MINERAÇÃO DE TEXTO E CLUSTERIZAÇÃO EM ESTUDOS BIBLIOMÉTRICOS: O MAPEAMENTO CIENTÍFICO DE TESES E DISSERTAÇÕES DE UM PROGRAMA DE PÓS-GRADUAÇÃO

Luciano Zamperetti Wolski[1];

Ricardo Pereira[2];

Alexandre Leopoldo Gonçalves[3];

Cristiano José Castro de Almeida Cunha[4];

Abstract: *Bibliometric studies have the particularity of analyzing the performance of a research field/area. The present study, following this premise, will map the works of a Postgraduate program at a Brazilian Federal University. For this purpose, text mining and clustering techniques are used to analyze 609 works, including theses and dissertations. The use of text mining techniques and visualization of similarities provided the indication of groupings of knowledge areas. The similarity analysis indicates the little interaction between the fields of knowledge, characterizing the existence of "conceptual islands". This result may have been impacted by the way the data was extracted in the similarity analysis. The analysis also made it possible to verify which themes arouse greater interest in the program in relation to the research being carried out.*

Keywords: *Mining texts; Bibliometry; Scientific mapping; VOSviewer®; Orange®*

Resumo: Os estudos bibliométricos possuem a particularidade de analisar o desempenho de um campo/área de pesquisa. O presente estudo, seguindo essa premissa, mapeará os trabalhos de um programa de Pós-graduação de uma Universidade Federal brasileira. Para tal, se utiliza de técnicas de mineração de textos e clusterização para analisar 609 trabalhos, entre teses e dissertações. A utilização das técnicas de mineração de textos e visualização de similaridades proporcionou a indicação de agrupamentos de áreas de conhecimentos. A análise de similaridades indica a pouca interação entre os campos de conhecimento, caracterizando a existência de "ilhas conceituais". Tal resultado pode ter sido impactado pela forma de extração dos dados na análise de similaridades. A análise, ainda, possibilitou verificar quais os temas que despertam maior interesse do programa em relação às pesquisas que estão sendo realizadas.

Palavras-chave: *Mineração de textos; Bibliometria; Mapeamento científico; VOSviewer®; Orange®*

[1] Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento –Universidade Federal de Santa Catarina (UFSC) Florianópolis – Brasil. ORCID: https://orcid.org/ 0000-0003-4683-1013. e-mail: lwolski@gmail.com
[2] Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento –Universidade Federal de Santa Catarina (UFSC) Florianópolis – Brasil. ORCID: https://orcid.org/ 0000-0003-4744-4891. e-mail: rikardop@gmail.com
[3] Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento –Universidade Federal de Santa Catarina (UFSC) Florianópolis – Brasil. ORCID: https://orcid.org/ 0000-0002-6583-2807. e-mail: a.l.goncalves@ufsc.br
[4] Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento –Universidade Federal de Santa Catarina (UFSC) Florianópolis – Brasil. ORCID: https://orcid.org/ 0000-0002-8459-6045. e-mail::01cunha@gmail.com

1. INTRODUÇÃO

Os estudos bibliométricos são técnicas quantitativas e estatísticas que tem o intuito de mensurar índices de produção e disseminação do conhecimento científico por meio da análise e investigação de textos científicos em um campo ou vários campos de pesquisa (Araújo, 2006; Roemer et al., 2015; de Bellis, 2009). É prática comum nesses estudos construir uma relação entre as referências citadas ou tópicos de palavras-chave, um método também conhecido como 'mapeamento científico' (Boyack & Klavans, 2010).

Esse tipo de estudo tem se popularizado entre os acadêmicos em função da grande quantidade de material bibliográfico que é produzido e disponibilizado atualmente. Ter uma visão resumida e sistematizada disso pode facilitar o entendimento e, até mesmo, apontar futuros caminhos de pesquisa. Os resultados de estudos bibliométricos podem auxiliar jovens pesquisadores ou mesmo aqueles mais experientes que se deparam com uma nova temática (Quevedo-Silva, Santos, Brandão, & Vils, 2016).

Com o advento de novos métodos, técnicas e ferramentas, o processo de análise bibliométrica têm sido facilitado, potencializando as análises por cobrir uma maior quantidade de dados e minimizar a intervenção humana na coleta e análise desses dados.

Uma destas técnicas/ ferramenta é a mineração de dados. Nos últimos anos, ela tem recebido muita atenção, devido ao grande volume de dados de texto criados diariamente das mais variadas formas. O volume de texto disponível é uma fonte inestimável de informação e conhecimento e se apresentam de forma estruturada e não estruturada (Allahyari, Pouriyeh, Assefi, Safaei, Trippe, Gutierrez, & Kochut, 2017). A descoberta de conhecimento em bancos de dados é um processo que envolve várias etapas a serem aplicadas ao conjunto de dados de interesse, a fim de extrair padrões úteis. Essas etapas são interativas e podem exigir que as decisões sejam tomadas pelo usuário. Várias técnicas são utilizadas na mineração de texto: processamento de linguagem natural (PLN), análise semântica, regras, função e redes neurais (Abbas, Zhang, & Khan, 2014).

A mineração de texto, como a mineração de dados ou a descoberta de conhecimento, é frequentemente vista como um processo para encontrar padrões implícitos, anteriormente desconhecidos e potencialmente úteis em um grande repositório de textos. Na prática, o processo de mineração de texto envolve uma série de interações do usuário com as ferramentas de mineração de texto para explorar o repositório e encontrar esses padrões (Tseng, Lin, & Lin, 2007).

