Articulo Original
vol. 1, nº 1
pp. 74-93 (2020)

Correlações entre grupos de pesquisa da Ciência da Informação no Brasil: uma abordagem baseada em palavras-chave

Correlations between Information Science research groups in Brazil: an approach based on keywords

Gisele de Felippe Schlogl (Departamento de Ciência da Informação, UFSC)
gischlogl@gmail.com - https://orcid.org/0000-0001-9657-6298

Moisés Lima Dutra (Departamento de Ciência da Informação, PGCIN/UFSC)
moises.dutra@ufsc.br - https://orcid.org/0000-0003-1000-5553

Resumo:
Analisar correlações entre grupos de pesquisa vem tendo um apelo crescente nos últimos anos. A identificação de proximidade entre diferentes projetos de pesquisa pode não apenas contribuir para desencadear novas parcerias, mas também para otimizar recursos e compartilhar resultados. No Brasil, o Sistema de Currículos Lattes do Conselho Nacional de Desenvolvimento Científico e Tecnológico é uma fonte rica de informações sobre a vida acadêmica e profissional de professores, pesquisadores e estudantes. Os currículos Lattes apresentam informações, boa parte delas atualizadas, em formato de texto semiestruturado. Este trabalho propõe identificar correlações entre grupos brasileiros de pesquisa em Ciência da Informação, por meio da análise de palavras-chave contidas nos resumos informativos e nas descrições dos projetos de pesquisa encontrados nos currículos Lattes dos participantes destes grupos. A análise apresentada a seguir foi feita com a aplicação de técnicas de mineração de texto nos currículos Lattes de pesquisadores vinculados a 27 programas de pós-graduação em Ciência da Informação de 24 instituições brasileiras de ensino superior, totalizando 399 currículos analisados. Entre os resultados obtidos, foi possível se identificar algumas tendências de pesquisa existentes entre os grupos e vinculá-las às áreas de Ciência da Informação, Arquivologia, Biblioteconomia e Museologia. Foi também possível se identificar os termos de pesquisas mais utilizados no momento. Além disso, a análise de ocorrência dos termos permitiu se identificar as áreas que concentram a maior parte da pesquisa em Ciência da Informação no Brasil, bem como perceber que existe uma propensão dos pesquisadores em utilizar certos termos para descrever suas pesquisas e seus resumos informativos.
Palavras-chave: Mineração de Texto; Currículo Lattes; Grupos de Pesquisa; Ciência da Informação.

Abstract:
Analyzing correlations between research groups has been increasingly appealing in recent years. The identification of proximity between different research projects can not only contribute to trigger new partnerships, but also to optimize resources and share results. In Brazil, the Lattes Curriculum System of the Brazilian National Council for Scientific and Technological Development is a rich source of information about the academic and professional life of professors, researchers, and students. Lattes curricula present information, much of it up-to-date, in a semi-structured text format. This paper intends to identify correlations between Brazilian research groups in Information Science through the analysis of keywords contained in the informative summaries and in the descriptions of the research projects found in the Lattes curricula of the participants of these groups. The analysis presented below was made with the application of text mining techniques to the Lattes curricula of researchers linked to 27 graduate programs in Information Science from 24 Brazilian institutions of higher education, totaling 399 curricula analyzed. Among the results obtained, it was possible to identify some existing research trends between the groups and link them to the areas of Information Science, Archivology, Library Science, and Museology. It was also possible to identify the most used research terms at the moment. In addition, the

AWARI: Revista de la Asociación Latinoamericana de Análisis de Redes Sociales
Presentado en: 20 de junio de 2020
Aceptado en: 02 de julio de 2020

analysis of the occurrence of the terms allowed to identify the areas that concentrate most of the research in Information Science in Brazil, as well as to realize that there is a propensity of researchers to use certain terms to describe their research and their informative summaries.
Keywords: Text Mining; Lattes Curricula; Research Groups; Information Science.

Resumen: Analizar correlaciones entre grupos de investigación ha sido cada vez más atractivo en los últimos años. La identificación de la proximidad entre diferentes proyectos de investigación no solo puede contribuir a desencadenar nuevas asociaciones, sino también a optimizar recursos y compartir resultados. En Brasil, el Sistema Curricular Lattes del Consejo Nacional Brasileño para el Desarrollo Científico y Tecnológico es una rica fuente de información sobre la vida académica y profesional de profesores, investigadores y estudiantes. Los currículos Lattes presentan información, gran parte actualizada, en formato de texto semiestructurado. Este trabajo pretende identificar las correlaciones entre los grupos de investigación brasileños en Ciencias de la Información a través del análisis de las palabras clave contenidas en los resúmenes informativos y en las descripciones de los proyectos de investigación encontrados en los currículos Lattes de los participantes de estos grupos. El análisis presentado a continuación se realizó con la aplicación de técnicas de minería de texto en los currículos Lattes de investigadores vinculados a 27 programas de posgrado en Ciencias de la Información de 24 instituciones brasileñas de educación superior, con un total de 399 currículos analizados. Entre los resultados obtenidos, fue posible identificar algunas tendencias de investigación existentes entre los grupos y vincularlas con las áreas de Ciencias de la Información, Archivología, Bibliotecología y Museología. También fue posible identificar los términos de investigación más utilizados en este momento. Además, el análisis de la aparición de los términos permitió identificar las áreas que concentran la mayor parte de la investigación en Ciencias de la Información en Brasil, así como darse cuenta de que existe una propensión de los investigadores a usar ciertos términos para describir su investigación y sus resúmenes informativos
Palabras clave: Minería de texto; Currículos Lattes; Grupos de investigación; Ciencias de la Información.

1 Introdução

O Sistema de Currículos Lattes foi desenvolvido pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) para ser uma base de dados de pesquisadores em Ciência e Tecnologia no Brasil. O Lattes é um sistema de informações integrado em que professores, pesquisadores e estudantes registram sua vida acadêmica e profissional (Lattes, 2007). O cadastro do currículo é obrigatório para docentes e pesquisadores do Brasil e deve ser constantemente atualizado. Além do registro e organização de informações, esta base de dados procura preservar a memória da pesquisa brasileira e é utilizada como base informativa para análise de mérito e competência por agências de fomentos e em processos seletivos. Pela sua credibilidade, abrangência e riqueza de informações, é adotado pela maioria das instituições brasileiras e atualmente é uma ferramenta de medição de análise e competência dos profissionais (Lattes, 2007).

Os currículos apresentam informações semiestruturadas de vínculo institucional, formação, atuação profissional, setor de atividade, campo do conhecimento, linhas de pesquisa, produção científica e tecnológica, entre outras, que compõem o diretório. Um dos principais campos do currículo Lattes, espécie de cartão de visita do profissional, é o resumo ou texto autoinformativo, que é descrito livremente pelo pesquisador. Sem possuir um limite específico de tamanho, o profissional define em poucas linhas neste resumo, conforme suas convicções, o que é melhor para o leitor que acessa seu currículo ficar conhecendo a respeito de sua vida profissional. Nele estão geralmente descritos de maneira bastante sucinta desde seu vínculo institucional e formação até as atividades e campos de conhecimento, de modo a realçar sua área de atuação e experiências vivenciadas.

O sistema Lattes é um grande manancial de informações relativas aos pesquisadores brasileiros. Porém, apesar disso, boa parte do que se pode extrair de informações e conhecimento não está explicitamente registrada nos currículos, já que cada um deles se limita tão somente a apresentar informações específicas de determinado pesquisador.

Este trabalho tem por objetivo analisar os currículos Lattes de pesquisadores de programas de pós-graduação em Ciência da Informação (CI), por meio da identificação e processamento de palavras-chave contidas nos resumos informativos e nas descrições dos projetos de pesquisa, com o intuito de se identificar correlações entre os grupos de pesquisa que abrigam estes profissionais. Para tal, são empregadas técnicas de mineração de texto e processamento de linguagem natural baseadas na sintaxe das palavras-chave descobertas. Não se trabalhou neste artigo com o componente semântico dos textos analisados. Ao final, espera-se contribuir para aumentar o corpus de conhecimento existente a respeito dos grupos de pesquisa em CI no Brasil, especialmente identificando os termos de pesquisa mais utilizados por estes grupos e as áreas potenciais às quais suas pesquisas estão relacionadas.

2 Mineração de Texto

A Mineração de Texto (em inglês: *Text Mining*, *Text Analytics* ou *Text Data Mining*) é um processo pelo qual se procura extrair informações de textos (Hearst, 1999; Feldman & Sanger, 2007; Rajman, 1998; Tan, 1999). Este processo busca por informações de alta qualidade, que se referem a uma combinação de relevância, novidade e interesse, e que são extraídas com o suporte da identificação de padrões e tendências, por meio da aplicação de métodos advindos da Estatística e da Inteligência Artificial, mais especificamente, do Aprendizado de Máquina ou *Machine Learning* (Machado *et al.*, 2010). A Mineração de Texto envolve a coleta automática ou semiautomática de dados textuais, seguida de sua estruturação, análise, adição e/ou retirada de elementos linguísticos, transformação e representação de texto como números, aplicação de técnicas estatísticas e de aprendizado de máquina, avaliação e interpretação dos resultados.

É considerada um campo multidisciplinar, que objetiva a extração de padrões, de conhecimento útil de dados em grandes quantidades de textos de linguagem natural, não-estruturados ou semiestruturados, que utiliza métodos para navegar, organizar, achar e descobrir informação em corpora textuais (Aranha, 2006). Para Hearst (1999), a mineração de texto objetiva a análise direta do texto para descobrir informações até então desconhecidas ou prover novas informações a partir dos dados encontrados em um conjunto de dados, separando o sinal do ruído, podendo levar a descoberta de informações quanto as respostas de perguntas cuja resposta ainda não é conhecida.

