UNIVERSIDADE DE SÃO PAULO
ESCOLA DE COMUNICAÇÕES E ARTES

EDUARDO CAETANO BARBOSA

**Extroversão e descoberta:
visualização aplicada em sistemas de recuperação da informação**

São Paulo
2019

EDUARDO CAETANO BARBOSA

**Extroversão e descoberta:
visualização aplicada em sistemas de recuperação da informação**

Versão Original

*Dissertação apresentada à Escola de Comunicações
e Artes da Universidade de São Paulo para a
obtenção do título de Mestre em Ciência da
Informação*

Área de Concentração: Cultura e Informação

Orientadora: Profa. Dra. Nair Yumiko Kobashi

*São Paulo
2019*

Autorizo a reprodução e divulgação total ou parcial deste trabalho, por qualquer meio convencional ou eletrônico, para fins de estudo e pesquisa, desde que citada a fonte.

Catálogo na Publicação
Serviço de Biblioteca e Documentação
Escola de Comunicações e Artes da Universidade de São Paulo
Dados Inseridos pelo autor

Caetano, Eduardo

Extroversão e descoberta: visualização aplicada em sistemas de recuperação da informação / Eduardo Caetano; orientadora, Nair Yumiko Kobashi. -- São Paulo, 2019.
178 p.: il. + anexo.

Dissertação (Mestrado) - Programa de Pós-Graduação em Ciência da Informação -
Escola de Comunicações e Artes / Universidade de São Paulo.

Bibliografia

Versão corrigida

1. Ciência da Informação 2. Big Data 3. Recuperação da Informação 4. Sistemas de
Recuperação da Informação 5. Visualização da Informação I. Kobashi, Nair Yumiko II.
Título.

CDD 21.ed. – 020

CAETANO, E. Extroversão e descoberta: visualização aplicada em sistemas de recuperação da informação. Dissertação apresentada ao Programa de Pós-graduação em Ciência da Informação da Escola de Comunicações e Artes da Universidade de São Paulo para obtenção do título de Mestre em Ciência da Informação

Aprovado em:

Banca Examinadora

Prof. Dr. _____ Instituição: _____

Julgamento: _____ Assinatura: _____

Prof. Dr. _____ Instituição: _____

Julgamento: _____ Assinatura: _____

Prof. Dr. _____ Instituição: _____

Julgamento: _____ Assinatura: _____

Prof. Dr. _____ Instituição: _____

Julgamento: _____ Assinatura: _____

DEDICATÓRIA

Lembrarei sempre com muito carinho da Gabriela Previdello, que, em meio às incertezas, pegou a minha mão e me encorajou a enfrentar essa jornada. Gostaria de ter tido a chance de retribuir. Dedico este trabalho a ela e às conversas que não chegamos a ter.

AGRADECIMENTOS

O ingresso, ou regresso, na academia é um desafio para aqueles que, já inseridos no mercado de trabalho, precisam dividir esforços para cumprir uma dupla jornada. Ainda que por vezes difícil e sofrido, o trajeto da pesquisa é muito enriquecedor e compensador. Agradeço a todo o apoio e paciência dos colegas, amigos e família que de alguma forma esbarraram nas aflições e empolgações do trabalho. Em especial, vale citar a participação de alguns.

À Profa. Dra. Nair Yumiko Kobashi, orientadora sempre presente e encorajadora, me trouxe de volta ao chão por diversas vezes em que tudo o que fazia era flutuar.

Ao Iron Neto, pela parceria, pela comida, pela indisciplina. Meu eterno relógio de perder as horas.

Aos meus pais, Eduardo e Ana, que me acompanham por todas as conquistas, do primeiro passo ao término deste trabalho, e a tantas outras que ainda quero passar ao lado deles. Junto a eles, Mariana e Maurício, irmãos para todas as horas, e à minha querida tia Lucia, que me apresentou ao universo erudito e esteve sempre tão presente.

Tive a sorte de ganhar alguns amigos, que compartilharam ideias, discussões e aflições, em especial Fernanda Ferreira, Tulio Dias Gomes, Emanuel Guedes, Laura Pimentel e Bianca Santana.

Por fim, agradeço ao Estúdio OKE, que proporcionou tantas oportunidades de colocar em prática parte do desenvolvimento deste trabalho.

Reforço meus sinceros agradecimentos a todos.

sem um numero
um numero
numero
zero
um
o
nu
mero
numero
um numero
um sem numero

Augusto de Campos, 1962

RESUMO

Em meio a um contexto Big Data, caracterizado por ondas de inovação a partir da disponibilidade de dados em volume, que podem ser armazenados e processados de forma cada vez mais eficiente, e que tem revolucionado estruturas culturais, sociais e econômicas, insere-se esta pesquisa, que busca estabelecer o paralelo entre os campos da Ciência da Informação e da Visualização da Informação e avalia o potencial da aplicação de visualizações em Sistemas de Recuperação da Informação (SRI). Para este fim, são explorados conceitos-base, como o próprio documento, a organização, a recuperação e, por fim, a visualização da informação. Uma série de exemplos e casos de uso são explorados, além de elencar os principais recursos e princípios da visualização na aplicação em interfaces de SRI. A fim de validar os conceitos de forma empírica, é apresentado um estudo de caso que detalha um sistema de análise de notícias e menções nas redes sociais que permitia a profissionais de comunicação acompanhar a percepção sobre os serviços prestados. Concluiu-se que o sucesso da utilização de visualizações em SRI está relacionado a aspectos ligados aos três paradigmas da informação, uma vez que são necessários requisitos técnicos e metodológicos que atendam às transformações de dados necessárias para gerar visualizações e, ao mesmo tempo, mantenham a integridade dos metadados de documentos digitais. Do ponto de vista cognitivo, é preciso levar em conta competências analíticas do usuário essenciais para a correta interpretação da leitura dos dados. Sob o paradigma social, coloca-se em discussão as políticas institucionais envolvidas em todo o processo de coleta, tratamento, organização e análise dos dados e que representam os contextos social e cultural da comunidade ou instituição que fará uso do sistema.

Palavras-chave: Ciência da Informação. Big Data. Recuperação de Informação. Sistemas de Recuperação de Informação. Visualização da Informação.

ABSTRACT

Amid a Big Data context, characterized by waves of innovation from the availability of bulk data that can be stored and processed more and more efficiently, it has revolutionized cultural, social and economic structures. It is in this context that the research seeks to establish the parallel between the fields of Information Science and Information Visualization and evaluates the potential of applying visualization in Information Retrieval Systems (SRI). To this end, basic concepts such as the document itself, the organization, the retrieval and, finally, the visualization of the information are explored. A series of examples and use cases are explored, as well as the main features and principles of visualization in application on SRI interfaces. In order to validate the concepts empirically, a case study is presented that details a system of analysis of news and mentions in social networks that allowed communication professionals to follow the perception about the services provided. It was concluded that the success of the use of visualizations in SRI is related to aspects related to the three information paradigms, since technical and methodological requirements are necessary to meet the data transformations necessary to generate visualizations and, at the same time, maintain the integrity of digital document metadata. From the cognitive point of view, it is necessary to consider the analytical skills of the user necessary for the correct interpretation of the reading of the data. Under the social paradigm, the institutional policies involved in the whole process of data collection, processing, organization and analysis are discussed and represent the social and cultural context of the community or institution that will make use of the system.

Keywords: Information Science. Big Data. Information Retrieval. Information Retrieval Systems. Information Visualization.

LISTA DE FIGURAS

Figura 1 – Volume de buscas pelo termo “Big Data” (2004 a 2018).....	24
Figura 2 – Disponibilidade de redes Wi-Fi no mundo (2001 a 2017).....	25
Figura 3 – Preço vs. capacidade de armazenamento de em discos rígidos (1956 a 2017).....	26
Figura 4 – Acurácia de reconhecimento de palavras por Aprendizado de Máquina..	27
Figura 5: Cronologia de autores citados no trabalho sob as lentes da Ciência da Informação e os paradigmas físico, cognitivo e social	53
Figura 6 – O Sistema de Recuperação da Informação e modelo simples de interação.	67
Figura 7 – Mapa dos caminhos trilhados por Yü, o Grande, gravado em pedra em 1137 a.C.	71
Figura 8 – Gráfico de todas as Importações e Exportações de e para a Inglaterra entre 1700 a 1782, de William Playfair.	72
Figura 9 – Gráfico de Charles Joseph Minard, que ilustra a trajetória geográfica e temporal da invasão da Rússia por Napoleão (1812-1813).....	72
Figura 10 – Escala de trens entre Paris e Lyon, de E. J. Marey (1880).	73
Figura 11 – Modelo simplificado de referência da Visualização da Informação.	75
Figura 12 – Representação hierárquica de registros sobre Organização da Informação: agrupamento Descritor x Primeiro autor.	78
Figura 13 – Mapa intelectual da Visualização da Informação (1993-2003).....	79
Figura 14 – Infográfico Interativo Como Votou Sua Vizinhança.	82
Figura 15 – Tela do sistema Film Finder, que mostra simultaneamente o ano de lançamento, a popularidade, o gênero e há ainda a capacidade de filtros por duração, ator, atriz, diretor ou diretora do filme.	85
Figura 16 – Interface de navegação na base de dados do IMDb “Quais os filmes que produzimos?”	87
Figura 17 – Interface gráfica da ferramenta on-line IMDb Explorer.....	89
Figura 18 – Ferramenta de análise de resultados de busca da Scopus.....	93

Figura 19 – Ao clicar em uma das visualizações menores, ela assume o topo da página e ganha destaque.....	94
Figura 20 – Ao clicar em uma área ou ponto do gráfico, uma nova janela é aberta com a lista atualizada e aplicado o filtro representado na visualização.	95
Figura 21 – A ferramenta WoS permite ao usuário selecionar entre dois tipos de gráficos: treemaps ou barras.....	96
Figura 22 – Ao clicar em uma área do gráfico, o usuário acessa a lista com o filtro aplicado. Ao contrário da Scopus, a atualização da página acontece na mesma janela.....	97
Figura 23 – Evolução no tempo de número de publicações que relacionam Visualização da Informação e Recuperação da Informação na base Scopus....	99
Figura 24 – Relação entre Organização da Informação e Visualização da Informação na aplicação de interface visual.	103
Figura 25 – Visualização simples em árvore.	105
Figura 26 – Três formas clássicas de representação de hierarquias, segundo Heer, Bostock e Ogievetsky (2010).	106
Figura 27 – Simulação do efeito da animação no clique do item no browser hiperbólico.....	107
Figura 28 – Simulação de Browser Hiperbólico com 3.418 nós – performance comprometida.....	108
Figura 29 – Treemap e a função de envólucro, evidenciando subcategorias.	109
Figura 30 – Exemplo de visualização “envólucro” com representações em bolhas.	110
Figura 31 – Sequência de navegação e detalhamento de categorias em uma visualização sunburst.	111
Figura 32 – Scatter-Gather Browser.	113
Figura 33 – Scatter-Gather Browser: detalhe de amostra de conteúdos de um mesmo grupo	113
Figura 34 – Rede de Retweets acerca dos termos #ContraTarifa, #PasseLivre e #ContraoAumento e Passe Livre	116
Figura 35 – Grafo gerado a partir de aproximação de termos utilizados nas fichas de trabalhos inscritos na Feira Brasileira de Ciências e Engenharia (Febrace) em 2016.	117

Figura 36 – Cartografia temática para o Programa de Pós-graduação em Direito..	119
Figura 37 – Tela da ferramenta ActiveGraph. Os controles na base da tela permitem atribuir metadados para cada atributo, como eixo X, Y, cor, tamanho, forma. .	120
Figura 38 – Visualização cone tree, em suas versões vertical e horizontal.	122
Figura 39 – Sistema Embedding Projector, parte do projeto TensorFlow	123
Figura 40 – Tela do experimento de visualização de dados multidimensionais em espaço imersivo.....	124
Figura 41 – Participante dos testes de análise de dados em ambiente imersivo de realidade virtual.	125
Figura 44 – Painéis ilustram a base de fontes de informação utilizadas no sistema de monitoramento.....	144
Figura 45 – Representação da categoria “Fatores externos”, em árvore e documentação.	146
Figura 46 – Painel utilizado para análise de notícias e mensagens de redes sociais.	149
Figura 47 – Tela inicial permite a comparação das quatro instâncias de análise propostas.	151
Figura 48 – Painéis de estrutura similar para atender à análise de tipos diferentes de documento.....	152
Figura 49 – Painel de análise entre três serviços em um município.....	155
Figura 50 – Painel comparativo de menções a empresas diferentes.....	156

LISTA DE SIGLAS

CI – Ciência da Informação

IHC – Interação Humano Computador

IMDb – Internet Movie Database

MEV – Modelo de Espaço Vetorial

RI – Recuperação da Informação

SRI – Sistemas de Recuperação da Informação

TIC – Tecnologias da Informação e Comunicação

VI – Visualização da Informação

VLDL – *Very Large Digital Libraries* (Bibliotecas Digitais muito Grandes)

WoS – *Web of Science*

SUMÁRIO

1. Introdução.....	23
1.1. Objetivos de pesquisa	34
1.2. Estrutura da dissertação	35
1.3. Metodologia	36
2. Referencial teórico.....	39
2.1. O contexto Big Data.....	39
2.2. As lentes da Ciência da Informação	43
2.2.1. Perspectiva histórica.....	43
2.2.2. Paradigmas da Ciência da Informação	46
2.3. Documento e organização da informação.....	55
2.3.1. Estrutura.....	57
2.3.2. Organização da informação.....	59
2.4. Sistemas de Recuperação da Informação	66
2.5. Visualização da Informação.....	70
2.5.1. Modelo simplificado da Visualização da Informação	75
2.5.2. A Visualização nas práticas da Ciência da Informação.....	76
2.6. Visualização em Sistemas de Recuperação de Informação.....	81
2.7. Recursos e técnicas de Visualização da Informação	101
2.7.1. Representação de hierarquias.....	104
2.7.2. Técnicas de agrupamento	111
2.7.3. Espaço 3D.....	121
2.7.4. Interatividade	125
2.7.5. Painéis de controle	127
2.7.6. Dispositivos de acesso	128
2.8. Extroversão e descoberta: uma discussão	131
3. Estudo de caso: Painéis de Análise de Comunicação para empresa prestadora de serviços públicos municipais.....	137
3.1. Caracterização e peculiaridades institucionais da empresa	139
3.2. Caracterização dos documentos.....	140
3.3. Estratégia de organização da informação	144
3.4. Plataforma de visualização e consulta	147
3.5. Exemplos práticos de uso.....	150
3.6. Resultados	156

4. Considerações finais.....	162
4.1. Desdobramentos e futuros trabalhos	165
REFERÊNCIAS	167
ANEXO 1: Análise de interdisciplinaridade a partir da recorrência de palavras-chave no histórico de projetos selecionados na FEBRACE (Feira Brasileira de Ciências e Engenharia)	178

1.Introdução

Nunca conseguimos produzir, coletar e processar tantos dados (informações) como atualmente. E é natural, por extensão, pensar que esse volume tende a crescer exponencialmente: os mesmos avanços conceituais, tecnológicos e metodológicos que nos trouxeram até o presente momento são também vivos e continuam em evolução, tanto na concepção de novas ideias como na capacidade de processamento, armazenamento e otimização das máquinas que auxiliam esse processo.

Não é, definitivamente, um cenário inusitado. O crescimento exponencial da produção e processamento de informações é abordado por teóricos de formação e atuação diversas, como nas tecnologias da informação, sociologia e comunicação (DE MAURO; GRECO; GRIMALDI, 2016; RAMOS, 2016). O próprio campo da Ciência da Informação (CI) nasce e evolui da necessidade de organizar e recuperar a informação em um contexto inevitável de crescimento exponencial de oferta e armazenamento, além do contínuo progresso das tecnologias e sistemas cada vez mais digitais e conectados (CAPURRO, 2003, 2007; FROHMANN, 2008; SVENONIUS, 2000).

Este trabalho parte da premissa de que estamos passando por uma onda de inovação tecnológica única, da mesma magnitude da expansão do acesso ao computador a partir das décadas de 1960 e 1970 e da internet nos anos 1990. Essa onda de inovação foi denominada popularmente, e no meio científico, como “Big Data”, ou “megadados” na tradução ao português.

Por onda de inovação, entende-se que estamos falando de um processo e não de um evento pontual. O Big Data acelerou as mudanças que já estavam em curso. Muitos discutem e planejam suas aplicações há tempos, mas os impactos práticos nas formas de produção, de como nos organizamos e nos relacionamos, começam a ficar mais evidentes a partir do início da década de 2010, quando barreiras tecnológicas importantes para lidar com grandes volumes de dados foram ultrapassadas. O aprimoramento do modelo de código aberto Hadoop, MapReduce e outros softwares e metodologias tornaram a aplicação mais acessível e viável fora do universo das

grandes corporações, minimizando também a necessidade de altos investimentos financeiros (FACELI et al., 2011).

Como é possível observar no gráfico a seguir (Figura 1), o volume proporcional de buscas no Google pelo termo “Big Data” dá indícios de sua popularização a partir de 2013.

Figura 1 – Volume de buscas pelo termo “Big Data” (2004 a 2018).

Fonte: Google Trends (2019).

Um dos marcos do ano de 2013 é o lançamento do livro de divulgação científica *Big Data, a revolution that will transform how we live, work, and think*, de Mayer-Schönberger e Cukier (2013). De forma simples e direta, o livro inspira ao leitor comum aplicações em diversas áreas, negócios e processos que nascem dessa dita *revolução*. Na definição dos autores, “Big Data é a capacidade de a sociedade aproveitar informações de novas maneiras de forma a trazer *insights* ou bens e serviços de significativo valor” (MAYER-SCHÖNBERGER; CUKIER, 2013, p. 12, tradução nossa).

Desde então, o termo Big Data tornou-se onipresente na prática e na pesquisa em Tecnologias da Informação. A evolução de infraestrutura e de abordagens analíticas foi contínua, e, hoje, grandes barreiras foram quebradas.

Todos os anos, a companhia de investimentos e consultoria empresarial Kleiner Perkins¹ realiza estudos nos quais indica as principais tendências de tecnologia e

¹ Informações disponíveis em: <<https://www.kleinerperkins.com/about>>. Acesso em: 23 jan. 2019.

negócios. Sua porta-voz, Mary Meeker é bastante conhecida no mercado como investidora e lidera todos os anos esta pesquisa baseada em dados coletados do mercado. O relatório da companhia, *Internet Trends 2018*², divulgado em maio do mesmo ano, é extenso e aborda questões sobre hábitos do consumidor, comportamento no uso da tecnologia, inovações em áreas como e-commerce, publicidade, transformações no mercado de trabalho e mudanças gerais no cenário das Tecnologias da Informação e Comunicação (TIC). O relatório também aponta a evolução de tecnologias que estão ligadas diretamente ao desenvolvimento do Big Data. A Figura 2 evidencia a disponibilidade e acesso a conexões de redes locais sem fio no mundo, e também podemos notar o crescimento exponencial a partir de 2013, acelerado após 2015.

Figura 2 – Disponibilidade de redes Wi-Fi no mundo (2001 a 2017).

Fonte: WiGLE.net as of 5/29/18 apud Kleiner Perkins (2018, p. 16).

² Disponível em: <<https://www.slideshare.net/joshsc/techcrunch-mary-meeker-2018-internet-trends-report>>. Acesso em: 23 jan. 2019.

O gráfico da Figura 3 ilustra, em verde, a capacidade crescente de armazenamento de informações em disco rígido, que, em 2017, chega a 10 mil gigabytes, o equivalente a 10 terabytes em um único disco. A relação de custo pelo armazenamento é inversamente proporcional e, no mesmo ano, chega a quase 1 centavo de dólar.

Figura 3 – Preço vs. capacidade de armazenamento de em discos rígidos (1956 a 2017).

Fonte: John McCallum @ IDC, David Rosenthal @ LOCKSS Program Stanford: Kryder's Law. Time + Ray Kurzweil análises de diversas fontes, incluindo Gwennap (1996), Kempt (1961) e outros referenciados por Kleiner Perkins (2018, p. 145).

Já a Figura 4 mostra a acurácia dos algoritmos de Aprendizado de Máquina do Google, em especial para o reconhecimento de voz, que atinge o limite de 95% de compreensão, similar à capacidade humana. O Aprendizado de Máquina é o principal recurso de métodos analíticos como a inteligência artificial, e é usado para, entre

outras funções, a recomendação de conteúdo de acordo com hábitos do usuário. Essa evidência ilustra a crescente capacidade de processamento de informações que, em conjunto com a capacidade de armazenamento e o acesso a ambientes conectados, alavancam as possibilidades e acessibilidade a aplicações que se beneficiam do contexto Big Data.

Figura 4 – Acurácia de reconhecimento de palavras por Aprendizado de Máquina.

Fonte: Google (5/17) apud Kleiner Perkins (2018, p. 25).

Os avanços foram muito evidentes, e empresas nativas do universo digital, como Google e Facebook, emergem como vetores de transformação. Através da correlação de dados de naturezas distintas, essas ferramentas mostram ter o poder de compreender características e indícios de comportamentos do usuário, para quem proporciona uma experiência cada vez mais personalizada de *recomendação de conteúdo*, seja na forma de artigo editorial, seja como publicidade ou novas conexões em redes sociais. O usuário passa a ter uma atividade passiva no consumo de informações. São muitas as implicações desse cenário. Por um lado, tal sistema aumenta as chances de um usuário encontrar rapidamente informações que possam

ser relevantes, assim como, entre outros exemplos, facilitar a interação com dispositivos, prevendo os próximos caracteres que está prestes a digitar. Por outro lado, como observa Pariser (2012), há um risco, ao qual chama de efeito bolha, que acontece quando o consumo recorrente de conteúdos filtrados em nível pessoal reafirma gostos, comportamentos, posicionamentos políticos e preconceitos, e priva a pessoa do contato com conteúdos que divergem de suas crenças. O contato com a diferença, segundo o autor, ajuda a construir uma compreensão mais abrangente e diversa sobre questões importantes ao exercício do indivíduo enquanto parte de uma sociedade.

Souza, Almeida e Baracho (2013) discutem, sob uma perspectiva crítica, como esses avanços refletiram – e refletem – no panorama atual da Ciência da Informação. Os autores destacam, em particular, quatro fenômenos informacionais – as redes sociais, o Big Data, a computação na nuvem (*cloud computing*) e a Websemântica – e como impactam o campo da CI, que passa a ter seus objetos de pesquisa adotados por áreas como linguística computacional e linguística de *corpus*; ciência da computação; sistemas de informação e engenharia, entre outras. Segundo os autores,

Há muito que o campo da ciência da computação deixou de ser considerado como aquele que provê tecnologias e sistemas; e tomou a frente de áreas nucleares de organização de informação e conhecimento, ensejando novos diálogos e repertórios de parte da CI. Exemplos de áreas relativamente novas e pujantes não faltam, como as citadas processamento de linguagem natural e a linguística computacional, além de uma recrudescida inteligência artificial, que vêm desenvolvendo pesquisas em classificação e indexação automáticas, construção automática de ontologias, redes neurais, análise do discurso, além de alimentar as pesquisas de ponta em *text mining*, *web mining*, *data mining* e assemelhados. (SOUZA; ALMEIDA; BARACHO, 2013, p. 163-164)

Do ponto de vista teórico, as redes sociais são estruturas que unem atores – indivíduos e organizações – através de laços, mas os autores analisam as redes sociais como ambientes que se constroem amparadas em “*arranjos sociotécnicos*” (SOUZA; ALMEIDA; BARACHO, 2013), exemplificadas pelas redes de relacionamentos nascidas na web, como os já citados Facebook e Twitter.

O Big Data é relacionado à explosão informacional, e o aumento vertiginoso de registros é, em parte, proveniente da troca de mensagens em redes sociais, mas não se limitando a elas. Uma característica é o extenso uso de informação não estruturada, principalmente de cunho textual. Já a computação na nuvem é utilizada para exemplificar a disponibilidade e acessibilidade a recursos de informação e documentos.

De forma conclusiva, os autores afirmam que “As redes sociais romperam limites geográficos de conectividade; a nova explosão de informações rompeu com a possibilidade do tratamento individual e intelectual aos registros; e as nuvens romperam, enfim, com a tangibilidade dos registros” (SOUZA; ALMEIDA; BARACHO, 2013, p. 169). Por fim, a websemântica é apresentada quase como um instrumento de navegação da Ciência da Informação no cenário dos avanços, e destaca as diferentes abordagens das áreas com relação a outro instrumento organizacional, ou seja, as ontologias aplicadas, por exemplo, na prática da organização do conhecimento, em que as terminologias de uma área são o ponto de partida. As ontologias são estudadas e utilizadas em diversos campos do conhecimento. Nas visões da Ciência da Computação, da Ciência da Informação ou da Filosofia, o elemento sempre presente da linguagem e da significação parece ser o ponto de contato que as une.

Ao longo da trajetória profissional, o autor deste trabalho teve a oportunidade de acompanhar o progresso desses conceitos aplicados ao mercado sob três ópticas diferentes.

A primeira, **editorial**, projetando e gerindo aplicações de conteúdo na Editora Abril de 1999 a 2009, onde, desde o início, lidava com números de audiência e aplicava esses conhecimentos para projetar interfaces, especialmente com foco em arquitetura da informação e experiência do usuário. Nesse período, destaco o acesso ao público leigo para publicação e distribuição de conteúdo on-line através dos blogues geridos por ferramentas abertas e de uso sistematizado, como o WordPress e Blogger (entre outros); e a apropriação desse meio de expressão pelos grandes e tradicionais grupos de mídia (jornais, revistas e publicações em geral e seus canais on-line).

A segunda perspectiva é a do **planejamento e de definições de estratégias digitais**, de 2010 a 2014, onde lidei com o mercado das marcas, comunicação institucional e marketing digital, criando metodologias de análise e relatórios para aferição de alcance, interação e engajamento nas redes sociais. Nessa época, o YouTube consolidou sua popularidade, e os brasileiros migravam do Orkut para Facebook e Twitter, que logo assumiram seus papéis de principais mediadores e recomendação de conteúdo on-line.

E a terceira, iniciada em 2015, é derivada da **Inteligência de Dados no contexto da comunicação**. Inteligência de Dados, no âmbito do mercado, é o processo de organizar informações, estruturadas ou não, e integrá-las a sistemas de softwares para que sejam mais bem analisadas e interpretadas e, com isso, contribuir para tomadas de decisões mais assertivas. Na comunicação, um dos desafios da área é detectar tendências e avaliar estratégias a partir da estruturação e classificação de documentos digitais disponíveis, publicamente ou corporativamente, como mensagens de redes sociais, e-mails, imagens, vídeos, páginas web e notícias, para citar apenas alguns exemplos.

Ao longo deste percurso, pude exercer papéis que envolveram a produção, organização, gestão e disponibilização de informações, sua promoção e o papel atual, além de análise quantitativa de evidências de comunicação.

As experiências de construção e gestão de websites e, mais recentemente, de painéis de inteligência (do inglês, *dashboards*), conectam as duas áreas relacionadas neste trabalho – Sistemas de Recuperação de Informação (SRI) e Visualização da Informação (VI). O termo “painéis de inteligência” também deriva da terminologia utilizada no mercado, e sua definição faz alusão aos painéis de controle de um carro ou maquinário, que mostra, simultaneamente, diversas informações. No caso dos carros, os painéis mostram a velocidade, nível do combustível, rotações e outros indicadores de eficiência e controle.

Segundo Turtle e Croft (1997), os SRI possuem três funcionalidades principais:

- a) identificação e representação de documentos (no caso, documentos digitais);
- b) aquisição e representação das necessidades de informação; e

- c) especificação da função de comparação que seleciona os documentos relevantes baseados em suas representações.

Já a VI, especificada como um método analítico, visa potencializar a cognição humana, de forma a descobrir novas informações a partir de grandes volumes de dados (TUFTE, 1983; CARD; MACKINLAY; SHNEIDERMAN, 1999; CHEN, 2004). Não à toa, as visualizações de dados se tornaram recursos ubíquos em sistemas que lidam com volumes cada vez maiores de informações. É uma área de estudo que tem origem na Ciência da Computação e, hoje, é bastante discutida no âmbito da Interação Humano Computador (IHC), que associa tanto questões matemáticas e estatísticas como aspectos ligados à cognição e interpretação humana. É preciso destacar também o recurso da visualização nas práticas da CI, muito presente em estudos bibliométricos, cientométricos e, de forma geral, na informetria.

Esta pesquisa se propõe a estudar a intersecção desses conceitos (SRI e VI), no contexto do Big Data e pelas lentes da CI, e ressaltar o potencial de aplicação em plataformas digitais.