Nesta pesquisa, utiliza-se algumas técnicas de mineração de texto, com o auxílio da ferramenta *Orange*®[5] (software para o desenvolvimento de modelos de mineração de dados), com destaque para o método hierárquico, representado por um dendrograma, com *bag of words* (*bow*) e *embeddings* de palavras. O método hierárquico utiliza algoritmos de agrupamento com base na distância, ou seja, usam função de similaridade para medir a proximidade entre documentos de texto (Allahyari et al., 2017).

A abordagem *bow* se baseia na contagem de palavras para medir a similaridade baseada em texto entre documentos, quanto mais dois textos se assemelham em suas frequências de palavras, mais semelhantes eles são (Le & Mikolov, 2014). Enquanto que o uso das *embeddings* de palavras, as relações sintáticas e semânticas entre as palavras/frases podem ser extraídas e representadas (Onan, 2019).

Os estudos bibliométricos podem ter propósitos variados - os mais comuns são retratar um campo e as tendências de pesquisas; autores, instituições, países e periódicos mais influentes. O presente trabalho tem por objetivo traçar um panorama das pesquisas do Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento (PPGEGC) da Universidade Federal de Santa Catarina (UFSC).

O PPGEGC objetiva à formação de mestres e doutores, bem como produzir conhecimento técnico-científico para a sociedade (PPGEGC, 2020). O PPGEGC foi lançado oficialmente em 2004, na Universidade Federal de Santa Catarina, como um programa interdisciplinar que, atualmente, está fundamentada em três áreas de concentração: Engenharia do Conhecimento (EC), Gestão do Conhecimento (GC) e Mídia do conhecimento (MC). As áreas de conhecimento são formadas por três identidades paradigmáticas: cognitivista (EC), autopoiética (GC) e conexionista (MC). A área da EC estuda a modelagem e o desenvolvimento de sistemas de conhecimento. A área de GC estuda o estabelecimento do ciclo estratégico de sistemas do conhecimento e a área de MC estuda a difusão e comunicação do conhecimento (PPGEGC, 2020).

Este artigo, então, tem sua gênese na necessidade de conhecer os trabalhos realizados por um programa de Pós-Graduação, retratando-os de modo a permitir o acompanhamento do que foi pesquisado por seus mestres e doutores e indicando as tendências de estudos aos futuros pós-graduandos.

Além da introdução, o artigo apresenta quatro seções. A seção 2 descreve os procedimentos metodológicos da pesquisa. Em seguida, na terceira e quarta seção, são

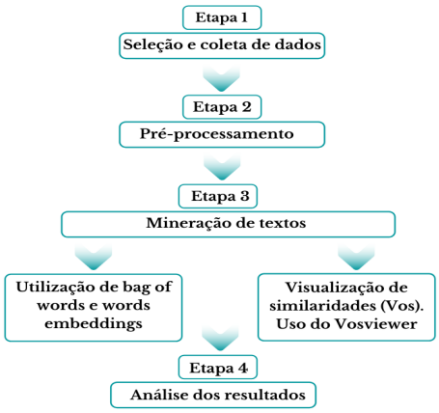
[5] http://Orange®.biolab.si/

apresentados e discutidos os resultados do estudo. Por fim, a última seção apresenta as considerações finais, limitações e recomendações para trabalhos futuros.

2. PROCEDIMENTOS METODOLÓGICOS

O estudo identificou os principais agrupamentos na base de teses e dissertações do PPGEGC (Programa de Pós-graduação em Engenharia e Gestão do Conhecimento) no período de 2006 a 2020. Para isso, coletou-se os dados da base de teses e dissertações do PPGEGC (http://btd.egc.ufsc.br/). O trabalho seguiu 4 etapas: (Figura 1): 1) seleção e coleta de dados; 2) pré-processamento; 3) mineração de textos; e 4) análise dos resultados.

Figura 1. Etapas da pesquisa.



Fonte: elaborado pelos autores (2021)

Na etapa 1, as teses e dissertações foram extraídas do BTD-EGC, totalizando 609 documentos no período de 2006 a agosto de 2020. Do total de documentos extraídos, 293 são teses e 316 dissertações defendidas no PPGEGC. Os documentos foram obtidos através do software *Octoparse*, utilizado para extração de dados visuais da web. O software utiliza a técnica *web scraping* que extrai dados de sites e transforma os dados não estruturados em formatos estruturados que podem ser armazenados em computadores pessoais ou na plataforma de nuvem. Os dados foram armazenados em uma planilha de cálculo, contendo a seguinte estrutura: autor, título, ano, resumo, palavras-chave, tipo (tese ou dissertação) e área do programa (EC, GC, MC) que serviram como *Corpus* para este estudo.

Na etapa 2, os dados foram pré-processados utilizando técnicas de mineração de texto para a limpeza desses dados: *Transformation*, *Tokenization*, *Filtering* e *N-grams Range*. A opção *Transformation*, removeu os acentos e todo texto foi transformado em letras minúsculas. Na *Tokenization* o texto foi dividido por palavras apenas por padrão, omitindo-se a pontuação.

O *Filtering* removeu as palavras irrelevantes utilizando o idioma português, o que resultou na remoção das palavras: “ano”, “crescente”, “numero”, “publicacoes” e “realizadas”. Após a aplicação dos filtros nos 609 documentos do BTD-EGC foram contabilizados 103.519 *tokens*. Já no *N-Grams*, o intervalo configurado foi de 1 a 2 *tokens*.