Uma atividade essencial a ser realizada num contexto de Mineração de Texto é o Processamento de Linguagem Natural (PLN). Sarkar (2016) define o PLN como um campo especializado de Ciência da Computação, Engenharia e Inteligência Artificial, com raízes na Linguística. As técnicas de PLN permitem que os computadores processem e entendam a linguagem humana natural e, desta forma, consigam extrair dela as informações de alta qualidade que buscamos, durante o processo de mineração textual. O PLN é a base da Mineração de Texto, é a técnica sem a qual o processamento de enormes quantidades de dados não-estruturados em tempo aceitável e com resultados profícuos seria humanamente impossível de ser realizado. Uma lista não-exaustiva de tarefas relacionadas ao PLN envolve, de maneira geral (Sarkar, 2016; Ingersoll, Morton & Farris, 2013; Weiss, Indurkhya & Zhang, 2010):

a) **Normalização do Texto:** Conversão para um formato de codificação padrão (por exemplo, UTF-8); "Limpeza" de caracteres inadequados e remoção de caracteres especiais; Expansão de contrações; Harmonização entre minúsculas e maiúsculas; Correção de caracteres repetidos; Correção de erros de digitação; Remoção de *stopwords*; Identificação das relações léxicas e semânticas da língua em questão (lexemas, morfemas, homônimos, homógrafos, homófonos, heterônimos, heterógrafos, termos polissêmicos, capitônimos, sinônimos, antônimos, hipônimos e hiperônimos; Radicalização dos termos; Lematização dos termos; "Tokenização" do texto.

b) **Compreensão do Texto:** Etiquetagem POS (*parts of speech*) dos *tokens*; Detecção de sentenças; Análise da sintaxe e da estrutura das sentenças; Modelagem de sentenças; Identificação da gramática de dependência; Construção do grafo de dependência; etc.

Entre todas estas técnicas, gostaríamos de destacar cinco que são relevantes para este trabalho. A tokenização divide o texto em palavras ou termos, denominados *tokens*, identificando nesse processo espaços em branco e pontuações que costumam delimitar os termos (Andrade, 2015). Cada *token* pode estar relacionado a mais de um termo. Para Neves (2013), o mecanismo percorre o texto e identifica cada termo entre os caracteres que, mesmo em menor unidade, podem ter significado se analisados de forma isolada. Segundo Andrade (2015), após a tokenização, outro mecanismo para o tratamento dos dados é a remoção de *stopwords*. *Stopwords* são termos que trazem pouca informação, tais como artigos, preposições, conjunções, bem como outras palavras auxiliares que não agregam valor ao texto. As palavras pouco frequentes são consideradas irrelevantes e excluídas do processo de análise, a fim de potencializar o processamento do texto (Neves, 2013).

A etiquetagem POS é a tarefa básica de rotular palavras de uma sentença com etiquetas morfossintáticas que as identificam como categorias gramaticais (substantivos, verbos, adjetivos, etc.) e podem, ainda, conter atributos refinados de cada categoria, por exemplo: gênero e número para um substantivo (Domingues, Faveiro & Medeiros, 2007). Outra técnica de mineração relacionada com a identificação de termos são os n-Gramas. Esta técnica que consiste em agrupar um conjunto de palavras que aparecem juntas com frequência no texto, com combinações diferentes (Trevisan, 2015). Ou seja, os n-gramas funcionam como termos compostos, que ampliam o escopo da mineração para além dos *tokens* simples. Finalmente, destacamos a técnica conhecida por Lematização. A lematização elimina as variações morfológicas e identifica o radical das palavras, retirando os prefixos, sufixos, número, grau e gênero e substituindo as palavras por suas formas canônicas (Madeira, 2015; Neves 2013). Essas formas canônicas podem ser um substantivo ou um verbo no infinitivo e representam conjuntos de palavras morfologicamente próximas.

3 Procedimentos Metodológicos

Este trabalho empreende uma análise por meio de mineração textual nos currículos Lattes de pesquisadores e professores que possuem vínculo permanente com programas de pós-graduação em Ciência da Informação de instituições de ensino superior no Brasil.

Estes grupos foram identificados na lista que consta na Plataforma Sucupira[1] da CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) de cursos avaliados e reconhecidos em instituições de ensino no país. Um total de 24 grupos foi identificado. Em três instituições, foram identificados mais de um programa de pós-graduação: Universidade de São Paulo (USP), Universidade Federal de Minas Gerais (UFMG) e Universidade Federal do Estado do Rio de Janeiro (UNIRIO), totalizando 27 programas de pós-graduação. A Tabela 1 apresenta todas as instituições identificadas na Plataforma Sucupira, com a correspondente quantidade de programas de pós-graduação das instituições, a nota de cada programa na última avaliação da CAPES e a quantidade de profissionais de vínculo permanente em cada programa.

Tabela 1: Instituições de Ensino e Programas de Pós-Graduação Analisados

Nome da IES	Sigla da IES	UF	Curso	Pós-Graduação			Quantidade de
			Sigla	ME	DO	MP	Currículos
FUNDAÇÃO CASA DE RUI BARBOSA	FCRB	RJ	PPGMA	-	-	3	18
FUNDAÇÃO UNIVERSIDADE FEDERAL DE SERGIPE	FUFSE	SE	PPGCI	-	-	3	11
UNIVERSIDADE DE BRASÍLIA	UNB	DF	PPGCINF	5	5	-	20
UNIVERSIDADE DE SÃO PAULO	USP	SP	PPGCI	4	4	-	18
UNIVERSIDADE DE SÃO PAULO	USP	SP	MPGCI	-	-	4	6
UNIVERSIDADE DO ESTADO DE SANTA CATARINA	UDESC	SC	PPGInfo	-	-	3	14
UNIVERSIDADE ESTADUAL DE LONDRINA	UEL	PR	PPGCI	4	4	-	12
UNIVERSIDADE ESTADUAL PAULISTA JÚLIO DE MESQUITA FILHO, MARÍLIA	UNESP-MAR	SP	POSCI	6	6	-	32
UNIVERSIDADE FEDERAL DA BAHIA	UFBA	BA	PPGCI	4	4	-	16
UNIVERSIDADE FEDERAL DA PARAÍBA, JOÃO PESSOA	UFPB-JP	PB	PPGCI	4	4	-	22
UNIVERSIDADE FEDERAL DE ALAGOAS	UFAL	AL	PPGCI	1	-	-	10
UNIVERSIDADE FEDERAL DE MINAS GERAIS	UFMG	MG	PPGDOC	5	5	-	15
UNIVERSIDADE FEDERAL DE MINAS GERAIS	UFMG	MG	PPGCI	5	5	-	16
UNIVERSIDADE FEDERAL DE PERNAMBUCO	UFPE	PE	PPGCI	4	4	-	13
UNIVERSIDADE FEDERAL DE SANTA CATARINA	UFSC	SC	PPGCIN	5	5	-	21
UNIVERSIDADE FEDERAL DE SÃO CARLOS	UFSCAR	SP	PPGCI	3	-	-	10
UNIVERSIDADE FEDERAL DO CARIRI	UFCA	CE	PPGB	-	-	3	18
UNIVERSIDADE FEDERAL DO CEARÁ	UFC	CE	PPGCI	3	-	-	11
UNIVERSIDADE FEDERAL DO ESPÍRITO SANTO	UFES	ES	PPGCI	1	-	-	9
UNIVERSIDADE FEDERAL DO ESTADO DO RIO DE JANEIRO	UNIRIO	RJ	PPGB	-	-	3	19
UNIVERSIDADE FEDERAL DO ESTADO DO RIO DE JANEIRO	UNIRIO	RJ	PPGARQ	-	-	3	10
UNIVERSIDADE FEDERAL DO PARÁ	UFPA	PA	PPGCI	3	-	-	8
UNIVERSIDADE FEDERAL DO RIO DE JANEIRO	UFRJ	RJ	PPGCI	4	4	-	20
UNIVERSIDADE FEDERAL DO RIO GRANDE DO NORTE	UFRN	RN	PPGIC	-	-	3	9
UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL	UFRGS	RS	PPGCIN	1	-	-	13
UNIVERSIDADE FEDERAL FLUMINENSE	UFF	RJ	PPGCI	4	4	-	18
UNIVERSIDADE FUMEC	FUMEC	MG	PPGSIGC	4	-	-	10
						Total de Currículos Analisados	399

Fonte: Elaborado pelos autores

A plataforma Lattes permite a extração individualizada dos currículos em formato XML. A coleta dos 399 arquivos XML foi realizada na plataforma entre 07/04/2020 e 29/04/2020. É importante ressaltar que alguns pesquisadores estão vinculados a mais de um programa. Por isso, os currículos que se encontravam duplicados em dois ou mais programas foram mantidos, de forma a preservar os dados referentes aos vínculos individuais com cada programa. Após a coleta, criou-se os corpora textuais que seriam utilizados na mineração. De cada arquivo XML, extraiu-se o texto do resumo autoinformativo do pesquisador, bem como todas as descrições de projeto de pesquisa registrados no currículo. Os pesquisadores registram estes projetos como coordenador ou como participante. Esta escolha foi feita porque tanto os resumos quanto as descrições de projeto de pesquisa são

[1] https://sucupira.capes.gov.br/sucupira/public/consultas/coleta/programa/quantitativos/quantitativoIes.jsf?areaAvaliacao=31&areaConhecimento=60700009

registrados como texto livre e são, portanto, dados não-estruturados. Para cada pesquisador, foi criado um arquivo CSV com o resumo do seu Lattes e outro CSV com as descrições dos seus projetos de pesquisa.