O sucesso dessas aplicações depende, no entanto, da qualidade das bases de dados e documentos a serem consultados. É preciso que os dados estejam organizados e estruturados, seja pela estrutura de seus metadados, seja pelos processos estatísticos e algoritmos inerentes às técnicas como o Aprendizado de Máquina, uma das formas de aplicação da Inteligência Artificial. Os dados estruturados são organizados em linhas e colunas. São geralmente encontrados em banco de dados relacionais e eficientes quanto à recuperação e processamento. Já os dados não estruturados referem-se a dados que não podem ser organizados em linhas e colunas, como vídeos, comentários em redes sociais, e-mails, entre outros. Geralmente são dados de difícil acesso e recuperação e não dispõem de componentes necessários para identificação de tipo de processamento e interpretação, tornando o seu uso um desafio (FACELI et al., 2011).

Através do estudo exploratório da literatura fundacional das áreas da CI e da VI, este trabalho identifica duas funções da visualização de dados aplicadas aos SRI, como aos acervos de documentos digitais. A saber: a extroversão e a descoberta.

A primeira função, a de **extroversão**, aponta para o uso das metodologias e técnicas de visualização de informação para caracterizar e compreender um ou um conjunto de repositórios de documentos digitais a partir da definição do problema que se deseja resolver, identificação das fontes de dados importantes para dar solução ao problema, busca de dados, tratamento estatístico dos dados e análise. Gráficos, grafos e representações visuais podem trazer à superfície características intrínsecas do acervo de um ou mais repositórios digitais, evidenciar relações entre documentos e, por conseguinte, entre temas, conceitos e áreas do conhecimento de um ou mais domínios.

“Extroversão” é um termo com origem na Psicologia, citado pela primeira vez na teoria dos Tipos Psicológicos de Carl Jung de 1921. A teoria distingue duas formas de *atitude*, ou *disposição*, atribuídas ao indivíduo – a extroversão e a introversão. A extroversão tem enfoque no objeto exterior, nos fatos e pessoas em que se observa a atenção para a ação, comunicabilidade, sociabilidade e facilidade de expressão. Já a introversão tem enfoque no objeto interior, na ação reflexiva, reservada. Os indivíduos assumem as duas atitudes ao longo da vida, mas aquela que se destaca determinará as características dominantes e, assim, caracterizará o indivíduo (JUNG, 2009). O aspecto da introversão não será explorado na pesquisa, pois denota o contrário do que podemos afirmar como comunicabilidade.

A pesquisa adota o termo “extroversão” por traduzir qualidades de ação atribuídas à função da visualização ligada à comunicação. A VI tem o potencial de expressar, traduzir e evidenciar, de forma clara e fácil, características de uma base de documentos digitais. Extroversão está aqui sendo utilizada de forma metafórica, para significar a capacidade que a Visualização da Informação tem de explicitar e comunicar a composição de uma base de documentos.

A segunda função, de **descoberta**, propõe que o usuário possa recuperar informações usando a visualização de dados como instrumento na interface de pesquisa em repositórios digitais. A partir da seleção de partes das representações visuais, deve ser possível ao usuário dar significado aos documentos que integram o conjunto representado, seja por pontos, áreas ou outras formas de representação que compõem a visualização gráfica a ser explorada.

No campo da Ciência da Informação, e mesmo da Computação, utiliza-se a expressão “descoberta de informação” em bancos de dados. As ontologias, por exemplo, são caracterizadas, entre outras coisas, como instrumentos de descoberta de conhecimentos (ALMEIDA, 2013). Outra expressão que pode ser usada é “mineração de dados”, comumente aplicados no campo da Ciência de Dados e Computação. Mineração e descoberta também definem, por fim, a capacidade de recuperar informações relevantes (WU et al., 2013).

Um dos desafios do presente trabalho é identificar casos de uso que encontrem denominadores comuns às duas funções (extroversão e descoberta), pois os dados relevantes de documentos digitais para recuperação de informações podem ser perdidos na aplicação de rotinas e algoritmos necessários à visualização. Para que uma visualização seja efetiva como recurso de interface de sistemas de informação, em uma realidade Big Data, caracterizada pelo volume e velocidade de produção de informações, é preciso manipular documentos estatisticamente e, ao mesmo tempo, manter íntegros seus metadados de forma a poder recuperá-los. Essa é uma ação que, em repositórios muito grandes, demandam infraestrutura de armazenamento e processamento de informações, assim como métodos analíticos específicos.

1.1. Objetivos de pesquisa

Objetivo geral

Verificar o potencial e os desafios da aplicação da Visualização de Informação em Sistemas de Recuperação da Informação, no contexto Big Data, caracterizado pela disponibilidade de informações em grande volume, velocidade e variedade.

Objetivos específicos

- Levantar e explorar o paralelo entre as bibliografias fundacionais da Ciência da Informação (CI) e Visualização da Informação (VI);
- Sistematizar as principais características que devem ser levadas em conta na aplicação de Visualização da Informação em Sistemas de Recuperação da Informação (SRI); e
- Identificar e analisar exemplos de aplicação de visualização no auxílio a buscas e recuperação de informações; e
- Documentar e analisar um estudo de caso de aplicação da visualização da informação em um sistema de recuperação de informações, de forma a complementar o trabalho de forma empírica.

1.2. Estrutura da dissertação

O trabalho é dividido em três partes. A primeira, esta introdução, apresenta o contexto e a justificativa do tema, hipóteses, objetivos e metodologia da pesquisa.

A segunda parte reúne o referencial teórico, resultado do levantamento bibliográfico e de exemplos de aplicação. Primeiro, na seção 2.1, é apresentado o conceito Big Data, que delinea os contextos técnico, social e cultural derivados da capacidade de armazenar, processar e disponibilizar de forma ágil grandes volumes de dados. A seção 2.2 estabelece o paralelo entre a Ciência da Informação e a Visualização da Informação, relacionando aspectos históricos e epistemológicos. Na seção 2.3, apresenta-se uma definição de documentos, em especial os documentos digitais, destacando sua caracterização pelo uso da linguagem e os impactos da forma como são organizados. Em sequência, na seção 2.4, há um panorama sobre definições e conceitos envolvidos em Sistemas de Recuperação da Informação, em especial bibliotecas digitais. Os princípios da Visualização da Informação são detalhados na seção 2.5, além da exposição de uma visão histórica e como é aplicada no contexto da Ciência da Informação. Na seção 2.6, são discutidas aplicações práticas da visualização em plataforma de recuperação de documentos a partir da análise de diversos exemplos. Para encerrar, a seção 2.7 resgata os principais aspectos dos capítulos anteriores e traz uma seleção de técnicas e tipos de visualização e seus potenciais usos na recuperação de documentos.

A terceira parte do trabalho traz a descrição do Estudo de Caso mencionado na metodologia, que auxilia a compreensão dos pontos levantados na segunda parte.

As considerações finais, assim como algumas perspectivas para sequência e desdobramento de pesquisas, estão na quarta parte.

1.3. Metodologia

Esta pesquisa tem, de um lado, perspectiva exploratória, de base bibliográfica e de levantamento de exemplos que busquem demonstrar as intersecções dos estudos de Visualização da Informação com a Ciência da Informação.

O estudo exploratório tem como objetivo proporcionar maior familiaridade com o problema de pesquisa, com vistas a torná-lo mais explícito ou a construir hipóteses (GIL, 2007). Para este trabalho, foram feitos levantamentos bibliográficos e análises de exemplos que permitam aprofundar a compreensão do tema tratado.

Para o levantamento bibliográfico, foram feitas pesquisas em diversos repositórios, combinando palavras-chave do universo da visualização e recuperação de informações em bases generalistas, como *Springer*, *Scopus* e *Web Of Science*; em bases de especializadas da área da Ciência da Informação, como ASIS&T e IBICT, repositório de teses e mestrados do IBICT, BRAPCI e na revista *Scientometrics*, assim como em bases de interesse da Ciência da Computação, destacando-se, como principais fontes, a biblioteca digital da Association for Computer Machinery (ACM) e o periódico *IEEE Transactions on Visualization and Computer Graphics*.

Foram usadas como ponto de partida as expressões derivadas das combinações dos termos de interesse, nos idiomas português e inglês:

- Visualização da informação / Information Visualization
- Recuperação da Informação / Information Retrieval

Foi uma busca difícil de ser realizada, pois a intersecção das palavras leva a diversas interpretações distintas que fogem ao objetivo deste trabalho. Dentre os artigos levantados, pôde-se observar tipos de interpretação diferentes para Visualização da Informação, o que demandou que esse conceito fosse mais bem definido, como veremos no referencial teórico. Em resumo, foram identificados estes enfoques:

- a) recuperação de objetos multimídia (busca de informações em acervos de fotografias, iconografia, arte, vídeos e outros formatos multimídia);

- b) uso da visualização em infografia (em especial na área de comunicação e design);
- c) disposição visual de informações na tela, ligadas aos estudos de interface e arquitetura da informação;
- d) visualização da informação como instrumento de análise exploratória, aplicada nas mais diversas áreas do conhecimento. A visualização não é o fim, mas o meio pelo qual se procuram respostas aos problemas de pesquisa; é usada para aferir resultados e exploração de dados. Nesse sentido, é muito recorrente encontrar trabalhos que utilizam a palavra-chave, mas não tratam diretamente do assunto, pontuando a aplicação de visualizações como parte da metodologia para encontrar resultados; e
- e) história e atributos da visualização da informação, assim como tipos de gráficos e modelos de aplicação.

A intenção deste trabalho foi encontrar a bibliografia que detalhasse especificamente a visualização da informação como foco de estudo e suas aplicações na compreensão de conjuntos de documentos. Dessa forma, privilegiou-se a literatura do **enfoque E**, que contempla a história, atributos, tipos e modelos de aplicação. O **enfoque D** também foi utilizado na demonstração de exemplos práticos aplicados na manipulação de documentos.

Como forma de concretizar os achados e questionamentos do levantamento bibliográfico, foi proposto um exercício empírico, baseado na metodologia de Estudos de Caso. Segundo Yin (2003), o estudo de caso é considerado uma importante estratégia metodológica para a pesquisa em Ciências Humanas, pois permite ao investigador um aprofundamento em relação ao fenômeno estudado, revelando nuances difíceis de serem enxergadas “a olho nu”. O autor afirma que o estudo de caso favorece uma visão holística sobre os acontecimentos da vida real, com destaque para seu caráter de investigação empírica de fenômenos contemporâneos.

Os critérios utilizados no estudo de caso são baseados nas conclusões derivadas do referencial teórico e, por isso, é conveniente ao leitor que sejam apresentados junto ao detalhamento da metodologia de estudo de caso adotada neste trabalho, apresentada na introdução do capítulo 3.

2. Referencial teórico

2.1. O contexto Big Data

O Big Data é um grande vetor de transformação na forma como lidamos com os crescentes volumes de informação, e sua popularização, na ciência e no mercado, acontece a partir de 2013. Os avanços só foram possíveis com a disseminação de métodos analíticos, mas também na capacidade das máquinas em armazenar e processar informações.

Desde esse momento, o termo “Big Data” tornou-se onipresente na prática e na pesquisa em Tecnologias da Informação e ganhou espaço rapidamente em outras disciplinas, como Sociologia, Medicina, Biologia, Economia, Administração e Ciência da Informação. Com a adoção do termo em maior escala, é comum se deparar com uma variedade de definições, e seu uso é, muitas vezes, confuso e inadequado.

Em busca de uma definição formal, De Mauro, Greco e Grimaldi (2016) coletaram 1.437 artigos científicos que mencionam “Big Data” em seus títulos, resumos ou palavras-chave e os analisaram a partir dos conjuntos de termos associados mais comuns³. Os autores identificaram quatro grandes atributos, que denominam “temas essenciais”, capazes de caracterizar o Big Data: informação, tecnologia, métodos analíticos e impacto.

O primeiro atributo, **informação**, é de especial interesse, pois é nele que os autores identificam a relação direta entre Big Data e bibliotecas digitais, aspecto importante de estudos realizados no domínio da Ciência da Informação. Segundo os autores,

Quando a magnitude de uma biblioteca digital em termos de volume, velocidade e variedade de conteúdo, base de usuários ou qualquer outro aspecto requer “tecnologias ou abordagens especializadas”, nós

³ O experimento consistiu na mineração de base de dados da *Scopus*, popular repositório científico de artigos que tivessem o termo “Big Data” no título, resumo ou palavras-chave. Depois de uma limpeza de dados incompletos, os autores chegaram a um conjunto de 1.437 artigos que foram analisados com relação à frequência de associações de termos que, ao final, foram a base para a definição dos temas descritos por eles (DE MAURO; GRECO; GRIMALDI, 2016).

entramos no domínio das VLDL [do inglês *Very Large Digital Library*].
(DE MAURO; GRECO; GRIMALDI, 2016, p. 125, tradução nossa)⁴

Os autores afirmam também que a hierarquia dados-informação-conhecimento-sabedoria (DICS) é uma visão alternativa que confere à informação *status* de dado estruturado de forma que seja útil a algum propósito. A identificação desse propósito é um desafio comum às aplicações de Big Data e, para eles, “é a informação, e não o dado, o combustível fundamental do fenômeno Big Data” (DE MAURO; GRECO; GRIMALDI, 2016, p. 124, tradução nossa)⁵.

O atributo **tecnologia** indica a frequente menção a termos específicos sobre questões tecnológicas. O desafio de trabalhar com grandes volumes de dados apresenta pré-requisitos importantes, como capacidade e velocidade de armazenamento e processamento. Inclui soluções que vão desde o hardware até softwares, e incluem os avanços em computação em nuvem, o aprimoramento dos modelos usando Hadoop e MapReduce e também a infraestrutura de transmissão e comunicação.

Métodos analíticos, o terceiro atributo destacado, diz respeito às técnicas de processamento. Manyika et al. (2011) mapearam, à época, os procedimentos analíticos mais comuns aplicados ao Big Data, que incluem: análises de grupos (*cluster analysis*), algoritmos genéticos, processamento de linguagem natural, Aprendizado de Máquina, redes neurais, modelos preditivos, modelos de regressão, análise de redes sociais, análise de sentimento, processamento de sinais e imagens e visualização de dados.

Métodos avançados necessitam de competências específicas e interdisciplinares, e novos perfis de pesquisadores e de profissionais devem acumular conhecimentos de diversas disciplinas, como Matemática, Estatística e Ciência da Computação, somados a outros domínios do conhecimento, como Sociologia, Comunicação, Medicina, Biologia, entre outros. Esses profissionais são os cientistas de dados, cuja

⁴ “When the magnitude of a DL in terms of volume, velocity and variety of content, user base or any other aspect requires ‘specialized technologies or approaches’, we enter the domain of very large DLs (VLDLs).” (DE MAURO; GRECO; GRIMALDI, 2016, p. 125)

⁵ “[...] we can conclude that information, not data, is the fundamental fuel of the current Big Data phenomenon.” (DE MAURO; GRECO; GRIMALDI, 2016, p. 124)

formação ainda é pouco estruturada e necessita de novos métodos e disciplinas de ensino.

O quarto e último atributo, **impacto**, descreve como o Big Data pode influenciar nossas vidas. Suas aplicações são adaptáveis aos requisitos inerentes a diferentes domínios científicos e usos industriais e comerciais. Otimização de processos, novos produtos e aplicações começam a fazer parte do cotidiano, mudando a forma como nos comunicamos, relacionamos, aprendemos e organizamos nossas vidas.

Questões éticas emergem e se tornam inevitáveis. Podem ser destacadas aquelas ligadas à privacidade, à disponibilização de dados de usuários de serviços digitais para uso comercial ou de recomendação de conteúdo e publicidade, interferindo no tipo de informação que cada usuário consome. Esse tipo de discussão tem sido cada vez mais abordado, e medidas diversas começam a ser tomadas para tornar os dados pessoais mais protegidos e menos disponíveis ao uso comercial ou de comunicação e política. O recente Regulamento Geral de Proteção de Dados (GDPR, da sigla em inglês)⁶, um processo iniciado em 2012 e que entrou em vigor em maio de 2018, obrigou empresas do mundo todo a adaptar seu funcionamento de forma a dar ao usuário maior controle sobre os dados que gera, o que pode ou não ser compartilhado ou armazenado. No Brasil, a pauta é abordada na Lei Geral de Proteção e Dados Pessoais, sancionada em agosto de 2018 e que deve entrar em vigor em 2020⁷.

A apropriação de dados para uso político e eleitoral é outro tema de debate, evidenciado pelos recentes casos envolvendo a empresa Cambridge Analytica a campanhas ligadas à extrema-direita, com atuação direta no Brexit e na eleição de Donald Trump⁸. Na disputa presidencial do Brasil de 2018, o controle de diversas contas e grupos do aplicativo WhatsApp fez com que notícias falsas e mensagens de

⁶ O site de promoção do General Data Protection Regulation está disponível em: <<https://eugdpr.org/>>. Acesso em: nov. 2018.

⁷ O texto completo da Lei n. 13.709, de 14 de agosto de 2018, pode ser acessado em: <http://www.planalto.gov.br/ccivil_03/_Ato2015-2018/2018/Lei/L13709.htm>. Acesso em: ago. 2019.

⁸ Mais informações sobre os casos e o possível envolvimento com o Brasil podem ser encontradas na matéria “O que a Cambridge Analytica, que ajudou a eleger Trump, quer fazer no Brasil”, de Paulo Flores, publicada no *Nexo Jornal* em 8 de dezembro de 2017. Disponível em: <<https://www.nexojornal.com.br/expresso/2017/12/08/O-que-a-Cambridge-Analytica-que-ajudou-a-eleger-Trump-quer-fazer-no-Brasil>>. Acesso em: 29 nov. 2018.

ódio fossem espalhadas de forma ágil e capilarizada, influenciando, dessa forma, os resultados eleitorais⁹.

A partir dos quatro atributos levantados, De Mauro, Greco e Grimaldi (2016, p. 131, tradução nossa) propõem uma definição formal: “Big Data é o ativo de Informação caracterizado por seu Volume, Velocidade e Variedade e tem como requisitos Tecnologias e Métodos Analíticos para que possa ser revertido em um Valor”¹⁰. Nesta definição, “Volume”, “Velocidade” e “Variedade” descrevem as características da informação. “Tecnologia” e “Métodos Analíticos” descrevem os requisitos necessário para o uso da informação; “Valor” descreve a transformação da informação em descobertas que podem criar valor econômico e científico para empresas e para a sociedade.

Pode-se interpretar, portanto, que o Big Data é um *contexto* em que esses quatro atributos (informação, tecnologia, métodos analíticos e impacto) compõem o cenário para a proposta deste trabalho. O Big Data proporciona uma nova forma de compreensão de grandes ou pequenos volumes de dados através de uma perspectiva analítica que propõe coleta, representação e interpretação de dados aplicados a diversos domínios do conhecimento.

O cenário posiciona e ajuda a entender o objeto de pesquisa deste trabalho, a Visualização da Informação, por vezes referenciada como Visualização de Dados (CHEN, 2006), como um método analítico que envolve procedimentos estatísticos e cognitivos e capaz de tornar ainda mais claros o estudo, a avaliação e a observação de padrões em grandes volumes de informações.

⁹ “Entenda as irregularidades envolvendo uso do WhatsApp na eleição”, *Folha de S.Paulo*, São Paulo, 22 out. 2018. Disponível em: <<https://www1.folha.uol.com.br/poder/2018/10/entenda-as-irregularidades-envolvendo-uso-do-whatsapp-na-eleicao.shtml>>. Acesso em: 29 nov. 2018.

¹⁰ “Big Data is the Information asset characterised by such a High Volume, Velocity and Variety to require specific Technology and Analytical Methods for its transformation into Value.” (DE MAURO; GRECO; GRIMALDI, 2016, p. 131)

2.2. As lentes da Ciência da Informação

O item anterior teve como objetivo identificar o cenário atual do Big Data, suas principais características e origem em grandes bases de dados, como bibliotecas digitais ou, mais especificamente, Bibliotecas Digitais Muito Grandes (*Very Large Digital Libraries* – VLDL) (DE MAURO; GRECO; GRIMALDI, 2016).

Essa abordagem facilita o paralelo com a Recuperação da Informação como parte fundamental da própria Ciência da Informação. Este capítulo pretende localizar o campo da Visualização da Informação (VI) nos processos que integram a Recuperação da Informação, os quais envolvem bases de documentos, sistemas e interação de usuários. Nesse sentido, é proposta uma leitura dos paradigmas da informação descritos por Rafael Capurro (2010) como físico, cognitivo e social, e sua aplicação prática para a análise das dimensões envolvidas na VI.

Da mesma forma, os atributos do Big Data, descritos no capítulo anterior, são refletidos na definição dos paradigmas, nos quais a informação é a principal fonte disponível, e exigem recursos tecnológicos e metodológicos e esforços de automação, de viés físico; recorrem a processos cognitivos para facilitar a compreensão humana, relativo ao paradigma cognitivo; para então construir valor, associado ao paradigma social, que apresenta impactos diretos na vida e na forma como se organiza e evolui a sociedade, trazendo consigo profundas discussões éticas.

2.2.1. Perspectiva histórica

Existem diversos exercícios de definição da área da Ciência da Informação. Uma definição clássica de Ciência da Informação é: tem como objeto a produção, seleção, organização, interpretação, armazenamento, recuperação, disseminação, transformação e uso da informação (GRIFFITH, 1980 apud CAPURRO, 2003).

Columbié (2005) atribui a origem do campo à Documentação e à Recuperação da Informação e atribui quatro publicações científicas fundacionais da área:

- a) o *Tratado da documentação* de Paul Otlet, de 1935, que trabalha a construção de modelos de organização para fins de recuperação através de processos de documentação;
- b) a Teoria Matemática da Comunicação, de Shannon e Weaver, de 1948, que cunha o termo “informação”, e modela os processos na tríade emissor, receptor e ruído existente em sua transmissão;
- c) o artigo “As we may think”, publicado em 1945 por Vannevar Bush, que explora os impactos para a ciência e a sociedade da transmissão de informações, através da idealização de um dispositivo, o Memex, que armazena, organiza e torna possível o compartilhamento de conhecimentos e documentos;
- d) por fim, “A recuperação da informação” de Calvin Mooers, texto publicado em 1950, no qual é cunhado o termo “Recuperação da Informação” baseado em modelos computacionais de organização e posterior recuperação de documentos.

A data de 1958 é assinalada como um dos marcos na formalização da nova disciplina, quando foi fundado, no Reino Unido, o Institute of Information Scientists (IIS). Mais tarde, Borko (1968) traduz objetivos e aspectos práticos do termo com a publicação do artigo “Information science, what is it?”. É um artigo curto, de caráter exploratório, com apenas duas referências, mas que sistematiza as discussões sobre a definição da área.

Através da análise de 655 projetos publicados na edição 14 do *Current Research and Development in Scientific Documentation*, Borko (1968, p. 9, tradução nossa) elenca nove categorias de interesse da área:

1. **Necessidades e usos de informação** (estudos comportamentais do usuário; estudos de citação; padrões de comunicação);
2. **Criação e reprodução de documentos** (criação assistida por computador; microformatos; registro e armazenamento; escrita e edição);
3. **Análise linguística** (linguagem computacional; lexicografia; processamento de linguagem natural em texto; psicolinguística; análise semântica);
4. **Tradução** (tradução por máquinas; facilitadores de tradução);

5. **Elaboração de resumos, classificação, codificação e indexação** (sistemas de classificação e indexação; análise de conteúdo; classificação assistida por máquinas; extração e indexação; estudos de vocabulários);
6. **Desenho de sistemas** (centros de informação; Recuperação da Informação; mecanização de operações bibliotecárias; disseminação seletiva de informações);
7. **Análise e avaliação** (estudos comparativos; qualidade de indexação, modelagem; testes de métodos e *performance*; qualidade de tradução);
8. **Reconhecimento de padrões** (processamento de imagens; análise de fala);
9. **Sistemas adaptativos** (Inteligência Artificial; resolução de problemas; sistema auto-organizados).¹¹

Os fatores apontados por Borko (1968) são desafios que se mantêm constantes desde então, ainda que a complexidade cresça de forma exponencial no contexto Big Data. Da mesma forma, a categorização proposta para dividir os trabalhos analisados reflete um campo bastante amplo e de muitas aplicações, que também estão no centro dos interesses de estudos atuais, como no campo do Aprendizado de Máquina e áreas correlatas.

Os anos anteriores à consolidação da Ciência da Informação como campo do conhecimento representam, para Columbié (2005), um período fundacional da Ciência da Informação que tem força até a década de 1970. Nesse período, pode-se dizer que

¹¹ 1. Information Needs and Uses (Behavioral studies of users; citation studies; communication patterns; literature use studies)
 2. Document Creation and Copying (Computer-assisted composition; microforms; recording and storing; writing and editing)
 3. Language Analysis (Computational linguistics; lexicography; natural language (text) processing; psycholinguistics; semantic analysis)
 4. Translation (Machine translation; translation aids)
 5. Abstracting, Classification, Coding and Indexing (Classification and indexing systems; content analysis; machine-aided classification, extracting and indexing; vocabulary studies)
 6. System Design (Information centers; information retrieval; mechanization of library operations; selective dissemination of information).
 7. Analysis and Evaluation (Comparative studies; indexing quality; modeling; test methods and performance measures; translation quality.)
 8. Pattern Recognition (Image processing; speech analysis).
 9. Adaptive System (Artificial intelligence; automata; problem solving; self-organizing systems). (BORKO, 1968, p. 4)

havia a preocupação com a Recuperação da Informação do ponto de vista do sistema e de sua eficiência. Não à toa, as medidas de precisão e revocação surgem como modelos de avaliação quantitativos. Em fins dos anos 1970, começam a ganhar força as interpretações sobre o sucesso dos sistemas sob o ponto de vista do usuário, dando início a uma fase crítica à tradição voltada ao sistema, e predomina um ponto de vista cognitivo, em que é analisada a interpretação da informação pelo indivíduo, influenciado por questões da Psicologia. À época, os estudos de VI se voltam também aos aspectos da cognição e da Psicologia.

O percurso histórico da Visualização da Informação tem suas raízes diretas na Ciência da Computação e ganha *status* de campo de estudos mais tarde, ao fim dos anos 1980 e início de 1990 (TUFTE, 1983; CARD; MACKINLAY; SHNEIDERMAN, 1999), coincidindo, portanto, com o período crítico da Ciência da Informação, descrito por Columbié (2005). O avanço da computação e das tecnologias da informação em geral somado à produção de bases de dados cada vez maiores criaram condições para a aplicação recorrente da Visualização da Informação.

2.2.2. Paradigmas da Ciência da Informação

Outra abordagem aproxima a VI do período crítico da CI, se analisados pelos paradigmas de Rafael Capurro (2003). Em parte, a gênese de cada paradigma é encontrada na visão histórica de Radamés, principalmente a relação entre a fase fundacional e o paradigma físico, e a fase crítica, relacionada ao paradigma cognitivo. O paradigma social é percebido também na fase crítica, em que são discutidas questões sobre o contexto do usuário e os impactos relativos à ética do uso da informação. A seguir, são detalhados os paradigmas físico, cognitivo e social e uma interpretação paralela da Visualização da Informação nestes contextos.

a) Paradigma físico

O paradigma físico tem relação direta com a fase fundacional apontada por Columbié (2005), em que a Teoria da Informação de Shannon e Weaver é precursora. Segundo Capurro (2003, p. 7-8),

Essa teoria [de Shannon e Weaver], tomada como modelo na ciência da informação, implica numa analogia entre a veiculação física de um sinal e a transmissão de uma mensagem, cujos aspectos semânticos e pragmáticos intimamente relacionados ao uso diário do termo informação são explicitamente descartados. [...] Torna-se evidente que, no campo da ciência da informação, o que esse paradigma (Físico) exclui é nada menos que o papel ativo do sujeito cognoscente ou, de forma mais concreta, do usuário, no processo de recuperação da informação científica, em particular, bem como em todo processo informativo e comunicativo, em geral.

A esse paradigma, portanto, podemos atribuir todos os processos computacionais que não necessitam de associações semânticas. São, por exemplo, questões tecnológicas em que o que está em jogo são aspectos práticos, como a capacidade de armazenamento, processamento, assim como métodos estatísticos que não consideram o significado. A ação semântica, sob esse paradigma, relaciona-se ao cumprimento de ordens por um computador.

A Visualização da Informação é calcada em métodos estatísticos, como o modelo proposto por Card, Mackinlay e Shneiderman (1999), que ilustra a sequência de transformação de dados brutos até sua forma visual, passando por fases de tratamento e transformações de dados e o mapeamento visual necessário para apresentação gráfica. Adicionam-se questões de infraestrutura à capacidade de armazenamento e processamento, que não dependem de qualquer dado semântico. Essas características podem ser compreendidas como pertencentes ao paradigma físico.

Uma forma de avaliação de visualização de dados que se encaixa sob este paradigma, é a medida Data-ink, definida por Tufte (1983), que registra a proporção de tinta utilizada pela quantidade de informação demonstrada na visualização. Esse método traz uma visão de eficiência, mas não mede a capacidade de compreensão do usuário ou a facilidade com que se apreende a informação contida no gráfico.