Na Etapa 3 utilizou-se três técnicas de mineração de texto *bag of words* e *word embeddings* utilizados com o auxílio da ferramenta Orange®® e visualização de similaridades com a ferramenta *VOSviewer*®®. No *bag of words* foram considerados como parâmetro a contagem do número de ocorrências de uma palavra no documento, enquanto que no *document embedding* é realizado o mapeamento de palavras em espaços vetoriais numéricos, onde utilizamos a configuração padrão do Orange®®.

Os passos a seguir foram realizados para as duas técnicas utilizadas: cálculo da distância e agrupamento hierárquico. O cálculo da distância foi realizado pelas colunas e a métrica utilizada foi o cosseno, isso resultou numa matriz de distâncias que foi utilizada para gerar o agrupamento hierárquico com a utilização do método de ligação *Ward*, para calcular as distâncias entre os agrupamentos.

Na visualização de similaridades utilizou-se o *VOSviewer*® - uma ferramenta de software que possibilita a criação e visualização de mapas bibliométricos com base em dados de rede (Eck & Waltman, 2010). Ele utiliza uma técnica de mapeamento com base na visualização de similaridades (VOS). A ferramenta *VOSviewer*®® constrói mapas bibliométricos de várias maneiras de modo a enfatizar diferentes aspectos da produção da literatura. O *VOSviewer*®® usa uma abordagem unificada para mapeamento e agrupamento e é baseado no coeficiente da matriz de ocorrência normalizada e uma medida de similaridade que calcula a força de associação entre os termos (van Eck & Waltman, 2013). Os termos relacionados são estruturados em *clusters*, agrupados pela mesma cor. A proximidade dos termos pode ser interpretada como uma indicação da semelhança do contexto em que ocorrem (Vosner, Kokol, Bobek, Zeleznik, & Zavrsni, 2016).

Por fim, na Etapa 4 será realizada a análise dos resultados através de nuvem de palavras e tabela de dados, visando auxiliar na comparação entre os resultados gerados com a utilização das técnicas de *bag of words, word embeddings* e visualização de similaridades.

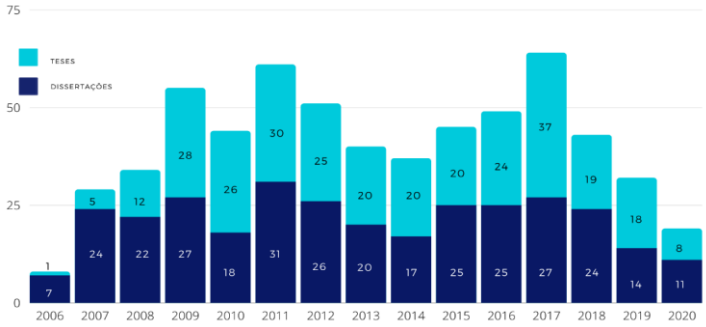
3. RESULTADOS E DISCUSSÃO

O trabalho teve como propósito realizar um estudo bibliométrico, do tipo mapeamento científico, dos trabalhos do PPGEGC, utilizando técnicas de mineração de texto. Neste sentido,

realizou-se a coleta de teses e dissertações publicadas no site do PPGEGC até agosto de 2020.

As técnicas de mineração de texto foram aplicadas utilizando o resumo dos documentos para a obtenção dos resultados, de acordo com a metodologia descrita na seção anterior. Os documentos estão divididos em três áreas de concentração do Programa de Pós-Graduação (Engenharia do Conhecimento, Gestão do Conhecimento e Mídia do Conhecimento) e possui dois tipos de documentos (teses de doutorado e dissertações de mestrado). Com este *corpus*, chega-se a um total de 609 documentos obtidos de 2006 até agosto de 2020, do qual 316 são dissertações e 293 teses e um total de 60 professores orientadores.

Figura 2. Distribuição das teses e dissertações de 2006 a 2020.

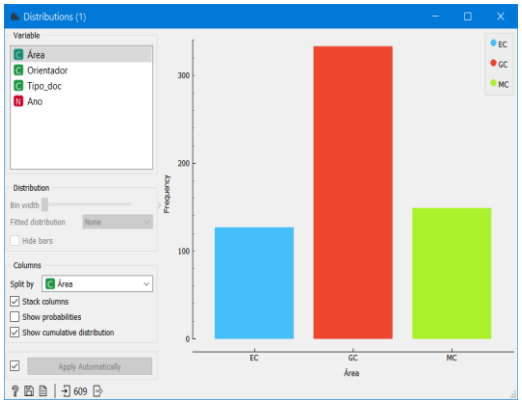


Fonte: Elaborado pelos autores (2021)

Na Figura 3 pode ser vista a distribuição dos documentos quanto às áreas de concentração: Engenharia do Conhecimento n=127 (20,85%) documentos, Gestão do Conhecimento n=333 (54,68%) documentos e a Mídia do Conhecimento n=149 (24,47%) documentos.

A partir dos documentos coletados, os agrupamentos foram gerados através do Orange®® com a utilização de duas abordagens: *bag of words* e *word embeddings*, para posterior comparação com a análise bibliométrica.