Posteriormente, os arquivos CSV foram organizados por programas/pastas. Também foram geradas versões gerais destes CSVs, ou seja, um CSV que concatenou todos os resumos dos 399 currículos e outro que concatenou todas as descrições de projetos de pesquisa. Esta estratégia permitiu realizar recortes analíticos por programa e pelo conjunto de todos os programas, de maneira a se obter um panorama geral da pesquisa em CI no Brasil. A seguir, foram aplicadas técnicas de higienização e normalização textual nos corpora gerados. Finalmente, executou-se os algoritmos de mineração textual (frequência de termos, identificação de n-gramas, lematização e etiquetagem POS).

Existem diversas iniciativas encontradas na literatura para efetuar cruzamentos abrangentes e personalizados com dados coletados na base Lattes (Mena-Chalco & Cesar-Jr, 2013; Dias, 2016; Dutra *et al.*, 2019). Este trabalho utilizou algoritmos proprietários, desenvolvidos na linguagem Python. Os grafos foram gerados na ferramenta *open-source* e gratuita Gephi. A quantidade de análises possível e resultados obtidos excedeu em muito o escopo deste artigo. Consequentemente, decidiu-se por apresentar a seguir um recorte das possibilidades existentes.

4 Resultados

Inicialmente, procurou-se identificar o termo mais frequente que ocorre no corpus geral (CG) dos 27 programas analisados. Sem surpresa, o termo “informação” é o termo mais frequente, com 8.283 ocorrências nos projetos de pesquisa e 9.386 nos resumos do CG. A seguir, o Quadro 1 apresenta os três n-gramas mais frequentes encontrados nas descrições dos projetos de pesquisa, para n=3 e n=4, por programa.

Quadro 1: n-gramas n=3 e n=4 mais frequentes em projetos de pesquisa

UNIVERSIDADE	N=3			N=4		
FCRB _ PPGMA	projeto de pesquisa	banco de dados	linha de pesquisa	casa de rui barbosa	revisão final dos textos	elaboração do inventário analítico
FUMEC _ PPGSIGC	captura de movimento	gestão do conhecimento	sistemas de informação	software de código aberto	movimento para a animação	projeto de natureza interdisciplinar
UDESC _ PPGINFO	ciência da informação	programa de extensão	projeto de pesquisa	didática e formação docente	ensino, pesquisa e extensão	modelo de comunicação científica
UEL _ PPGCI	ciência da informação	gestão da informação	organização da informação	universidade estadual de londrina	contruçao de vocábulários controlados	a consecução dos objetivos
UFAL _ PPGCI	ciência da informação	tecnologias de informação	gestão da informação	informação e do conhecimento	micro e pequenas empresas	o retrato da criminilidade
UFBA _ PPGCI	ciência da informação	difusão de conhecimento	sistemas de saúde	laboratório de tecnologias intelectuais	transferência de informações gerenciais	difusão de inovações gerenciais
UFC _ PPGCI	ciência da informação	prontuários do paciente	competência em informação	informação e do conhecimento	artigos científicos em eventos	unidade curricular de recursos
UFCA _ PPGB	ciência da informação	curso de biblioteconomia	arquitetura da informação	território local de atuação	arquitetura da informação pervasiva	unidade curricular de organização
UFES _ PPGCI	ciência da informação	competência em informação	curso de biblioteconomia	instituições de ensino superior	representação dos arquivos manuscritos	campo da competência leitora
UFF _ PPGCI	ciência da informação	gestão de documentos	análise de domínio	gestão eletrônica de documentos	imagens de lâminas histopatológicas	eletrônica de documentos ged
UFMG _ PPGCI	ciência da informação	organização da informação	gestão de documentos	arquivo público da cidade	divisão de coleções especiais	memória intelectual da ufmg
UFMG _ PPGGOC	ciência da informação	recuperação da informação	projeto de pesquisa	informação e do conhecimento	recursos de computação gráfica	projeto de inclusão digital
UFPA _ PPGCI	ciência da informação	projetos de extensão	projeto de pesquisa	poder nas estruturas informacionais	rede transamazônica de cooperação	informação e a comunidade
UFPB _ PPGCI	ciência da informação	gestão da informação	arquitetura da informação	laboratório de tecnologias intelectuais	informação científica e tecnológica	acesso livre à informação
UFPE _ PPGCI	ciência da informação	produção de indicadores	gestão da informação	indicadores científicos e tecnológicos	capacidade do laboratório multiusuário	produção científica e tecnológica
UFRGS _ PPGCIN	ciência da informação	dados de pesquisa	direito à informação	acervo digital da pesquisa	documentação e acervo digital	promoção do acesso aberto
UFRJ _ PPGCI	ciência da informação	dados de pesquisa	projeto de pesquisa	competência crítica em informação	povos e comunidades tradicionais	integração pragmática de dados
UFRN _ PPGIC	ciência da informação	gestão da informação	curso de biblioteconomia	núcleo temático da seca	informação e do conhecimento	organização internacional do trabalho
UFSC _ PGCIN	ciência da informação	revisão por pares	banco de dados	estabelecimento de relações semânticas	tratamento temático da informação	profissional de informação bibliotecário
UFSCAR _ PPGCI	ciência da informação	análise de indicadores	política de indexação	produção científica e tecnológica	tratamento temático da informação	indicadores sobre a produção
UFSE _ PPGCI	ciência da informação	mediação da informação	unidades de informação	história do português brasileiro	informação em bibliotecas universitárias	curso subsequente de manutenção
UNB _ PPGCINF	ciência da informação	gestão da informação	gestão do conhecimento	programa de gestão documental	análise de redes sociais	programa de gestão documental
UNESP MAR _ POSCI	ciência da informação	organzação do conhecimento	gestão do conhecimento	textos narrativos de ficção	tratamento temático da informação	redes de bibliotecas escolares
UNIRIO _ PPGARQ	documentos de arquivo	gestão de documentos	curso de arquivologia	tratamento técnico e arquivístico	informações no arquivo virtual	arquivologia e análise social
UNIRIO _ PPGB	organização do conhecimento	ciência da informação	sistemas de organização	projeto de iniciação científica	ditames legais da acessibilidade	bibliotecas públicas no Brasil
USP _ PPGCI	ciência da informação	organização da informação	recuperação da informação	rede transamazônica de cooperação	avaliação da produção científica	rede de colaboração científica

Fonte: Dados da pesquisa

Observa-se que os termos n=3 apresentam mais recorrência entre os programas e apontam muitos campos de estudo na Ciência da Informação. Os termos n=4 apresentam mais variedade e tendem a ser mais determinantes para a identificação de campos mais específicas de estudo, sendo possível identificar não apenas a grande área da Ciência da Informação, Arquivologia, Biblioteconomia, mas também outras subáreas associadas à CI como Gestão do Conhecimento, Gestão da Informação, Arquitetura da Informação, entre outras.

O Quadro 2 repete a análise acima para n=5 e n=6. É interessante se observar o maior detalhamento dos n-gramas encontrados aqui, que se apresentam quase como um funil de conteúdo. É possível especular que essas descobertas possam servir para motivar o interesse de pesquisadores por um determinado programa de pós-graduação, em razão dos resultados por ele apresentados.

Quadro 2: n-gramas n=5 e n=6 mais frequentes em projetos de pesquisa

UNIVERSIDADE	N=5		N=6	
FCRB _ PPGMA	redação da introdução e revisão	banco de dados da fundação	redação da introdução e revisão final	elaboração do inventário analítico do arquivo
FUMEC _ PPGSIGC	sistema de captura de movimento	gestão do conhecimento e inovação	captura de movimento para a animação	gestão da informação e do conhecimento
UDESC _ PPGINFO	contribuir nas discussões da didática	reflexão sobre as contribuições teórica	discussões da didática e formação docente	grupo de pesquisa didática e formação
UEL _ PPGCI	mercado de trabalho em informação	trabalho em informação e documentação	organização da informação e do conhecimento	gestão da informação da produção intelectual
UFAL _ PPGCI	tecnologias da informação e comunicação	uso das tecnologias da informação	gestão da informação e do conhecimento	aquisição de informação e de conhecimento
UFBA _ PPGCI	mecanismos de difusão do conhecimento	arquivos e repositórios em saúde	documentos arquivos e repositórios em saúde	difusçao do conhecimento para as inovações
UFC _ PPGCI	recursos e serviços de informação	apresentação e publicação de artigos	apresentação e publicações de artigos científicos	gestão e visibilidade da informação científica
UFCA _ PPGB	biblioteconomia e ciência da informação	organização e tratamento da informação	unidade curricular de organização e tratamento	empresa junior do curso de biblioteconomia
UFES _ PPGCI	conectada por redes de colaboração	programa de competência em informação	**sociedade conectada por redes de colaboração**	implantar ações culturais em instituições públicas
UFF _ PPGCI	cursosde graduação em arquivologia	grupos de trabalho dos enancibs	método para alcance dos objetivos	denominados trajetos da produção intelectual nacional
UFMG _ PPGCI	escola de ciência da informação	memória dos projetos de extensão	direitos da criança e do adolescente	gestão da informação e do conhecimento
UFMG _ PPGGOC	organização e tratamento da informação	biblioteconomia e ciência da informação	produtos para a organização da informação	ferramentas automatizadas para gestão de planilhas
UFPA _ PPGCI	microfísica do poder nas estruturas	cooperaão em informação e conhecimento	crítica sobre a tecnologia da informação	informação e conhecimento para o desenvolvimento
UFPB _ PPGCI	acesso livre a informação científica	departamento de ciência da informação	gestão da informação e do conhecimento	facilitem o acesso livre à informação
UFPE _ PPGCI	área de ciência da informação	pesquisa em ciência da informação	ampliação da capacidade do laboratório multiusuário	produção científica da ciência da informação
UFRGS _ PPGCIN	busca e uso da informação	centro de documentação e acervo	centro de documentação e acervo digital	promoção do acesso aberto a dados
UFRJ _ PPGCI	instituições de ensino e pesquisa	tecnologias de informação e comunicação	modelo de análise de periódicos eletrônicos	dados de pesquisa em formatos digitais
UFRN _ PPGIC	tecnologias de informação e comunicação	departamento de ciência da informação	gestão da informação e do conhecimento	acessibilidade e inclusão oferecidas aos estudantes
UFSC _ PGCIN	sistemas de organização do conhecimento	revisão pelos pares na aprendizagem	relações semânticas em sistemas de organização	estudos e pesquisas em competência informacional
UFSCAR _ PPGCI	biblioteconomia e ciência da informação	indicadores sobre a produção científica	compreensão, elaboração e análise de indicadores	análise bibliométrica para apoio à gestão
UFSE _ PPGCI	mediação da informação em bibliotecas	curso de bibliotconomia e documentação	mediação da informaç˜eo em bibliotecas universitárias	dimensão estética da mediação da informação
UNB _ PPGCINF	segurança da informação e comunicação	tabela de temporalidade de documentos	sistema informatizado de gestão de documentos	bibliotecas, museus e centros de documentação
UNESP MAR _ POSCI	tecnologias de informação e comunicação	organização e representação do conhecimento	ciências da informação e da documentação	fatores críticos de sucesso da gestão
UNIRIO _ PPGARQ	gestão de documentos e arquivos	classificação de documentos de arquivo	inserção das informações no arquivo virtual	análise social, categorias, conceitos e classificações
UNIRIO _ PPGB	sistemas de organização do conhecimento	curso de engenharia de produção	desenvolver as competências exigidas pela ciência	prática numa perspectiva crítica a integração
USP _ PPGCI	tecnologias da informação e comunicação	mercado de trabalho em informação	estudos sobre meio ambiente e sustentabilidade	tecnologias da informação e comunicação TICS

