, 2010; Johnston, 1976). Esta medida pode ser descrita como medidas de associação indicada na literatura como medida largamente utilizada para agrupar documentos (Willett, 1988).

Para descrição dos agrupamentos obtidos, gerou-se tabela descritiva dos clusters com base nos códigos CPC utilizados para obtenção dos clusters. Parte desta tabela é apresentada em Tabela 2. E os agrupamentos foram denominados conforme indicado na Figura 3.

Próteses	2,3,4,6	*Joints* para estruturas tubulares, filtros e próteses
	5	Materiais de coating para estruturas tubulares, filtros, de próteses
	10	Catéteres, *probes* ocos para próteses, estruturas tubulares, acessórios podendo focar em *joints*
	11	Preparações medicinais contendo peptídeos (maioria maior que 20 aminoácidos), alguns para cobrir próteses
	9	Prótese auditivo
Outros	8	Dental
	1	Métodos ou dispositivos ortopédicos não cirúrgicos
	7	Outros
Cobertura ou liberação	11	Preparações medicinais contendo pepítdeos
	5	Materiais de *coating*

Figura 3: denominação dos agrupamentos obtidos

A partir dos agrupamentos obtidos, gerou-se figuras para demonstrar a evolução temporal das linhas de investimento em P&D expresso pelos agrupamentos, apresentados na Figura 4.

Observa-se um decréscimo na quantidade de patentes nos últimos anos, possivelmente em decorrência da demora até a publicação. No entanto, observa-se que o grupamento 5 referente a materiais de cobertura apresentam este decréscimo antes do decréscimo no conjunto global.

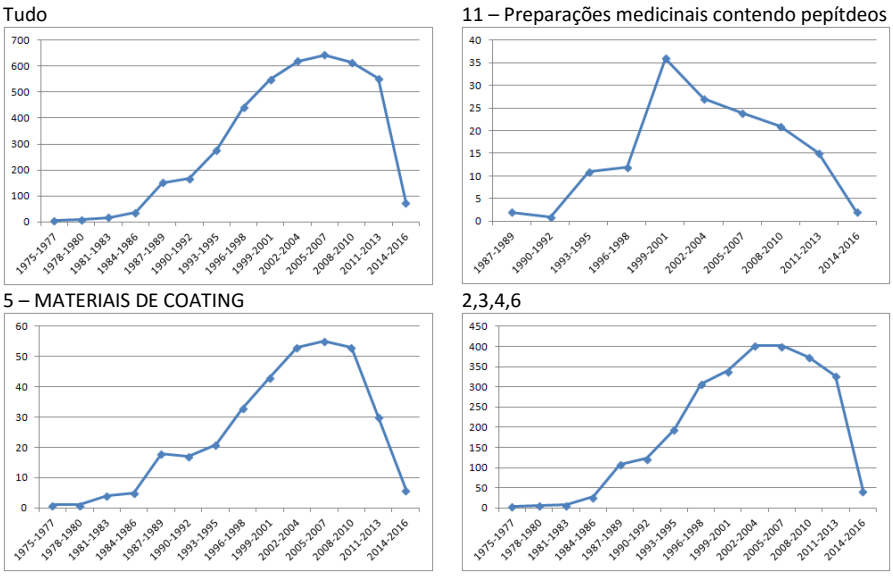

Figura 4: evolução temporal dos agrupamentos

6 obtained otherwise than by reactions only involving carbon-to-carbon unsaturated bonds
1 **Polyesters** derived from hydroxy carboxylic acids, obtained otherwise than by reactions only involving carbon-to-carbon unsaturated bonds;
2- outros
3 – **Polyurethanes**, obtained otherwise than by reactions only involving carbon-to-carbon unsaturated bonds,
4 Macromolecular materials
5 Materials characterised by their function or physical properties
7, 9 Biologically active materials,
10 characterised by the human or animal origin of the biological material; subjected to a specific treatment prior to implantation.
11 .Other specific inorganic materials not covered by A61L 27/303 to A61L 27/32}

Figura 5: parte dos grupos obtidos por agrupamento de patentes do grupo 5 utilizando CPC no nível subgrupo4

O cluster 5 foi reanalisado, conduzindo nova análise de cluster para este subconjunto de patentes, utilizando códigos de CPC no nível subgrupo, isto é, com maior detalhamento do

conteúdo tecnológico da patente. Os procedimentos similares às etapas anteriores geraram subagrupamentos nomeados na Figura 5.

3.5. Discussão

Os resultados foram discutidos com grupo multidisciplinar composto por analistas de tecnologia e coordenador do projeto de avaliação tecnológica de uma universidade pública. A discussão visou a adequação dos resultados obtidos com a literatura da área e consequente adequação ao uso pretendido.