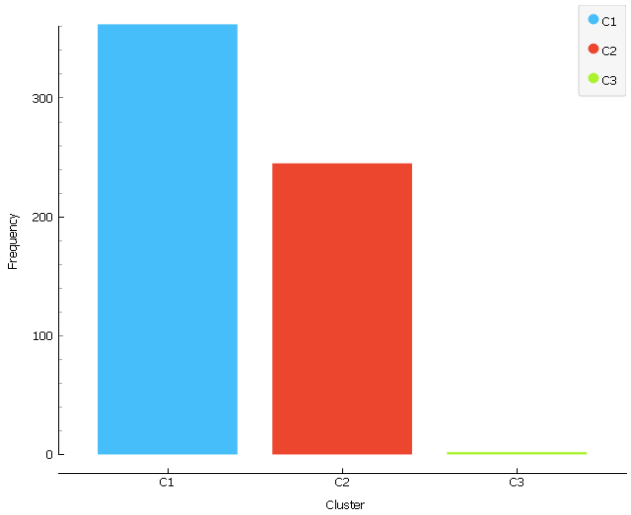
Figura 3. Distribuição por áreas de concentração.



Fonte: Elaborado pelo autores por meio do *Orange*® (2021)

A seguir, o cálculo das distâncias foi realizado utilizando as linhas do conjunto de dados e a distância métrica usada foi o cosseno que resultou em uma matriz de distância. Por fim, utilizou-se o agrupamento hierárquico para visualizar os objetos da matriz de distância através de um dendrograma com o método de ligação *Ward*, para calcular as distâncias entre os agrupamentos. Na Figura 4 podemos observar o resultado do cluster após a análise do dendrograma.

Figura 4. Distribuição dos agrupamentos com *bow*.



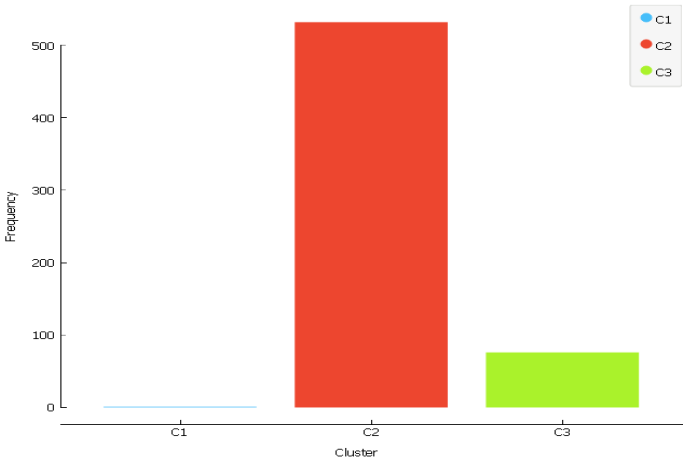
Fonte: Elaborado pelos autores por meio do Orange® (2021)

Como pode ser visto na Figura 4, o resultado com base na matriz de distância foi de 3 agrupamentos representados pelas cores azul, vermelho e verde, com o corte feito na maior distância entre os agrupamentos em 0,0005. O agrupamento azul possui 362, o agrupamento vermelho resultou em 245 (40,23%) documentos e o agrupamento verde possui 2 documentos (0,33%).

Na abordagem *word embeddings* analisa-se as palavras de cada documento no *corpus*, obtém *embedding* para cada palavra usando um modelo pré-treinado para o idioma escolhido e obtém-se um vetor para cada documento agregando palavras (ORANGE®, 2020).

Os cálculos das distâncias foram os mesmos utilizados no *bow*, o cálculo das distâncias feito por colunas e a distância métrica usada foi o cosseno. Também utilizou-se o agrupamento hierárquico para visualizar os objetos da matriz de distância através de um dendrograma (Figura 5) com o método de ligação *Ward*.

Figura 5. Distribuição dos agrupamentos com *embeddings*.

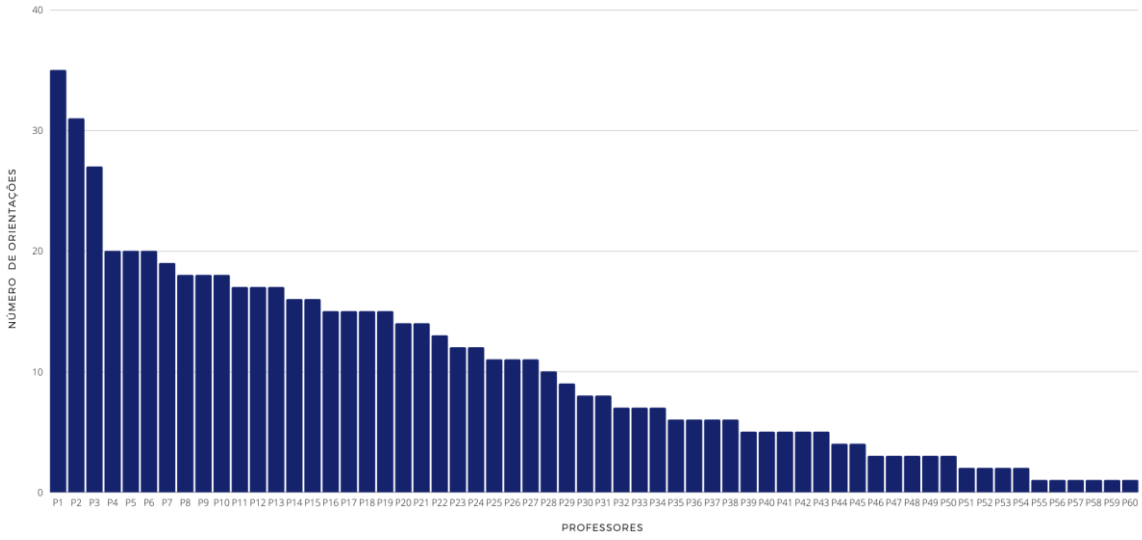


Fonte: Elaborado pelos autores por meio do Orange® (2021)

A Figura 5 apresenta a distribuição dos agrupamentos utilizando *embeddings*, onde verificou-se, também, 3 agrupamentos, representados pelas cores, verde, vermelho e azul, com o corte feito na maior distância entre os agrupamentos. O agrupamento azul possui apenas 1 documento (0,16%) quase imperceptível na figura, o agrupamento vermelho resultou em 532 (87,36%) documentos e o agrupamento verde com 76 documentos (12,48%).