Fonte: Dados da pesquisa

A análise dos n-gramas mais frequentes encontrados nos resumos dos currículos apresentou resultados muito similares aos ilustrados acima. Desta forma, decidiu-se por não especificar estes resultados por programa, tal como foi feito anteriormente. O Quadro 3 apresenta uma síntese desta análise para o CG.

Quadro 3: n-gramas mais frequentes em resumos do Corpus Geral

nGramas Resumos dos Currículos			
n=3	ciencia da informação	organização do conhecimento	gestão do conhecimento
n=4	meio ambiente e sustentabilidade	laboratório de tecnologias intelectuais	textos narrativos de ficcção
n=5	sistemas de organização do conhecimento	biblioteconomia e ciência da informação	escola de ciência da informação
n=6	o acesso livre a informação científica	estudos sobre meio ambiente e esustentabilidade	centro de documentação e acervo digital

Fonte: Dados da pesquisa

4.1 Frequência de Adjetivos

A partir da aplicação da etiquetagem POS, foi possível se identificar os adjetivos existentes nos corpora analisados. Para além da simples curiosidade, esta identificação permitiu observar o uso de adjetivos no contexto dos projetos de pesquisa e resumos, na tentativa de se encontrar possíveis correlações entre as áreas. O Quadro 4 apresenta os cinco adjetivos mais frequentes encontrados nas descrições dos projetos de pesquisa, agrupados por categorias *ad-hoc*, para facilitar a visualização.

Quadro 4: Adjetivos mais utilizados em projetos de pesquisa

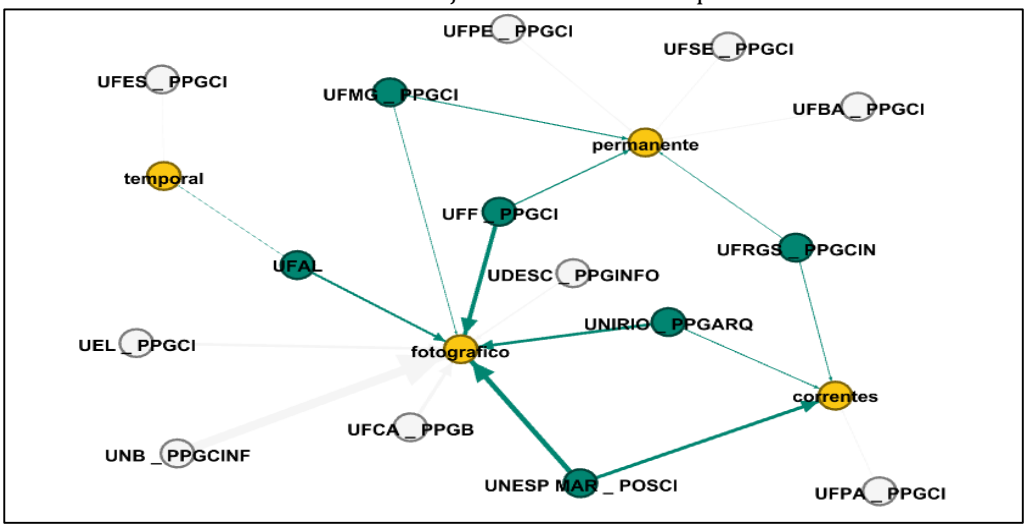
Programa	Relativo a Pesquisa	Representação da Informação		Produção Documental		Tecnologias
FCRB _ PPGMA	científico	digital	virtual	documental	histórico	ciência
FUMEC _ PPGSIGC	científico	digital	virtual	documental	histórico	tecnológico
UDESC _ PPGINFO	científico	digital	eletrônico	documental	histórico	tecnológico
UEL _ PPGCI	científico	digital	eletrônico	documental	histórico	tecnológico
UFAL _ PPGCI	científico	digital	eletrônico	documental	temporal	tecnológico
UFBA _ PPGCI	científico	digital	virtual	documental	histórico	tecnológico
UFC _ PPGCI	científico	digital	eletrônico	documental	histórico	tecnológico
UFCA _ PPGB	científico	digital	eletrônico	documental	histórico	ciência
UFES _ PPGCI	científico	digital	virtual	documental	temporal	tecnológico
UFF _ PPGCI	científico	digital	eletrônico	documental	histórico	tecnológico
UFMG _ PPGCI	científico	digital	eletrônico	documental	histórico	tecnológico
UFMG _ PPGGOC	científico	digital	eletrônico	documental	histórico	tecnológico
UFPA _ PPGCI	científico	digital	eletrônico	documental	histórico	tecnológico
UFPB _ PPGCI	científico	digital	eletrônico	documental	histórico	tecnológico
UFPE _ PPGCI	científico	digital	virtual	documental	permanente	tecnológico
UFRGS _ PPGCIN	científico	digital	eletrônico	documental	histórico	tecnológico
UFRJ _ PPGCI	científico	digital	eletrônico	documental	histórico	tecnológico
UFRN _ PPGIC	científico	digital	eletrônico	documental	histórico	tecnológico
UFSC _ PGCIN	científico	digital	eletrônico	documental	histórico	tecnológico
UFSCAR _ PPGCI	científico	digital	virtual	documental	histórico	tecnológico
UFSE _ PPGCI	científico	digital	virtual	documental	histórico	tecnológico
UNB _ PPGCINF	científico	digital	eletrônico	documental	histórico	tecnológico
UNESP MAR _ POS	científico	digital	eletrônico	documental	histórico	ciência
UNIRIO _ PPGARQ	científico	digital	virtual	documental	histórico	tecnológico
UNIRIO _ PPGB	científico	digital	eletrônico	documental	histórico	tecnológico
USP _ PPGCI	científico	digital	eletrônico	documental	histórico	tecnológico

Fonte: Dados da pesquisa

Os termos que apresentaram maior recorrência são os seguintes, a partir do mais frequente: “científico”, “digital” “documental”, “eletrônico”, “histórico” e “tecnológico”. Sugere-se uma análise sobre os conceitos a partir do significado semântico dos adjetivos, que avalie múltiplas interpretações e determine semelhanças e diferenças entre áreas. Na análise dos projetos de pesquisa, foram encontrados três adjetivos recorrentes em todos os programas de pós-graduação, sendo eles, por ordem de frequência: “científico”, “digital” e “documental”. O termo

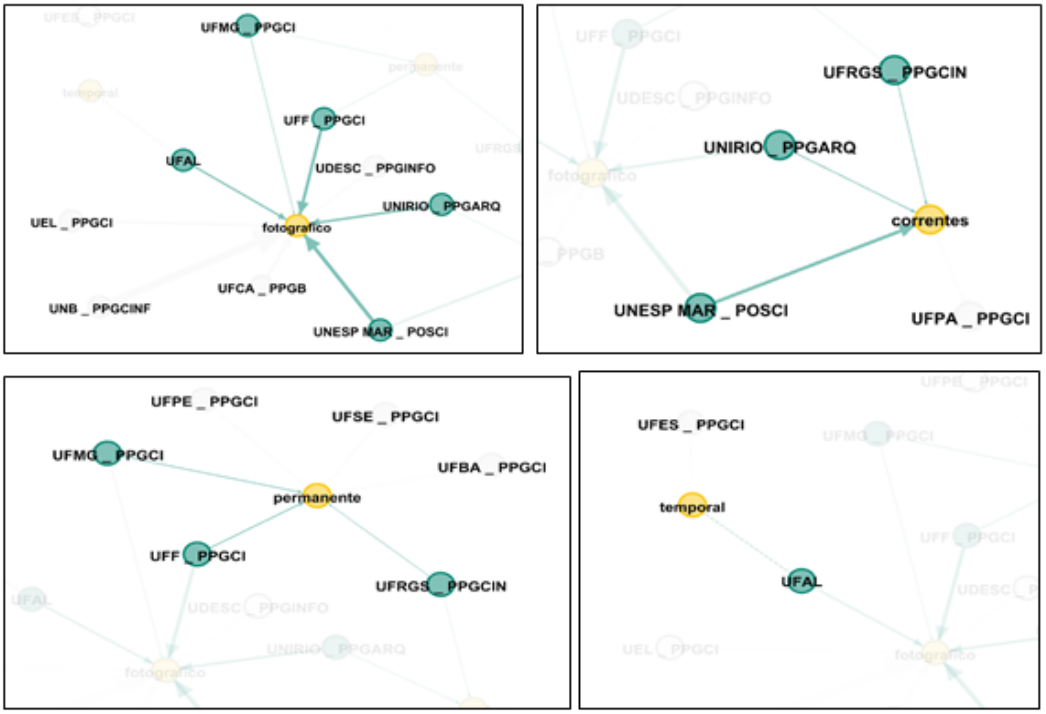
"científico" cabe bem aos projetos de pesquisa, já que está atrelado à pesquisa científica na área. O termo "digital", assim como "eletrônico" que aparece em seguida, mas não ocorre de forma significativa em todas as instituições, podem se referir ao meio da representação da informação com registro de documentos e a informação no formato eletrônico ou digital. Já o termo "documental" possivelmente está relacionado ao meio da representação da informação na produção de documentos. E o termo "tecnológico" pode representar as diferentes tecnologias, seus uso, transferências e aquisições tecnológicas.