Outro aspecto da VI que a aproxima do paradigma físico são os experimentos ligados à anatomia da visão humana explorados por Card, Mackinlay e Shneiderman (1999), que estudam a velocidade com que o aparelho humano da visão compreende uma visualização em contraponto a textos ou outras formas de representação. Consideram

as limitações físicas inerentes ao sistema da visão, calculando a velocidade com que vemos as ondas de luz. Novamente estão sendo considerados aspectos físicos, que não influem na compreensão semântica da visualização e atribuição de significado e valor.

A perspectiva semântica e a compreensão do usuário em seu contexto são abordadas nos segundo e terceiro paradigmas.

b) Paradigma cognitivo

Capurro (2003) reforça que a visão de paradigmas deve evitar a sequência cronológica como definidora. No entanto, em um paralelo com a visão histórica de Columbié (2005), a corrente ligada ao paradigma cognitivo evidencia-se no período considerado de fase crítica da Ciência da Informação.

Se o paradigma físico lidava com a informação do ponto de vista do sistema e sua eficiência, sem considerar a semântica ou a apreensão do usuário, o paradigma cognitivo traz concretamente questões relacionadas ao *indivíduo* que busca informação. Um exemplo expressivo que evidencia a diferença com o paradigma físico é a comparação de trabalhos fundacionais: a teoria da informação de Shannon e Weaver, de 1948, menciona o papel de um **receptor**, enquanto Belkin, Oddy e Brooks (1982) referem-se a um sujeito cognoscente e, mais tarde, ao *usuário* – a informação, à qual chamam “informação objetiva”, é percebida apenas nos *espaços mentais*.

Os autores, no contexto do paradigma cognitivo, definem a teoria do “estado anômalo do conhecimento” (do inglês *anomalous state of knowledge* – ASK) e a concepção do “espaço problema”. As duas concepções versam sobre a necessidade de informação quando o conhecimento ao alcance do usuário, para resolver o problema, não é suficiente (CAPURRO, 2003, p. 9).

Peter Ingwersen (1992) utiliza elementos dessas teorias para construir um modelo de Recuperação da Informação que contempla todo o processo comunicacional da informação, sob o ponto de vista do usuário e sua interação em sistemas, digitais/computacionais ou dispositivos, como bibliotecas e museus. O autor aponta ainda outros sujeitos cognoscentes no processo de Recuperação da Informação e

seus respectivos “*aboutness*”, expressão que usa para se referir ao conjunto de experiências e repertório dos papéis exercidos pela figura do autor, do indexador e então do usuário no processo de Recuperação (INGWERSEN, 1992). São fortes fatores de incertezas que permeiam a Recuperação da Informação.

Modelos de avaliação e *performance* tornam-se mais subjetivos. Pode-se pontuar o conceito de Relevância, que Ingwersen denomina “matéria negra” da Ciência da Informação. Saracevic (2007) faz uma detalhada revisão da literatura sobre o tema. Além disso, os experimentos e pesquisas com usuários tornam-se mais frequentes com a adoção de heurísticas e testes.

O reflexo do paradigma cognitivo para a Visualização da Informação tem relação com os aspectos da Psicologia na compreensão de como os usuários percebem as informações contidas num gráfico.

Um exemplo propício para a compreensão da avaliação da VI é trabalhado no conceito de ergonomia cognitiva, como tratam as pesquisadoras cubanas Deborah Ponjuan e Gloria Dante (2010). As autoras utilizam a definição da International Ergonomics Association (IEA), que se refere à ergonomia cognitiva como:

[...] a área da Ergonomia que estuda os processos mentais, tais como percepção, cognição, memória, razão, emoção e como afetam as interações entre pessoas e os produtos dos sistemas, e ambientes. Seus tópicos de estudo mais relevantes incluem: carga de trabalho mental, cognição, tomada de decisões, análise de desempenho, interação humano-computador, fiabilidade humana, estresse de trabalho, diferenças culturais, atitudes, prazer e motivação. (INTERNATIONAL ERGONOMICS ASSOCIATION, 2003 apud PONJUAN; DANTE, 2010, p. 23, tradução nossa)¹²

¹² “[...] el área de la Ergonomía que estudia los procesos mentales, tales como percepción, cognición, memoria, razonamiento y emoción, y cómo afectan las interacciones entre las personas y los productos, los sistemas, y los ambientes. Sus tópicos de estudio más relevantes incluyen: carga de trabajo mental, cognición, toma de decisiones, desempeño experto, interacción humano-computadora, fiabilidad humana, estrés de trabajo, entrenamiento, diferencias culturales, actitudes, placer y motivación.” (INTERNATIONAL ERGONOMICS ASSOCIATION, 2003 apud PONJUAN; DANTE, 2010, p. 23)

Avaliações que buscam precisão, como o Data-ink, atribuíveis ao paradigma físico, perdem espaço para avaliações heurísticas ligadas à compreensão e utilização final pelo usuário, com métodos de avaliação mais subjetivos, exemplificados pelos aspectos da ergonomia cognitiva aplicada à VI.

c) Paradigma social

Capurro (2003) diz que os trabalhos de Ingwersen e Vakkari, assim como outros que adaptam os modelos mentalistas ao estudo da Recuperação da Informação, podem ser considerados na intersecção entre os paradigmas cognitivo e social, pois também tratam o contexto em que o usuário está envolvido em sua busca de informação.

Para Bernd Frohmann (2008), o paradigma cognitivo é idealista e, principalmente, associal. Considera a informação separada do usuário, em um mundo numênico¹³, ideal e inatingível, ou coloca o usuário no centro das discussões, deixando de lado os condicionamentos sociais e materiais.

Capurro (2003) também se posiciona sob o paradigma social e chama a atenção para as questões éticas que envolvem todas as áreas que lidam com a informação, citando a ética da robótica, ética dos dados, entre outras (CAPURRO, 2003; 2007). Frohmann (2008) também chama a atenção para os impactos sociais da disponibilização e comercialização de dados e informações e usa como exemplo a base de fichas de presidiários. Legalmente, nos Estados Unidos, pessoas que já cumpriram pena há mais de um determinado período de anos têm direito a anular o registro de seu histórico. No entanto, por ser uma base de dados aberta, empresas colecionam os dados históricos e os disponibilizam no mercado, onde decisões podem ser tomadas baseadas em um registro que já deveria ter sido “esquecido”, diminuindo as chances de alguém conseguir emprego, empréstimos e outros serviços que considerem o histórico de antecedentes criminais.

¹³ Numênica se refere às colocações de Immanuel Kant, que, em sua *Crítica da razão pura* (1781), divide o significado das coisas em dois mundos: o fenomênico, em que as coisas são assim como nos parecem (como as “percebemos”). O mundo numênico, em oposição, trata as coisas num modo ideal, das coisas tal como realmente são. Esse mundo, pela mesma razão, não é acessível, pois está em plano “transcendental” (MORUJÃO, 2013).

Ao nos referirmos a dados, e dentre eles, a documentos ou informações, é preciso ter em mente que sua construção está condicionada ao contexto social em que é produzido e as causa e efeitos posteriores à sua aplicação. Podem trazer benefícios claros para o usuário ou instituição (empresas, organizações etc.), mas podem impactar de forma negativa aspectos sociais e culturais.

A Visualização da Informação, ou visualização de dados, por sua função de potencializar a cognição humana e tornar mais fácil a compreensão de grandes volumes de dados, pode ser considerada pelas suas duas abordagens, tanto como fruto de uma construção social/organizacional quanto como potencial de impacto na tomada de decisões que implicam efeitos sobre a mesma sociedade/organização.

Formatos de visualização com barras, linhas e pizzas são populares e usados de maneira mais recorrente, porém outros formatos precisam de grau de construção de conhecimento maior para se saber ler as informações neles representadas. O estudo de xenográficos proposto por Maarten Lambrechts, por exemplo, reúne gráficos diversos em seu website¹⁴ e avalia níveis de dificuldade para um usuário humano analisar e interpretar informações. Estudos atuais refletem também a mudança de suporte e funcionalidades que podem ser exploradas na interação entre diferentes dispositivos. Uma visualização deve funcionar tanto horizontal como verticalmente se considerarmos as telas de celulares e tablets. Portanto, é preciso que o usuário possua repertório visual e que ele aprenda a ler as informações representadas em uma visualização. Isso geralmente acontece dentro do que se chama Regimes de Informação.

Bezerra et al. (2016) afirmam que, dentro do contexto de Ciência da Informação, o regime de informação se configura como uma formação social conjunta de elementos em rede – como atores sociais (sujeitos, dispositivos e tecnologias), regras de poder, a organização e a gestão política da informação que se operacionalizam em práticas sociais com produtos e serviços. Os autores constroem a interpretação sobre Regimes da Informação que vêm levantando o debate sobre as ações de informação que compõem o regime de informação em suas diversas formações (instituições públicas ou privadas) sociais.

¹⁴ Disponível em: <<https://xeno.graphics/>>. Acesso em: 27 jan. 2019.

Toma-se por pressuposto que o repertório visual de gráficos está inserido numa realidade institucional e que as visualizações têm efeitos variados, de acordo com o Regime de Informação em que está inserido.

Ainda sob o paradigma social, além da construção social, é necessário considerar o efeito que a visualização da informação pode ter na sociedade e na cultura, reforçado ainda mais no contexto Big Data pela caracterização de volume, diversidade, velocidade de produção e processamento e impacto na sociedade – visando benefícios ou amplificando a propagação de informações erradas, falsas ou manipuladas.

Uma visualização reflete a seleção ou transformação de dados, e todas as questões éticas e de impacto inerentes ao paradigma social têm reflexos nos efeitos da visualização.

Tufte (1983) já alertava sobre os erros que comprometem a leitura correta dos dados, como falhas em proporções, falta de contrastes, ou mesmo a manipulação de escala do gráfico, que evidenciam ou suavizam picos e padrões que se deseja reforçar ou eliminar.

Como forma de ilustrar a cronologia dos principais autores citados neste trabalho, a figura 05 demonstra, de forma esquemática, o paralelo entre os campos da Ciência da Informação e Visualização da Informação. Os paradigmas físico, cognitivo e social são representados em três cores com transições em degradê, indicando que a compreensão de seus conceitos não deve ser meramente cronológica, ou demarcada por datas específicas. É importante ressaltar que seus efeitos são cumulativos, ou seja, não devem ser encarados como fases, mas como sobreposições e inclusão de novas camadas de complexidade ao se analisar o impacto e as formas de produção e percepção da informação. Por motivos de otimização de espaço da imagem, as primeiras décadas demonstradas não seguem escala exata e a disposição de autores indicam os períodos em que as obras analisadas neste trabalho foram publicadas, sem compromisso de apontar datas específicas.

Figura 5: Cronologia de autores citados no trabalho sob as lentes da Ciência da Informação e os paradigmas físico, cognitivo e social

2.3. Documento e organização da informação

Ao apontar a relevância das novas mudanças no campo da CI, é necessário lançar vistas ao próprio conceito de informação. Para isso, utilizamos como referencial a abordagem dos neodocumentalistas, movimento que resgata a discussão sobre documento e a documentação, com base nos estudos de Paul Otlet, de 1934, e Suzanne Briet, de 1951. Adotamos o termo “documento” para nos referirmos a unidades de informação, utilizado como sinônimo nas bibliotecas e centros de documentação etc., ou ainda objetos informacionais, como tratam Svenonius (2000) e Gilliland (2008). No âmbito deste trabalho, documento é definido como a representação da informação, em especial no formato digital.

Ao tratarmos de recuperação e visualização de informações, estamos, portanto, discorrendo sobre recuperação e visualização de “documentos”.

Bernd Frohmann (2008, p. 21-22), considerado parte do movimento neodocumentalista, aborda o documento como a parte material da informação:

Mas se “documento” nomeia a materialidade da informação, e se a materialidade é importante para os aspectos públicos e sociais da informação, então os estudos da documentação tornam-se importantes para os estudos da informação. A documentação se torna o meio de materialização da informação. Estudar a documentação é estudar as consequências e os efeitos da materialidade da informação.

O autor compara a materialidade de documentos da mesma forma à que Michel Foucault atribuiu materialidade aos seus “enunciados” – ela é medida pela “massa, inércia e resistência” (FROHMANN, 2008, p. 22), que são recursos teóricos para compreender a complexidade de efeito potencial de um documento, seu impacto social e poder de relação.

O sentido é amplo e dá espaço para uma interpretação mais diversa e extremamente fragmentada de documento. Nossas mensagens trocadas via celulares ou rede sociais são *enunciados* em potencial, documentos que têm a força individual de um “*statement*”, mas são compostos por um punhado de palavras que, sem contexto, podem não ter significado, mas em conjunto passam a ser *fontes de massa, energia*

e *força* (FROHMANN, 2008). A metáfora das propriedades físicas deixa clara a visão derivada do paradigma social, em que os parâmetros que definem o que é ou não um documento têm consequências sociais, políticas e culturais.

Capurro (2007, p. 20) se refere à CI no sentido de uma Teoria da Mensagem:

A “mensagem” é uma “oferta de sentido”, enquanto “informação” se refere à seleção em um sistema e a “compreensão” à possibilidade que o receptor integre a seleção com seu pré-conhecimento, que está constantemente aberto à revisão, ou seja, à comunicação de acordo com as intenções do emissor. O receptor se transforma, cada vez, também em um emissor.

Tanto a relação que Frohmann (2008) estabelece entre documento e enunciado quanto a abordagem da CI por uma teoria da mensagem compõem indícios de uma visão fragmentada da informação, ou seja, o sentido emerge de um conjunto, e não apenas de uma unidade de documento – os registros de uma conversa escrita só possuem sentido se as mensagens trocadas são analisadas em conjunto. Em se considerando o contexto Big Data, é preciso reforçar também que estamos tratando de documentos digitais: bases de dados científicas, páginas da web, arquivos pdf, imagens, áudios, vídeos, fichas de produtos, e-mails, mensagens trocadas via rede social, entre tantas outras possibilidades de informações registradas em meios computacionais.

Ao tratar de documentos digitais, Frohmann (2008) se refere à capacidade de processá-los em grau e escala únicos na história. Uma abordagem que pode ser compreendida no contexto Big Data.

Svenonius (2000) fala a respeito da dificuldade de definir os limites do documento digital e o descreve como instável, dinâmico e suscetível a constantes mudanças, o que os torna mais difíceis de organizar e, conseqüentemente, de visualizar ou de recuperar.

Resgatando a discussão sobre documento como materialidade da informação, e dentro de um arcabouço histórico da Teoria da Documentação de Lund (2009), pode-se discorrer sobre duas de suas características. Smit (2012) as coloca como

características centrais para que ocorra a informação: o registro e a institucionalização.

Na lógica de transformação contemporânea, em que observamos abundância, variedade e capacidade de processamento para que utilizemos o meio digital cada vez mais para qualquer ação de comunicação, a criação do documento, da mensagem, do enunciado já é, por si só, um registro. A informação *nasce registrada*.

Da mesma forma, quando usamos aplicações on-line proprietárias, como as redes sociais e programas de edição on-line, como o Google Drive e outras aplicações que armazenam dados pessoais de usuários na nuvem, sujeitamos documentos pessoais a formatos, políticas, metadados e mesmo a disponibilidade de acesso por terceiros aos termos de uso e políticas da instituição detentora da ferramenta.

Acesso é outra questão relevante. As informações armazenadas, cada vez mais conectadas, abrem portas para apropriações diversas, seja pela simples disponibilidade, por formas de invasão ou de comercialização de conjuntos de dados. A mesma informação pode estar armazenada em servidores diferentes, registrada de formas distintas, relacionada a outros conjuntos de dados, sob a administração de instituições diversas. Ainda assim, é “a mesma informação”, mas que, por conta dos milhares de contextos aos quais pode estar relacionada, tem uma capacidade infinita de interpretações.

No contexto Big Data, as informações passam por processos infinitos de registro (reprodutibilidade) e por uma *re-institucionalização*, uma vez que transitam e servem simultaneamente a diversas instituições, empresas e outras atividades legais – ou não. As dimensões do *registro* e da *instituição* são reafirmadas e amplificadas, e o caráter material da informação parece ganhar ainda mais *massa, energia e força*.

2.3.1. Estrutura

Do ponto de vista do paradigma físico, em que a informação perde seu sentido e vira pura forma, pode-se dizer que os documentos são itens, observações ou linhas em tabela ou base de dados, e cada item é composto por um conjunto de variáveis. As variáveis podem ser divididas em dois tipos: quantitativas e qualitativas.

As quantitativas são aquelas consideradas numericamente mensuráveis. Podem ser *discretas*, que atribuem a uma contagem, sempre em números inteiros – por exemplo, número de páginas de uma publicação –, ou *contínuas*, que podem assumir qualquer valor entre dois pontos – como os indicadores científicos, que atribuem notas ponderadas para a aferição de níveis de produtividade, ou indicadores de popularidade, como aqueles envolvidos no Altmetrics (FACELI et al., 2011).

Já os dados qualitativos se baseiam em qualidades, que não podem ser mensuradas numericamente. Os dados podem ser **ordinais**, que podem ser ordenados – como o grau de instrução e classes sociais –, e **nominais**, aqueles que não são ordenáveis – como palavras-chave que classificam um documento, ou números que representam categorias. Dados qualitativos também são chamados de categóricos, aqueles aos quais podem ser atribuídas categorias ou classificações. Podem ter origem na organização de dados quantitativos, por agrupamentos resultantes de algoritmos, por exemplo, e qualitativos, como a definição de gêneros literários (FACELI et al., 2011).

As variáveis, como descritas acima, podem ser diretamente relacionadas aos metadados, no contexto da CI e da organização de documentos digitais. Guilliland (2008) define que os metadados descrevem três características do documento:

- a) a primeira, **conteúdo**, é relacionada ao que faz parte do objeto e do que se trata, é intrínseca ao documento;
- b) o **contexto** indica o quem, o quê, o porquê e os aspectos associados à criação do objeto, e são extrínsecos a ele; e
- c) em terceiro, **estrutura** é o conjunto formal de associações dentro ou entre documentos individuais. Podem ser extrínsecos, intrínsecos ou ambos.

Em um contexto Big Data é comum mencionar o desafio de lidar com “dados não estruturados” ou “semiestruturados”. Na compreensão de documento utilizada, podemos elaborar que o documento “não estruturado” pode possuir conteúdo, contexto (podendo ser considerado semiestruturado), mas lhe faltam as associações formais necessárias para sua caracterização e relacionamento a outros documentos.

O que torna uma base de documentos um tipo distinto de base de dados é a presença inevitável de variáveis qualitativas e categóricas. Não obstante, essas variáveis – ou metadados – envolvem linguagem, via essencial para a compreensão de documentos

ou conjuntos de documentos por um usuário humano. Mesmo os agrupamentos de dados ou documentos derivados de algoritmos e cálculos complexos só podem ser compreendidos por um usuário humano quando tais agrupamentos podem ser descritos e comunicados.

Em resumo, para um processo bem-sucedido de recuperação e visualização de informação, é necessário ter o máximo de conhecimento sobre as características dos documentos disponíveis, sua *massa*, *energia* e *força*, assim como a composição de seus metadados, principalmente aqueles gerados através de algoritmos. Apenas com este conhecimento é possível aferir a qualidade e a consistência da base a ser visualizada e recuperada.

Não é intenção deste trabalho se aprofundar sobre o tema da linguagem, porém é importante para compreender diversas metodologias que emergiram no contexto Big Data, como aquelas que usam o processamento de linguagem natural, recurso muito utilizado em aplicações de Aprendizado de Máquina e outros métodos analíticos.

2.3.2. Organização da informação

Taylor (1999) e Alvarenga (2003) definem que a organização de informação tem um objetivo específico: possibilitar a recuperação de informação por alguém que dela necessite. Deste postulado, pode-se entender o sentido inverso: a Recuperação da Informação só é possível após a Organização da Recuperação. Mooers (1951), assim como Rijsbergen (1979) e Ingwersen (1992), trata da Organização da Informação como parte do processo de recuperação, sendo ela também uma área de interesse da CI.

Taylor (1999) e Svenonius (2000) sugerem, de forma mais sucinta, que aquilo que não pode ser organizado, não pode ser recuperado. No contexto deste trabalho, podemos sugerir que, antes da impossibilidade de recuperação, também não é possível, sem organização da informação, visualizá-la. Ou seja, tudo aquilo que não pode ser organizado, não pode ser visualizado.

Ainda no campo da CI, faz-se necessário pontuar a distinção entre Organização da Informação (OI) e Organização do Conhecimento (OC). De forma resumida, a OI tem como objeto a informação registrada passível de ser armazenada e organizada, possibilitando sua recuperação quando necessário. Já a OC se encontra no mundo dos conceitos que visam à construção de representações do conhecimento (CORRÊA, 2011; DE LIRA; ASSIS, 2015).

De Lira e Assis (2015), baseados em Café e Sales (2010), dizem que se organiza para poder compreender o mundo e se comunicar melhor, e dão como exemplo que, ao explicar a alguém o significado de determinada palavra, normalmente recorre-se à sua classificação, incluindo a palavra em um grupo maior que a define. Assim, ao explicar o significado da palavra “mesa”, seria dito que é um móvel (classe) que possui uma superfície e quatro pés. Nesse momento, organiza-se no pensamento o conceito dessa palavra, focando a classe dos móveis e excluindo outras classes (CAFÉ; SALES, 2010 apud DE LIRA; ASSIS, 2015).

No contexto dos sistemas de informação não é diferente. Os sistemas organizam um acervo, cuja função é armazenar documentos, compreendendo melhor do que tratam, dispondo de forma ordenada, a fim de que as informações neles registradas possam ser recuperadas com facilidade.

A esse processo, podemos atribuir o papel da classificação. A noção de classificar como uma construção cultural e forma de exercício de poder já bastante difundida e discutida, e parte dessa construção pode ser observada em Olson (2002), assim como em García Gutierrez (2006) ou Frohmann (2008).

Os autores apontam diversas questões sobre a classificação e os efeitos na sociedade e em nossa cultura, mas, para este trabalho, gostaríamos de propor uma divisão mais pragmática com relação à sua aplicação, tendo sempre em vista a noção de *instituição* como parte central e reguladora de um sistema de classificação que será resgatado e aprendido por usuários em uma relação cotidiana, específica a um campo do conhecimento ou não.

A classificação é a ferramenta que vai proporcionar e moldar o contexto da informação, agora fragmentada, em duas abordagens:

Classificar para recuperar: a classificação tem extrema importância na construção do conhecimento, pois é através dela que se consegue aproximar semelhantes, compreender um conjunto. “A materialidade da informação pode também ser estudada através da investigação do papel da documentação na criação de tipos ou categorias” (FROHMANN, 2008, p. 28).

A classificação possui uma vertente que surge do ato da documentação. Nesta, mais ligada às práticas da biblioteconomia e outras áreas-base da CI, a instituição define a forma de organizar seus próprios estoques de informação e, como consequência, as formas de recuperar documentos.

Classificar para compreender: nesta outra abordagem, partimos do princípio que temos uma massa de informação (ou dados) não explorada, vasta (quase infinita) e em parte desconhecida. A classificação então é uma ferramenta para ajudar a compreensão do conjunto, de dissecação do conhecimento. Pode não haver uma relação direta com a necessidade de recuperação, do ponto de vista da necessidade/desejo de informação por um usuário humano (INGWERSEN, 1992).

Um exemplo que materializa essa abordagem é o estudo da linguagem. O aprofundamento da investigação revela o quão vasta e mutante é a compreensão da língua, e, para entender suas estruturas, foi sendo feita uma desconstrução e estabelecendo-se classes. De uma unidade de texto, vamos adentrando em classificações de frases, palavras, sílabas, gramemas, fonemas. Todas categorias ou classes de elementos que compõem esse conjunto complexo e infinito da linguagem.

A língua e o estudo da linguagem são uma analogia possível para discutir os impactos do Big Data. A língua é mutável, dinâmica, variada, com incontáveis formas de agrupamento em enunciados que podem ou não fazer sentido, dependendo da comunidade. É algo vivo, comporta-se como um sistema, e sua origem e as formas como se desenvolve são inesgotáveis temas de estudo.

No universo do grande volume de dados é possível o paralelo aos fenômenos da linguística, pois apresenta características similares de possibilidades – nem sempre as combinações de gramemas, fonemas (etc.) de dados serão compreensíveis mesmo no universo da língua. É preciso encontrarmos uma sequência lógica na leitura

dos dados para que possamos encontrar as palavras, sujeitos, verbos e predicados nas informações que os dados podem nos dar.

O desafio é encontrar sequências semânticas que contribuam para a compreensão de algo ou para a realização de uma tarefa. No Big Data, pela grande presença de dados não estruturados, é preciso lançar mão de métodos que possam, de alguma forma, organizar e encontrar sentidos e interpretações ocultas na análise de grandes volumes de informações.

Uma abordagem cognitiva e tecnológica pode sugerir que cada usuário veja uma organização diferente da mesma base de documentos. Os processos de organização podem ser derivados do trabalho de um agente indexador humano, de técnicas automatizadas do mesmo trabalho ou ainda de abordagens totalmente calculadas por máquinas, de lógicas algorítmicas complexas.

Muitos atores são envolvidos no processo de classificação e impactam as diferentes formas de interpretação e mesmo do sucesso posterior na recuperação de documentos relevantes. Ao processo de classificação, sob uma perspectiva do paradigma cognitivo, Ingwersen (1992) se refere às diferentes instâncias de *aboutness*, que em português é traduzido como “de que trata o texto”, ou seja, à questão do sentido, ou significado que o documento, texto ou imagem carrega.

No contexto Big Data, onde encontramos volume e variedade de dados, a classificação – a partir do autor ou de um indexador, manual ou automática – é essencial e intrínseca à construção de estrutura e unidade de uma base de documentos, que só podem ser analisados quando contextualizados frente a outros.

Como mencionado anteriormente, dados estruturados ou não, em conjunto, podem ser contextualizados em situações e usos distintos simultaneamente. O mesmo conjunto de documentos pode ser organizado de formas diferentes e simultâneas. Os avanços em métodos automáticos de classificação, em grande parte proporcionados pelo aprimoramento da área de Processamento de Linguagem Natural (PNL) e sua aplicação em análises linguísticas, tradução automática, a produção automática de resumos, classificação, codificação e indexação, assim como avaliações e análises automáticas de *corpus* de texto (SOUZA; ALMEIDA; BARACHO, 2013).

Os métodos de automação aumentam o poder e a agilidade da classificação, permitindo que o exercício epistemográfico de Garcia Gutierrez (2006). Epistemografia, para o autor, “é uma configuração transdisciplinar que tem como objeto a organização horizontal e interativa dos conhecimentos”, e representa, neste trabalho, a possibilidade e o ato de classificar e reclassificar continuamente, simultaneamente, um mesmo conjunto de documentos. Nas bases de dados bibliográficas ou factuais, a atualização da forma de tratamento é ininterrupta. No contexto Big Data, esse ato é potencializado e acontece instantaneamente, ampliando toda a capacidade de interpretação de uma mesma informação em muitos contextos diferentes.

O risco que isso representa são as más associações. A classificação automática não resolve a questão de veracidade e relevância da categorização e posterior compreensão do conjunto de documentos pelo usuário.

A instituição tem um papel fundamental nesse processo, pois é quem define a estrutura e regula o acesso às informações, assim desenhando o limite de possibilidades de organização dos dados. A definição da estrutura de metadados, por exemplo, é evidência do controle que uma instituição pode ter nesse processo.

O desenvolvimento de metodologias cada vez mais avançadas no processamento de linguagem natural, derivadas do Aprendizado de Máquina e de outros recursos de Inteligência Artificial mais avançados, como Deep Learning e Redes Neurais, reforçam o papel inevitável do algoritmo e do processamento estatístico da informação, tanto para fins de recuperação como para uso em visualizações (WU et al., 2013).

Duas abordagens computacionais merecem ser mencionadas no contexto deste trabalho: a Indexação Automática e as Técnicas de Agrupamento (ou *clustering*, em inglês). Pode-se dizer que a diferença entre as duas abordagens é a forma de organização da informação.

O poder e a efetividade de algoritmos para a automação dos processos de classificação e descrição de documentos são bastante questionados, como observado em Kobashi e Fernandes (2008). À época da publicação do artigo referido, a infraestrutura e a capacidade de armazenamento e processamento da informação ainda não eram suficientemente desenvolvidas de forma a substituir a análise e a

descrição de um humano. Com a evolução e expansão do contexto Big Data, caracterizado pelo imenso volume, variedade e produzido em velocidade cada vez mais exponencial, a tarefa de organização se torna quase inviável sem auxílio das máquinas.

Enquanto a indexação automática analisa a linguagem natural para encaixar documentos em uma estrutura previamente pensada por humanos e no contexto das instituições (classificar para recuperar), as técnicas de agrupamento geralmente partem do caminho inverso, em que algoritmos encontram padrões e correlações entre documentos para definir classes, ou grupos, que compõem determinada coleção de documentos. Essas classes, então, são analisadas e descritas pela ação humana (classificar para compreender).

Independentemente da visualização, os sistemas desenvolvidos para apresentar coleções de documentos precisam realizar procedimentos e cálculos voltados a dois objetivos específicos: o agrupamento de documentos considerados similares; e a separação dos grupos de documentos dissimilares entre si.