A Figura 6 traz a distribuição do número de orientações realizadas no período da pesquisa. Omitiu-se os nomes dos professores orientadores, indicando rótulos de P1 até P60, que os caracterizaram. P1 representa o professor orientador com maior número de orientações e P60 o professor com o menor número de orientações. O professor P1 realizou um total de 35 orientações no período, sendo 14 dissertações de mestrado e 21 teses de doutorado.

Figura 6. Quantidade de teses e dissertações orientados por professor do programa



Fonte: elaborado pelos autores (2021)

A próxima análise considerou os principais assuntos abordados pelos professores com maior número de orientações de dissertações (quadro 1) e teses (quadro 2) no período de 2016 a agosto de 2020.

Quadro 1. Dez palavras mais frequentes nas dissertações dos professores com maior número de orientações de 2016 a 2020.

	P2	P17	P13	P27	P18	P12	P7	P11	P8	P10
1	social	conhecimento	ideias	livro	comunicacao	design	dados	gestao	gestao	midia
2	inovacao	projetos	processo	digital	gastronomica	jogos	conhecimento	setores	produtoras	publico
3	pesquisa	auditoria	selecao	livro didatico	florianopolis	desenvolvimento	taxonomias	administracao	unidades	publico alvo
4	competencias	modelo	organizacoes	didatico	gestao	gdd	metodo	maturidade	unidades produtoras	alvo
5	inovacao social	auditoria conhecimento	pesquisa	didatico digital	pesquisa	conhecimento	abertos	gc	maturidade	surdos
6	empreendedores	gestao	criterios	web	rede	educativos	dados abertos	convenios	cidades	prototipo
7	sociais	gestao conhecimento	inovacao	caracteristicas	unesco	industria	pesquisa	grau	modelos	hipervideo
8	empreendedoras	modelos	forma	pesquisa	acoes	conhecimentos	construcao	grau maturidade	nivel	solucoes
9	competencias empreendedoras	ideias	selecao ideias	avaliacao	area	jogo	implementacao	municipal	avaliacao	linguagem
10	inovacoes	processo	analise	instrumento	estudo	digitais	data	organizacao	meio	diferentes

Fonte: os autores (2021)

Em seguida, realizou-se a mesma análise considerando os temas por área de concentração. Foram extraídas os termos de dissertações e teses por área de conhecimento no período de 2016 a agosto de 2020. A coleta dos dados ocorreu em agosto de 2020. Logo, algumas teses e dissertações incluídas após esta data no site/base do PPGEGC não foram contempladas neste estudo. Foram analisados os últimos 5 anos de pesquisa no PPGEGC a fim de identificar o que se está pesquisando no programa nas áreas de concentração.

Quadro 2. Dez termos mais frequentes nas teses dos professores com maior número de orientações de 2016 a 2020.

	P1	P13	P21	P5	P10	P4	P25	P11	P2	P17	P3	P27	P28
1	inovacao	inovacao	rea	teoria	alunos	portais	ideias	internacionalizacao	conhecimento	inovacao	ca	livro	web pragmatica
2	inovacao aberta	mtf is	educacao	equipe	aluno	mecanismos	modelo	superior	criacao	parcerias	praticas	digital	web
3	aberta	mtf	pesquisa	funcao	professores	usuarios	avaliacao	ensino	estrategias	social	gc	livro didatico	verificar consistencia
4	capacidades dinamicas	is	participantes	mulheres	pesquisa	gc	contexto	ensino superior	dados	inovacao social	praticas gc	didatico	verificar
5	dinamicas	universidade	recursos	praticas	visuais	gestao	organizacional	estrategias	criacao conhecimento	intersetoriais	organizacoes	didatico digital	valores
6	capacidades	adocao	processo	fatores	deficientes	framework	contexto organizacional	instituicoes	bibliografica	parcerias intersetoriais	desempenho	web	urbano baseado
7	modelo	adocao mtf	modelo	complexa	deficientes visuais	realizada	indice	ies	meio	dinamica	gestao	caracteristicas	urbano
8	capital	empreendedorismo	estudo	barreiras	afetividade	judiciario	cenario	instituicoes ensino	processo	iniciativas	estudo	pesquisa	unem agentes
9	capital intelectual	processo	conhecimento	organizacoes	professor	analise	potencial	internacionalizacao instituicoes	revisao	dinamica parcerias	influencia	avaliacao	unem
10	intelectual	inovacao empreendedorismo	brasil	software	processo	forma	potencial implementacao	analise	entrevistas	analise	potencial	instrumento	torna

Fonte: os autores (2021)

Quanto às áreas de concentração, extrai-se os documentos no período de 2016 a agosto de 2020. Nas dissertações, observa-se que a área da EC possui 10 documentos (18,52%) com um total de 3827 palavras. Os termos foram extraídos dos resumos dos documentos no período indicado. A área da GC possui 23 documentos (42,59%) e um total de 8600 termos. Já a área de MC possui 21 documentos (38,89%) e um total de 9458 termos.