Grafo1: Outros adjetivos relevantes frequentes



Fonte: Dados da pesquisa

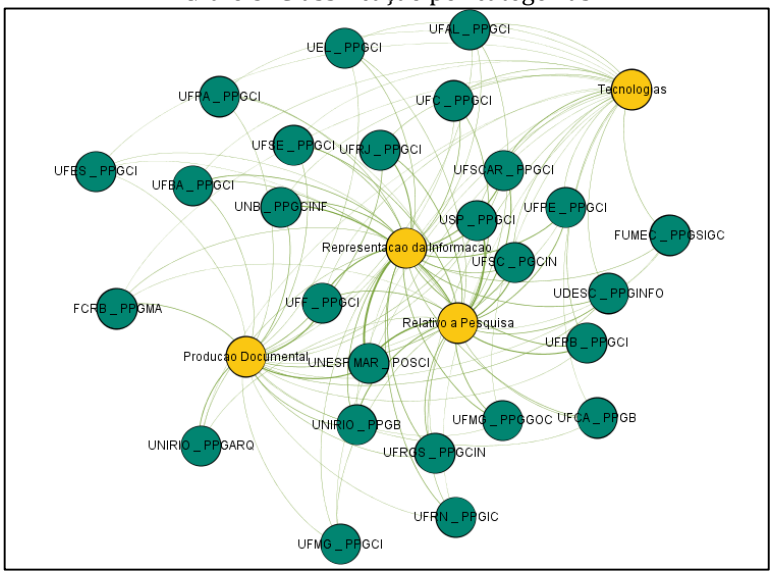
Grafo 2: Adjetivos frequentes e relações entre programas



Fonte: Dados da pesquisa

Indo além, outros adjetivos relevantes pela distribuição entre os grupos foram identificados, apesar de não terem sido tão recorrentes quantos os demais anteriormente citados (Grafos 1 e 2). Observa-se a presença dos adjetivos “fotográfico”, “permanente”, “corrente” e “temporal”, fortemente atrelados a área da Arquivologia, e os programas que os utilizam mais frequentemente. A partir desta análise, é possível estabelecer importantes correlações entre os grupos de pesquisa, corroborando o objetivo principal deste trabalho. Estas correlações permitem identificar projetos de pesquisas relacionados, conforme apresentado nos grafos acima. Além disso, a partir das categorias apresentadas no Quadro 4, elaborou-se um grafo para identificar a proximidade das instituições em relação a frequência dos adjetivos utilizados, agrupados por categorias.

Grafo 3: Classificação por categorias



Fonte: Dados da pesquisa

A análise do Grafo 3 segue na mesma linha dos demais. Quanto mais próxima é a relação entre um programa e uma categoria, maior é a frequência existente entre elas, isto é, mais o programa em questão utiliza adjetivos desta categoria. A seguir, o Quadro 5 apresenta a frequência de adjetivos encontrada nos resumos dos currículos.

Quadro 5: Adjetivos mais frequentes nos resumos dos currículos

Programa	Adjetivo 1	Adjetivo 2	Adjetivo 3	Adjetivo 4	Adjetivo 5
FCRB _ PPGMA	cultural	social	docente	histórico	permanente
FUMEC _ PPGSIGC	acadêmica	digital	social	cultural	digital
UDESC _ PPGINFO	docente	digital	científica	tecnológica	humanas
UEL _ PPGCI	docente	digital	industrial	permanente	eletrônico
UFAL _ PPGCI	documental	humanas	científico	permanente	adjunto
UFBA _ PPGCI	permanente	científica	digital	social	periódicos
UFC _ PPGCI	adjunto	permanente	tecnológica	adjunto	docente
UFCA _ PPGB	periódicos	adjunto	acadêmica	digital	sustentável
UFES _ PPGCI	cultural	permanente	docente	social	digital
UFF _ PPGCI	social	digital	permanente	docente	documental
UFMG _ PPGCI	social	cultural	digital	tecnológica	científico
UFMG _ PPGGOC	social	científica	digital	interdisciplinares	cultural
UFPA _ PPGCI	permanente	digital	adjunto	científica	docente
UFPB _ PPGCI	científica	social	permanente	docente	digital
UFPE _ PPGCI	científica	digital	permanente	adjunto	periódicos
UFRGS _ PPGCIN	científica	digital	permanente	adjunto	social
UFRJ _ PPGCI	social	coordenadora	científica	permanente	docente
UFRN _ PPGIC	adjunto	digital	administrativa	competitiva	científica
UFSC _ PGCIN	adjunto	científica	documental	coordenadora	digital
UFSCAR _ PPGCI	tecnológico	digital	científica	docente	permanente
UFSE _ PPGCI	social	docente	científica	permanente	acadêmica
UNB _ PPGCINF	cultural	científica	digital	documental	docente
UNESP MAR _ POSCI	docente	coordenadora	permanente	periódicos	digital
UNIRIO _ PPGARQ	tecnológico	permanente	documental	coordenadora	fotográficos
UNIRIO _ PPGB	tecnológico	científica	coordenadora	social	docente
USP _ PPGCI	docente	tecnológica	científica	ambiental	social

Fonte: Dados da pesquisa

Quando se faz um comparativo entre os Quadros 4 e 5, observa-se que os adjetivos “digital” e “científico” são predominantes nos dois casos (resumos e projetos de pesquisa). Com relação ao CG, os adjetivos mais frequentes encontrados em todos os resumos são, pela ordem de frequência, “social” e “digital”, seguidos por “coordenadora” (indicativo da forte presença feminina na área da CI), “científica”, “docente”, “permanente”, “cultural”, “adjunto”, “tecnológico” e “periódicos. O Quadro 6 apresenta esta informação, especificada por programa.

Quadro 6: Adjetivos mais frequentes nos resumos dos currículos

Adjetivo x Frequência / Programa	social 235	digital 178	coordenadora 146	científica 130	docente 128	permanente 101	cultural 75	adjunto 67	tecnológico 69	periódicos 50
FCRB _ PPGMA	0%	1%	3%	0%	0%	3%	27%	1%	0%	0%
FUMEC _ PPGSIGC	1%	2%	0%	1%	1%	0%	4%	3%	3%	4%
UDESC _ PPGINFO	1%	2%	3%	3%	6%	1%	1%	1%	0%	0%
UEL _ PPGCI	0%	2%	7%	1%	5%	3%	3%	3%	3%	0%
UFAL _ PPGCI	1%	1%	3%	2%	1%	2%	1%	6%	0%	2%
UFBA _ PPGCI	3%	3%	5%	7%	8%	9%	7%	3%	6%	8%
UFC _ PPGCI	3%	2%	3%	2%	2%	2%	1%	4%	1%	2%
UFCA _ PPGB	5%	2%	2%	3%	0%	2%	0%	9%	6%	8%
UFES _ PPGCI	2%	3%	5%	0%	4%	5%	15%	3%	0%	2%
UFF _ PPGCI	26%	8%	3%	4%	4%	5%	3%	0%	3%	2%
UFMG _ PPGCI	43%	2%	3%	3%	1%	3%	12%	4%	6%	0%
UFMG _ PPGGOC	1%	3%	3%	4%	1%	1%	3%	4%	0%	0%
UFPA _ PPGCI	1%	3%	3%	3%	2%	7%	0%	4%	3%	0%
UFPB _ PPGCI	7%	6%	3%	10%	5%	7%	5%	3%	4%	6%
UFPE _ PPGCI	5%	4%	2%	9%	2%	3%	3%	4%	3%	6%
UFRGS _ PPGCIN	5%	4%	3%	11%	1%	5%	4%	4%	3%	4%
UFRJ _ PPGCI	5%	6%	8%	12%	5%	9%	3%	1%	1%	10%
UFRN _ PPGIC	5%	2%	1%	2%	1%	0%	0%	4%	0%	2%
UFSC _ PGCIN	5%	2%	3%	5%	2%	2%	0%	9%	1%	2%
UFSCAR _ PPGCI	5%	5%	5%	3%	3%	3%	3%	3%	17%	6%
UFSE _ PPGCI	5%	1%	3%	3%	5%	3%	1%	1%	0%	2%
UNB _ PPGCINF	5%	6%	5%	3%	4%	4%	7%	4%	4%	4%
UNESP MAR _ POSCI	5%	7%	10%	7%	21%	12%	3%	3%	4%	20%
UNIRIO _ PPGARQ	5%	0%	1%	2%	1%	5%	3%	3%	14%	6%
UNIRIO _ PPGB	5%	3%	6%	9%	5%	4%	5%	3%	14%	6%
USP _ PPGCI	5%	4%	6%	5%	8%	1%	32%	4%	7%	4%