Além da separação dos diferentes grupos, os sistemas devem ser capazes de efetuar cálculos que indiquem graus de similaridade entre os documentos componentes de cada grupo a ser visualizado. Esse cálculo determina a proximidade entre os itens individuais que compõem cada grupo a ser apresentado. A capacidade para realizar tais cálculos, de modo preciso, é fundamental em sistemas em que os itens da coleção são visualizados em grupos e individualmente. Uma das metodologias para relacionar documentos é representá-los de forma vetorial, disponibilizados em um espaço bi ou tridimensional (CORRÊA, 2011).

A análise de grupos aproxima objetos baseada nas informações encontradas nos dados que os descrevem e em seus relacionamentos. Nessa situação, o objetivo é obter grupos cujos objetos que o compõem sejam similares entre si, e diferentes ou não relacionados aos objetos que compõem os outros grupos. Quanto maior a similaridade dentro do grupo e maior a diferença entre os diversos grupos, melhor o agrupamento (CORRÊA, 2011).

A visualização da informação é um recurso muito utilizado no auxílio de análise de grupos, uma vez que consegue mostrar, visualmente, a distribuição entre os

elementos próximos, assim como a distância entre os agrupamentos que os distinguem.

Somado ao contexto da classificação automática, é importante acrescentar à dinâmica as condições de coleta de documentos que compõem o conjunto, como falhas ou incompletude no preenchimento de campos; definições de categorias, por humanos ou máquinas; e mesmo disponibilidade de metadados provida na coleta ou mineração de dados.

No contexto Big Data, manter a qualidade e confiabilidade dos dados é um desafio ainda maior. O tratamento dos dados é um passo essencial para qualquer sistema funcionar de forma eficaz e eficiente. Seja por métodos manuais, como a ação de um indexador humano, ou automáticos, a qualidade pode ser afetada tanto na definição inicial de quais os metadados necessários à função quanto na clara compreensão entre os usuários que vão buscar, recuperar ou interpretar o conjunto dos documentos.

A clara definição sobre o tipo de documento que está sendo considerado é um ponto fundamental para que, posteriormente, possa ser avaliada a consistência da gestão de seus metadados, seja em processos automáticos de recuperação de informações, seja no tratamento dos dados para o objetivo da visualização da informação. As visualizações são recursos que têm o poder de evidenciar falhas que, se não percebidas, podem levar o usuário a uma interpretação equivocada.

2.4. Sistemas de Recuperação da Informação

Sob o contexto do paradigma físico, Calvin Mooers, que cunhou o termo Recuperação da Informação em 1951, o define como uma disciplina que “trata dos aspectos intelectuais da descrição da informação e sua especificação para busca e de qualquer sistema, técnicas ou máquinas que são empregadas para realizar a operação” (MOOERS, 1951 apud FERNEDA, 2003, p. 11).

Nessa mesma linha, pode ser citado Shera que sob uma epistemologia social, enfatiza que o sucesso da Recuperação da Informação depende de duas frentes teóricas: uma que compreenda a linguagem e os processos comunicacionais e seus papéis na transmissão do conhecimento, e outra que compreenda os padrões do pensamento humano. Shera traduz as frentes teóricas a partir de quatro questões fundamentais (CORRÊA, 2011):

- a) Como a linguagem influencia a representação e organização cognitiva da informação nos indivíduos?
- b) Quais os efeitos que a necessidade de comunicação tem sobre a representação simbólica e/ou organização cognitiva de informação dos indivíduos?
- c) Quais padrões ou estruturas de organização cognitiva são compartilhados pelos indivíduos ou por grupos de indivíduos?
- d) Como os mecanismos de recuperação podem ser projetados para facilitar, aumentar ou imitar esses padrões naturais de organização cognitiva?

Ao longo do tempo, foram criados muitos modelos de recuperação da informação que podem ser observados tanto no campo da Ciência da Informação como no campo da Ciência da Computação.

Ingwersen (1992) apresenta um modelo simplificado que envolve os três principais componentes associados à recuperação de informação: os estoques de informação, representados ao lado esquerdo da Figura 5, envolvendo as representações de texto e as técnicas de recuperação, os usuários, suas necessidades de informação – inseridas num espaço-problema, no lado direito da figura – e, por fim, o que é denominado “funções intermediárias”, que envolve o sistema propriamente dito, sua

interface, os mecanismos de comparação, as estruturas de armazenamento, entre outros.

Figura 6 – O Sistema de Recuperação da Informação e modelo simples de interação.

Fonte: Ingwersen (1992, p. 53).

Esse modelo é oportuno, por considerar todo o processo de comunicação envolvido na informação. Nele estão implícitas questões dos paradigmas físico, referentes às técnicas; cognitivo, ao interpretar o espaço problema e as necessidades do usuário; e social, com a inclusão das influências externas expressadas, por exemplo, na compreensão do “*aboutness*” (INGWERSEN, 1992; RIJSBERGEN, 1979).

O lado superior esquerdo da Figura 5 identifica “*Author Texts*”. No modelo de Ingwersen (1992, p. 53), é descrito como Transformações do Estado de Conhecimento do Autor, que, com uma camada bibliográfica, forma uma base identificada como “*Text Representation*” e considerada o estoque de informação em si. Por Representação de Textos, podemos nos referir a “Documentos”, como estão sendo chamados neste trabalho.

No canto inferior esquerdo da mesma imagem, “*IR Technique(s) in systems*” é onde está descrita a *lógica da busca* do sistema, por exemplo, o uso da lógica booleana, muito comum em sistemas comerciais ou institucionais, que usam operadores de

proximidade, como “AND”, “NOT”, “OR”, entre tantos outros. A construção de técnicas de recuperação da informação é um estudo em constante evolução, reforçado ainda mais no contexto Big Data, onde é possível trabalhar com diversos formatos de documentos, produzidos em quantidade e velocidade cada vez maiores.

A evolução de processos computacionais cada vez mais inteligentes buscam simular a cognição humana através do uso do Aprendizado de Máquina e de outras formas de Inteligência Artificial que possibilitam ao sistema se adaptar a cada usuário individualmente (FACELI et al., 2011).

Ao lado direito do modelo, são descritos os processos do usuário, e o que o leva a realizar uma busca, a usar um sistema de recuperação da informação. O “*Problem Space*”, referente ao contexto em que está o usuário, e que pode ou não gerar uma necessidade de informação, ou “*Information Problem*” (INGWERSEN, 1992; BELKIN, 1993).

“*Intermediary Functions*”, no centro da Figura 5, pode ser traduzido como “Funções Intermediárias” e contempla toda a capacidade do sistema em entender e auxiliar o problema ou necessidade de informação do usuário, assim como viabilizar as possibilidades de busca e evidenciar a lógica do sistema-fonte. As Funções Intermediárias podem ser aplicadas como parte das competências de um intermediário humano – um bibliotecário, por exemplo –, ou como na determinação de funções e de interface para o usuário final de um sistema digital (INGWERSEN, 1992, p. 56).

Por meio de olhar mais específico sobre os sistemas de recuperação da informação, Turtle e Croft (1997) identificam a RI como um campo de estudos preocupado em identificar documentos potencialmente úteis ou relevantes em uma dada coleção. Segundo os autores, um SRI apresenta três componentes principais: a identificação e representação dos conteúdos dos documentos; a aquisição e representação das necessidades de informação; e a especificação da função de comparação que seleciona os documentos relevantes baseados nas representações (TURTLE; CROFT, 1997).

Bibliotecas digitais são, talvez, a principal e mais evidente materialização de SRI na área da CI. Não à toa, os primeiros projetos de SRI partem da necessidade de se

organizar e recuperar digitalmente documentos ou representação de documentos (BUSH, 1945; MOOERS, 1949). Falar sobre a trajetória das bibliotecas digitais é, portanto, refletir sobre a trajetória da Recuperação da Informação. Segundo Sayão (2008, p. 11):

Quando se percorre toda a linha temporal da evolução técnica das bibliotecas digitais, se torna claro que as suas bases teóricas e práticas estão fortemente vinculadas às pesquisas desenvolvidas pela área da computação denominada de recuperação da informação.

O autor destaca a evolução histórica de um imaginário sobre bibliotecas universais, pontuando marcos diversos, como a proposta da biblioteca de Alexandria e referências a Jorge Luis Borges, e, numa perspectiva tecnológica e processual, destacando pensadores como Licklider, Lancaster, Paul Otlet (1934), Vannevar Bush (1945), Ted Nelson e Tim Berners Lee (SAYÃO, 2008). Todos contribuíram para uma visão utópica de biblioteca universal à qual nos aproximamos graças, também, aos avanços das tecnologias de informação e comunicação.

Sobre as diferentes abordagens, estudos e aplicações de Bibliotecas Digitais, Sayão (2008, p. 22) aponta que:

A comunidade de Biblioteconomia e Ciência da Informação visualiza a biblioteca digital menos como um sistema de computação – uma máquina – e mais como uma instituição; como uma extensão lógica do que as bibliotecas vêm fazendo desde os tempos imemoriais, ou seja, adquirindo, organizando e disseminando conhecimento usando as tecnologias correntes.

Já os profissionais da área de Ciência da Computação, “enxergam as bibliotecas digitais como uma extensão dos sistemas de computadores em rede [...]; para pesquisadores da área de base de dados, a biblioteca digital é tão somente uma ampla base de dados” (SAYÃO, 2008, p. 23).

O autor ainda aponta a abordagem de outros atores envolvidos com bibliotecas digitais, como a indústria editorial, educadores e professores, arquivistas, pesquisadores, cada um com diferentes objetivos e requisitos de projeto. Destaca

ainda o grande valor econômico e de exploração comercial de bibliotecas digitais, em especial a apropriação por sistemas de comércio eletrônico (SAYÃO, 2008).

2.5. Visualização da Informação

O que deve ser procurado em projetos de visualização de informações é o claro retrato da complexidade. Não a complicação do simples; em vez disso, a tarefa do designer é dar acesso visual ao sutil e ao difícil – isto é, a revelação do complexo. (TUFTE, 1983, p. 15)

Sob uma perspectiva histórica, Edward Tufte (1983) destaca que mostrar informações de forma gráfica é uma prática presente em toda a história da sociedade moderna. O mapa ilustrado na Figura 6, tido como um exemplo primordial da representação de informações no mapa, foi produzido na China no século XI. Conhecido como Yü Chi Thu (Mapa dos caminhos trilhados por Yü, o Grande, tradução nossa), apresenta escala em *grid*, o desenho da costa e o traçado dos rios de forma muito semelhante ao real, o que o torna único para o seu tempo e para a história da cartografia chinesa (NEEDHAM, 1959 apud Tufte, 1983).

Foi no século XVIII, no entanto, que práticas de visualização começaram a ser sistematizadas no Ocidente. Os trabalhos de três personagens são destacados por aliarem quantidade e complexidade de informações, assim como limitações de produção de sua época.

Figura 7 – Mapa dos caminhos trilhados por Yü, o Grande, gravado em pedra em 1137 a.C.

Fonte: Tufte (1983, p. 21).

O britânico **William Playfair** (1759-1823), economista e político influente, deu início ao uso sistemático da visualização de dados econômicos do país. A primeira visualização de série temporal utilizando dados econômicos foi publicado em seu livro *The commercial and political atlas* (Atlas Comercial e Político, 1786). A Figura 7 mostra, por exemplo, o fluxo de importações vs. exportações da Inglaterra. Para Playfair, gráficos eram preferíveis a tabelas, pois a representação gráfica mostrava a forma dos dados em uma perspectiva comparativa. Foi ele, também, que utilizou os gráficos de barra pela primeira vez (TUFTE, 1983).

Figura 8 – Gráfico de todas as Importações e Exportações de e para a Inglaterra entre 1700 a 1782, de William Playfair.

Fonte: Tufte (1983, p. 32).

Figura 9 – Gráfico de Charles Joseph Minard, que ilustra a trajetória geográfica e temporal da invasão da Rússia por Napoleão (1812-1813)

Fonte: Tufte (1983, p. 41).

Charles Joseph Minard é o autor de um famoso exemplo da história da visualização de dados representado na Figura 8, que descreve a trajetória da invasão da Rússia por Napoleão em 1812. Na mesma imagem é possível compreender uma narrativa completa envolvendo a evolução no tempo, o percurso e estratégia de rota, as constantes baixas no número de soldados, o relevo e a queda de temperatura que comprometeram a ação de Napoleão (TUFTE, 1983).

Outro exemplo de destaque está representado na Figura 9. Trata-se da Escala de Trens de Paris a Lyon de **E. J. Marey** (1880). Nele, cada linha é uma estação e estão separadas de forma proporcional a distância entre cada uma delas; chegadas e partidas das estações estão representadas no comprimento da seção entre a linhas inclinadas. A inclinação da linha reflete a velocidade do trem – quanto mais vertical a linha, mais rápida a viagem. Por fim, o cruzamento entre as linhas inclinadas indica exatamente quando dois trens em direções opostas vão se cruzar no caminho. Embora seja uma visualização confusa no primeiro momento, uma vez que ficam claras as variáveis e a forma de cada uma ser apresentada, torna-se fácil a compreensão sobre tempo de viagem, tempos de parada e tantos outros detalhes da trajetória de trens entre duas cidades.

Figura 10 – Escala de trens entre Paris e Lyon, de E. J. Marey (1880).

Fonte: Tufte (1983, p. 31).

Limitados a recursos de impressão e métodos manuais de produção, a visualização da informação ganha espaço em veículos de imprensa e trabalhos científicos sem apresentar, no entanto, padrões claros de qualidade e função.

O desenvolvimento dos computadores e da Ciência da Computação provocam uma transformação na capacidade de armazenar e processar informação, e os recursos computacionais abrem novas possibilidades e aplicações da visualização da informação na compreensão de grandes volumes de dados.

A partir dos anos 1990, a visualização começa a ser compreendida como uma área do conhecimento de caráter interdisciplinar, mas que necessariamente envolve processos computacionais. Um dos livros mais referenciados da área é *Readings in information visualization, using vision to think* (CARD; MACKINLAY; SHNEIDERMAN, 1999), uma coletânea que reúne artigos sobre uma grande diversidade de aplicações e técnicas, divididos em capítulos temáticos que definem as principais características do campo.

No livro de Card, Mackinlay e Shneiderman (1999), **visualização** é definida como o uso de representações de dados de forma visual, interativa e assistida por computador, para amplificar a cognição. A área então é dividida em duas frentes, a **visualização científica**, que lida com dados científicos, tipicamente embasados em evidências físicas – registros de experimentos; e a **visualização da informação**, cuja natureza dos dados pode ser composta por elementos abstratos e não embasada em evidências físicas de laboratório (CARD; MACKINLAY; SHNEIDERMAN, 1999, p. 6-7).

A adoção do termo Visualização da Informação (VI) neste trabalho, em lugar de Visualização de Dados (VD), é justificada pela abertura à complexidade do registro da atividade humana, da comunicação e da linguagem, expressões da natureza do abstrato e dos tipos de dados que podem gerar. É uma abordagem que permite, também, estabelecer paralelos com a Ciência da Informação (CI). Importante ressaltar a diferença de uma compreensão comum em que a VI tem a função de tornar a informação visual, o que aproxima o termo da infografia, e formas gráficas e metafóricas de representação.

2.5.1. Modelo simplificado da Visualização da Informação

Se considerarmos o modelo simplificado da RI, como proposto por Ingwersen (1992), a VI pode ser entendida como uma “função intermediária”, ou seja, parte da complexidade envolvida em questões de interface e auxílio na conexão entre requisições e buscas. Ela é parte do espaço de interação entre o usuário e as bases de documentos (os estoques de informação).

Card, Mackinlay e Shneiderman (1999), assim como Ingwersen (1992), propõem um modelo simplificado que mapeia os processos envolvidos na transformação de dados em visualizações que permitam ao usuário a realização de tarefas. A Figura 10 ilustra o modelo, que é muito recorrente na literatura da VI.

Figura 11 – Modelo simplificado de referência da Visualização da Informação.

Fonte: Card, Mackinlay e Shneiderman (1999, p. 17).

No modelo, as setas mostram o percurso entre **dados brutos**, à esquerda; e o usuário humano indica a série de transformações de dados que acontecem nesse processo. O sentido inverso, da direita para a esquerda, indica a ligação do usuário com os processos de transformação pelos quais passam os dados. A **transformação de dados** mapeia os dados brutos que, no exercício empírico deste trabalho, apresentado na terceira parte, considera como bases de documentos, variados e em constante produção, as **tabelas de dados**, que descrevem relações entre os documentos através do enriquecimento de metadados, de forma a estruturá-los. Os **mapeamentos visuais** transformam as tabelas de dados em **estruturas visuais**, que combinam medidas espaciais, marcações e propriedades gráficas. Por fim, as **transformações nas visualizações**, que as criam a partir das estruturas visuais especificando parâmetros gráficos como posição, escala, cor, volume. A **interação do**

usuário controla os parâmetros para três instâncias de transformação, restringindo a visualização para certos intervalos de dados, por exemplo, ou mudando a natureza da transformação. A visualização e seu controle são aplicados no auxílio de realização de uma **tarefa**, ou um **problema de informação**, se feita a correlação com o modelo simplificado de RI de Ingwersen (1992).

2.5.2. A Visualização nas práticas da Ciência da Informação

A contribuição da Visualização da Informação para a Ciência da Informação é um tema que já foi explorado em outros trabalhos, sob diferentes óticas e práticas.

Dias (2007) faz uma descrição detalhada de diversos recursos da Visualização da Informação e mostra possíveis aplicações no contexto da Ciência da Informação, incluindo a recuperação de informações.

Segundo o autor (DIAS, 2007, p. 111),

[...] os profissionais de CI devem atentar para a contribuição e o benefício que a área de VI pode trazer quando aplicada na área da CI. Tal contribuição é a de oferecer meios alternativos para a transmissão de informações para que sejam apropriadas pelo usuário de forma mais rápida, intuitiva e agradável.

Também podem ser aplicados em diversas áreas relacionadas à Ciência da Informação, para fins de gestão, compreensão de bases de dados e descoberta de novos conhecimentos e relações entre documentos.

A Visualização da Informação é destacada por Santos e Kobashi (2009; 2012), como um recurso importante para os campos da bibliometria, cientometria e infometria. Segundo os autores (SANTOS; KOBASHI, 2012, p. 63),

[...] recursos de visualização são alternativas promissoras para melhor perceber e compreender dados manipulados por meios estatísticos. Com efeito, é possível, por meio de mapas da ciência e da técnica, fazer múltiplas inferências sobre os fatos que ocorrem na pesquisa científica.

A infometria pode ser definida como “O estudo da aplicação de métodos matemáticos para os objetos da Ciência da Informação” (BARROS, 2015). Pode ser considerado o campo mais geral, capaz de envolver todos os tipos de informações, independentemente de sua forma ou origem (BARROS 2015).

Com objetos de informação mais específicos, a cientometria trata o desenvolvimento da ciência como um processo informacional social, a Bibliometria, cujo termo fora proposto por Paul Otlet em 1934, descrita como “a aplicação da matemática e métodos estatísticos para livros e outras mídias de comunicação” (PRITCHARD, 1969, p. 349 apud BARROS, 2015). No contexto deste trabalho, os livros e outras mídias de comunicação podem ser também compreendidos como documentos.

Ainda podem ser citadas a webmetria, que se baseia nas abordagens bibliométricas e infométricas para análise de páginas da web como se fossem documentos, e a recente Altmatria, que “usa medidas de impacto acadêmico com base na atividade em ferramentas e ambientes on-line” (PRIEM, 2014, p. 266). Dentre as medidas da altmetria, podem ser consideradas métricas como visualizações de página, downloads de arquivos, citações em sites, blogues e envolvimento dos usuários em plataformas como Research Gate, Facebook, Twitter e bases populares com recursos de interação com usuários.

Santos e Kobashi (2012) apresentam diversos tipos de visualizações e as descobertas possíveis através de sua análise no contexto da bibliometria, cientometria e infometria. As cartografias, por exemplo, são recursos que, por meio da análise de *corpus* de publicações,

[...] promovem a apreensão de diferentes tipos de ligações entre dados: pesquisadores e instituição, inventores e empresas, redes de colaboração entre empresas depositantes de patentes, entre inventores, pesquisadores e temas, inventores e inovações, entre outros. (SANTOS; KOBASHI, 2012, p. 66)

O dendrograma, resultado de análises estatísticas de agrupamentos, é o recurso de visualização de cartografias utilizado na Figura 11, que mostra as relações de proximidade entre Primeiro Autor (em vermelho) e descritores (em preto) ligados à área da Organização da Informação, em uma base especializada. Observa-se que os

nós estão agrupados em duplas (Maria Luiza Almeida Campos, Categorização) ou trios (Comunicação Científica, Carlos Henrique Marcondes, Publicações Eletrônicas) e, em princípio, pode-se verificar nos agrupamentos a coerência temática entre pesquisadores e referências de trabalhos publicados.

Figura 12 – Representação hierárquica de registros sobre Organização da Informação: agrupamento Descritor x Primeiro autor.

Fonte: Santos e Kobashi (2012, p. 66-67).

Outro tipo de estudo recorrente da bibliometria e da cientometria é a análise de cocitação. Uma cocitação é a ocorrência simultânea de referências a dois artigos feitos por um terceiro artigo. Chen (2004), utilizou um mapa de rede de cocitações num histórico de 10 anos de artigos sobre a campo da Visualização da Informação (de 1993 a 2003) para identificar os trabalhos mais influentes da área (Figura 12). Cada nó é um artigo, e as arestas coloridas representam relações de cocitações. O tamanho do nó é proporcional ao número de vezes em que é citado ou referenciado por outros artigos. Quanto maior o nó, maior a frequência de citações e referências ao artigo. O comprimento e a espessura das linhas são proporcionais à força de uma medida de cocitação. Uma linha curta e espessa indica uma relação de cocitação mais forte do que outra mais longa e fina. A cor identifica a época quando a relação se torna forte o suficiente pela primeira vez (CHEN, 2004).

Em formato eletrônico/digital, a visualização pode, com um simples recurso de interação, permitir um hiperlink nas arestas para a lista de artigos que formam aquela relação. Em grandes volumes de informação, esse recurso pode não ser suficiente, afinal, mesmo uma seleção dentre os documentos utilizados ainda é grande demais para se obter maior valor sobre a informação acessada. A esse desafio, ainda deve ser considerado o acesso remoto e on-line, que depende de infraestrutura e velocidade de conexão e processamento, do dispositivo que está sendo utilizado, e da prontidão de resposta à necessidade imediata de um usuário. Em outras palavras, o desafio é fazer com que a VI se comporte como um SRI.

2.6. Visualização em Sistemas de Recuperação de Informação

À medida que crescem exponencialmente o volume, a variedade e a velocidade na produção de informações, e que avanços tecnológicos e metodológicos tornam possível o contexto Big Data, a visualização da informação torna-se um recurso ubíquo, aplicado em diversos campos científicos e de mercado, como parte de ferramentas, softwares e aplicativos adaptados para uma diversidade crescente de casos de uso. A visualização da informação é, portanto, assim como parte dos métodos praticados na Ciência da Informação, um conjunto de metodologias que só faz sentido quando aplicado em outros campos do conhecimento.

Na comunicação, por exemplo, a visualização da informação é muito utilizada como recurso do Jornalismo de Dados, também referenciado no meio acadêmico como Jornalismo Digital Baseado em Dados (JDBD). Modelos interativos, visuais, construídos em plataforma digitais têm como função a investigação e disponibilização de dados de forma pública e acessível.

Daniela Ramos (2017) formula a hipótese do JDBD, que, para a autora, se estende a outros meios de comunicação como a televisão, rádio, jornais e revistas, pois as bases de dados integram, ou deveriam integrar, a estrutura organizacional das empresas de comunicação, e pode ser observada a relação da informática, para processos de busca, classificação e filtro de informações.

O JDBD, por sua característica de potencial amplitude de impacto na sociedade, exerce um papel importante na construção de um repertório visual no consumidor de informações. Veículos de imprensa mantém canais digitais dedicados a esse tipo de investigação. No Brasil, vale citar *O Estado de São Paulo*¹⁵ e o *Nexo Jornal*¹⁶ como referências.

¹⁵ Disponível em: <<http://blog.estadaodados.com/>>. Acesso em: 2 fev. 2019.

¹⁶ Disponível em: <<https://www.nexojornal.com.br/grafico/>>. Acesso em: 2 fev. 2019.

O exemplo apresentado na Figura 13 é um infográfico interativo do jornal *O Estado de São Paulo*¹⁷, publicado em outubro de 2018, que permite acessar as informações de resultado das urnas em cada colégio eleitoral. A ferramenta oferece dados sobre as votações de 1º e 2º turno das eleições aos cargos de presidente e governador, além de apontar senadores e deputados federais mais votados. É possível explorar os colégios eleitorais de todo o país, que ainda podem ser filtrados por faixas de renda, raça, idade e gênero. Através da visualização em mapa, é possível avaliar os resultados eleitorais em diversas escalas – cidades, estados regiões e país, permitindo a recuperação de informações de cada seção eleitoral. Recursos interativos como *zoom*, filtros e cliques garantem ao usuário auxílios para a busca de informações sobre seu próprio colégio eleitoral ou outros de seu interesse.

Figura 14 – Infográfico Interativo Como Votou Sua Vizinhança.

Fonte: Imagem capturada pelo autor. O Estado de São Paulo, out. 2018.

Um debate frequente nas práticas da JDBD é que alguns sistemas despendem grandes esforços na construção de interfaces para navegação e recuperação da informação em granularidade, mas que tem baixa audiência – poucos usuários fazem uso efetivo do recurso. Nota-se uma nova abordagem em que são oferecidas análises

¹⁷ Disponível em: <<https://www.estadao.com.br/infograficos/politica,como-votou-sua-vizinhanca-explore-o-mapa-mais-detalhado-das-eleicoes,935858>>. Acesso em: 2 fev. 2019. A metodologia utilizada pela equipe é detalhada na reportagem. Créditos: Texto: Rodrigo Menegat e Cecília do Lago / Dados: Rodrigo Menegat / Design: Ariel Tonglet e Bruno Ponceano / Desenvolvimento: Ariel Tonglet / Concepção: Daniel Bramattim.

descritivas e seleções de destaques como complemento ao sistema, ou mesmo dispensando a disponibilização dos dados em interfaces interativas.

No universo das corporações, é frequente o uso de softwares e aplicativos que possuem como um dos principais recursos a visualização. As plataformas OLAP (*On-Line Analytical Processing*), ou Processamento Analítico On-line, são exemplos de ferramentas para apoiar empresas na análise *ad hoc* de suas informações para auxílio à tomada de decisão. A VI é o principal recurso de análise que torna acessíveis a descoberta de padrões e comportamentos de dados para os usuários de uma mesma organização, ou de uma comunidade de interesses, que não necessariamente possuem competências de processamento estatísticos de informações. São ferramentas potentes para a distribuição de conhecimentos por permitirem a customização de interfaces e conjuntos de visualizações de informação de acordo com a necessidade específica de organizações e de diferentes usuários (ANZANELLO, 2007).

De forma resumida e simples, as OLAP permitem às organizações relacionar bases de dados de fontes e naturezas distintas, disponíveis em formato digital, com o potencial de atualizações em tempo real. A organização de diferentes visualizações agrupadas e disponibilizadas em interfaces facilitam sua leitura em conjunto. As interfaces customizadas são comumente chamadas painéis, ou dashboards. É uma transformação tecnológica que, somada a metodologias analíticas e ao processamento estatístico, cada vez mais recorrentes no contexto Big Data, amplificam a geração de valor a partir da descoberta de novos conhecimentos pelas organizações.

São inúmeras as opções no mercado, e exemplos comuns são as plataformas Tableau¹⁸, Power BI¹⁹, Qlik²⁰ e Microstrategy²¹. Grandes empresas multinacionais também possuem suas próprias soluções, como IBM²² e Oracle²³. De forma geral, são ferramentas pagas, mas permitem o uso livre ou oferecem planos gratuitos para

¹⁸ Disponível em: <<https://www.tableau.com/>>. Acesso em: 2 fev. 2019.

¹⁹ Disponível em: <<https://powerbi.microsoft.com/pt-br/>>. Acesso em: 2 fev. 2019.

²⁰ Disponível em: <<https://www.qlik.com/pt-br/>>. Acesso em: 2 fev. 2019.

²¹ Disponível em: <<https://www.microstrategy.com/br/>>. Acesso em: 2 fev. 2019.

²² Disponível em: <<https://www.ibm.com/analytics/cognos-analytics>>. Acesso em: 2 fev. 2019.

²³ Disponível em: <<https://www.oracle.com/br/solutions/business-analytics/>>. Acesso em: 2 fev. 2019.

estudantes e aplicações científicas. O Google Data Studio²⁴ é uma opção gratuita e compartilhável, porém com recursos mais limitados.