As três áreas de concentração do PPGEGC, possuem visões de mundo e consequentemente objetivos diferentes. A Engenharia do Conhecimento tem como objetivo a "pesquisa e o desenvolvimento de métodos, técnicas e ferramentas para a construção de

modelos e sistemas de conhecimento em atividades intensivas em conhecimento" (PPGEGC, 2020).

Já a área de Gestão do Conhecimento (GC) "estuda as bases conceituais e metodológicas para implantação da gestão organizacional baseada no conhecimento. Portanto, visa à transformação dos conhecimentos individuais em conhecimentos coletivos e organizacionais. Por meio da visão autopoiética, os conteúdos ministrados e as pesquisas realizadas enfocam o conhecimento organizacional, a economia, e o trabalhador do conhecimento." (PPGEGC, 2020).

A área de "Mídia do Conhecimento" estuda "o compartilhamento e disseminação do conhecimento, desenvolvimento e avaliação das mídias voltadas a catalisar a habilidade de grupos para pensar, comunicar, disseminar, preservar, apreender e criar conhecimento. São abordadas as questões relacionadas à filosofia da ciência, à epistemologia e à sociologia da comunicação; aos processos de inclusão e inovação; às teorias da cognição; às técnicas e equipamentos de produção desse tipo de mensagens e às teorias que as estudam" (PPGEGC, 2020).

Na área da EC as palavras estão relacionadas com alguns temas como: gestão de ideias, seleção de ideias, descoberta de conhecimento, sistema de conhecimento, serviços de conhecimento, dados abertos e *front end* da inovação.

Os temas que se destacam na GC são: gestão do conhecimento, auditoria do conhecimento, maturidade em GC, inovação social, organizações inovadoras, *front end* da inovação, competências empreendedoras e incubadoras sociais.

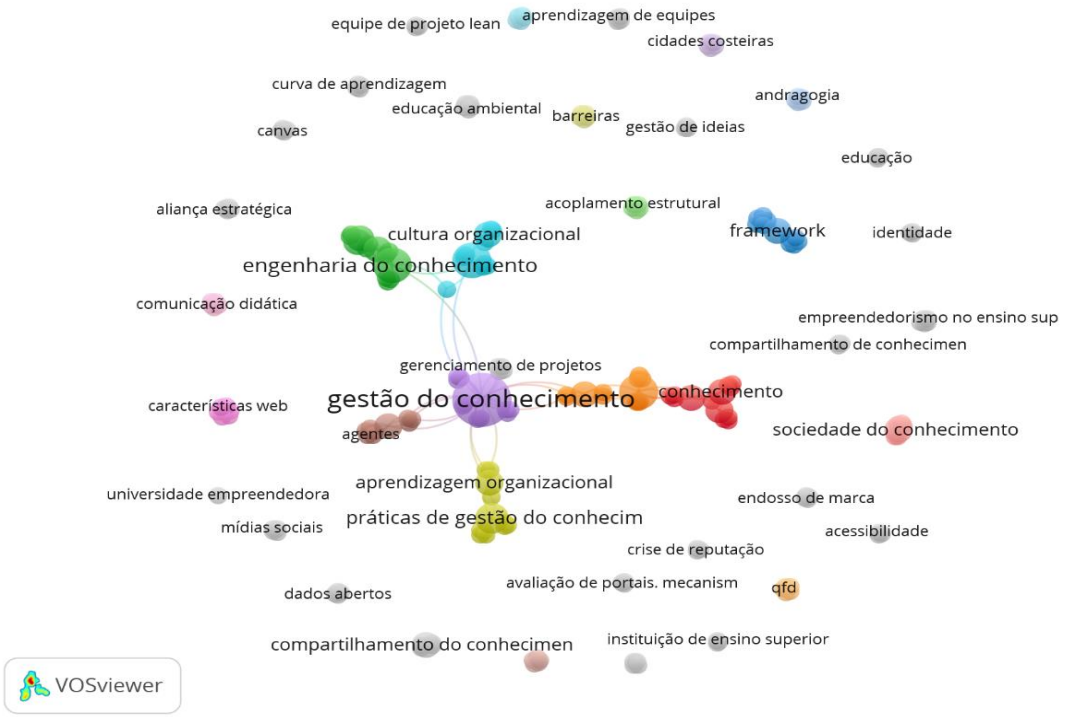
Já na área de MC alguns temas relacionados com as palavras encontradas nos documentos: mídia do conhecimento, visualização do conhecimento, disseminação do conhecimento, mídia social, inovação social, rede social, comunicação organizacional, marca de gestão e comunicação da marca.

Em uma etapa seguinte, os dados coletados foram analisados como o apoio da ferramenta *VOSviewer*®. A funcionalidade de mineração de texto do *VOSviewer*® fornece suporte para a criação de mapas de termos com base em um *corpus* de documentos. Seu mapa é bidimensional e os termos são localizados de forma que a distância entre eles pode ser interpretada como uma indicação da relação entre esses termos. Em geral, quanto menor a distância entre os termos, mais fortes as relações entre eles. As relações entre os termos é determinada com base em coocorrências em documentos (Van Eck & Waltman, 2010)

A análise de coocorrência de palavras-chave tem por fundamento a suposição de que quando dois itens aparecem em um mesmo contexto, eles estão relacionados de certa forma. A

coocorrência tende a ser empregada para explorar mudanças de temas em um campo de pesquisa, como forma de medir a frequência da ocorrência de itens (ou seja, palavras ou sintagmas nominais) em um determinado corpo de literatura em um campo selecionado. A coocorrência também possibilita a determinação das disciplinas que estão envolvidas na construção da base intelectual de um campo de pesquisa específico (Liu et al., 2015).