Fonte: Dados da pesquisa

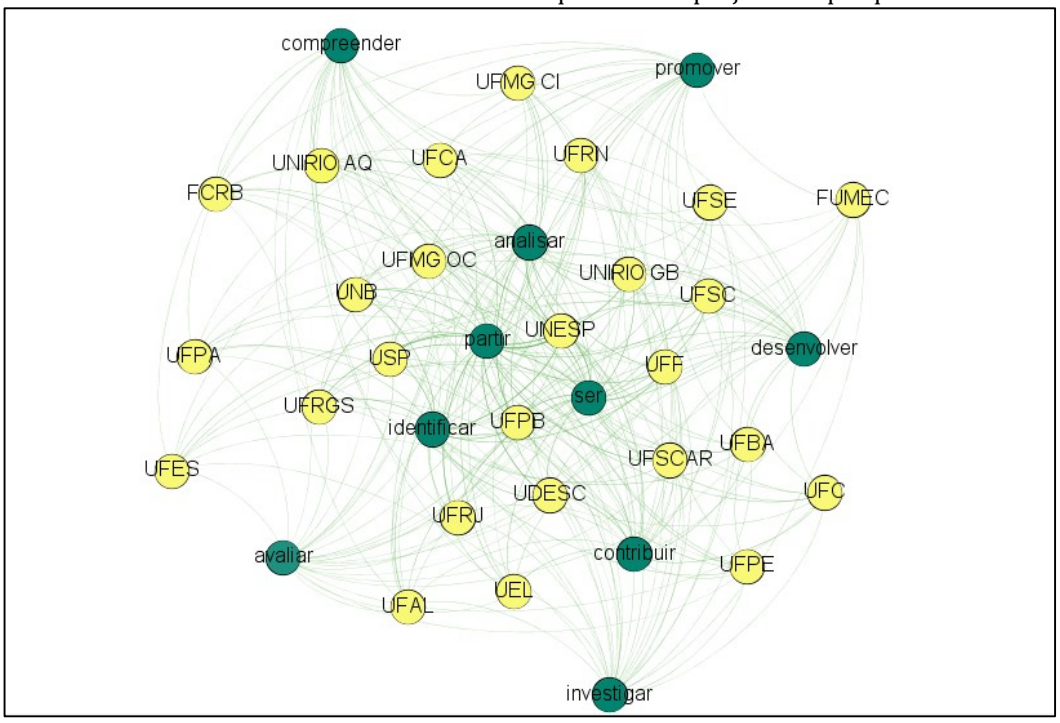
O quadro acima apresenta os 10 adjetivos mais frequentes encontrados em todos os resumos de currículos. Efetuou-se um cálculo de porcentagem quanto à ocorrência de cada adjetivo em relação ao total de ocorrências de todos adjetivos dentro dos programas, de maneira que se possa dar um "peso" a cada um. É possível se observar que os adjetivos têm valor distinto em cada programa, e que um adjetivo pode ser mais ou menos utilizados pelos pesquisadores de um programa.

4.2 Frequência de verbos no infinitivo

Verbos constituem itens de articulação no texto e exercem a função de núcleo do predicado nas sentenças. Mesmo que se apresentem de forma aparentemente padronizada, os verbos apresentam características distintas conforme são empregados em cada contexto.

A próxima análise é novamente um subproduto da etiquetagem POS. Ela procurou identificar os verbos no infinitivo que mais ocorrem (com uma frequência maior que dez) nas descrições dos projetos de pesquisa dos programas de pós-graduação. O Grafo 2 apresenta os dez verbos mais frequentes. A lista, do mais frequente para o menos frequente é: "partir", "ser", 'identificar", "analisar", "contribuir", "desenvolver", "discutir", "compreender", "conhecer", "estudar".

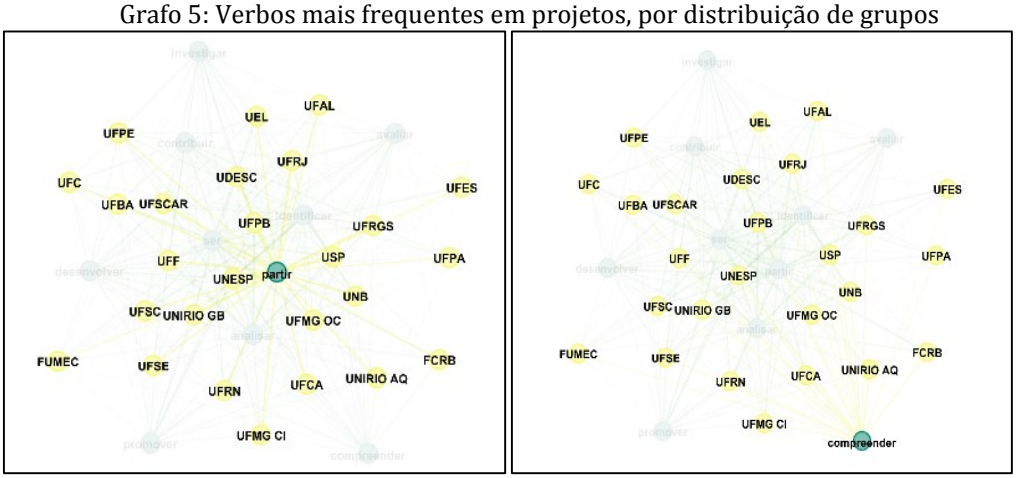
Grafo 4: Verbos no infinitivo mais frequentes nos projetos de pesquisa



Fonte: Dados da pesquisa

O verbo mais frequente em todos os projetos de pesquisa dos programas de pós-graduação (CG) com seus respectivos números de ocorrências foi “partir”, com 1.283 ocorrências. O verbo “compreender” obteve 260 ocorrências e ficou na última posição entre os dez verbos mais frequentes. Observa-se que para cada verbo as instituições são posicionadas quanto à frequência de seu uso. Nesta análise, todos os verbos são recorrentes em todas instituições com maior ou menor frequência.

Grafo 5: Verbos mais frequentes em projetos, por distribuição de grupos

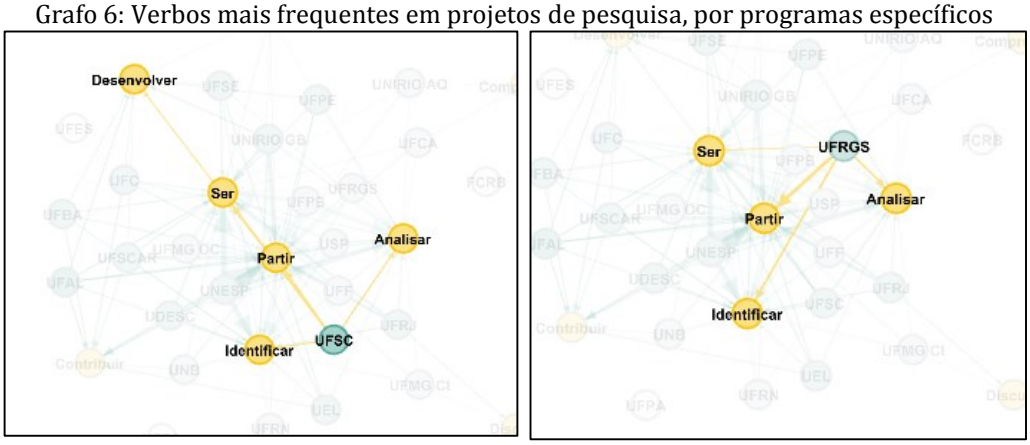


Fonte: Dados da pesquisa

Ao se examinar de forma mais minuciosa, é útil observar o posicionamento dos programas em relação aos verbos. Se analisar a frequência do verbo “partir”, o programa mais próximo dele é o da UNESP, e o mais distante é o da FUMEC. Isto se dá em razão da frequência dos verbos. Nos projetos de pesquisa da UNESP o verbo “partir” é encontrado 169 vezes. Nos da FUMEC, este verbo é encontrado apenas 17 vezes. Ou seja, apesar do grafo não apresentar o número exato de ocorrência dos

verbos no infinitivo, compreende-se que quanto mais próximo do verbo, maior a frequência em que o mesmo é utilizado por determinado programa. No sentido inverso, quanto mais próximo das extremidades, menor é a frequência em que o verbo é utilizado pelo programa.

Grafo 6: Verbos mais frequentes em projetos de pesquisa, por programas específicos



Fonte: Dados da pesquisa

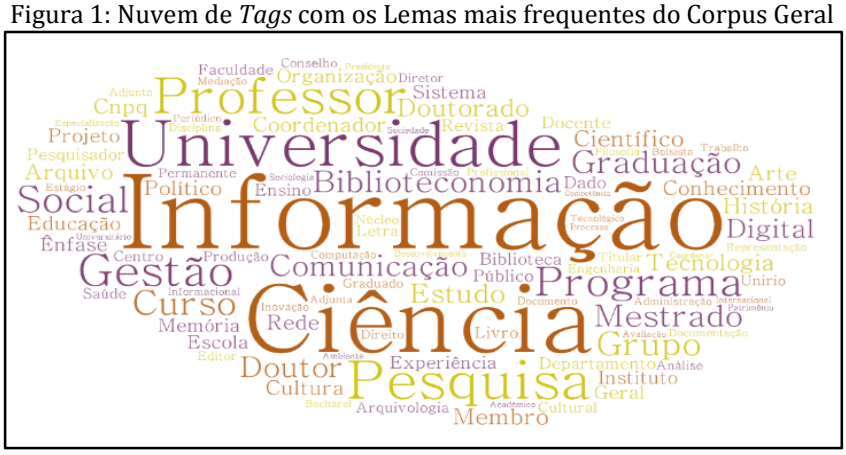
Na análise apresentada no Grafo 6 foram considerados os verbos com frequência superior a 10 ocorrências. O programa da UFSC está representado com os 5 verbos mais frequentes. Enquanto o verbo "identificar", mais próximo do nodo do programa, indica maior frequência, o verbo "desenvolver", que está mais distante, indica menor frequência. A outra análise refere-se ao programa da UFRGS. Aqui, o verbo "analisar" é o mais frequente e "identificar" é o menos frequente. A partir deste tipo de análise, pode-se considerar uma perspectiva diferente de reflexão a respeito das relações entre grupos de pesquisa e instituições e de que forma cada um destes constrói seus pilares.