A depender das funcionalidades de cada ferramenta, é possível criar uma interface interativa que alie visualizações de informações e a capacidade de recuperação de grandes conjuntos de documentos. No entanto, o desafio não depende exclusivamente da plataforma, pois existem fases anteriores de tratamentos específicos e estruturação de dados que permitam manter simultaneamente a integridade estatística, fundamental à construção das visualizações, assim como o acesso aos detalhes e conteúdos dos documentos finais. A qualidade da estrutura dos documentos, os metadados, também são essenciais para garantir a veracidade e o valor da informação recuperada.

Já no campo da Visualização de Informação, a Visualização de Documentos é um desafio antigo, presente desde a consolidação do próprio campo da VI, que reúne e especifica técnicas próprias voltadas ao processamento de texto em arquivos digitais. Exemplo bastante citado na literatura é o *Film Finder* (AHLBERG, SHNEIDERMAN, 1999), ilustrado na Figura 14, que permite a compreensão de um conjunto de documentos – no caso, fichas catalográficas de filmes –, organizados de forma que evidenciam informações valiosas para a compreensão da produção cinematográfica da época. A abordagem aplicada nessa interface é percebida como influente referência em trabalhos atuais, reforçando seu papel fundacional na visualização de documentos.

O trabalho citado permite, simultaneamente, tirar conclusões e caracterizar a produção do cinema desde sua origem até o início da década de 1990. É possível identificar diretores de produção mais expressiva, atrizes e atores mais populares, épocas de atuação, e outras nuances que seriam muito mais difíceis de compreender apenas observando uma lista. Simultaneamente, a ferramenta também dá acesso a fichas catalográficas que contêm informações mais completas sobre cada filme²⁵.

A partir desse sistema, Ahlberg e Shneiderman (1999) reforçam três princípios fundamentais ao utilizar recursos gráficos em interfaces como o *Film Finder*. Os

²⁴ Disponível em: <<https://datastudio.google.com/u/0/>>. Acesso em: 2 fev. 2019.

²⁵ Uma demonstração do sistema *Film Finder* pode ser encontrada em vídeo no link: <<https://www.youtube.com/watch?v=g9JadyYUyK8>>. Acesso em: 26 jan. 2019.

autores destacam, em primeiro lugar, a consulta dinâmica (ou “*dynamic queries*”), que possibilita ao usuário a aplicação de filtros dinâmicos ao longo do processo de busca, como os controles deslizantes presentes na interface do *Film Finder*, que tornam possível a seleção de filmes por duração, grau de popularidade, ator, atriz ou diretor. O segundo princípio é a exibição em *starfield*, que considera duas variáveis ordinais (tempo e popularidade), e organiza os filmes em pontos coloridos, de forma relacional, isso dá ao usuário a capacidade de ver toda a base de dados disponível em uma área reduzida da tela. Por fim, o autor cita o termo “*tight coupling*” (*acoplamento estreitado*, em português), que descreve a ação de “acoplar” informações complementares através de hiperlinks. No caso da Figura 14, observa-se que a informação do ator Sean Connery está associada ao filme original.

Figura 15 – Tela do sistema Film Finder, que mostra simultaneamente o ano de lançamento, a popularidade, o gênero e há ainda a capacidade de filtros por duração, ator, atriz, diretor ou diretora do filme.

Fonte: Ahlberg e Shneiderman (1999, p. 250).

A compreensão do que temos como documento, aqui, é simples: são fichas catalográficas, com poucos campos de metadados e de estrutura fácil de ser aplicada. Com os avanços já descritos como parte do contexto Big Data, o mesmo documento tem outro nível de complexidade e formas de estruturação. Uma iniciativa on-line ativa desde outubro de 1990 é a Base de Dados de Filmes na Internet (*Internet Movie Database*), conhecida pela sigla IMDb. Pode-se afirmar, por analogia, que a plataforma é uma evolução da base trabalhada no *Film Finder* de Ahlberg e Shneiderman (1999). O documento, antes uma ficha catalográfica reduzida, ganha outros níveis de complexidade, caracterizados pelo grande volume já existente de produção cinematográfica, a variedade de formatos, que inclui, além de filmes, seriados, curtas e outros produtos audiovisuais.

O *Internet Movie Database* (IMDb)²⁶ é uma base de dados on-line que contém filmes, programas de TV e vídeo games e inclui elenco, equipe de produção, personagens fictícios, biografias, resenhas, e outras diversas informações. É uma base rica em informações e é muito grande para ser explorada através da navegação manual por cada conteúdo. Em abril de 2018, o IMDb tinha em sua base mais de 4 milhões de títulos de filme, curtas, médias e capítulos de séries, mais de 7 milhões de personalidades e por volta de 75 milhões de usuários registrados. É um universo rico para exploração de dados e o desenvolvimento de ferramentas de visualização (BURCH et al., 2018).

O exemplo apresentado na Figura 15 mostra uma interface de exploração da base de dados do IMDb, de autoria de Yu Hua Lim (2014), disponível como exemplo de aplicação do software OLAP Tableau²⁷. A base de filmes que está sendo considerada é a produção cinematográfica entre os anos de 1889 e 2012. É possível buscar pelo nome do filme, definir o intervalo de anos de produção que se deseja analisar e o filtro de notas do público. Abaixo, a interface mostra os países com maior produção e, como última opção de filtro, os gêneros de filmes mais frequentes. Na área inferior da interface, é possível visualizar a produção cinematográfica distribuída por data no eixo horizontal, e as notas de avaliação no eixo vertical. Os valores acima de cinco são

²⁶ Disponível em: <<https://imdb.com>>. Acesso em: 27 jan. 2019.

²⁷ Plataforma disponível em: <<https://public.tableau.com/en-us/s/gallery/imdb-movies-visualized>>. Acesso em: 27 jan. 2019.

considerados favoráveis, em verde; os valores abaixo, como avaliações baixas, desfavoráveis.

Figura 16 – Interface de navegação na base de dados do IMDb “Quais os filmes que produzimos?”

Fonte: Lim (2014).

Cada ponto representa um filme, e as informações básicas podem ser recuperadas ao passar o mouse sobre o gráfico, mas o acesso final ao documento (uma página on-line do site IMDb, por exemplo), não é disponibilizado. Ainda na visualização, é possível notar a linha que representa a média das avaliações em cada ano, o que facilmente nos mostra, por exemplo, quando foram lançados mais filmes de sucesso, e que essa média foi crescendo ao longo do tempo. Informações mais gerais do conjunto de documentos também estão disponíveis de forma gráfica, apontando os gêneros e países de maior relevância em volume de produção.

Podem-se identificar os princípios descritos por Ahlberg e Shneiderman (1999) em seu *Film Finder*. O usuário participa dinamicamente da busca através dos filtros, os dados são apresentados de forma multidimensional, e as informações, encadeadas, acopladas como complemento à busca.

Uma outra abordagem é oferecida por Burch et al. (2018), que propõem o *IMDb Explorer*, uma visualização disponível on-line, composta por dois gráficos, ilustrados na Figura 16. O primeiro, ao qual os autores chamam de “Movie Cosmos”, à esquerda, mostra o conjunto de filmes e sua organização, e, à direita, gráfico de linhas denominado “Career Line”, que exibe a trajetória de um filme, ator ou personagem ao longo do tempo. Ambos os gráficos são interativos e, ao passar o mouse sobre os elementos, é possível recuperar detalhes dos filmes e atores apontados.

O “Movie Cosmos” é apresentado numa visualização de bolha, ou *Circle Packing*, do inglês, que tem a capacidade de organizar hierarquicamente os filmes. Um filtro na parte superior permite a escolha da hierarquia a ser visualizada. É possível ordenar a hierarquia visual a partir do país, em seguida por gênero e, então, por diretor. Essa função é editável, e uma nova organização pode ser visualizada utilizando o gênero como classificação principal, representada nas bolhas maiores, seguindo por tamanho e relação conter-contido, a divisão por países e, então, por diretores.

O “Career Line”, gráfico à direita, mostra o ano de produção dos filmes no eixo horizontal, enquanto o eixo vertical indica a receita de determinado filme, ator, atriz, diretor ou diretora. Filmes seriados, por exemplo, estão conectados, assim como os diretores ou atores, mostrando a receita envolvida para cada ente analisado.

As duas visualizações são interativas e permitem ações do usuário. Ao selecionar um determinado filme ou gênero no gráfico “Movie Cosmos”, os detalhes são destacados no “Career Line”, onde podem ser compreendidos aspectos ligados ao rendimento do filme e atores envolvidos.

Figura 17 – Interface gráfica da ferramenta on-line IMDb Explorer.

IMDb Explorer: Visual Exploration of a Movie Database

VINCI '18, August 13–15, 2018, Växjö, Sweden

Fonte: Burch et al. (2018).

A base de dados do IMDb já serviria como uma expressão possível a partir da consolidação do contexto Big Data. Outros documentos poderiam ser incluídos e novas investigações podem ser feitas, tais como menções em notícias ou em postagens de redes sociais de cada filme, diretor e ator, suas presenças sociais, envolvimento em processos judiciais e outros rastros digitais disponíveis, públicos ou privados, possíveis de serem rastreados, acessados, coletados e estruturados.

Outros tipos de investigação são possíveis, com camadas mais complexas de cocitação que revelem convívios sociais entre pessoas. A Visualização da Informação certamente é um facilitador para a compreensão dessas relações de forma mais ágil. Como, no entanto, manter a integridade do acesso aos documentos, em toda a sua variedade e ainda possíveis de serem interpretados e recuperados por usuários humanos?

Não há um propósito específico para uma ferramenta que cumpra a função descrita no exercício acima, mas imaginá-la serve como exemplo para ilustrar o desafio. Ao realizar os processos estatísticos envolvidos na criação de visualizações de grandes bases de documentos, o acesso ou vínculo aos documentos utilizados nos cálculos pode ser perdido, limitado à capacidade de processamento necessário para manter as duas funções no uso de visualizações em interfaces de Sistemas de Recuperação da Informação.

Não à toa, a ferramenta *IMDb Explorer* (Figura 16) pode ter a usabilidade comprometida em máquinas que não tenham o poder de processamento gráfico suficiente para uma navegação fácil. A experiência, em uma situação de conexão e de infraestrutura de acesso, pode ser truncada e lenta.

O contexto Big Data, considerando seus atributos de tecnologia e métodos analíticos, avança de forma que os requisitos para que o uso da visualização de informação em interfaces de SRI possa ser cumprido.

Outro limitador à adoção de gráficos na interface de recuperação da informação envolve o usuário e o grau de abstração necessários para compreender o que, exatamente, uma visualização está comunicando. Na visualização de citações proposta por Chen (2004) abordada anteriormente (Figura 12), uma leitura correta demanda competências analíticas do usuário humano, que permitam a interpretação de uma visualização de rede em grafo, seus nós, arestas e os atributos de cor, tamanho e espessura.

Certamente, são duas questões relevantes que podem explicar por que grande parte das soluções de uso de visualização da informação em interfaces não são adotadas em escala, limitando-se, geralmente, a usos muito específicos ou em bases de dados de tamanhos suportados por tecnologias anteriores ao Big Data. A aplicação desses recursos pode encontrar limitações tecnológicas, como as de armazenamento e de processamento que, atualmente, impedem seu uso em VLDL, assim como ter de adotar tipos complexos de visualização que só funcionam para determinado tipo de aplicação, específica a um campo do conhecimento ou atividade profissional.

Iniciativas mais abrangentes e de maior impacto acontecem em grandes repositórios científicos. A *Web of Science* (WoS), gerida pela empresa Clarivate Analytics, e a *Scopus*, administrada pela Elsevier, possuem uma grande oferta em recursos analíticos que visam facilitar o trabalho do pesquisador. É possível, por exemplo, fazer análises de citações, a produção de autores e instituições, comparar estatísticas e produções de revistas especializadas.

Dentre os recursos oferecidos, ambas disponibilizam uma funcionalidade de especial interesse para este trabalho. Ao realizar uma pesquisa e acessar a lista de resultados, o usuário pode acionar o recurso “Análise dos resultados”, na WoS, ou “Análise dos

resultados de busca”, na *Scopus*, onde encontrará determinados dados bibliográficos dos documentos recuperados organizados em visualizações simples e interativas. Os tipos de gráfico são recorrentes no repertório de usuários comuns, como linhas, barras e pizza, que apresentam um intervalo limitado de itens mais relevantes. A interação com os gráficos mostra informações complementares e, no clique, dão acesso a um novo resultado de busca com a lista de documentos que compõe a variável escolhida no gráfico. Esse novo resultado pode ser explorado mais uma vez pelo modo de análise, evidenciando as características da nova seleção. O mesmo processo pode ser feito diversas vezes, podendo explorar o refinamento de buscas ou mudanças de estratégia de busca de forma mais ágil.

A prática é rotina nos estudos da bibliometria e da cientometria através da exportação de dados de busca e transformação utilizando outros softwares para explorar e gerar visualizações. Segundo as empresas, a funcionalidade torna mais acessível a análise bibliométrica básica, permitindo a pesquisadores de todos os níveis de experiência e repertório a identificação de tendências de pesquisa, principais autores, instituições ou países que estão produzindo sobre determinado assunto de interesse. Aos usuários mais experientes, os recursos colaboram no processo de refinamento das buscas e reduzem o tempo para descobrir novos padrões e afunilar os resultados.

Identificamos ao menos cinco características comuns aos dois repositórios citados valem ser mencionadas:

1. são generalistas e atendem a todas as áreas do conhecimento científico;
2. o grande volume de itens no acervo;
3. a variedade de formatos;
4. a necessidade de atender públicos com repertórios distintos e casos de uso específicos; e
5. a formalidade do processo e controle de vocabulários na gestão dos metadados, o que garante fiabilidade na organização, indexação e gestão de documentos em sua base.

A *Scopus* lançou sua funcionalidade de análise de resultados de busca em 2012²⁸, resultado do trabalho de Michael Habib²⁹, à época líder de experiência de pesquisa da plataforma, e, desde então, a funcionalidade ganhou popularidade. Ao longo dos anos, as formas de interação com o usuário e a interface evoluíram, mas seu propósito e funcionamento básico permanecem o mesmo.

A ferramenta permite oito tipos de análise baseados na contagem de documentos: evolução por ano, evolução por ano e por fonte, documentos por autor, por afiliação, país, tipo, áreas do conhecimento e agência financiadora. Antes disponibilizadas em abas e analisadas em separado, hoje as visualizações são dispostas em um painel que apresenta um *preview* de todos os gráficos e permite uma visão mais ampla de contexto. Um filtro é disponibilizado no painel para definição de um período a partir dos campos de data de início e fim. Ao aplicar o filtro, todas as visualizações são atualizadas.

Na interação do clique em uma miniatura, a vista escolhida é destacada na área superior da página, e os dados de apoio atualizados na lista lateral (Figura 18).

No gráfico destacado, ao passar o cursor pelos pontos ou áreas de interesse, o nome da categoria e o número de documentos que o compõem são mostrados (Figura 18). O clique abre uma nova aba ou janela do navegador, com a lista de documentos considerados (Figura 19).

Já a ferramenta de análise de resultados da *Web of Science* teve grandes avanços recentes e desponta como uma aposta da empresa, que tem disponibilizado recentemente treinamentos e tutoriais on-line. As atualizações do sistema da WoS 5.28 e 5.29, de março e abril de 2018, anunciaram evoluções importantes com a inclusão de novas visualizações interativas e a possibilidade de alternar o tipo entre TreeMaps e gráfico de barras que mostram até 25 principais categorias de resultados. A lista completa, em tabela, fica acessível na parte inferior da página, como mostra a Figura 20.

²⁸ HABIB, M. Sneak Preview Scopus Analyze Results: Overview and use case. **Michael Habib**, 8 dez. 2011. Disponível em: <<http://www.mchabib.com/2011/12/08/sneak-preview-scopus-analyze-results-overview-and-use-case/>>. Acesso em: 22 maio 2019.

²⁹ O currículo de Michael Habib pode ser consultado em seu site. Disponível em: <<http://www.mchabib.com/resume/>>. Acesso em: 22 ago. 2019.

Figura 18 – Ferramenta de análise de resultados de busca da Scopus

Fonte: O autor. Captura de tela do site da Scopus. Disponível em:

<<https://www.scopus.com/search/form.uri?display=basic>>. Acesso em: 12 de maio de 2019.

Figura 19 – Ao clicar em uma das visualizações menores, ela assume o topo da página e ganha destaque.

Fonte: O autor. Captura de tela do site da Scopus. Disponível em:

<<https://www.scopus.com/search/form.uri?display=basic>>. Acesso em: 12 de maio de 2019.

Figura 20 – Ao clicar em uma área ou ponto do gráfico, uma nova janela é aberta com a lista atualizada e aplicado o filtro representado na visualização.

Scopus Search Sources Alerts Lists Help SciVal Eduardo Caetano

500 document results

View secondary documents View 359 patent results View 12 Mendley Data

TITLE-ABS-KEY ("Information visualization" AND "Information Retrieval") AND (LIMIT-TO (SUBJAREA, "COMP"))

Edt Save Set alert Set feed

Search within results...

Analyze search results Show all abstracts Sort on: Date (newest)

AB Export Download View citation overview View cited by Save to list

	Document title	Authors	Year	Source	Cited by
1	Visualization of information retrieval process in the theses section of academic libraries: A proposed model	Mohammadzadeh, S., Mohammadzadeh, M., Biglu, M.-H.	2019	Iranian Journal of Information Processing Management 34(2), pp. 817-840	0
View abstract View at Publisher Related documents					
2	An architecture of open-source tools to combine textual information extraction, directed search and information visualisation	Sorring, D., Proflich, H.-J.	2019	Artificial Intelligence in Medicine 93, pp. 13-28	1
View abstract View at Publisher Related documents					
3	Visualization of Geographic Information Retrieval Results Supporting User Participation	Xiaohu, X., Zhang, Y.	2018	Proceedings - 9th International Conference on Information Technology in Medicine and Education, ITME 2018 8589469, pp. 1077-1080	0
View abstract View at Publisher Related documents					
4	Pictorial visualization of EMR summary interface and medical information extraction of clinical notes	Buan, W., Appawani, N., Kim, K., Kim, H., Lee, W.-S.	2018	CIWEMSA 2018 - 2018 IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications, Proceedings 8439958	3
View abstract View at Publisher Related documents					
5	3D visualization of legislative relationships	Elm, A., Socher, G., Khosroud, F.	2018	ACM International Conference Proceeding Series a111	0
View abstract View at Publisher Related documents					
6	Emerging trends and new developments in information sciences: a document co-citation analysis (2009-2018)	Hou, J., Yang, X., Chen, C.	2018	Scientometrics 115(2), pp. 869-892	8
View abstract View at Publisher Related documents					
7	Research on the Design Strategy of Smartphone Shopping App for the Older Adults	Hu, Q., Fang, X.	2018	Proceedings - 2017 International Conference on Industrial Informatics - Computing Technology, Intelligent Technology, Industrial Information Integration, ICIICI 2017 2017-December, 8328652, pp. 346-350	0
View abstract View at Publisher Related documents					
8	Information visualization for interactive information retrieval	Hocher, O.	2018	CHIIR 2018 - Proceedings of the 2018 Conference on Human Information Interaction and Retrieval 2018-March, pp. 371-374	1
View abstract View at Publisher Related documents					
9	Automatic Annotation of Text with Pictures	Poojs, J.K., Bagheri, E.	2018	IT Professional 20(1), pp. 36-44	1
View abstract View at Publisher Related documents					
10	Graicion: A graphical tool for the visualization of web search engines results with multi-agent based modeling (Book Chapter)	Bouramoud, A.	2018	Information Retrieval and Management: Concepts, Methodologies, Tools, and Applications 3, pp. 1338-1359	0
View abstract View at Publisher Related documents					
11	Knows: A visual search interface	Schlossinger, M., Hinkel, M.	2018	Communications in Computer and Information Science 850, pp. 191-198	0
View abstract View at Publisher Related documents					
12	FAB LAB: The knowledge and uncertainty research	Wallace, M.	2018	CEUR Workshop Proceedings	0

Refine results

Limit to Exclude

Access type

- Open Access (14)
- Other (486)

Year

- 2019 (2)
- 2018 (11)
- 2017 (18)
- 2016 (13)
- 2015 (16)

View more

Author name

- Shrederman, B. (10)
- Hocher, O. (9)
- Takama, Y. (7)
- Chen, C. (6)
- North, C. (6)

View more

Subject area

- Computer Science (500)
- Mathematics (87)
- Engineering (85)
- Social Sciences (46)
- Decision Sciences (20)

View more

Document type

Source title

Publication stage

Keyword

Affiliation

Funding sponsor

Country/territory

Source type

Language

Limit to Exclude

Restore original settings

Export refine

Fonte: O autor. Captura de tela do site da Scopus. Disponível em:

<<https://www.scopus.com/search/form.uri?display=basic>>. Acesso em 12 de maio de 2019.

Além das oito possibilidades de análise oferecida pela *Scopus*, a WoS disponibiliza outras oito –, isto é, 16 no total: categorias do Web of Science, anos de publicação, tipos de documento, organizações (aprimorada), agências financiadoras, autores, títulos da fonte, títulos de séries de livros, títulos de encontros, países/regiões, editores, autores grupo, idiomas, áreas de pesquisa, números do subsídio, organizações.

Figura 21 – A ferramenta WoS permite ao usuário selecionar entre dois tipos de gráficos: treemaps ou barras

Fonte: O autor. Captura de tela do site da Web of Science. Disponível em:

<<https://apps.webofknowledge.com/>>. Acesso em: 12 de maio de 2019.

O funcionamento é bastante similar ao da *Scopus*. Ao passar o mouse, o nome da categoria aparece completo, assim como o número de documentos associados. Um botão disponibiliza o acesso à lista de documentos, e o resultado pode ser acessado na mesma página do navegador, não havendo a necessidade de abrir uma nova aba ou janela como na outra plataforma analisada (Figura 21).

Figura 22 – Ao clicar em uma área do gráfico, o usuário acessa a lista com o filtro aplicado. Ao contrário da Scopus, a atualização da página acontece na mesma janela.

The screenshot displays the Web of Science interface. At the top, there's a navigation bar with links to various services like Web of Science, InCites, Journal Citation Reports, etc. The main header shows 'Web of Science' and 'Clarivate Analytics'. Below this, a search bar is present. The left sidebar contains filters for 'Resultados: 24' and 'Refinar resultados'. The main content area shows a list of search results with titles, authors, and publication details. The results are sorted by 'Data' (Date). The first result is 'A Semantic and Content-Based Search User Interface for Browsing Large Collections of Foley Sounds' by Urbain, Gabriel; Moinet, Alexis; Frisson, Christian; et al. The second result is 'Evaluation on Geospatial Information Extraction and Retrieval: Mining Thematic Maps from Web Source' by Dewandaru, Agung; Supriana, Iping S.; Akbar, Saiful. The third result is 'PHENIX: INNOVATING THE CLASSICAL MUSIC EXPERIENCE' by Liem, Cynthia C. S.; Gomez, Emilia; Schedl, Markus. Each result includes a checkbox for selection, an 'Exportar...' button, and an 'Adicionar à Lista marcada' button. The right sidebar shows 'Analisar resultados' and 'Criar relatório de citações'.

Fonte: O autor. Captura de tela do site da Web of Science. Disponível em: <<https://apps.webofknowledge.com/>>. Acesso em: 12 de maio de 2019.

Cada plataforma possui suas especificidades com relação à organização, interface e usabilidade. A ferramenta *Scopus* apresenta o conjunto de visualizações em uma só tela como um dashboard, permitindo uma visão de contexto mais ampla do que a oferecida na WoS. Outro ponto positivo é a disponibilização do filtro por período, que permite a atualização das visualizações sem a necessidade de voltar à busca. Na ferramenta da Clarivate, não é possível aplicar filtros na tela de exibição dos gráficos. Para qualquer edição na busca, é necessário retornar à página de resultados, refinar a pesquisa, para então visualizar a tela de análise com a atualização.

Pontos positivos da WoS são a quantidade de variáveis que podem ser utilizadas nas visualizações, e o uso de treemaps que, para mostrar proporções, são alternativas aos gráficos de pizza utilizados na *Scopus*.

Com o intuito de explorar as características de cada um dos acervos, realizamos buscas nas duas plataformas usando a expressão [“information visualization” AND “Information Retrieval”]. Pela especificidade do assunto, o volume de documentos recuperados não é grande: na *Scopus*, foram identificados 631 documentos; na WoS, 182.

Não é a intenção nos aprofundarmos na análise do acervo de cada repositório, mas algumas características valem ser destacadas. Naturalmente, as filiações e autores são bem diferentes e evidenciam uma maior autoridade sobre o tema na base *Scopus*, se considerada a abrangência em volume de documentos e autores. Essa base se mostrou mais completa sobre o assunto, não apenas pelo número de documentos recuperados, mas os autores presentes são reconhecidos como referências no campo da visualização da informação, com Ben Shneiderman, largamente citado em trabalhos científicos da área. Não à toa, a instituição de maior destaque é a Universidade de Maryland, onde Shneiderman atua desde os anos de 1990, tida como um dos centros mais importantes da produção acadêmica sobre VI. As similaridades também reforçam outras características do campo – em ambas as bases, a Ciência da Computação é, com destaque, o campo onde ocorre a maior parte da pesquisa sobre o tema. Um pico na produção é percebido entre os anos de 2006 a 2008, como pode ser observado na Figura 22.

O intuito primário das duas funcionalidades é permitir ao pesquisador tomar decisões práticas sobre o seu trabalho, como quais os periódicos mais adequados e reconhecidos para determinado tema, ou quem são outros autores com quem poderia iniciar uma colaboração. Os critérios são quantitativos, baseados no volume de documentos que cada categoria visualizada possui. Não consideram, por exemplo, qualidade ou relevância. Para isso, as empresas dispõem de outras ferramentas analíticas que permitem a análise mais detalhada de citações e avaliações sobre a produção de autores, periódicos e instituições.

Figura 23 – Evolução no tempo de número de publicações que relacionam Visualização da Informação e Recuperação da Informação na base Scopus.

Fonte: O autor. Captura de tela do site da Scopus. Disponível em:
<https://apps.webofknowledge.com/>. Acesso em: 12 de maio de 2019.

A recuperação de documentos é um objetivo secundário e apresenta limitações para o uso com esse determinado fim. Independentemente das diferentes escolhas de visualização e interface, outros recursos poderiam ser adicionados para facilitar o processo de recuperação sem ser necessário migrar da navegação visual para a lista tradicional.

Algumas escolhas colaboram na universalização do serviço. A escolha por gráficos populares e comuns ao repertório de um público abrangente – gráficos de barra, linha e pizza – colabora para a fácil apreensão e disseminação da funcionalidade, e demandam menos requisitos técnicos para serem gerados, assim como a necessidade de instalação de extensões de browser ou softwares específicos. Limitar a profundidade da informação disponibilizada poupa a carga visual para o usuário e o processamento, porém elimina uma granularidade que pode ser valiosa à pesquisa,

como as sub-hierarquias de especializações dentro de uma mesma área de conhecimento.

Exemplos de funções que poderiam ser agregadas para o aprimoramento da função de Recuperação de Documentos são os atributos detalhados por Ahlberg e Shneidermann (1999) aplicados ao *Film Finder*. Podemos observar claramente a função “*tight coupling*”, com a apresentação de dados complementares ao passar o cursor sobre as partes do gráfico e os links para aprofundamento sobre o conhecimento de citações e outros dados. O recurso “*dynamic queries*”, ou seja, a disponibilização de filtros e outras formas de refinamento de pesquisa, poderia ser mais bem explorado. A WoS não possui recursos de filtro nas visualizações, e a *Scopus* permite apenas o filtro por ano de publicação. A ausência desses recursos prejudica o uso das visualizações como interface, pois quebram o fluxo do usuário na percepção da mudança das visualizações.

O “*starfield display*”, outro recurso observado no *Film Finder*, ou soluções similares de visualização, permitiria cruzar mais de uma variável na mesma visualização. Ao utilizar dois eixos, ganhamos uma noção espacial e de localização, adicionando uma camada de relação entre os documentos representados por pontos, por exemplo. Outras variáveis ainda cabem se considerarmos atributos como cor e tamanho dos pontos.

2.7. Recursos e técnicas de Visualização da Informação

Os exemplos mencionados no item anterior ilustram algumas das aplicações possíveis da Visualização da Informação no processo de recuperação de documentos. Buscamos casos de fácil acesso, e que cobrem interesses de públicos diferentes e amplos, mas que possam atender a interesses específicos e detalhados de recuperação.

A literatura sobre o tema costuma trazer exemplos de técnicas, tipos de visualização e softwares especializados, mas com aplicações voltadas para um público especialista em uma área específica de conhecimento, ou método analítico.