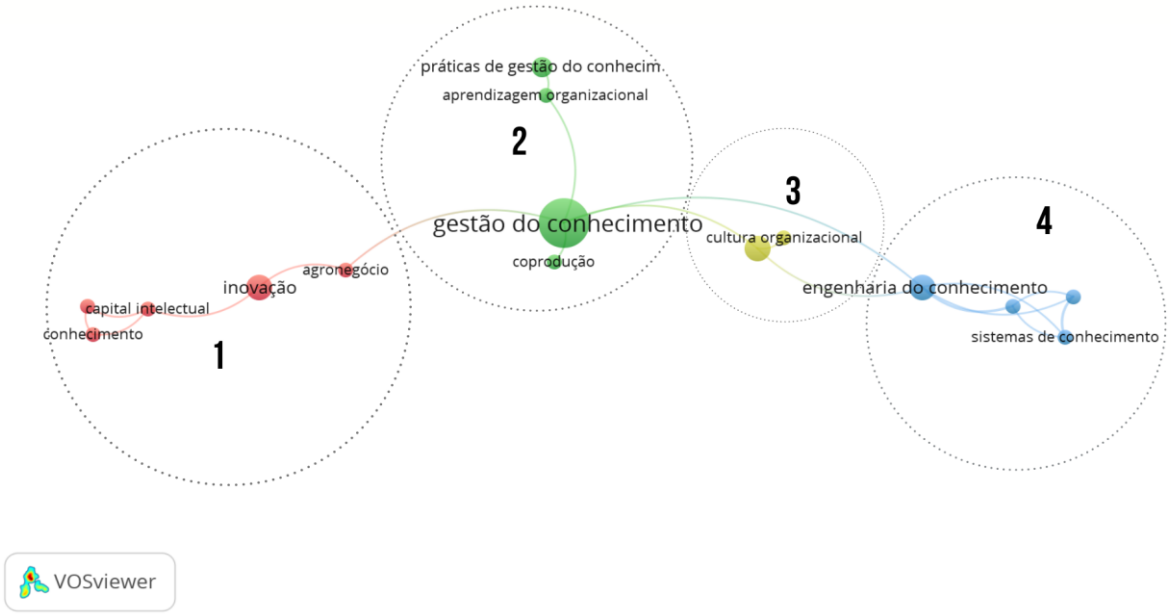
Figura 7. Mapa de palavras-chave



Fonte: elaborado pelos autores utilizando a ferramenta *VOSviewer*® (2021)

O mapa apresentado na Figura 7 mostra a rede de palavras-chave com maior ocorrência com base nos dados analisados. A rede está subdividida em quatro agrupamentos (*clusters*), circundados por outros termos que não apresentaram ligação, e por isso não foram aglutinados em algum dos agrupamentos *(clusters)* criados. Considerou-se como condição a ocorrência mínima de duas palavras no conjunto dos documentos. Dezoito palavras preencheram esse requisito (*threshold*), conforme verifica-se na figura 8.

Figura 8. Mapa de palavras-chave 2



Fonte: elaborado pelos autores utilizando a ferramenta *VOSviewer*® (2021)

Cada agrupamento *(cluster*) agrega as palavras-chave que apresentam similaridades e interações entre temas correlatos. Por exemplo, um conjunto de trabalhos X aborda um tema específico; esses documentos apresentam ligações estruturais que se dão por linhas de ligação. Quanto mais forte for essa linha, maior a interação entre esses temas. A ligação acontece porque um tema pode derivar de outro ou porque mesmos documentos abordam temas diferentes.

O agrupamento (*cluster)* 1 (figura 8), de cor vermelha, agrega palavras-chave relacionadas com capital intelectual, conhecimento, inovação e perda do conhecimento. O agrupamento *(cluster* 2), de cor verde, é o principal cluster do mapa disponibilizado pelo *VOSviewer*®. Isso é verificado pelo tamanho e a centralidade do nó, que aglutina as palavras-chave gestão do conhecimento, coprodução, aprendizagem organizacional e práticas de gestão do conhecimento. O agrupamento *(cluster)* 3, na cor amarela, aglutina palavras-chave referentes à inovação social e cultura organizacional. Por fim, o agrupamento (*cluster)* 4, de cor azul, reúne palavras-chave associadas à engenharia de conhecimento, ontologias, web semântica e sistemas de conhecimento.

Quadro 3 - Frequência de palavras na análise de similaridades

Peso /Ocorrência	Força do *link*	Palavras
9	5	Gestão do conhecimento

4	6	Engenharia do conhecimento
4	3	Inovação Social
4	2	Inovação
4	3	Práticas de GC
2	4	Ontologias
2	4	Sistemas de conhecimento
2	4	Web Semântica
2	3	Capital intelectual
2	2	Agronegócio
2	2	Aprendizagem organizacional
2	2	Conhecimento
2	2	Perda do conhecimento
2	1	Coprodução
2	1	Cultura organizacional
2	0	Compartilhamento do conhecimento
2	0	Framework

Fonte: autores (2021) utilizando o *VOSviewer*®

A partir dos resultados, destaca-se alguns pontos sobre a pesquisa:

Em primeiro lugar, constatou-se que a área da Gestão do Conhecimento (GC) possui o maior número de teses e dissertações. A quantidade de dissertações da área de Mídia do Conhecimento (MC) aproxima-se dos números da GC, mas nas publicações de teses, GC supera o número de publicações das demais áreas.