Com relação à frequência de verbos no infinitivo nos resumos dos currículos, esta não apresentou resultados significativos para análise. Os verbos "partir" e "ser" apresentaram maior recorrência.

4.3 Frequência de Lemas

Lemas representam a redução de um subconjunto de termos que possuem proximidade gramatical a uma forma canônica (termo) em comum que os representa. Após o processo de lematização, identificou-se os lemas mais frequentes para o CG, a partir da combinação de todos os resumos e descrições dos projetos de pesquisa. A Figura 1 apresenta em forma de nuvem de *tags* os lemas mais frequentes encontrados no CG, sendo os cinco mais frequentes, a partir do primeiro: "informação" "pesquisa", "projeto", "ciência" e "professor".

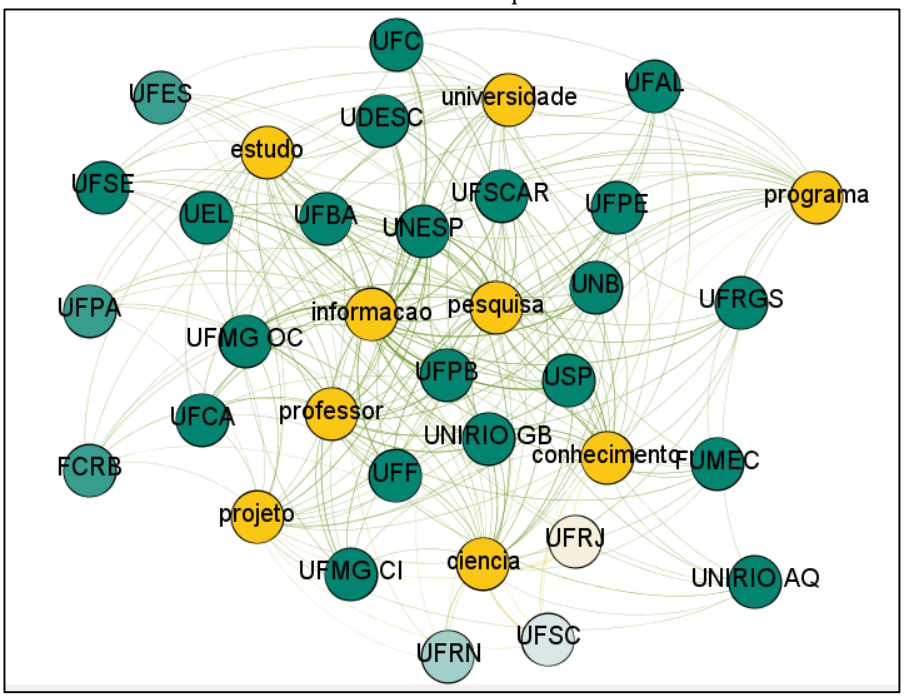
Figura 1: Nuvem de *Tags* com os Lemas mais frequentes do Corpus Geral



Fonte: Dados da pesquisa

O Grafo 7 apresenta a informação acima expandida (9 lemas mais frequentes no CG) e em forma de grafo, na qual procura destacar a proximidade com a qual os programas se relacionam com cada lema.

Grafo 7: Lemas mais frequentes no CG



Fonte: Dados da pesquisa

Nos resumos dos currículos, outros termos importantes definem a forma de apresentação dos profissionais como “professor” com “conhecimento” específico que atua em determinada “universidade”. E mesmo que possa ser outra interpretação, são lemas significantes que podem apresentar outros significados para além do vínculo institucional.

A seguir, a Tabela 2 apresenta lemas que, apesar de não estarem no topo da lista dos dez mais frequentes, definem campos de atuação e linhas de pesquisa dentro da grande área da Ciência da Informação. Verifica-se que alguns destes lemas de menor frequência são comuns entre as áreas afins da CI e, se organizados, podem ser diretamente relacionados com a Ciência da Informação, Arquivologia,

Biblioteconomia e Museologia. Isso não significa, no entanto, que estes lemas sejam necessariamente relevantes pelo foco que possuem dentro dos programas. Esta última característica não possui relação direta com a frequência do lema.

Os lemas foram agrupados de maneira *ad hoc* por área afim, de maneira a deixar clara a sua influência no escopo da CI como um todo. Verifica-se que, enquanto programas possuem boa ocorrência destes lemas, de maneira quase geral, outros acabam por se "especializar" em subconjuntos destes lemas. Isto é um bom indício do rumo que determinados programas estão tomando com relação às suas pesquisas, quantitativamente falando. É possível se observar, por exemplo, que o programa UFF/PPGCI dá um enfoque maior para pesquisas da área da Arquivologia e que o foco do UFMG/PPGGOC se concentra primordialmente em tópicos da Ciência da Informação. Além destes, outros exemplos podem ser visualizados. Observa-se que o UFMG/PPGCI está atrelado às quatro áreas, pois possui uma ocorrência equilibrada de lemas entre elas. Por sua vez, o UFC/PPGCI está mais atrelado às áreas de Ciência da Informação e Biblioteconomia, e o UFRGS/PPGCIN está mais atrelado às áreas de Ciência da Informação e Arquivologia. Este tipo de análise favorece a identificação de linhas de pesquisa, por meio da identificação de lemas, e pode ainda ser usado como um facilitador no momento de se analisar potenciais parcerias em projetos de pesquisa entre os programas e/ou instituições.

Tabela 2: Lemas do CG que representam campos de atuação na CI e suas ocorrências

Programa	Ciência da Informação			Arquivologia			Biblioteconomia			Museologia		
	dado	tecnológico	eletrônico	arquivo	acervo	arquivístico	biblioteca	bibliotecário	memória	patrimônio	preservação	museu
FCRB _ PPGMA	25	0	0	84	75	23	0	0	30	0	19	31
FUMEC _ PPGSIGC	46	40	0	0	0	0	0	0	0	0	0	0
UDESC _ PPGINFO	98	0	0	0	0	0	122	66	0	0	0	0
UEL _ PPGCI	39	0	33	0	0	0	49	29	61	0	31	0
UFAL _ PPGCI	28	38	0	0	0	0	0	0	30	0	0	0
UFBA _ PPGCI	43	0	0	48	60	55	43	0	48	0	42	0
UFC _ PPGCI	34	43	43	0	0	0	70	31	0	0	0	0
UFCA _ PPGB	28	0	0	0	27	0	103	0	28	0	0	0
UFES _ PPGCI	0	0	0	60	15	18	48	18	15	0	0	0
UFF _ PPGCI	73	0	43	176	67	235	0	0	0	67	99	0
UFMG _ PPGCI	29	0	0	69	66	0	33	0	70	42	0	62
UFMG _ PPGGOC	129	45	37	0	41	0	84	0	0	0	0	18
UFPA _ PPGCI	15	0	0	33	18	47	39	0	0	0	24	19
UFPB _ PPGCI	74	0	0	0	0	0	0	0	59	46	0	0
UFPE _ PPGCI	39	35	0	0	48	0	49	0	44	0	28	0
UFRGS _ PPGCIN	92	29	0	29	34	36	0	0	0	0	0	0
UFRJ _ PPGCI	142	35	0	0	0	0	0	0	57	0	0	36
UFRN _ PPGIC	23	0	0	0	24	0	38	0	0	0	0	0
UFSC _ PGCIN	166	38	0	46	0	0	49	0	0	0	0	0
UFSCAR _ PPGCI	80	75	0	0	41	0	116	0	37	0	0	0
UFSE _ PPGCI	25	0	0	0	32	0	52	0	0	0	0	28
UNB _ PPGCINF	110	0	0	127	88	90	0	0	102	0	0	49
UNESP MAR _ POSCI	235	0	0	0	107	0	294	0	0	0	0	0
UNIRIO _ PPGARQ	0	0	0	157	125	107	0	0	60	31	30	0
UNIRIO _ PPGB	93	0	0	0	0	0	117	0	86	49	0	0
USP _ PPGCI	97	0	0	0	0	0	107	0	52	0	0	0

Fonte: Dados da pesquisa

Estas informações também podem ser úteis na comparação e avaliação de linhas de pesquisa entre diferentes grupos, na busca por se identificar maiores concentrações da produção científica em determinadas áreas da informação e seus produtores. Nesta mesma linha de raciocínio, é factível se identificar a escassez de estudos em algumas áreas por determinados programas. Finalmente, é possível também se definir correlações entre os programas de maneira a se estabelecer graus de proximidade ou distanciamento entre eles.

5 Conclusões

Este trabalho buscou identificar correlações entre grupos brasileiros de pesquisa em Ciência da Informação por meio da aplicação de técnicas de mineração

textual em currículos Lattes dos participantes destes grupos. Como insumo principal, foram usadas as palavras-chave de maior ocorrência observadas nos textos dos resumos informativos e nas descrições dos projetos de pesquisa encontrados nos currículos, ambos preenchidos livremente pelos pesquisadores.