Segundo Dix (2012), ao criar visualizações, é preciso ter em mente dois tipos de audiências. O primeiro tipo diz respeito aos analistas de dados, técnicos que podem ser cientistas, estatísticos ou profissionais da informação e que podem estar envolvidos no desenvolvimento das estratégias e projeto de SRI. São necessidades específicas que colaboram na compreensão dos processos destacados anteriormente, de definição e estruturação de documentos, organização da informação e, então, a visualização. É um público com repertório mais aprofundado, distinto do segundo tipo de audiência, aqui denominado “usuários finais”, ou consumidores de informação. Esses usuários não participam da concepção e do projeto de um SRI, mas precisam tomar decisões rápidas a partir das informações contidas no sistema (DIX, 2012). Podem ser, por exemplo, os pesquisadores usando as plataformas WoS e *Scopus*, funcionários e gestores de uma corporação em plataformas de gestão de conhecimento, assim como audiência de um veículo de informação que não possui habilidades estatísticas ou analíticas aprofundadas, com repertório limitado sobre tipos de visualização e como interpretá-los.

Também é importante lembrar que a visualização ocupa um lugar dentro de um contexto mais abrangente, onde os dados visualizados são provenientes da produção de documentos. A visualização ajuda indivíduos e organizações a tomarem decisões, e essas decisões direcionam ações que reconfiguram a produção e, conseqüentemente, os dados que estão sendo visualizados.

Além de considerar a audiência e seus casos de uso, é importante ter conhecimento aprofundado sobre o conjunto – ou conjuntos – de documentos a serem disponibilizados. São muitas as adversidades que podem comprometer não apenas a visualização, mas toda a eficiência de um SRI.

Como vimos anteriormente, o documento, e, em específico, o documento digital, para cientistas de dados, são itens em uma tabela de dados, e a integridade precisa ser mantida. Os problemas mais comuns são inconsistências na estrutura, incompletude no preenchimento dos metadados, erros de preenchimento, padronização de termos, isso para trazer alguns poucos exemplos (FACELI et al., 2011). A resolução dessas inconsistências é, muitas vezes, a fase mais lenta e trabalhosa do processo, mas também a mais importante, pois nela são estabelecidas as estratégias de organização da informação.

Sejam automáticas ou manuais, as técnicas de organização da informação constroem a forma como o estoque da base de dados será estruturada. Podem ser manuais, que passam pelo crivo de um indexador humano, ou automática, utilizando algoritmos e funções estatísticas computacionais para processamento em massa de documentos e encontrar relações que não são percebidas ou executáveis por um usuário humano.

Uma análise sobre a relação da organização com a Visualização da Informação em Sistemas de Recuperação da Informação pode ser encontrada em Vieira e Correa (2011) e De Lira e Assis (2015). Os autores analisaram seis bibliotecas digitais cujas interfaces são baseadas em VI. Os exemplos apresentados pelos autores mostram a aplicação de mapas conceituais, cartografias temáticas, desenho de árvores, nuvem de tags e browser hiperbólico. A sistematização entre as opções de visualização da informação e o método de organização da informação podem ser conferidos na Figura 23.

Figura 24 – Relação entre Organização da Informação e Visualização da Informação na aplicação de interface visual.

Quadro 1 – Relação entre OI e VI na aplicação de interface visual			
Site de SRI	Técnica ou modelo de VI	Interface Visual	Instrumentos de OI para o emprego de visualização
1- AquaBrowser da biblioteca da Universidade de Edimburgo	Desenho de Grafos		Adoção de um tesouro ou taxonomia como instrumento de OI que permita uma representação e organização adequada de termos e que estabeleçam as relações existentes entre conceitos.
2- Mecanismo de busca YIPPY	Desenho de Árvores		Adoção de uma taxonomia como instrumento de OI que permita a categorização de conteúdos semelhantes, evidenciando as relações hierárquicas existentes entre eles.
3- Infoteca-e da Embrapa	Browser Hiperbólico		Adoção de uma taxonomia como instrumento capaz de organizar termos em categorias da mais genérica a mais específica, estabelecendo relações hierárquicas entre eles.
4- Website do IHMC CampTools	Mapa Conceitual		Adoção de ontologia como instrumento que possibilite congrega conjuntos de conceitos padronizados, termos e definições, evidenciando a forma com que esses conceitos estão relacionados entre si.
5- WEBSOM - método para construção de mapas visuais.	Cartografia Temática		Adoção da ontologia enquanto instrumento que permite a representação e a categorização de documentos, bem como o estabelecimento de temáticas relacionadas; adoção de critérios para identificar a coocorrência entre termos.
6- Sistema de busca da BRAPCI	Nuvem de Tags		Adoção de listas de cabeçalhos de assunto para auxiliar na organização, estruturação e representação das tags que servirão como pontos de acesso para localização de documentos.

Fonte: De Lira e Assis (2015, p. 123).

Infelizmente, as interfaces analisadas não estão mais disponíveis on-line e não puderam ser analisadas novamente, ainda que os repositórios de documentos ainda existam. De fato, muitos dos exemplos discutidos e analisados em artigos científicos deixam de evoluir e acabam sendo abandonados e se tornam indisponíveis para avaliação posterior.

Nesta seção, selecionamos algumas técnicas e tipos de visualização que podem ter sua aplicação em sistemas de recuperação da informação. São exemplos de forma de organização visual de documentos e algumas características básicas que devem ser observadas no planejamento e desenho de um SRI que utilize visualizações como ponto relevante no seu processo de navegação e descoberta.

A disponibilização de determinadas técnicas tem limitações claras que podem envolver a instalação de extensões ou o uso de browsers específicos, o que já é um obstáculo na decisão sobre investir numa funcionalidade similar. São recursos que ainda não são aplicados em larga escala como alternativa de navegação em grandes repositórios, mas que vemos aplicado em soluções de mercado ou experimentos científicos.

2.7.1. Representação de hierarquias

Árvores, taxonomias e hierarquias são, talvez, as formas mais onipresentes de estrutura de dados após a disposição de tabelas. Árvores são encontradas em tesauros, taxonomias biológicas e partes de ontologias formais, organogramas, dados XML ou sistemas de arquivos. É também uma das formas de representação mais antigas, observada pela primeira vez na Árvore de Porfírio (234-305 a.C.), que organizou visualmente em uma árvore as Categorias do Mundo de Aristóteles (MAZZOCCHI, 2013).

Uma das formas mais comuns de visualização deste tipo são gráficos que organizam caixas e linhas, como exemplificado na Figura 24, muito utilizadas para representar mapas de navegação de sites, organogramas e genealogias.

Figura 25 – Visualização simples em árvore.

Fonte: O autor.

Alguns problemas surgem decorrentes dessa visualização, como a limitação de caracteres ou a representação de grandes estruturas de hierarquia. Certamente, é um tipo de visualização bastante limitado para um contexto de grande volume de informações e níveis de hierarquia.

Gusmão, Delazeri e Matos (2016) identificaram 18 Técnicas de Visualização Hierárquica (em inglês *Hierarchical Visualization Techniques* – HVT), e as classificaram com relação à capacidade de representar bases de dados de diferentes tamanhos e estrutura. A árvore representada na Figura 24, por exemplo, é recomendada para estruturas com poucos elementos (até 100). Dentre as técnicas apontadas pelos autores, vale destacar algumas que são exemplos recorrentes ao citar a visualização de documentos.

Visualizações bastante comuns são aquelas que possuem como estrutura básica nós e links. Os exemplos mostrados na Figura 25 representam a mesma base hierárquica em três formatos diferentes, uma clássica linear, na qual conseguimos ver a expansão total da árvore, mas com problemas de sobreposição; a segunda disposta de forma radial, em que o problema de sobreposição é resolvido; e numa estrutura clássica de pastas de arquivo (HEER; BOSTOCK; OGIEVETSKY, 2010).

Figura 26 – Três formas clássicas de representação de hierarquias, segundo Heer, Bostock e Ogievetsky (2010).

Fonte: Flare Data Visualization for the Web (s.d.). Disponível em: <<http://flare.prefuse.org/>>. Acesso em: 24 maio 2019.

Quanto maior o tamanho da árvore – em níveis de hierarquia e de elementos – mais complexa fica sua representação pelos formatos já mencionados, e outras abordagens trazem novas alternativas de representação. A seguir destacamos algumas delas.

2.7.1.1. Browser hiperbólico

A visualização em browser hiperbólico utiliza um conceito da matemática chamado geometria hiperbólica, em que a circunferência de uma sequência de círculos cresce exponencialmente na medida em que vão se expandindo, distorcendo os círculos para manter a estrutura inicial ou mesmo tempo em que mostra a expansão de um novo ramo da árvore (DIX, 2012). Ele otimiza a disposição dos elementos aproveitando de maneira mais eficiente o espaço, a depender do número de itens de uma subcategoria. De acordo com a classificação de Gusmão, Delazeri e Matos (2016), browsers hiperbólicos são recomendados para hierarquias de até 500 elementos.

Segundo Lamping e Rao (1995, p. 33-34, tradução nossa), pesquisadores que documentaram a técnica,

O navegador hiperbólico inicialmente exibe uma árvore com sua raiz no centro, mas a exibição pode ser suavemente transformada para trazer outros nós em foco [...]. Em todos os casos, a quantidade de espaço disponível para um nó cai como uma função contínua de sua distância na árvore a partir do ponto no centro.³⁰

Figura 27 – Simulação do efeito da animação no clique do item no browser hiperbólico

Fonte: Lamping e Rao (1995, p. 35).

³⁰ "The hyperbolic browser initially displays a tree with its root at the center, but the display can be transformed smoothly to bring other nodes into focus [...] In all cases, the amount of space available to a node falls off f as a continuous function of its distance in the tree from the node in focus." (LAMPING; RAO, 1995, p. 33-34)

Allen (2002) viu grande potencial de uso do browser hiperbólico para busca de assuntos em bibliotecas digitais, mas reportou problemas de usabilidade e deficiências do sistema à época, pois a transição entre os nós em foco era lenta e truncada, além da necessidade de instalação de plug-ins para o uso. Os exemplos indicados pela autora, softwares da empresa Inxight, não estão mais disponíveis na rede e não puderam ser revisitados. Uma simulação da técnica está disponível on-line no site da universidade austríaca Technische Universität Graz, e o comportamento da visualização apresentada é simples, interativo e com as animações características do browser hiperbólico. São oferecidos dois casos, um com 205 nós³¹, em que podemos ver etiquetas com o nome das classes e a resposta ao movimento é imediato, deixando a interação fluida. No segundo caso, a simulação é feita com 3.418 nós³² e a *performance* cai drasticamente – as etiquetas não são mais exibidas, e a velocidade lenta de processamento compromete a fluidez.

Figura 28 – Simulação de Browser Hiperbólico com 3.418 nós – performance comprometida.

Fonte: Technische Universität Graz (2014b).

³¹ Technische Universität Graz, Hyperbolic Browser: 205 nodes. **FluidDiagrams**, 2014a. Disponível em: <<https://projects.isds.tugraz.at/fluiddiagrams/fd/examples/Hyperbolic/Hyperbolic-small.html>>. Acesso em: 24 maio 2019.

³² Technische Universität Graz, Hyperbolic Browser: 3.418 nodes. **FluidDiagrams**, 2014b. Disponível em: <<https://projects.isds.tugraz.at/fluiddiagrams/fd/examples/Hyperbolic/Hyperbolic.html>>. Acesso em: 24 maio 2019.

2.7.1.2. Treemaps e visualizações de invólucros (*enclosure*)

Uma técnica alternativa e difundida é a visualização em treemaps, adequados para estruturas com até 1.000 elementos (GUSMÃO; DELAZERI; MATOS, 2016), que explora o espaço em duas dimensões. É uma técnica particularmente útil quando acrescentamos uma dimensão de tamanho ou volume – por exemplo, a quantidade de documentos presentes sob um mesmo ramo da árvore. O treemap divide o espaço verticalmente e horizontalmente em retângulos, e a hierarquia é expressada pelo tamanho da área de cada retângulo interno. Esta é a visualização presente no exemplo da Análise de Resultados, da WoS (Figura 20).

As divisões de um treemap podem funcionar também como invólucros – níveis inferiores de categorias podem ser representados simultaneamente, numa relação de conter e estar contido em outras áreas. Essa visão permite a comparação direta de duas subdivisões internas e a rápida conclusão sobre os temas específicos mais relevantes de um conjunto. A interatividade no clique, revelando camadas inferiores de hierarquia também aumentam o potencial de uso dessa visualização.

Figura 29 – Treemap e a função de invólucro, evidenciando subcategorias.

Fonte: Heer, Bostock e Ogievetsky (2010).

Uma vista alternativa ao treemap com a função de envólucro é a visualização de bolha, utilizada no exemplo do *IMDb Explorer* ilustrado na Figura 16. Por usar círculos em vez de retângulos, podemos produzir um tipo diferente de envólucro com uma aparência mais orgânica. Embora não utilize o espaço da forma otimizada como treemaps, o espaço vazio entre os elementos facilita a percepção da hierarquia e contagem de elementos.

Figura 30 – Exemplo de visualização “envólucro” com representações em bolhas

Fonte: Heer, Bostock e Ogievetsky (2010).

2.7.1.3. Sunburst

Ainda em espaços de duas dimensões, podemos destacar outra técnica, o sunburst. Nesta visualização, a área que representa o tamanho ou volume de um ramo é circular. O topo da hierarquia está localizado no centro, e os subníveis aparecem em uma segunda camada, ocupando o limite do raio ocupado pelo ramo-pai. Chen (2004) indica que a técnica sunburst resolve alguns problemas de navegação dos treemaps, pois é mais fácil observar a ordenação de classes. Retângulos muito finos, em conjunto com outras formas irregulares de área tornam comparações mais difíceis, enquanto o sunburst permite a comparação regular das áreas ocupadas radialmente. Assim como no treemap, a interação com o clique pode isolar uma área para que seja detalhada.

Figura 31 – Sequência de navegação e detalhamento de categorias em uma visualização sunburst.

Fonte: Vasco Veturiano, s.d. Disponível em: <<https://github.com/vasturiano/sunburst-chart>>. Acesso em: 20 maio 2019.

2.7.2. Técnicas de agrupamento

Frequentemente, a Recuperação da Informação lida com bases em que não há uma estrutura fixa, um conjunto coeso de metadados ou variáveis. São grandes bases de resultados de busca, com atributos comuns, mas sem uma estrutura hierárquica. Nesses casos, embora não exista uma estrutura dada, outras formas de estrutura podem ser induzidas através de métodos de agrupamento (*clustering*, em inglês).

Hierarquias são centradas em nossa compreensão lógica do mundo e, sendo uma construção social, precisam de maior percepção sensorial desenvolvida. Em

contraste, agrupamentos visuais de elementos similares tendem a corresponder de forma mais direta à percepção humana, sendo possível a detecção de padrões para além de uma estrutura predefinida. É mais simples reconhecer grupos que possuem aspectos semelhantes, assim como perceber aqueles que não fazem parte dele (DIX, 2012).

Ao tratarmos de estoques de documentos, presumimos que sejam, ao menos em parte, compostos de dados categóricos e comuns à parte ou a todo o grupo. É o caso quando tratamos de textos, imagens, áudio ou vídeo em que, em vez de tentar calcular alguma forma de média, muitas vezes é melhor dar um ou mais exemplos reais que representem o grupo.

Esses modelos podem ser deliberadamente escolhidos como “típicos”, usando alguma medida. Por exemplo, ao usarmos texto, pode-se calcular as medidas de similaridade baseadas na co-ocorrência de palavras e então apontar os documentos mais centrais e relevantes para cada grupo.

Dix (2012) aponta um método de visualização chamado Scatter-Gather Browser, que agrupa documentos que utilizam textos. A Figura 31 mostra a tela principal com os resultados de agrupamentos do navegador, formado por duas colunas com cinco módulos cada. Cada um desses módulos representa um agrupamento, e um detalhe do agrupamento pode ser visto na Figura 32.

O Scatter-Gather Browser foi concebido para ajudar os usuários a encontrarem documentos com base em critérios difíceis de definir ou apenas “reconhecidos quando vistos” – um claro contraponto à navegação hierárquica. O sistema funciona a partir de uma coleção de documentos, podendo ser um grande repositório único, como as VLDL, ou aplicado ao resultado de busca de uma palavra-chave. O sistema então utiliza um algoritmo automático, como a detecção de co-ocorrência de palavras, para organizar os documentos em 10 grupos distintos. Os usuários selecionam um ou mais clusters que lhes parecem mais relevantes, e então o sistema divide o grupo de documentos selecionado em outros 10 novos grupos. Quando os grupos são pequenos o suficiente, o usuário pode então acessar a lista em formato convencional para então recuperar a seleção final de documentos.

Figura 32 – Scatter-Gather Browser.

Fonte: Dix (2012, p. 16).

Figura 33 – Scatter-Gather Browser: detalhe de amostra de conteúdos de um mesmo grupo

Documentos "típicos", selecionados pelo número de palavras-chave encontradas

Fonte: Adaptado de Dix (2012, p. 16).

O problema é que, quanto maior a biblioteca, maior o tamanho dos agrupamentos originais. Na Figura 32, que ilustra o detalhe de um grupo, podem ser observadas duas áreas principais. No topo, são especificados o número de palavras-chave utilizadas para caracterizar o grupo de acordo com o volume de ocorrências e uma linha com aquelas mais recorrentes. Na área abaixo, são mostrados três documentos

únicos representados pelo título ou fragmento do texto, ou seja, uma forma visual de observar elementos típicos do grupo em questão.

As possibilidades de agrupamento são praticamente infinitas, dependendo de um conjunto de regras determinadas pela aplicação de algoritmos. Um dos modelos mais utilizados para esse fim, em diferentes sistemas de visualização, é o Modelo de Espaço Vetorial (MEV), que adota uma abordagem matemática vetorial. Isso significa que o documento pode ser localizado no espaço tridimensional, ou multidimensional. Ao localizar diversos documentos em uma mesma visualização, poderão ser observadas, de forma mais precisa, aproximações e distâncias. O MEV foi desenvolvido no âmbito do sistema SMART (*System for the Manipulation and Retrieval of Text*) (SALTON; MCGILL, 1983).

As medidas derivadas do MEV e de outros modelos de agrupamento podem ser usadas em visualizações de diferentes formas, o dendrograma apresentado anteriormente na Figura 11 é resultado de outra técnica estatística que estabeleceu a relação entre autores e assuntos. A seguir, destacamos formatos de visualização que têm grande potencial de evidenciar agrupamentos.

2.7.2.1. Grafos

Outro formato de visualização que pode evidenciar grupos e relações são grafos, usados para representar matematicamente relações entre elementos, bastante comuns, por exemplo, nos estudos de cocitação, como o gráfico representado na Figura 12, no qual Chen (2004) ressalta os artigos mais relevantes e mais citados no campo da Visualização da Informação.

Uma abordagem comum e intuitiva do layout de rede é modelar o gráfico como um sistema físico: os nós são partículas carregadas que se repelem, e os links são “molas” que unem os nós relacionados. Uma simulação física dessas forças determina as posições dos nós; técnicas de aproximação e distanciamento dos elementos evitam sobreposições, o que permite ao layout que represente um grande número de nós. Além disso, a interatividade permite ao usuário que direcione o foco e detalhamento de acordo com o clique em cada elemento (HEER; BOSTOCK; OGIEVETSKY, 2010).

A representação e o estudo das redes incluem – mas não se limitam a – áreas como a Antropologia, Biologia, Comunicação, Economia, Geografia, Ciência da Informação, Psicologia Social, Sociologia e Sociolinguística (SOUZA; ALMEIDA; BARACHO, 2013). Os grafos são recursos frequentes nesse contexto, utilizados por estudos acadêmicos, artigos e reportagens de análises em veículos de comunicação. Estudos que envolvem fluxos de comunicação, como a análise de troca de mensagens em redes sociais, troca de e-mails etc., se beneficiam do formato. O Twitter³³ é uma rede social que permite a mineração de dados estruturados de uma forma que seja possível criar laços entre as publicações e seus autores e identificar grupos com temas de interesse. É uma base muito utilizada na análise da repercussão de movimentos políticos, manifestações e debates públicos.

O grafo representado na Figura 33 é parte de um estudo realizado por Malini (2016), coordenador do Laboratório de Estudos sobre Imagem e Cultura (LABIC) da Universidade Federal do Espírito Santo, que analisa as mobilizações contra o aumento das tarifas de ônibus e trens em São Paulo em janeiro de 2015, denominadas #ContraTarifa e lideradas pelo coletivo Movimento Passe Livre.

³³ Disponível em: <<http://twitter.com>>. Acesso em: ago. 2019.

Figura 34 – Rede de Retweets acerca dos termos #ContraTarifa, #PasseLivre e #ContraoAumento e Passe Livre

Fonte: Malini (2006, p. 19).

É possível observar na imagem agrupamentos de discursos ou temas. Atribuídos por cores diferentes, a posição de cada um deles no grafo explicita também distâncias entre os perfis-membros, evidenciando alinhamentos de discurso entre diferentes emissores. Segundo o autor, basicamente três perspectivas aglutinam a hegemonia da narrativa do ato: a do Movimento Passe Livre (laranja), a da Polícia Militar (verde) e a do petismo (vermelha). Para cada discurso, o autor ainda consegue analisar o léxico usado por cada grupo e características de comportamento (MALINI, 2006).

Grafos também podem ser usados para descobrir relações não óbvias sobre documentos que já tenham alguma estrutura formada. A Figura 34 mostra os resultados da análise dos documentos de inscrição de trabalhos para a Feira Brasileira

de Ciências e Engenharia (Febrace)³⁴ (ANEXO 1). A base é composta por 340 documentos. As cores indicam a área de conhecimento à qual cada trabalho pertence. A aproximação de documentos de cores distintas evidencia relações temáticas não óbvias. A investigação pôde evidenciar, por exemplo, que trabalhos de diferentes áreas do conhecimento estavam abordando o tema “água”, identificado no aglomerado interconectado por arestas, à direita, na imagem. A identificação dessas relações pode, por exemplo, sinalizar possíveis colaborações entre equipes e abordagens diferentes para um mesmo problema.

Figura 35 – Grafo gerado a partir de aproximação de termos utilizados nas fichas de trabalhos inscritos na Feira Brasileira de Ciências e Engenharia (Febrace) em 2016.

Fonte: Caetano e Martinazzo (2017, no prelo, p. 6). Verificar anexo 1.

³⁴ Estudo realizado pelo autor deste trabalho em parceria com Alexandre Martinazzo (Poli-USP).

Nota-se que cada abordagem apresentada utiliza técnicas distintas de agrupamento e evidenciam diferentes tipos de relações. As possibilidades são diversas e, mais uma vez, é importante ter conhecimento claro sobre o objetivo do que se deseja observar, a natureza do conjunto de documentos a ser analisado e método de agrupamento a ser utilizado.

2.7.2.2. Gráficos de dispersão

Estruturas visuais de duas dimensões são tipicamente utilizadas para gráficos ou dados geográficos, mas também têm sua aplicação para visualização de documentos (CARD; MACKINLAY; SHNEIDERMAN, 1999). Os gráficos de dispersão (*scatterplots*, em inglês) são visualizações em duas dimensões e valores projetados em eixos cartesianos. Os pontos localizados no gráfico ganham coordenadas de acordo com as variáveis representadas nos eixos, localizando no espaço os elementos representados e evidenciando proximidades, grupos e características de distribuição.

Além da posição derivada dos valores nos eixos, outras variáveis associadas ao ponto podem ser expressas visualmente com o uso, por exemplo, de sua cor, tamanho ou formato e, dessa forma, incluir outras dimensões simultâneas para análise de padrões.

O gráfico de dispersão é uma das visualizações mais utilizadas na aplicação da estatística para análise de dados científicos. Tufte (1983) estimou que 70 a 80% das visualizações usadas em publicações científicas eram *scatterplots*. Friendly e Denis (2005) atribuem ao gráfico o *status* de segunda grande inovação em visualização de dados, depois das representações de William Playfair (1759-1823) – barras, linhas e pizza. Os autores consideram o formato a “mais versátil, polimórfica e mais útil invenção em toda história dos gráficos estatísticos” (FRIENDLY; DENIS, 2005, p. 103, tradução nossa)³⁵.

³⁵ “[...] most versatile, polymorphic, and generally useful invention in the history of statistical graphics.” (FRIENDLY; DENIS, 2005, p. 103)

O formato é utilizado como instrumento de análise de documentos e seus metadados. O *Film Finder* (Figura 14) utiliza um gráfico de dispersão em seu *starfield display*, por exemplo. Correa e Vieira (2013) exploram gráfico de dispersão para analisar relações de proximidade temática entre cabeçalhos de assunto – termos de um vocabulário controlado. A Figura 35 traz a visualização dos termos em espaço reduzido, criada através da escala multidimensional (MDS) sobre a matriz de coocorrência dos termos.

Apenas os descritores com frequência maior ou igual a 20 aparecem no mapa, o que evidencia um risco do uso de muitos elementos na visualização.

A proximidade dos termos no mapa indica a relação de associação entre eles, isto é, os conceitos mais próximos no mapa são os que possuem uma proximidade temática maior ou maior co-ocorrência.

Figura 36 – Cartografia temática para o Programa de Pós-graduação em Direito

Fonte: Correa e Vieira (2013, p. 31).

Aplicado ao contexto de bibliotecas digitais, Marks et al. (2005) propõem uma ferramenta de visualização chamada *ActiveGraph* (Figura 36) que, segundo os autores, oferece ao usuário uma visão concisa e customizável de uma biblioteca digital. O conjunto de documentos é representado como um dataset em gráficos de

dispersão em duas ou três dimensões. O conjunto pode representar quaisquer documentos em qualquer mídia – livros, periódicos, artigos, imagens, páginas e recursos de web ou toda a base unificada. Os autores ressaltam que, uma vez que a ferramenta é voltada para documentos de bibliotecas digitais, as variáveis são constituídas em grande parte por dados categóricos e metadados, como título, autor, data de publicação, números de citações.

Camadas de interação são disponibilizadas e os metadados de cada documento podem ser acessados no painel à direita. Recursos de gestão permitem ao usuário editar e criar metadados.

Figura 37 – Tela da ferramenta ActiveGraph. Os controles na base da tela permitem atribuir metadados para cada atributo, como eixo X, Y, cor, tamanho, forma.

Fonte: Marks et al. (2005).

A Figura 35 traz a representação de documentos por número de citações dentre um conjunto de artigos publicados sobre visualização da informação entre 2002 e 2003

pelo Los Alamos National Laboratory (LANL). O indicador de citações está no eixo Y. O eixo X indica periódicos. Cada ponto é um documento, e rapidamente podemos identificar aquele de maior número de citações, que aparece selecionado. A aplicação de uma escala logarítmica permite ver faixas de publicação, em que a grande maioria tem poucas ou nenhuma citação (MARKS et al., 2005).

Infelizmente, assim como outros exemplos deste trabalho, a aplicação não está disponível. Embora o uso de gráficos de dispersão para recuperação de documentos pareça ser uma abordagem promissora, não foram identificadas plataformas ativas e acessíveis que usem gráficos de dispersão para recuperar documentos em suas interfaces.

2.7.3. Espaço 3D

O uso de 3D aumenta o espaço disponível, e é um recurso explorado em formatos que demandam estruturas de base bem maiores, acima de 5 mil elementos. A técnica chamada “*cone tree*” (Figura 37) representa uma árvore em três dimensões e consegue disponibilizar uma quantidade de ramificações com capacidade de representar hierarquias bem maiores e ramificadas, uma vez que os representa em anéis verticais ou horizontais, conectados aos seus pais, criando visualizações em cone. Quando o usuário seleciona um nó, um anel se expande dele, e o coloca em primeiro plano, sem perder o contexto geral da hierarquia. A interação de rodar o cone faz com que seja possível a seleção entre itens do mesmo nível.

Figura 38 – Visualização cone tree, em suas versões vertical e horizontal.

Fonte: Dix (2012, p. 13).

Segundo Chen (2004), um mundo em 3D nos dá um grau a mais de liberdade, mas, algumas vezes, esse grau de liberdade também nos conduz a confusões e complicações. Shneiderman (2003) questiona se a riqueza de gráficos em 3D é ou não um bom modelo de análise.

É preciso notar, entretanto, que na medida em que cresce o volume de documentos, assim como a quantidade de variáveis e possibilidades de organização, as visualizações 1D e 2D acabam não sendo suficientes para refletir o novo grau de complexidade. No contexto Big Data, a quantidade de informações demonstradas em uma única visualização fica mais difícil de ser interpretada, ao mesmo tempo em que a capacidade de interação no ambiente digital se torna mais ampla.

Os avanços no armazenamento e processamento de informações tornam as interfaces 3D mais plausíveis, e novos recursos são desenvolvidos e tornados acessíveis. O *TensorFlow*³⁶, por exemplo, é uma biblioteca de código aberto para Aprendizado de Máquina aplicável a uma ampla quantidade de tarefas, como a criação e treinamento de redes neurais para detectar e decifrar padrões e correlações entre textos, análogo à forma como humanos aprendem ou raciocinam. A visualização de dados é um recurso essencial para a compreensão de como a máquina se comporta.

³⁶ Disponível em: <<http://tensorflow.org>>. Acesso em: 17 fev. 2019.