3.5.1. Quanto à adequação dos resultados

Para apresentação de definições de códigos CPC para interpretação pelo analista de tecnologia, utilizou-se uma tabela de definições obtida a partir de dados disponíveis, estruturada para leitura textual por ser humano, não uma matriz de dados para integração em sistemas. Desta forma, as definições apresentadas em tabelas de resultados não eram passíveis de interpretação precisa. Por exemplo, a Figura 6 ilustra que o código CPC de nível subgrupo C07K 7/06, definida como "*having 5 to 11 amino acid*", expressa um detalhe de algo mais específico do *main group* C07K 7/00 "*Peptides...*", já desdobrado previamente em C07K 7/04, definido como o "*Linear peptides containing only normal peptide links*".

C07K 7/00 Peptides having 5 to 20 amino acids in a fully defined sequence; Derivatives thereof NOTE
In this subgroup cyclic compounds related to specific compounds which are classified in a specific group, e.g. C07K 7/062, are classified in this specific group only
C07K 7/02 **.** Linear peptides containing at least one abnormal peptide link
C07K 7/04 **.** Linear peptides containing only normal peptide links
C07K 7/06 **. .** having 5 to 11 amino acid

Figura 6: trecho do arquivo de definições C07K

Embora a leitura textual no arquivo de definições permita a identificação da finalidade restritiva ou de especificação do C07K 7/06, a transcrição do conteúdo para uma matriz de dados e posterior integração para geração de tabela descritiva dos clusters não permite a interpretação adequada. Tais correções foram realizadas, no caso descrito, pela consulta manual dos arquivos de definições disponíveis, limitando a otimização do processo de análise por fornecedores de tecnologia os quais necessitam avaliar diversas tecnologias para oferta-los no mercado para tecnologias.

3.5.2. Quanto a aspectos metodológicos da análise de cluster

Ao utilizar a transformação para analisar as codificações de conteúdo tecnológico para o nível de grupos, obteve-se uma relação entre variáveis de agrupamento x quantidade de objetos a serem agrupados mais adequada segundo a literatura (Everitt, Landau, Leese, & Stahl, 2011). Os agrupamentos obtidos no nível de subgrupo (codificações originais não transformadas), extrapolam a proporção ideal. Estes valores estão muito distantes dos valores indicados na literatura como ideais (Mooi & Sarstedt, 2011), embora usualmente referindo-se a análise de cluster de variáveis contínuas. Os resultados demonstraram-se razoáveis devido à significativa muticolinearidade e, pelo fato de representarem subgrupos mais homogêneos previamente isolados pela obtenção de clusters utilizando categorias de conteúdo tecnológico (CPC) no nível de grupo. Uma saída viável é omitir da matriz para obtenção de clusters os códigos de pouca ocorrência (dados não computados, mas se considerar 5% das patentes como ponto de corte, o critério seria facilmente atendido). Retomando as definições CPC, as categorias de nível de subgrupo na realidade apresentam relacionamentos entre si por ainda existirem estrutura hierárquica com outros níveis hierárquicos. Tal fato não pode ser considerado no procedimento utilizando as ferramentas disponíveis no MS Excel, mas poderia ser trabalhada eficientemente em programas dedicados, com grande possibilidade de sanar estas deficiências.

3.5.3. Quanto a limitações, riscos e possibilidade de falhas dos procedimentos envolvidos

Quanto à reprodutibilidade do procedimento (manutenção do uso dos componentes implementados), deve-se mencionar que o *Crawler* funciona apenas enquanto a estrutura da página de busca e apresentação de resultados da USPTO for mantida e for permitida pela instituição.

A análise da evolução de quantidade de patentes pode sofrer influência de alguns fatores alheios à evolução dos paradigmas tecnológicos ou do investimento na indústria em Pesquisa e/ou Desenvolvimento. Por exemplo, Simmons (2005) indica: (i) possibilidade alteração de políticas de registro de patentes pelas empresas; (ii) influência de leis nacionais de proteção de propriedade intelectual que alteram as estratégias de registro de patentes das empresas. Tais questões não são considerados no processo de análise ou apresentação dos resultados, o que pode induzir a interpretações enviesadas, caso o analista de tecnologia (que irá interpretar os resultados para tomar a decisão) não estiver claramente avisado.