A utilização de ferramentas automatizadas permitiu empreender diversas análises, entre as quais se destaca a identificação de subáreas de conhecimento, campos de atuação e linhas de pesquisa dentro da grande área da Ciência da Informação. Procurou-se fazer um recorte significativo nos resultados a serem apresentados, de maneira a dar uma ideia do potencial existente na proposta. Entretanto, é preciso se enfatizar que a lista de resultados apresentada anteriormente não é de maneira nenhuma exaustiva, pois o resultado da mineração executada é passível de ser analisado por um número muito maior de vertentes, e outras abstrações poderão daí advir.

Entre o que foi apresentado, gostaríamos de destacar a identificação dos n-gramas mais frequentes, por meio da qual foi possível relacionar a grande área da CI com as áreas afins de Arquivologia, Biblioteconomia e Museologia. Vertentes específicas de pesquisa também vieram à tona nestas análises, tais como "dimensão estética da medição da informação", "análise de redes sociais", "rede transamazônica de cooperação", "revisão por pares na aprendizagem", entre outras. A abrangência desses resultados possibilita ainda uma análise mais específica dos campos de atuação discriminados por programa analisado e/ou instituição.

A identificação dos adjetivos mais utilizados proporcionou traçar correlações entre os grupos de pesquisa, na medida em que os adjetivos podem ser associados a diferentes áreas de estudo, configurando-se esta como uma perspectiva de análise que possibilita identificar projetos de pesquisas relacionados a objetivos específicos. Além disso, pode-se visualizar desdobramentos disso no sentido de se identificar perfis de profissionais e de grupos a partir da análise dos adjetivos usados por estes e pelo "peso" destes adjetivos no corpus do programa ou no corpus geral. A partir desse "peso", pode-se ainda abstrair categorias de maneira a se observar tendências de cada grupo a usar mais este ou aquele adjetivo, o que, no final das contas, se configura na observação de um comportamento de grupo e pode ensejar diversas perspectivas no sentido de caracterizar e entender este grupo e seus elementos. Nessa mesma linha, a identificação dos verbos mais recorrentes apresentou uma forma de reflexão sobre as relações entre os grupos e de que forma cada um constrói seus pilares, a partir do uso mais frequente de determinados verbos. Um olhar atento aqui conseguirá identificar diversos fenômenos para além da ordem estrutural.

Os lemas identificados nos resumos dos currículos e nos projetos de pesquisa permitiram a identificação de linhas e campos de pesquisa semelhantes, o que pode ser potencialmente benéfico para profissionais e estudantes da área da CI, do ponto de vista do estabelecimento de parcerias entre programas de pós-graduação, a fim de agregar e difundir o conhecimento, produtividade mais abrangente, maiores possibilidades de inovação tecnológica, entre outros.

De maneira geral, os resultados aqui apresentados são preliminares, pois este estudo será continuado de forma mais aprofundada. Novos testes serão feitos com corpora maiores e novas versões dos algoritmos de mineração de texto serão testadas. O componente semântico será levado em consideração e técnicas como vetores embutidos de palavras (*word embeddings*) e proximidade semântica de termos serão exploradas. Igualmente, o estabelecimento de *clusters* automáticos (*clustering*), identificação de tópicos (*topic modelling*) e o reconhecimento de entidades nomeadas (*named entity recognition*) são técnicas de mineração textual que se pretende explorar.

Finalmente, é possível também efetuar a mineração de maneira mais granularizada, focando o processo não mais em grupos, mais sim em indivíduos, de maneira a se estabelecer correlações num nível mais micro. Acreditamos que o estabelecimento de correlações, seja entre grupos, seja entre programas/instituições, seja entre pesquisadores contribui para avançar a compreensão das pesquisas em CI no Brasil, atentando-se para a forma como estas estão sendo desenvolvidas, além de identificar tendências e prever novos rumos para esta área de conhecimento no país.

REFERÊNCIAS

Andrade, P. H. M. A. D. (2015). Aplicação de Técnicas de Mineração de Textos para Classificação de Documentos: um Estudo da Automatização da Triagem de Denúncias na CGU. Brasília, 2015. 65p. Dissertação (Mestrado Profissional em Computação Aplicada). Disponível em: https://repositorio.unb.br/handle/10482/21004. Acesso em: 26 jun. 2020.

Aranha, C., & Passos, E. (2006). A tecnologia de mineração de textos. Revista Eletrônica de Sistemas de Informação, 5(2). Disponível em: http://www.periodicosibepes.org.br/index.php/reinfo/article/view/ 17. Acesso em: 30 jun. 2020. doi: https://doi.org/10.21529/RESI.2006.0502001.

Dias, T. M. R., & Moita, G. F. (2016). Um estudo da produção científica brasileira a partir de dados da Plataforma Lattes. 2016. 181 p. Tese (Doutorado em Modelagem Matemática e Computacional) - Centro Federal de Educação Tecnológica de Minas Gerais, Belo Horizonte, 2016. Disponível em: https://sig.cefetmg.br/sigaa/verArquivo?idArquivo=2033874&key=d8d1d2008e1ebe20f0f136527af3a222. Acesso em: 30 jun. 2020.

Domingues, M. L., Favero, E. L., & Medeiros, I. P. (2007). Etiquetagem de Palavras para o Português do Brasil. In Proceedings of the 5th Workshop in Information and Human Language Technology (TIL'2007), Rio de Janeiro, Brazil (pp. 1721-1724). p. 1721-1724. Disponível em: http://www.nilc.icmc.usp.br/til/til2007/arq0179.pdf. Acesso em: 20 jun. 2020.

Dutra, S. T., Lezana, Á. G. R., Dutra, M. L., & Pinto, A. L. (2019). A Bibliometric Analysis of the Scientific Production and Collaboration between Graduate Programs in Manufacturing Engineering in Brazil. *Informação & Sociedade*, *29*(1). Disponível em: https://periodicos.ufpb.br/index.php/pbcib/article/view/47991. Acesso em 20 jun. 2020.

Feldman, R., & Sanger, J. (2007). The text mining handbook: advanced approaches in analyzing unstructured data. Cambridge university press. Disponível em: https://www.researchgate.net/publication/200504395_The_text_mining_handbook_Advanced_approaches_in_analyzing_unstructured_data. Acesso em 22 jun. 2020.

Hearst, Marti A. (1999). "Untangling text data mining". Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics. pp. 3–10. doi:10.3115/1034678.1034679. ISBN 978-1-55860-609-8.

Ingersoll, G., Morton, T., & Farris, A. (2013). Taming text. How to Find, Organize, and Manipulate It, Shelter Island, NY/London. Disponível em: https://www.aclweb.org/anthology/P99-1001.pdf. Acesso em: 25 jun. 2020. doi: https://dl.acm.org/doi/10.3115/1034678.1034679

Lattes, P. (2007). Currículo Lattes. Disponível em: http://lattes.cnpq.br/. Acesso em: 30 Mai 2020.

Machado, A. P., Ferreira, R., Bittencourt, I. I., Elias, E., Brito, P., & Costa, E. (2010). Mineração de texto em Redes Sociais aplicada à Educação a Distância. Revista Digital da CVA-RICESU, 6(23). Disponível em: https://www.semanticscholar.org/paper/Mineração-de-texto-em-Redes-Sociais-aplicada-à-a-Machado-Ferreira/60a045db477689ddd00997ef18d30381fe2ee34c. Acesso em: 12 jun. 2020.

Madeira, R. D. O. C. (2015). Aplicação de técnicas de mineração de texto na detecção de discrepâncias em documentos fiscais (Doctoral dissertation). Disponível em: https://bibliotecadigital.fgv.br/dspace/handle/10438/14593. Acesso em 12 jun. 2020.

Mena-Chalco, J. P., & Júnior, C. (2013). Prospecção de dados acadêmicos de currículos Lattes através de scriptLattes. Bibliometria e Cientometria: reflexões teóricas e interfaces. São Carlos: Pedro & João, 109-128. Disponível em: https://www.researchgate.net/profile/Jesus_Mena-Chalco/publication/280113692_Prospeccao_de_dados_academicos_de_curriculos_Lattes_atraves_de_scriptLattes/links/55aa9a8f08aea3d086827791.pdf. Acesso em 20 jun. 2020.

Neves, P. I., Corrêa, D. A., & Cavalcanti, M. C. (2013). Uma análise sobre abordagens e ferramentas para Extração de Informação. Seção de Engenharia e Computação–Instituto Militar de Engenharia (IME). Departamento de Informática–Universidade Federal Rural do Rio de Janeiro (UFRRJ). Laboratório Nacional de Computação Científica (LNCC). Disponível em: http://rmct.ime.eb.br/arquivos/RMCT_3_tri_2013/RMCT_123_E8A_13.pdf. Acesso em: 10 jun. 2020.

Rajman, M., & Besançon, R. (1998). Text mining: natural language techniques and text mining applications. In Data mining and reverse engineering (pp. 50-64). Springer, Boston, MA. Disponível em: https://link.springer.com/chapter/10.1007/978-0-387-35300-5_3. Acesso em: 15 jun. 2020.

Sarkar, D. (2016). Text analytics with Python: A practical real-world approach to gaining actionable insights from your data. New York: Apress; 2016.

Tan, A. H. (1999). Text mining: The state of the art and the challenges. In Proceedings of the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases (Vol. 8, pp. 65-70). sn. Disponível emhttp://: http://www.ntu.edu.sg/home/asahtan/papers/tm_pakdd99.pdf. Acesso em 05 jun. 2020.

Trevisan, A. C. (2015). Mineração de textos no Twitter (Bachelor's thesis, Universidade Tecnológica Federal do Paraná). Disponível em: http://repositorio.roca.utfpr.edu.br/jspui/bitstream/1/6659/1/CT_COSIS_2015_1_01.pdf. Acesso em: 20 jun. 2020.

Weiss, S. M., Indurkhya, N., & Zhang, T. (2015). Fundamentals of Predictive Text Mining. New York: Springer, 2010. 226 p. (Texts in Computer Science).