Em segundo lugar, apesar do *bow* e *embeddings* serem técnicas diferentes da visualização de similaridades (VOS), para obtenção dos agrupamentos, as palavras encontradas, em ambas análises, são muito próximas.

Terceiro, verificou-se, por meio da mineração de texto ou na análise de similaridades, as áreas de pesquisa mais ativas do PPGEGC, no período de tempo selecionado, destacando-se as palavras por área de concentração.

Quarto, a análise de similaridades indica a pouca interação entre os campos de conhecimento, caracterizando a existência de “ilhas conceituais”. O resultado pode ter sido impactado pela forma de extração dos dados na análise de similaridades, que não considera os

trabalhos em sua integralidade, levando em conta, apenas os campos de resumo e palavras-chave das teses e dissertações analisadas.

Quinto, a análise possibilitou verificar quais os temas que despertaram maior interesse do programa em relação às pesquisas que estão sendo realizadas.

CONSIDERAÇÕES FINAIS

Os estudos bibliométricos possuem a particularidade de analisar o desempenho de um campo/área de pesquisa, mapeando os trabalhos mais relevantes, indicando autores, países, instituições e periódicos proeminentes. O presente trabalho, por se tratar de um estudo bibliométrico, mapeou os estudos desenvolvidos por um programa de pós-graduação de uma Universidade Federal brasileira. Foram utilizadas técnicas de mineração de textos e clusterização para analisar 609 trabalhos, entre teses e dissertações. O trabalho seguiu quatro etapas: seleção e coleta de dados; pré-processamento; mineração de textos e análise dos resultados.

Para tal, utilizou-se o banco de teses e dissertações do PPGEGC, evidenciando as palavras com maior frequência nos documentos, nos últimos cinco anos, utilizando os métodos *bow* e *embeddings*. Também se utilizou a análise de coocorrência de palavras para determinar os *clusters* com maior ocorrência no *corpus* analisado.

A coleta de dados e a preparação para o pré-processamento demandou grande esforço para extração do *corpus*, os dados, nem sempre seguiam um padrão. Por exemplo, a grande maioria dos documentos no site do PPGEGC, não possuíam as palavras-chave do documento, somente os documentos a partir de 2019 receberam esse campo.

Mesmo assim, os dados encontrados são de grande relevância para o PPGEGC, dando uma visão das pesquisas realizadas, principalmente nos últimos cinco anos. As ferramentas *Orange*® e *VOSviewer*®, foram primordiais para a obtenção dos resultados.

E ainda, em relação aos resultados obtidos, os métodos aplicados ajudaram a revelar as pesquisas que estão sendo realizadas e os orientadores mais proeminentes, em especial nos últimos cinco anos.

Porém, há muito o que extrair do *corpus*, sugere-se como trabalhos futuros a verificação da interação entre professores orientadores e coorientadores, a análise comparada dos trabalhos desenvolvidos pelo PPGEGC e outros programas similares, a comparação de produção entre professores de áreas do PPGEGC e de outros programas nacionais e internacionais, dentre

outras análises. E ainda, complementação do corpus com os últimos trabalhos desenvolvidos no EGC após agosto de 2020, data da última coleta de dados.

AGRADECIMENTOS

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.

REFERÊNCIAS

Abbas, A., Zhang, L., & Khan, S. U. (2014). A literature review on the state-of-the-art in patent analysis. World Patent Information, 37, 3–13. https://doi.org/10.1016/j.wpi.2013.12.006

Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., & Kochut, K. (2017). A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques. Undefined, 13. Retrieved from http://en.wikipedia.org/wiki/Statistics

Araújo, C. (2006) Bibliometria: evolução histórica e questões atuais. Em questão, 12(1), 11-3.

Boyack, K., &Klavans, R. (2010). Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately? Journal of the American Society for Information Science and Technology.

De Bellis, N. (2009). Bibliometrics and citation analysis: from the science citation index to cybermetrics. Plymouth (UK): Scarecrow Press.

Onan, A. (2019). Two-Stage Topic Extraction Model for Bibliometric Data Analysis Based on Word Embeddings and Clustering. IEEE Access, 7, 145614–145633. https://doi.org/10.1109/ACCESS.2019.2945911

PPGEGC. Programa de Pós-graduação em Engenharia e Gestão do Conhecimento, 2020. Página inicial. Disponível em: <https:// https://ppgegc.paginas.ufsc.br/>. Acesso em: 30 de set. de 2020.

Roemer, R.; Borchardt, R. (2015). Meaningful Metrics: A 21st Century Librarian's Guide to Bibliometrics, Altmetrics, and Research Impact. Chicago: Association of College & University Libraries.

Quevedo-Silva,F., Santos,E., Brandão,M. & Vils, L. (2016). Estudo bibliométrico: orientações sobre sua aplicação. REMark – Revista Brasileira de Marketing

Tseng, Y.-H., Lin, C.-J., & Lin, Y.-I. (2007). Text mining techniques for patent analysis. Information Processing & Management, 43(5), 1216–1247. https://doi.org/10.1016/j.ipm.2006.11.011

Van Eck, N.J., & Waltman, L. (2010). Software survey: VOSviewer®, a computer program for bibliometric mapping. Scientometrics, 84(2), 523-538.

Vosner, H., Kokol, P., Bobek, S., Zeleznik, D., & Zavrsni, J. (2016).