Tais sistemas complexos lidam com dados multivariados de grande dimensionalidade, impossíveis de se compreender em gráficos de uma ou duas dimensões.

A Figura 38 mostra a ferramenta *Embedding Projector*³⁷, disponibilizada pela organização que mantém o *TensorFlow* e visa transparecer ao usuário humano o resultado dos algoritmos utilizados no processo de Aprendizado de Máquina. A figura mostra um mapa em três dimensões de representações de palavras em vetores, que aproximam umas às outras a depender de suas combinações em grandes massas de texto. O modelo calcula medidas de similaridade em vetores 3D que aproximam palavras de forma semântica. Uma evolução MEV. Não é intuito deste trabalho detalhar a técnica, mas, sim, evidenciar os recursos do sistema representado.

Figura 39 – Sistema Embedding Projector, parte do projeto TensorFlow

Fonte: O autor.

Na ferramenta, as palavras são distribuídas no espaço tridimensional, e recursos interativos são disponibilizados ao usuário – por exemplo, zoom e controles de navegação. Ao clicar em um ponto, que na imagem representa palavras, são

³⁷ Disponível em: <<http://projector.tensorflow.org/>>. Acesso em: 17 fev. 2019.

ressaltadas na rede 3D aquelas que possuem aproximação semântica e usos recorrentes em frases e encadeamento de parágrafos. Ainda é possível ao usuário carregar sua própria base de dados para visualização do comportamento dos algoritmos criados pelo *TensorFlow*.

Um passo além ao 3D é explorar plataformas imersivas para navegação e análise de dados multidimensionais. Wagner Filho (2018) propõem uma aplicação em realidade virtual que permite ao usuário que explore dados em um gráfico de dispersão 3D. O usuário é equipado com dois controladores em formato de mãos que, como seletores, oferecem a interação com cada ponto de dado.

Figura 40 – Tela do experimento de visualização de dados multidimensionais em espaço imersivo.

Fonte: Alberto e Filho (2018, p. 43).

Os dados utilizados para o experimento são provenientes do histórico de votações da Câmara dos Deputados Federais de mais de vinte partidos, e o posicionamento oficial de cada um deles. No resultado, as cores representam partidos; bolas são deputados; e quadrados, o posicionamento do partido.

Por meio de técnicas de redução de dimensionalidades e cálculos estatísticos, podem ser vistos a proximidade de deputados e de partidos de acordo com o posicionamento histórico de suas votações.

Figura 41 – Participante dos testes de análise de dados em ambiente imersivo de realidade virtual.

Fonte: Wagner Filho (2018, p. 48).

Nos testes realizados com trinta usuários com idades e experiências distintas (Figura 40), os autores concluíram que o esforço para encontrar as informações foi menor e demandou menos navegação, além de prover percepções subjetivas de precisão e engajamento muito maiores (WAGNER FILHO, 2018).

2.7.4. Interatividade

Seja qual for a escolha de projeto para visualização, uma característica que deve ser intrínseca ao uso de gráficos e diagramas como parte da interface de SRI é a interatividade, ou a capacidade do usuário de alterar os parâmetros que estão sendo visualizados para que possa encontrar informações e conexões adicionais.

A interação do usuário pode suavizar dificuldades intrínsecas na compreensão e utilidade da visualização, permitindo escolhas que podem ser alteradas dinamicamente (DIX, 2013). Isso pode ser usado em visualizações complexas, como no *Film Finder*, que permite a edição de intervalos e parâmetros usados como filtros para atualização da visualização estática. No entanto, diversas situações mais simples de visualização podem ser potencializadas com o uso de interações menos complexas.

Shneiderman (1996) aglutina diversos métodos de interação em três componentes, referenciados como o “Mantra da Visualização”:

1. **Overview** (Panorama): ter uma visão panorâmica de toda a coleção;
2. **Zoom and Filter** (Zoom e Filtros): refinar a busca e identificar itens ou grupos de itens de maior interesse;
3. **Details-on-Demand** (Detalhes sob demanda): selecionar um item ou um grupo de itens e obter mais detalhes quando necessário.

De acordo com Card, Mackinlay e Shneiderman (1999), a sequência Overview-Details é responsável por oferecer ao usuário um mecanismo em que ele possa navegar, sistemicamente, na estrutura, na relação todo-parte, quando se pretende comparar uma variável com outras e, também, quando se deseja observar uma variável exclusiva.

Dix (2013) apresenta um detalhamento dessas funções em diversos possíveis recursos que podem adicionar uma rica camada de interação do usuário com as visualizações e servem como organização de diversos exemplos já apresentados neste trabalho. São eles:

Realce e foco: No exemplo cone tree (Figura 36), o usuário seleciona um dos nós, fazendo com que as categorias-pai sejam iluminadas, além de o item rodar para a frente, na tela. O realce é útil até mesmo para simples elementos, mas são mais relevantes nesses casos em que há necessidade de realçar trajetos de hierarquia ou reforçar arestas ligadas a um determinado nó de um grafo.

Detalhamento e hyperlinks: nos exemplos interativos de treemaps e sunburst, é possível adentrar a uma categoria e abrir suas subdivisões e ainda detalhar informações sobre determinado documento. O detalhamento pode vir na mesma visualização ou estar conectado via hiperlink com recursos e visualizações próprias.

Panorama e contexto: assim como pode ser interessante detalhar algum conjunto de documentos, é importante também termos uma visão de panorama, uma visão completa que ofereça a compreensão de um contexto. O exemplo *IMDb Explorer* (Figura 16) oferece as duas visões – uma global, mostrando hierarquias de gêneros, país e diretor de filme; e, ao lado, detalhes sobre a seleção. O Browser Hiperbólico,

grafos e gráficos de dispersão também são formatos ricos para trabalhar panorama e contexto.

Parâmetros editáveis: ao fazer uma visualização estática, o autor deve fazer as escolhas de acordo com os parâmetros que quer mostrar. Ao torná-los editáveis pelo usuário, é possível transformar a busca pelo foco de interesse mais facilmente. Os filtros dinâmicos do *Film Finder*, ou a definição do intervalo de anos de publicação na busca analítica da *Scopus*.

Troca de representações: outra forma simples de interação é permitir ao usuário trocar completamente o tipo de visualização, como for mais útil ao tipo de análise que se pretende fazer. A função de análise de resultados de busca da WoS, por exemplo, permite ao usuário escolher entre a visualização em treemap ou gráficos de barra. Esse é um recurso usado também para adaptar a visualização ao contexto de aplicação do usuário, ao identificar, por exemplo, o tipo de dispositivo que está sendo usado para ver e interagir com a visualização.

Fusão temporal: visualizações presentes em uma mesma interface, como em painéis de controle, podem ter pontos de conexão e mostrar aspectos diferentes de uma mesma base de documentos. Essas visualizações são sincronizadas, ou seja, são afetadas simultaneamente por filtros, ou destaque de um mesmo elemento, pontos de conexão entre as visualizações. Um exemplo de fusão temporal pode ser observado no *IMDb Explorer* (Figura 16), em que a interação em um gráfico afeta o outro.

2.7.5. Painéis de controle

Dashboard, ou painel de controle, é um termo bastante utilizado na área de negócios. São interfaces que devem apresentar, de forma rápida, os principais indicadores de sucesso de uma empresa ou monitoramento de determinada operação. Frequência e facilidade de atualização também são fatores importantes que determinam a funcionalidade. O termo foi adotado por analogia ao painel do carro, onde diversas medidas do funcionamento do veículo são disponibilizadas simultaneamente num *painel de controle*.

Para além da definição e uso pelo mercado, os dashboards são ferramentas que agrupam visualizações, de uma ou mais fontes de dados, em uma mesma tela, em um mesmo contexto. Dessa forma, é possível ao usuário interpretar as informações de forma conjunta. Essa combinação de visualizações pode auxiliar, por exemplo, a compreensão do resultado de buscas, como foi aplicado na ferramenta de análise de resultados de busca da *Scopus*. A ferramenta, ao disponibilizar as visualizações possíveis num mesmo contexto, pode ser considerada um dashboard simples, no qual analisamos várias dimensões simultaneamente, agilizando a compreensão do resultado geral da busca. No modelo adotado na WoS, não há a possibilidade de ver e comparar duas variáveis diferentes, sendo necessária a análise separada de cada instância.

As ferramentas OLAP, citadas anteriormente, são recursos populares para o desenvolvimento de dashboards. A interface de buscas para a base de dados de filmes IMDb é um exemplo de painel de controle direcionado à recuperação de informações, utilizando diversos gráficos simultâneos que atualizam de acordo com a busca ou edição dos parâmetros.

A Fusão Temporal, tipo de interação que pode ser destacada nesse contexto, uma vez que a sincronização de interoperação entre as visualizações pode enriquecer o tipo de informação que se deseja encontrar. Esse tipo de disposição permite, também, a visualização de mais de uma base de dados simultaneamente, com uma seleção de visualizações de cada uma delas, reforçando interações de panorama e contexto.

2.7.6. Dispositivos de acesso

Um grande desafio ao uso de VI em interfaces de SRI é a crescente necessidade de uso nos mais variados tipos de dispositivo, em especial celulares. O menor tamanho da tela e sua mudança de proporção e a menor resolução podem inviabilizar o uso desses recursos.

Chittaro (2006) publicou um artigo sobre visualização da informação em celulares. O autor focou em restrições dos dispositivos como tamanho reduzido, baixa resolução, diferentes proporções de tela ou hardware menos potente. O tamanho limitado da tela

significava que o modelo clássico do desktop, overview + detail não funcionava bem em celulares.

Alternativas passam por simplesmente deixar a visualização aparecer pequena na tela, dificultando a leitura, incluindo recursos de zoom, ou simplesmente ocultar o gráfico. Outras soluções, geralmente mais bem-sucedidas, mudam o tipo de visualização e como são apresentadas na tela, privilegiando os dados mais relevantes para os casos de uso proporcionados pelo acesso via celular.

Blumenstein, Niederer e Wagner (2016) realizaram um levantamento de publicações dos principais repositórios de visualização da informação relacionados a dispositivos celulares com o objetivo de chegar a modelos de avaliação mais precisos. Dentre os apontamentos, reforçam três questões – a dificuldade em reproduzir testes e ter mais rigor no controle de avaliação dos usuários; a diversidade de dispositivos, e seus diferentes tamanhos e resoluções de tela, além de sistemas operacionais e suas versões; e, por fim, uma compreensão melhor das necessidades do usuário para cada caso de uso, considerando o momento e dispositivo de acesso (BLUMENSTEIN; NIEDERER; WAGNER, 2016).

2.8. Extroversão e descoberta: uma discussão

Para concluir a seção de referencial teórico, resgatamos os termos usados no título desta dissertação – extroversão e descoberta - como descritores das funções a serem cumpridas por SRIs que utilizam visualizações em suas interfaces.

A seleção dos termos buscou resumir os aspectos levantados nas seções anteriores, e são listados a seguir:

- a) devem ser compreendidos a partir das lentes da CI dos paradigmas físico, cognitivo e social da informação;
- b) servem ao objetivo de um usuário de recuperar documentos, que são unidades de dados semi ou não estruturados, caracterizados, necessariamente, pelo uso da linguagem e a presença de variáveis categóricas; e
- c) referem-se ao sucesso de SRIs que utilizam visualizações em suas interfaces e a avaliação dos recursos utilizados.

Como abordado na introdução, “extroversão” é um termo com origem na Psicologia, citado pela primeira vez na teoria dos Tipos Psicológicos de Carl Jung de 1921, e descreve características de pessoas em que se observa a atenção para a ação, comunicabilidade, sociabilidade e facilidade de expressão. A pesquisa adota o termo “extroversão” por traduzir qualidades de ação atribuídas à função de comunicação de um SRI. A metáfora é conveniente por aproximar aspectos cognitivos e sociais, da psicologia, ao contexto técnico-físico de objetivos de um SRI.

Já o termo “descoberta” deriva da expressão “descoberta de informação” em bancos de dados, utilizada com frequência no campo da CI e mesmo da computação. As ontologias, por exemplo, são caracterizadas, entre outras coisas, como instrumentos de descoberta de conhecimentos (ALMEIDA, 2013). Outra expressão que pode ser usada é “mineração de dados”, comumente aplicados no campo da Ciência de Dados e Computação. Mineração e descoberta também definem, por fim, a capacidade de recuperar informações relevantes (WU et al., 2013).

Em busca de uma síntese dos conceitos apresentados, sugerimos as seguintes definições:

Extroversão: a capacidade e o preparo do sistema para gerar visualizações de dados

Descoberta: a capacidade de otimização do sistema em facilitar a recuperação de informações de acordo com a necessidade de um usuário.

Na área da VI, Mackinsey (1986) propõe dois critérios de avaliação para linguagens gráficas, a *expressividade* e a *efetividade*. Nascimento e Ferreira (2005) resumem:

Uma visualização pode ser considerada *expressiva* se ela é capaz de mostrar todos os dados de interesse do usuário e nada mais que isso. Já a *efetividade* está relacionada com a facilidade de se compreender os dados apresentados. Para ser efetiva, uma visualização deve ser de rápida percepção e induzir a uma quantidade menor de erros de interpretação do que outras formas de se visualizar os mesmos dados. (p. 1271)

É preciso reforçar que Mackinsey (1986) utiliza os termos *expressividade* e *efetividade* como critério de avaliação de visualizações em qualquer caso de uso. Este trabalho apresenta os termos *extroversão* e *descoberta* no contexto de sistemas de recuperação da informação. Os critérios de Mackinsey não deixam de ser relevantes, afinal podem ser utilizados para avaliar os recursos de visualização aplicados na composição de interfaces de SRI, mas é a capacidade de extroversão e descoberta que podem caracterizar o sistema todo.

Os critérios aqui propostos devem ser objeto de estudo posterior, que priorize o levantamento de dados empíricos de forma a detalhar e aprofundar de forma propositiva as definições. Ainda assim, o resultado do levantamento e análise bibliográfica realizada neste trabalho permite apontar algumas questões que podem nortear a avaliação segundo as capacidades de extroversão e descoberta.

Em primeiro lugar, antes da avaliação através destes critérios, é preciso compreender o contexto em que documentos a serem visualizados e recuperados foram produzidos. Qual é o contexto institucional e social de produção dos documentos? Quais são as tarefas a serem atendidas e que fazem parte do propósito desta instituição? Essas questões são provenientes de uma interpretação do paradigma social da informação e ajudam a esclarecer propósitos, impactos e características dos documentos que compõe a base de dados.

Com o cenário e contexto do sistema claramente identificados, outras questões suscitam de cada critério de avaliação das funções das visualizações em interfaces de SRI e que podem ser associadas aos paradigmas físico e cognitivo.

A capacidade de **extroversão** de um sistema pode ser avaliada a partir de questões como as exemplificadas abaixo:

- A base de dados está preparada para lidar com dados semi ou não estruturados?
- Os documentos possuem metadados e variáveis próprias destinadas à visualização?
- O sistema é potente o suficiente para gerar visualizações dinâmicas e de rápida atualização?
- O sistema pode ser acessado por qualquer tipo de dispositivo?
- A performance do sistema nestes dispositivos é satisfatória em todos os casos de uso?
- Qual a performance de cada visualização aplicada na interface segundo os critérios de expressividade e efetividade de Mackinsey (1986)?
- Qual é o nível mínimo de conhecimento analítico e repertório visual necessários ao usuário para a rápida leitura dos dados?
- O usuário teve sucesso em compreender as visualizações propostas e navegar pela interface do sistema?
- Os dados disponibilizados na interface são relevantes às tarefas a serem realizadas?

Já o critério **descoberta** deve responder a perguntas de natureza diferente:

- A necessidade de recuperação do usuário foi atendida?
- Quanto tempo e passos de navegação o usuário levou até efetivamente cumprir sua tarefa de recuperação de informação?
- O usuário entende o método de organização da informação utilizado? É familiar ao contexto institucional de produção dos documentos?
- O sistema possui recursos de interatividade? Permite aplicar filtros, detalhar informações e descobrir relações entre documentos?

- Na percepção dos usuários, as visualizações que compõe a interface o ajudaram a cumprir sua tarefa?
- O sistema permite acesso aos detalhes de cada documento da base de dados?

Ao longo deste trabalho foram apresentados diversos exemplos de SRI que utilizam visualizações como parte de sua interface e recursos de navegação e que podem ser avaliados usando os critérios propostos.

A comparação dos sistemas de análise de buscas dos repositórios científicos WoS e Scopus é um exemplo conveniente por terem características institucionais similares, assim como a mesma natureza de documentos e forma de estrutura. As visualizações disponíveis nas duas plataformas utilizam dados bibliográficos e podem ser considerados de fácil interpretação, comuns ao repertório ao usuário acadêmico científico de forma geral.

Num rápido exercício, pode-se concluir que a base Scopus tem uma capacidade de descoberta maior que a base da WoS, pois permite a aplicação de um filtro.. A disposição em formato de painel de controle permite a análise simultânea de informações, dando um melhor contexto ao usuário.

Já a WoS tem uma capacidade de extroversão pode ser considerada maior, afinal oferece mais categorias a serem visualizadas, além da opção de troca do tipo de visualização.

A comparação entre os sistemas Film Finder (figura 14, p. 83), “What movies have we produced” (figura 15, p. 85) e IMDb Explorer (figura 16, p. 87) também nos é conveniente. As 3 bases são compostas por documentos similares – dados de filmes, diretores, atores e outros agentes da produção cinematográfica mundial. O contexto tecnológico é diferente para cada uma das plataformas. O Film Finder, por exemplo, é datado de 1999 e os recursos visuais utilizados são claramente mais rudimentares e simples, o que pode limitar sua capacidade de extroversão se comparado aos outros dois sistemas. Sua capacidade de descoberta não é prejudicada e pode ser destacada pela disponibilização de filtros dinâmicos e recursos de interatividade que facilitam observar e recuperar as fichas de cada ator, diretor e demais agentes presentes na base.

A interface “What movies have we produced?” (LIM, 2014) utiliza recursos já presentes no Film Finder, como os filtros dinâmicos que permitem ajustar o período e as notas de ranking a seres analisadas nas visualizações abaixo. Claramente, a base de dados tem um volume de informações maior, pois considera ao menos 15 anos a mais de filmes, atores e outros agentes da cadeia cinematográfica. Sua estrutura de organização, de fácil compreensão e bastante detalhada em seus metadados e variáveis, tem um bom potencial para extroversão. Lim (2014) seleciona apenas alguns dos metadados comuns ao repertório de um usuário médio, como país de origem, gênero, ano de produção e nota de avaliação. A curadoria dos metadados, somados aos recursos de filtro e a visualização em que é possível visualizar e acessar cada título de filmes, garante também uma boa performance de descoberta.

Ainda utilizando a mesma base, a plataforma IMDb Explorer (BURCH, 2018) tem caráter experimental e tenta evidenciar padrões e relação entre os dados que mostrem trajetórias. Para isso, utiliza duas visualizações interativas – a ação em uma interfere em tempo real na outra visualização, com um comportamento de filtro. Ao escolher uma unidade de informação na visualização *circle parking* nomeada como movie cosmos, o mesmo elemento é destacado na visualização *carrer line*, mostrando a evolução e faturamento estimado ao longo dos anos. É possível observar quando um ator, diretor ou filme foi lançado e quão bem-sucedido foi seu retorno financeiro. Essas características evidenciam uma grande capacidade de extroversão, porém alguns fatores interferem negativamente em sua capacidade de descoberta. Ainda que a interface possibilite a aplicação de filtros e diferentes ordenações hierárquicas dos dados, os recursos de visualização utilizados são menos frequentes no repertório do usuário médio, o que aumenta as chances de insucesso. Além disso, são poucos os detalhes oferecidos sobre cada informação e não há link ao conteúdo original.

Último exemplo a ser analisado, o sistema Scatter-Gather Browser (figura 31, p. 111) possui uma alta capacidade de descoberta, com algoritmos rápidos que organizam o documento através da similaridade entre eles e apontados quais os termos típicos de cada conjunto identificado pelo algoritmo. No entanto, sua capacidade de extroversão pode ser interpretada como baixa – a visualização dos grupos depende da leitura atenta das palavras e documentos típicos para se ter uma compreensão do todo e identificar padrões e diferenças entre grupos.

Esta seção buscou sintetizar o referencial teórico do trabalho como base para o levantamento de uma hipótese e de uma proposta de método de avaliação que considere critérios denominados como extroversão e descoberta. É preciso reforçar que esta pesquisa carece de aprofundamento, mas, neste trabalho, nos limitamos a uma introdução e convite para estudos posteriores.

3. Estudo de caso: Painéis de Análise de Comunicação para empresa prestadora de serviços públicos municipais

Este capítulo visa concretizar os achados e questionamentos do levantamento bibliográfico a partir de uma perspectiva empírica, baseada na metodologia de Estudos de Caso. Segundo Yin (2003), o estudo de caso é considerado uma importante estratégia metodológica para a pesquisa em Ciências Humanas, pois permite ao investigador um aprofundamento em relação ao fenômeno estudado, revelando nuances difíceis de serem enxergadas “a olho nu”. O autor afirma que o estudo de caso favorece uma visão holística sobre os acontecimentos da vida real, com destaque para seu caráter de investigação empírica de fenômenos contemporâneos.

Este trabalho adotou o Estudo de Caso do modelo causal/exploratório que, segundo Yin (2003), embora não se resuma à exploração, permite ao pesquisador elencar elementos que lhe permitam diagnosticar um caso com perspectivas de generalização naturalística.

Pode-se dizer que o estudo de caso apresentado é um estudo organizacional e gerencial, no qual são discutidos parâmetros de planejamento, coleta, organização e análise de dados não estruturados previamente. É um estudo de caso único, de sentido revelador – a situação que ocorre quando o pesquisador tem a oportunidade de observar e analisar um fenômeno previamente inacessível à investigação científica (YIN, 2003).

O caso analisado é o projeto e manutenção de uma plataforma de monitoramento e análise de comunicação de uma empresa de prestação de serviços públicos municipais em diversas cidades espalhadas pelo Brasil. A plataforma tem como objetivo a prevenção e mitigação de crises, além da análise de percepção da população de cada cidade sobre a marca e os serviços prestados. Os usuários-alvo são, em grande parte, profissionais ligados à comunicação e de diversos níveis gerenciais – pessoas ligadas às realidades locais em que atuam (responsáveis pela gestão da comunicação nos níveis municipal e regional), assim como gestores e a

diretoria de comunicação que devem ter uma visão mais ampla e comparativa entre cada localidade.

Ao longo de três anos, no período de 2016 a 2018, foram coletadas mais de 30 mil notícias, 320 mil mensagens de redes sociais, 30 mil registros de atendimento ao público das comunidades locais, analisados no contexto do cruzamento com dados e indicadores públicos e dados operacionais da empresa.

Esses conjuntos de documentos permitiram a realização de testes e estudos *ad hoc*, admitindo também a leitura de conjunturas temáticas de interesses diversos, explorando recursos de visualização da informação, não só com o intuito de evidenciar padrões e tendências, mas também uma forma de recuperação dos documentos digitais para análise. O problema apresentado, portanto, explora como os dados, coletados e organizados em visualizações, são incorporados no trabalho dos comunicadores, questionando e elencando questões da apropriação do sistema e porque foram ou não efetivamente adotados pelos profissionais.

O projeto gerou dados estratégicos sigilosos que não são fornecidos publicamente. Para manter a privacidade desses dados, foram omitidos nomes de categorias utilizados na estruturação dos dados, atendo-se ao funcionamento e seu papel na organização da informação e análise, porém evitando o detalhamento do significado de cada uma. Da mesma forma, são omitidas informações de localidade e nome de pessoas. São discutidas especificamente os passos de desenvolvimento do sistema e a aplicação de visualizações neste processo. Ao longo da explanação estão incluídas diversas evidências de telas geradas para interpretação dos documentos.

São consideradas variáveis importantes as condições de disponibilidade de dados, técnicas de coleta, ferramentas utilizadas, aspectos da realidade corporativa do caso estudado, assim como o preparo e a compreensão dos usuários-alvo para a utilização efetiva da plataforma construída. O projeto foi encerrado no fim do ano de 2018, e os motivos estão expostos nos resultados do estudo.

O arcabouço teórico discutido ao longo da segunda parte desta dissertação fornece a este estudo de caso os parâmetros de análise e funciona como matriz exploratória e planejamento do estudo.

O caso analisado tem sua existência diretamente relacionada ao contexto Big Data apresentado no capítulo 2.1. As propriedades “volume”, “velocidade” e “variedade” descrevem as características da informação – os documentos e a sua forma de coleta. “Tecnologia” e “métodos analíticos” são observados como princípios dos recursos das ferramentas utilizadas ao longo do projeto, e “valor” pode ser relacionado ao uso estratégico das análises das informações na definição de ações práticas.

As lentes da Ciência da Informação, em especial as perspectivas apresentadas pelos paradigmas físico, cognitivo e social da informação, aprofundados no capítulo 2.2, auxiliam-nos a ter uma compreensão contextualizada da área ao qual o estudo é aplicado.

Neste capítulo, adotamos como linha condutora a estrutura a seguir:

- a) caracterização do contexto do projeto e impactos na disponibilidade de dados;
- b) caracterização dos documentos coletados, aspectos de disponibilidade e ferramentas utilizadas;
- c) estratégias de organização da informação adotadas e os parâmetros utilizados para análise;
- d) como as informações foram dispostas para consulta e o formato de dashboard;
- e) exemplos de aplicação e visualizações; e
- f) avaliação de resultados.

3.1. Caracterização e peculiaridades institucionais da empresa

A instituição para a qual a plataforma foi desenvolvida, como mencionado anteriormente, é uma empresa privada que presta serviços de caráter público municipal. Até meados da década de 1990, esses serviços eram prestados exclusivamente por organizações públicas estaduais, com atuação em grande parte dos municípios limitados ao Estado de atuação.

Através de marcos regulatórios, o mercado de prestação de serviços municipais foi aberto à iniciativa privada, que pôde firmar contratos diretamente com os municípios, em diversos modelos jurídicos, como concessões completas ou parcerias público-privadas. Não serão detalhados os instrumentos de contratação e execução dos serviços.

O que importa à análise é que, por ser uma empresa privada, esta não está limitada por fronteiras estaduais e pode atuar de forma mais abrangente em todo o território nacional. Essa condição permite explorar prospecções e implementação de serviços da mesma natureza, porém condicionados a uma variedade quase infinita de contextos legislativos, jurídicos e culturais inerentes a cada município.

Em cada local há uma realidade geográfica distinta na disponibilidade de recursos naturais. Por ser um serviço público e de larga escala, os contratos preveem prazos extensos, de 30 a 50 anos. Essa característica também tem reflexos nas relações políticas locais, voláteis de acordo com o posicionamento do mandato na prefeitura de cada município. Por ter abrangência nacional, estão sujeitos não apenas às leis municipais, mas também às diferentes instâncias estaduais e federais.

Uma condição para a coleta dos dados é a disponibilidade das informações em formato digital passíveis de serem rastreadas. É natural que cidades com melhor infraestrutura e maior cultura de uso dos meios digitais pela população forneçam um maior número de documentos digitais a serem analisados, enquanto cidades com menor população e acesso digital mais limitado criam um cenário de escassez de dados para análise. Tais variáveis podem ser consideradas para contextualização de acordo com o paradigma social discutido no capítulo 2.2.

3.2. Caracterização dos documentos

O projeto estabeleceu a coleta de diferentes fontes de documentos. Foram coletadas mensagens em redes sociais, notícias de canais digitais, e-mails corporativos enviados em ações de assessoria de imprensa e, ocasionalmente, cadastros, dados

transacionais de contas e canais de atendimento (telefone, chat, e-mail). Cada um desses formatos possui características próprias e diferentes estruturas de metadados.

Os documentos provenientes de redes sociais estão sujeitos às regras e ao formato disponível em cada uma das plataformas. O Facebook foi, sem dúvida, a fonte mais relevante de informações, por ser a rede social mais popular e de maior abrangência no país. Segundo dados da empresa We Are Social, em parceria com a Hootsuite (2019), a rede é utilizada com frequência por mais de 130 milhões de pessoas no Brasil, que corresponde a 75% da população que tem acesso à internet (WE ARE SOCIAL; HOOTSUITE, 2019).

No entanto, a coleta de dados na plataforma só é permitida em páginas públicas, ou seja, aquelas com caráter institucional, geralmente ligadas à comunicação aberta com o público. Exemplos claros são páginas de empresas, instituições, associações civis, políticos e pessoas de renome que têm necessidade de reverberar atualizações e relacionamento com o público em geral. O regulamento da plataforma especifica que os dados compartilhados nessas páginas são abertos e possíveis de serem rastreados, enquanto os chamados “perfis privados”, que são as páginas de pessoas físicas, têm acesso restrito e a coleta de dados não é possível. Esse tipo de página, de perfil privado, permite ao usuário o controle da visibilidade do conteúdo compartilhado e é voltado para a comunicação entre amigos, parentes e conhecidos de forma geral.