A obtenção de agrupamentos visou à identificação de paradigmas ou linhas de (combinação de) conteúdos tecnológicos aos quais ocorre investimento em P&D (que resulta em algo patenteável). Paradigmas tecnológicos emergentes se manifestam, por definição, em quantidades de patentes menores, nos anos mais recentes, o que, considerando os procedimentos algorítmicos envolvidos na análise de cluster, tenderia a ficar mascarado na formação dos agrupamentos pela análise de cluster, se esta quantidade de patentes for muito pequena, mesmo que se utilizasse de toda a base de dados. Patentes como fontes de dados não permitem representar o conteúdo de projetos de pesquisa e desenvolvimento em andamento, mas uma invenção já consolidada e cuja proteção seja comercialmente interessante. Desta forma, o método apresenta a limitação de acabar dificultando e retardando a detecção de linhas emergentes pouco expressivos em termos de quantidade de patentes.

O procedimento empregado neste caso necessitou lançar mão de amostragem devido à grande quantidade de dados envolvidos e limitação dos softwares utilizados. A questão da definição de tamanho de amostra não é baseado em considerações inferenciais na análise de cluster, por não apresentar um fundamentos estatísticos, mas numéricos que permitem obter as características estruturais do conjunto de observações baseadas em propriedades matemáticas (Hair, Black, Babin, Anderson, & Tatham, 2009). Como o objetivo da análise era obter uma descrição do panorama de evolução tecnológica de determinado setor, amostras seriam suficientes para prover esta descrição e os resultados mostraram-se adequados, no entanto, conceitualmente, a amostragem, dependendo do corte realizado e a diversidade dos objetos a serem agrupados, possui um risco significativo da amostra acabar omitindo este novo paradigma, por exemplo. Embora paradigmas ou designs alternativos abandonados e pouco expressivos no passado não prejudiquem os objetivos da análise, seria desejável que os mais recentes fossem passíveis de serem identificados. Como não há nenhuma estimativa da diversidade das patentes e suas classificações nas categorias de CPC não é disponível na metodologia atual, não há como estimar o erro amostral e/ou representatividade da amostra nas diferentes situações ou problemas de análises que o método pode ser empregado. Até o momento, a revisão de literatura referente a metodologia de análise de cluster não identificou preocupações ou índices para a medida de diversidade e ou variedade de objetos para definição de tamanhos de amostra ou estimativa de erro amostral e/ou representatividade da amostra, embora algumas considerações quanto a multicolinearidade sejam mencionadas. No entanto, deve-se mencionar que a multicolinearidade, embora seja a base para a obtenção dos clusters, medidas para mensuração não são usualmente disponíveis em softwares ou utilizadas por analistas.

4. Considerações finais

O artigo demonstrou aplicabilidade da mineração de dados e análise de cluster sem uso de método de processamento de linguagem natural na gestão tecnológica para *technology sourcers*, o que seria inviável por análise individual por seres humanos devido a grandes quantidades de dados de patentes. A análise de cluster permitiu a obtenção de grupos de patentes

como mecanismo para identificação dos denominados designs alternativos. A evolução temporal destes grupos demonstrou aplicabilidade para análise do panorama tecnológico de segmento da indústria, no contexto de processo de *business intelligence* de uma instituição de pesquisa atuante como *technology sourcer*.

Em termos operacionais, algumas melhorias necessárias para o procedimento e incorporação no sistema compreendem: (i) trazer os dados localmente para agilizar o processamento; (ii) Incorporar sistemas de monitoramento de oscilações na quantidade global de patentes registradas em decorrência a mudanças no sistema de proteção de propriedade intelectual e incorporar na visualização dos resultados. Além disso, como, embora a migração para um sistema dedicado e especificamente desenvolvido para a análise possa ser mais eficiente e possuir menor limitação quanto a processamento de grande quantidade de dados, dependendo do tema e da seleção das palavras-chaves, a condução de análise a partir de amostragem de dados é inevitável. Desta forma, há a necessidade de (iii) definir um índice de estimativa da variabilidade dos dados (ou diversidade das patentes da amostra) para estimativa do erro amostral; (iv) e critérios de definição de tamanho de amostra (parâmetro binomial p).

Agradecimentos

Os autores agradecem a CAPES, FAPERGS (Proc. SPI 2842-25.51/12-0), CNPq (Proc.460785/2014-1), SEDETEC (Secretaria de Desenvolvimento Tecnológico da UFRGS) e Parque Científico e Tecnológico Zenit da UFRGS pelo apoio financeiro e infraestrutura para viabilização do projeto de pesquisa.

Referências

Abbas, A., Zhang, L., & Khan, S. U. (2014). A literature review on the state-of-the-art in patent analysis. World Patent Information.

Abernathy, W. J., & Clark, K. B. (1985). Innovation: Mapping the winds of creative destruction. Research Policy, 14(1), 3–22.

Abernathy, W., & Utterback, J. (1978). Patterns of industrial innovation. Technology Review, 80(7), 40–47.

Abraham, B. P., & Moitra, S. D. (2001). Innovation assessment through patent analysis. Technovation, 21, 245–252.

Arora, A., Fosfuri, A., & Gambardella, A. (2001). Specialized technology suppliers , international spillovers and investment : evidence from the chemical industry, 65.

Arora, A., & Gambardella, A. (2010). Ideas for rent: An overview of markets for technology. Industrial and Corporate Change, 19(3), 775–803.

Bodas Freitas, I. M., Marques, R. A., & Silva, E. M. D. P. E. (2013). University-industry collaboration and innovation in emergent and mature industries in new industrialized countries. Research Policy, 42(2), 443–453.

Bonino, D., Ciaramella, A., & Corno, F. (2010). Review of the state-of-the-art in patent information and forthcoming evolutions in intelligent patent informatics. World Patent Information, 32, 30–38.

Calof, J. L., & Wright, S. (2008). Competitive intelligence: A practitioner, academic and inter-disciplinary perspective. European Journal of Marketing, 42(7/8), 717–730.

Chang, S.-B. (2012). Using patent analysis to establish technological position: Two different strategic approaches. Technological Forecasting & Social Change, 79, 3–15.

Chesbrough, H., & Rosenbloom, R. S. (2002). The role of the business model in capturing value from innovation: evidence from Xerox Corporation's technology spin-off companies. Industrial and Corporate Change, 11(3), 529–555.

Chesbrough, H. W. (2003). The Era of Open Innovation. MIT Sloan Management Review, 44(3), 35–41.

Choi, S.-S., Cha, S.-H., & Tappert, C. C. (2010). A Survey of Binary Similarity and Distance Measures. Journal of Systemics, Cybernetics & Informatics, 8(1), 43–48.

Cleland, D. I., & King, W. R. (1975). Competitive business intelligence systems. Business Horizons, 18(6), 19–28.
Ernst, H. (2003). Patent information for strategic technology management, 25.
Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). Cluster Analysis. John Wiley & Sons.
Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). The KDD process for extracting useful knowledge from volumes of data. Communications of the ACM, 39(11), 27–34.
Gangadharan, G. R., & Swami, S. N. (2004). Business intelligence systems: design and implementation strategies. Information Technology Interfaces, 2004. 26th International Conference on, 139–144 Vol.1.
Gans, J. S., & Stern, S. (2003). The product market and the market for "ideas": commercialization strategies for technology entrepreneurs. Research Policy, 32(2), 333–350.
Gilad, B., & Gilad, T. (1985). A systems approach to business intelligence. Business Horizons, 28(5), 65–70.
Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (2009). Análise multivariada de dados - 6ed. Bookman.
Johnston, J. W. (1976). Similarity indices I: What do they measure? Battelle: Pacific Northwest Laboratories, BNWL-2152(NRC-1 Addendum).
Kang, I.-S., Na, S.-H., Kim, J., & Lee, J.-H. (2007). Cluster-based patent retrieval.
Moehrle, M. G.; Walter, L.; Bergmann, I.; Bobe, S.; Skrzipale, S. (2010). Patinformatics as a business process: A guideline through patent research tasks and tools. World Patent Information, 32, 291-299.
Montecchi, T., Russo, D., & Liu, Y. (2013). Searching in Cooperative Patent Classification: Comparison between keyword and concept-based search. Advanced Engineering Informatics, 27, 335–345.
Mooi, E., & Sarstedt, M. (2011). Cluster Analysis. In A concise Guide to Market Research: the proces, data, and methods using IBM Statistics (pp. 237–284). Springer-Verlag Berlin Heidelberg.
Natalicchio, A., Messeni Petruzzelli, A., & Garavelli, A. C. (2014). A literature review on markets for ideas: Emerging characteristics and unanswered questions. Technovation, 34(2), 65–76.
Negash, S. (2004). Business intelligence. The Communications of the Association for Information Systems, 13, 177-195.
Simmons, E. S. (2005). Trends disrupted—patent information in an era of change. World Patent Information , 27, 292–301
Teece, D. J. (1998). Capturing value from knowledge assets: The new economy, markets for know-how, and intangible assets. California Management Review, 40(3), 55–80.
Trappey, C. V, Trappey, A. J. C., & Wu, C.-Y. (2010). Clustering Patents Using Non-Exhaustive Overlaps. J Syst Sci Syst Eng, 19(2), 162–181.
Utterback, J. M., & Abernathy, W. J. (1975). A dynamic model of process and product innovation. Omega, 3(6), 639–656.
Willett, P. (1988). Recent trends in hierarchic document clustering: A critical review. Information Processing and Management.
Yang, Y., Akers, L., Klose, T., & Yang, C. B. (2008). Text mining and visualization tools – Impressions of emerging capabilities. World Patent Information , 30, 280–293.