Além disso, não são permitidas buscas automáticas em todo o estoque de páginas públicas da ferramenta. Para realizar a coleta, é preciso especificar exatamente quais as páginas em que se deseja realizar buscas, o que demanda um trabalho amplo de pesquisa anterior para localizar as páginas abertas de interesse. No âmbito deste projeto, foram mapeadas as páginas de veículos de imprensa, políticos, celebridades, instituições públicas e associações civis de todas as cidades e regiões atendidas.

Os documentos decorrentes da coleta de conteúdo das redes sociais geralmente resultam em uma estrutura preestabelecida pela plataforma, e são caracterizadas por corpos de texto mais curtos, e sem rigor ortográfico ou gramatical, o que dificulta, e em determinados momentos inviabiliza, a aplicação de métodos automáticos de compreensão de texto.

Somados ao conteúdo em texto, imagens e vídeos, ainda é possível coletar os dados de reações dos usuários como curtidas (e suas variáveis em coração, risadas, e outras representações de sentimentos), comentários e compartilhamentos.

Para a coleta de notícias, eram utilizadas duas fontes diferentes, que contemplavam mecanismos de buscas distintos e complementares disponíveis no mercado. A ferramenta *KnewIn*³⁸ possui em seu cadastro mais de 5 milhões de sites e blogues do mundo todo e, ainda assim, não cobre todas as localidades de interesse. Cidades menores possuem veículos de menor porte e de baixa relevância nessa ferramenta e não são captados. Como fonte alternativa, portanto, eram utilizados serviços de *clipping*, em que uma equipe de jornalistas identifica manualmente as notícias sobre a empresa nesses veículos de menor porte e as digitalizam.

O tratamento das duas fontes de notícia envolvia fases de desambiguação para que dois registros iguais vindos de fontes diferentes não fossem somados duas vezes. Novamente, há a dificuldade de identificação da origem e foi realizado um grande trabalho de levantamento de veículos locais.

Esses conteúdos são caracterizados por serem mais extensos. A dificuldade na aplicação de métodos automáticos deriva da complexidade do volume de palavras, que pode ser interpretado de forma pouco assertiva com o uso de métodos booleanos adotados.

Outros documentos utilizados foram registros de atendimento realizados nos canais da empresa nas redes sociais. São conjuntos de pequenas interações, registradas em formato de diálogo. Cada atendimento é, portanto, composto por um conjunto de mensagens.

A coleta de todos os documentos era feita diariamente e com a frequência era possível monitorar e identificar mensagens e notícias que pudessem apontar problemas operacionais e questões políticas.

As fontes de coleta eram os hiperlinks de sites, perfis de redes sociais, blogues e todo o tipo de canal digital.

³⁸ Disponível em: <<http://www.knewin.com.br>>. Acesso em: ago. 2019.

A gestão dessas fontes de informação é essencial para o sucesso da coleta e análise dos dados – afinal, é preciso saber de onde as informações estão sendo extraídas. Diversas fontes podem pertencer a um mesmo ator. Cada *ator* é classificado de acordo com uma característica de grupo social, ou tipo de ator. Foram utilizadas categorias como “empresas”, “imprensa”, “políticos”, “influenciadores”, “órgãos públicos”, entre outras.

Na ficha – ou documento – do ator, são especificados os canais digitais que possui (websites, perfis em redes sociais), e outras medidas quantitativas eram acrescentadas, como a quantidade de seguidores e a data da última atualização. Outras medidas poderiam ser acrescentadas sob demanda ou necessidade de análise.

A identificação do ator na coleta dos outros documentos e o cruzamento com estas fichas eram recursos poderosos na análise conjunta dos diferentes tipos de documento, além de criar medidas que ajudam a entender o poder de influência de cada ator a partir da presença dele em um grande volume de documentos.

A Figura 40 mostra o painel utilizado para compreensão da base de atores disponíveis e aponta quais os canais mais relevantes e de onde chegam mais documentos. Pela interface é possível compreender números de documentos, a divisão pelas classes de grupos de interesse, o hiperlink para todas as fontes associadas a cada ator. A associação à outras bases de documentos (notícias e postagens de sociais), permite também o acesso ao conteúdo mais recente produzido por um ator.

Pela mesma ferramenta, era possível ao comunicador editar e incluir novos atores, de forma a enriquecer continuamente o mapeamento de fontes de informação, além de permitir a consulta ao conteúdo gerado em cada canal.

Figura 42 – Painéis ilustram a base de fontes de informação utilizadas no sistema de monitoramento.

Fonte: O autor.

3.3. Estratégia de organização da informação

Uma vez coletados os dados, ainda não era possível associá-los de forma clara para análises quantitativa e qualitativa, pois não havia relação direta entre eles. De forma a construir uma compreensão unificada dos documentos de diversas naturezas foi construído um vocabulário controlado que pudesse ser utilizado em todos os tipos de documentos.

O vocabulário foi concebido de forma hierárquica, com base nos termos técnicos já utilizados na empresa. Foram estabelecidos três níveis hierárquicos – categorias, tags primárias e tags secundárias. “Tag” (etiqueta, em português) é um termo associado ao documento, que descreve seu conteúdo, e permite classificá-lo. Trata-se, portanto, de um processo de indexação de conteúdos com base em termos específicos.

As categorias compõem o primeiro nível da hierarquia, e cada documento pode ser associado a várias delas. Cada categoria era composta por tags primárias, um nível a mais de detalhamento em cada categoria que, na maior parte dos casos, já eram suficientes para a descrição do documento completo. As tags secundárias eram subordinadas às tags primárias e utilizadas na classificação como descritores, aplicadas apenas quando era necessário pontuar detalhes importantes para a análise correta dos dados.

Os termos utilizados para definir categorias, tags primárias e tags secundárias, foram documentadas e especificadas para referência interna do projeto. Para ilustrar a composição do vocabulário, exploramos algumas categorias. “Fatores externos” (Figura 42), que reunia, em sua definição, assuntos ou organizações fora do domínio da empresa, mas que possuem relações com ela. Dentre as tags primárias subordinadas à categoria, podem ser citadas “Entidades” – que indicavam menções a outras organizações ligadas aos serviços prestados, concorrentes, órgãos regulatórios, cada um especificado por seu nome ou sigla em tags secundárias. Outra tag primária subordinada à categoria “Fatores externos” era “Gestão pública”, assuntos relacionados a atores da cadeia executiva, legislativa e judiciária na política. As tags secundárias definiam os tipos de atores envolvidos, como prefeitura, vereadores, deputados estaduais, federais, polícia, judiciário, entre outros.

Figura 43 – Representação da categoria “Fatores externos”, em árvore e documentação.

Fonte: O autor.

Como exemplo complementar, podemos citar a categoria “Propaganda e marketing”, que organiza documentos derivados de ações de comunicação da empresa. Dentre as tags primárias, pode ser mencionado o termo “Campanhas”, cujas tags secundárias definiam os nomes de ações institucionais como grandes campanhas de conscientização, saúde e melhorias nos serviços prestados. Outra tag primária da mesma categoria é “Sustentabilidade”, que agrupava menções às ações junto à comunidade local. Cursos para a comunidade, concursos, ações com escolas são exemplos das tags secundárias.

Cada um dos serviços prestados eram categorias diferentes, e os mais eficientes para integração com os documentos digitais de outras naturezas da empresa, como os dados transacionais de atendimento. Sobre um mesmo serviço, era possível analisar quantitativa e qualitativamente a percepção negativa ou positiva da população. Uma grande quantidade de notícias e menções na rede pode não ser refletido em volume de atendimentos ao cliente, o que pode ser entendido como um indício de que determinado assunto, benefício ou problema operacional ligado ao serviço não impacta a operação, e ações de comunicação podem ser suficientes para mitigar possíveis problemas. Da mesma forma, o aumento de atendimentos sobre determinado serviço ainda não identificado em notícias e menções nas redes sociais

pode representar um problema que deve ser solucionado antes que domine noticiários e temas de debate da sociedade.

Parte da classificação era feita automaticamente, utilizando regras booleanas de busca que, ao associar determinadas palavras ou a ausência de outras, marcava a ocorrência com o termo ao qual a busca booleana apontava.

Tais regras eram construídas manualmente, e sua eficácia testada para aferição de resultados práticos. Após o método automático, era feita uma verificação manual, humana, para garantir a qualidade dos dados a serem analisados e minimizar o efeito de análises erradas.

O vocabulário construído teve como ponto de partida termos que já eram utilizados por áreas internas da empresa para denominar instâncias de seus serviços, nomes técnicos de problemas de operação e categorização utilizada na organização do atendimento de outros canais.

Esse ponto de partida foi crucial para obter conclusões associadas diretamente aos objetivos do projeto e a dados de outras áreas operacionais. Somados a esses termos, também eram marcados pessoas e locais de interesse citados em cada conteúdo. Não foram utilizados métodos automáticos mais complexos como a clusterização e o MEV.

O uso da mesma estrutura para classificar os diversos tipos de documento foi fundamental para realizar as análises globais. Um mesmo tema pode ter comportamentos diferentes em cada canal e fonte de informação. A interpretação, levando em conta o contexto local, pode gerar hipóteses sobre a intensidade de uma discussão e os argumentos que estão sendo utilizados.

3.4. Plataforma de visualização e consulta

Uma vez coletados em ferramentas distintas, os dados de redes sociais, notícias e atendimentos eram consolidados em bases de dados armazenadas na nuvem, e

disponibilizadas em uma plataforma OLAP, cujas características já foram abordadas no capítulo 2.6.

A plataforma se chama *Cortex Intelligence*³⁹, de origem e atuação no Brasil. A ferramenta permite que diferentes fontes de documento sejam cruzadas e disponibilizadas em formato de dashboards – conjuntos de visualizações de informação que permitem realizar análises mais aprofundadas sobre um conjunto de dados.

Para fins de análise de comunicação, foram criados dashboards específicos, de acordo com os objetivos que se desejava cumprir, além de tornar fácil o compartilhamento interno entre as equipes e disponibilizar para consulta.

Além das visualizações, a ferramenta dispõe recursos como filtros e campos de seleção. A interatividade é outra funcionalidade essencial e qualquer visualização permite o acesso a uma lista dos documentos que estão sendo considerados. Um clique a mais e os detalhes do documento podem ser acessados na íntegra.

Os recursos são disponibilizados on-line, o que permite o acesso remoto ao sistema controlado por senhas de acesso. A Figura 43 destaca um painel utilizado na análise de notícias e postagens de redes sociais.

³⁹ Disponível em: <<http://www.cortex-intelligence.com>>. Acesso em: ago. 2019.

Figura 44 – Painel utilizado para análise de notícias e mensagens de redes sociais.

Fonte: O autor.

3.5. Exemplos práticos de uso

Ao longo dos três anos de duração do projeto, foram realizados diversos estudos quantitativos e qualitativos da informação coletada.

A estruturação dos documentos digitais coletados permitiu o desenvolvimento de um modelo de análise comparativo que utilizava recursos qualitativos para atribuir pesos e relevância aos documentos coletados.

O modelo sugeria um **indicador de repercussão**, uma métrica derivada de um cálculo ponderado de medidas de engajamento de um conteúdo, que evidenciava o potencial de impacto de determinados assuntos em cada canal analisado. Para conteúdos de redes sociais, por exemplo, eram consideradas as reações (curtidas, amor, raiva, risos), comentários e compartilhamentos. A polarização das reações e comentários, assim como a avaliação do conteúdo original, permitia atribuir indicadores positivo e negativo para cada conteúdo, tema ou serviço a ser analisado. A repercussão de notícias envolvia, além dos compartilhamentos em redes sociais, a quantidade de vezes em que era replicada.

A comparação era feita entre quatro instâncias:

- a) monitoramento de imprensa;
- b) monitoramento de redes sociais;
- c) *performance* de conteúdo digital próprio; e
- d) interações com o público nos perfis sociais.

A primeira visualização oferecida na plataforma mostrava as principais medidas relacionadas a cada instância, evidenciando o volume de menções e o valor do Indicador de Repercussão positivo e negativo, a depender da análise de polarização dos documentos. O volume de atendimentos ajuda a compreender o tom das interações. Na Figura 43, podemos observar, por exemplo, que 17 mensagens em redes sociais tiveram maior repercussão do que as 30 notícias capturadas. A empresa realizou 42 postagens nas redes sociais e obteve uma repercussão muito mais alta que os demais monitoramentos, mostrando que a comunicação da marca teve um potencial de impacto e alcance maior em comparação com o volume de menções orgânicas ao nome da empresa em notícias e redes sociais. No mesmo período, pode

ser observado que 85% das 378 interações de atendimento são reclamações, que reforçam as repercussões negativas nas outras instâncias.

Figura 45 – Tela inicial permite a comparação das quatro instâncias de análise propostas.

Fonte: O autor.

As informações apresentadas são dados consolidados de um mês e, no painel, um novo período pode ser selecionado com a aplicação de filtros.

As informações também são classificadas por assuntos e polarização, e listas de apoio dão os detalhes dos documentos que estão sendo analisados. Como mostra a Figura 44, os painéis de detalhamento de cada instância possuem estrutura similar, buscando familiarizar o usuário a um modelo de interpretação único. Há ainda recursos de filtros para analisar as diferentes cidades e regiões analisadas pelo projeto. As visualizações são interativas e podem, também, funcionar como filtros ou dar acesso à lista de documentos que compõe cada área do gráfico disponibilizado.

Figura 46 – Painéis de estrutura similar para atender à análise de tipos diferentes de documento.

Fonte: O autor.

Os recursos de visualização utilizados são simples – gráficos de barra horizontais e verticais, destaque de números, listas e gráfico de áreas acumuladas para representar a evolução no tempo. A escolha teve como critério a familiaridade dos usuários com os formatos, mas é preciso destacar que a escolha da ferramenta restringe os tipos

de recurso disponíveis para composição dos painéis. Dos recursos de interação listados no capítulo 2.7, é possível identificar os três componentes do “Mantra da Visualização” (SHNEIDERMAN, 1996), em que o panorama (*overview*) é identificado no painel inicial; zoom e filtros (*zoom and filters*) são mais evidentes nos painéis de detalhamento de cada instância; e os detalhes sob demanda (*details on demand*) de cada documento são resgatados sob demanda no clique em uma área da visualização ou números na tabela. Da lista de interações proposta por Dix (2013), podemos destacar os recursos de detalhamento e hiperlinks, ao disponibilizarmos a lista de documentos e seus registros, além do link para o conteúdo original; panorama e contexto, com o aprofundamento por instância analisada; parâmetros editáveis, com a aplicação de filtros e funcionalidade de zoom nas visualizações; e, por fim, a fusão temporal, em que a interação em um elemento da visualização altera todo o conjunto das visualizações da página.

O modelo apresentado acima foi adotado ao longo do ano de 2018, e um histórico valioso foi acumulado. Os dados relativos a 2016 e 2017 estão armazenados, porém seguiram outros modelos de estrutura, o que dificulta a análise histórica entre os anos. Uma forma para tornar viável a análise histórica desses dados seria utilizar a base acumulada de 2018 para treinar modelos de Aprendizado de Máquina que pudessem classificar os dados retroativos utilizando parâmetros mais assertivos e consolidando modelos automáticos. O trabalho foi interrompido antes que os modelos pudessem ser construídos.

O mesmo conjunto de dados e tipos de documentos foram utilizados em outros modelos, de forma *ad hoc*, ou sob demanda. A Figura 46 representa um painel em que foram analisadas as menções em três tipos de serviços públicos municipais diferentes e suas respectivas empresas cessionárias, e visava compreender quais eram mais frequentes no discurso da imprensa e das redes sociais, além de uma avaliação da percepção do público sobre cada um dos serviços. A coleta selecionava apenas os documentos relacionados a uma cidade específica, o que tornava possível a análise de uma conjuntura peculiar àquele local.

Foi possível avaliar a evolução do diálogo sobre os três temas-alvo – serviços de saneamento, saúde e transporte – e a participação de cada um no total de documentos

coletados, os principais emissores, e o uso do indicador de repercussão para entender quantitativamente o impacto das mensagens na imprensa e nas redes sociais.

As visualizações são simples e utilizam barras empilhadas como principal recurso para perceber proporções de volume de ocorrências sobre os temas. Nuvens de palavras auxiliam na interpretação do uso recorrente de termos específicos.

Outro recurso de visualização utilizado foi o gráfico de dispersão. A Figura 47 mostra um painel que analisa a distribuição do volume de notícias produzidas em um determinado período, com dois nomes de empresas diferentes. Cada círculo representa uma localização. O tamanho indica o volume de menções à empresa publicadas no local. O eixo horizontal posiciona o local de acordo com o total de notícias analisadas, e o eixo vertical calcula a diferença de volume de menções das empresas A e B. O resultado do cálculo posiciona o círculo no quadrante superior se o volume de menções à empresa A for maior que a empresa B, e no inferior caso sejam mais frequentes as menções à empresa B. Gráficos de linha na lateral indicam o volume de ocorrências coletadas ao longo dos últimos 30 dias, e a tendência de buscas por cada termo segundo a fonte Google Trends.

Figura 47 – Painel de análise entre três serviços em um município.

Fonte: O autor.

Figura 48 – Painel comparativo de menções a empresas diferentes.

Fonte: O autor.

3.6. Resultados

O projeto passou por diversas fases em seus três anos de duração. Os dois primeiros anos foram dedicados a atender demandas *ad hoc* de análise e a disponibilização de recursos de monitoramento automático. O trabalho foi bastante relevante na prevenção e mitigação de problemas com o usuários locais dos serviços prestados pela imprensa, identificação de posicionamento contra e a favor da imprensa no âmbito político institucional. A resposta a essas análises eram ações de comunicação e relações institucionais que tinham sua eficiência aferida também pelo trabalho.

As análises *ad hoc* eram realizadas por profissionais habilitados na coleta, transformação e visualização dos dados, e os painéis produzidos serviam como forma de compartilhamento dos dados estudados e recomendações direcionadas para a equipe central da empresa, que direcionava a ação para seus comunicadores locais.

No último ano do projeto, os esforços foram direcionados ao desenvolvimento de painéis que permitissem aos comunicadores locais acompanhar mais de perto a percepção sobre a empresa e os serviços prestados na cidade e, com isso, ganhassem autonomia para análises e tomadas de decisão.

Os acessos aos painéis e o engajamento das equipes em sua utilização, no entanto, não foram satisfatórios. Um controle interno permitia à equipe de desenvolvimento acompanhar quais usuários estavam efetivamente acessando o serviço e utilizando-o de forma eficiente. Da equipe com cerca de 30 pessoas, apenas duas ou três acessaram com maior regularidade, e as demais apresentaram acessos bastante esparsos e alguns motivos podem ser destacados.

Concepção da plataforma sem a participação dos usuários finais: julgamos este um dos principais motivos da falta engajamento no uso dos painéis, que não tiveram o envolvimento dos funcionários da linha de frente no momento de planejamento da ferramenta. Apenas os gestores e coordenadores corporativos participaram na definição do projeto. Esses coordenadores não lidam com o dia a dia das operações, e essa característica causou um distanciamento entre as necessidades reais dos usuários da linha de frente das unidades com o que estava sendo construído. O processo foi centralizado por uma equipe corporativa ligada a outras áreas da empresa, e limitou o acesso aos comunicadores que atuam nas diversas cidades, que eram efetivamente os usuários finais.

Esse distanciamento da equipe de produção fez com que o valor trazido pelos recursos não fosse reconhecido. Ainda que treinamentos e manuais tenham sido feitos, ficou evidente a falta de aderência da proposta com as necessidades reais do usuário final.

Uma hipótese levantada é que se houvesse uma participação mais ativa do usuário final na construção da ferramenta, seus objetivos e valor seriam mais bem percebidos e, conseqüentemente, haveria maior engajamento no uso da plataforma.

Falta de repertório no uso de dados: ainda do ponto de vista do usuário, notou-se a dificuldade em compreender o que estava sendo medido. Os recursos disponibilizados de filtros e a interação com os gráficos geravam dúvidas sobre o funcionamento geral da plataforma, e o usuário acabava gastando mais tempo tentando entender o que estava sendo mostrado do que efetivamente analisando e tomando decisões. Somadas à variedade de recursos disponibilizados numa mesma interface, foram notadas dificuldades em compreender as visualizações, principalmente aquelas mais sofisticadas, como gráficos de dispersão, e outras que permitem a análise simultânea de muitas variáveis numa mesma visualização.

A participação do usuário no processo de planejamento e desenvolvimento da plataforma poderia ter colaborado na solução deste ponto, avaliando quais eram as métricas e informações mais relevantes para o seu cotidiano e, dessa forma, eliminar os elementos de filtro e visualizações que não eram efetivamente necessários para atender aos objetivos reais dos comunicadores.

Além das questões do usuário, outros pontos contribuíram para a falta de percepção de valor e insucesso da ferramenta.

Disponibilidade de dados: como mencionado anteriormente, a coleta de dados é uma fase extensa e trabalhosa. Dependia do levantamento de fontes de dados digitais dispersas em muitas cidades de médio e pequeno portes no país. Para muitas localidades não havia fontes disponíveis em abundância, e a empresa e seus serviços eram pouco citados em situações estáveis da prestação dos serviços. Dessa forma, muitas das cidades monitoradas não geravam dados em quantidade suficiente para uma análise que pudesse ser generalizada. Essa escassez de dados digitais locais colaborava para a não percepção de valor da ferramenta. Muitas vezes, o comunicador já tinha uma compreensão mais ampla do que a que os dados coletados mostravam. O contexto local, o contato com comunidade, veículos impressos e rádio eram instrumentos mais ágeis na rotina de trabalho.

A ferramenta para entender a qualidade e abrangência da base de fontes de coleta foi o conjunto de painéis ilustrados na Figura 40. A gestão das fontes foi particularmente trabalhosa. Além da adição de novas fontes, era necessária a verificação constante da validade das informações disponíveis e do funcionamento dos algoritmos de

captura dos dados. A mudanças de hiperlinks, o cancelamento ou mudança de nome em perfis sociais e sites que saem do ar, por exemplo, impactam diretamente na qualidade da base de atores e fontes. Essa atualização e gestão despendia muitas horas e, por conta disso, por vezes era negligenciado. Uma alternativa para contornar ou suavizar a rotina seria a aplicação de métodos automáticos de verificação, não viabilizados ao longo do projeto.

Tempo de coleta e tratamento: uma vez coletados os dados (as menções de notícias, postagens e interações de redes sociais e demais documentos), havia a necessidade de estruturá-los de acordo com o mapa de categorias e tags mencionadas anteriormente. Parte do trabalho era feita de forma automatizada, utilizando regras baseadas em fórmulas booleanas, porém, um profissional avaliava a qualidade e complementava a classificação dos documentos coletados. Como o volume no total da coleta era alto, a classificação ocorria ao longo do dia e disponibilizado com até dois dias de atraso. As notícias, por necessitarem de processos de agregação de fontes e remoção de duplicatas, eram disponibilizadas estruturadas apenas ao fim de cada mês. Por se tratar de uma área dinâmica, as decisões relacionadas à comunicação precisam ser tomadas de forma ágil para que ações imediatas possam ser implementadas. O tempo de coleta e tratamento, portanto, impactou no tipo de análise que podia ser feita – era possível ter uma visão histórica e estudar fatos passados, porém não refletia uma visão do presente.

Limitações e falhas da ferramenta OLAP: a ferramenta utilizada para gerar os painéis, embora seja poderosa em recursos de visualização e gestão de bases de dados, apresentou deficiências que impactaram diretamente na qualidade e precisão dos dados analisados. Para sua melhor *performance*, era necessário um planejamento técnico minucioso e que não incluía apenas a gestão da ferramenta OLAP, mas também outros recursos utilizados nos processos de tratamento, transformação e estruturação dos dados. A inclusão de uma nova funcionalidade ou visualização depende de adaptações e melhorias nos métodos e ferramentas utilizados em todos os processos.

Outro fator era a limitação no cruzamento de categorias de uma mesma classe. Havia a necessidade de calcular o volume de documentos marcados com duas ou mais categorias. O cálculo oferecido automaticamente era inconsistente e distorcia a

volumetria dos dados. Esse problema foi um grande impasse para a exploração correta dos dados e comprometia a fiabilidade das informações apresentadas para o usuário final.

Para contornar o problema, os dados deviam passar por uma outra camada de transformação e estruturação da base de dados, adicionando novas medidas e variáveis. O resultado era uma tabela pouco otimizada e pesada, o que também impactava no desempenho geral da ferramenta e tornava o tempo de tratamento ainda mais longo.

Sob a análise dos critérios propostos no trabalho, extroversão e descoberta, algumas hipóteses podem ser levantadas. Pontos colaboram para uma avaliação positiva da capacidade de extroversão e descoberta das iniciativas descritas neste estudo de caso, mas evidencia também outros fatores que tornam a análise mais complexa e impactam negativamente na interpretação.

A gestão das fontes de informação, as premissas de organização da informação e a associação a variáveis numéricas e qualitativas conferem ao sistema possibilidades de agrupamento e aplicação de diversas técnicas de visualização, reforçando sua capacidade de **extroversão**. Por outro lado, a baixa disponibilidade de dados e dificuldade na coleta, as falhas em cálculos oferecidos na plataforma OLAP, e o repertório analítico usuário-alvo são fatores que contribuem como aspectos que impactam negativamente na avaliação deste critério.

Para o critério **descoberta**, os recursos de interação, filtros, fusão temporal, organizados de forma a analisar panoramas, relações entre documentos, análises quantitativas, e o acesso ao documento final indicam para uma avaliação positiva. No entanto, fatores negativos podem ser reforçados como a falta de clareza sobre a real necessidade de informação do usuário e a morosidade na coleta, tratamento e organização da informação.

Como observado ao longo do trabalho, muitas interfaces mencionadas na produção acadêmica se tornam rapidamente obsoletas e são interrompidas, não permitindo o posterior estudo, reprodução ou aplicação de melhorias. O caso apresentado neste capítulo é um exemplo similar. O projeto foi descontinuado e, embora existam cópias

de segurança, não poderá acompanhar as constantes evoluções técnicas que poderiam colaborar no desenvolvimento e futuro da plataforma estudada.

Embora não esteja disponível, espera-se que o conhecimento adquirido nas fases de coleta, tratamento e estruturação dos dados sejam valiosos para a estruturação de novos projetos que pretendam analisar documentos provenientes do ambiente digital.

Por fim, podemos destacar que o caso analisado pode ser inserido no contexto dos paradigmas da informação. São evidentes fatores físicos, relativos aos métodos de programação e estatística utilizados; cognitivos, se observados os pontos relacionados ao usuário e suas dificuldades de interação com a plataforma; sociais, se levados em conta os aspectos institucionais que determinam desde o tipo de fontes que serão utilizadas à forma como os dados serão organizados, e mesmo ao valor que a empresa, e não o indivíduo, pode obter através da plataforma.

4. Considerações finais

O objetivo geral da pesquisa foi estudar o potencial de uso de visualizações em sistemas de recuperação da informação. Para isso, foi apresentada uma visão de contexto tecnológico, social e cultural representado pelo Big Data, e estudado o paralelo entre as áreas da Ciência da Informação e Visualização da Informação em aspectos históricos e epistemológicos, apresentados sob a perspectiva dos paradigmas físico, cognitivo e social da informação (CAPURRO, 2010). Esse cenário foi apresentado nas seções 2.1 e 2.2.

Em seguida, foram apresentados conceitos fundamentais que devem ser considerados ao projetar interfaces de SRI que componham visualizações no processo de busca e recuperação.

A seção 2.3 definiu documento como a materialização da informação (FROHMANN, 2008). O tipo de documento que tratamos é o digital, que, em base de dados, possui necessariamente variáveis, ou metadados, categóricos e textuais. Podem ser mensagens trocadas em redes sociais, páginas da web, artigos científicos, registros de atendimento ao consumidor, documentos transacionais, e-mails, imagens, entre outros.

Esses documentos são, em grande parte, semi ou não estruturados, ou seja, não podem ser organizados em linhas e colunas como em uma tabela normal de base de dados. A organização da informação é apresentada como outro conceito a ser levado em conta na aplicação de visualizações de SRI, uma vez que o que não pode ser organizado, não pode ser visualizado ou recuperado. A tecnologia e os métodos que despontam no contexto Big Data contribuem para um aprimoramento dos métodos automáticos de classificação, cada vez mais precisos e potentes. Ainda no âmbito da organização da informação, é destacado o papel da instituição e da adoção de vocabulários coerentes para o usuário que deverá visualizar e recuperar a informação.

A recuperação da informação, em especial os sistemas de recuperação da informação, foi apresentada na seção 2.4 e partiu do modelo simplificado da recuperação da informação (INGWERSEN, 1992, que encara a interação como um

processo comunicacional, que envolve o usuário, seu contexto e as características técnicas do sistema. As bibliotecas digitais são um tipo de sistema de recuperação da informação e ilustram a complexidade envolvida no projeto e gestão desse tipo de sistemas.

Na seção seguinte, foi apresentada a visualização da informação e seu modelo simplificado (CARD; MACKINLAY; SHNEIDERMAN, 1999), que apontou as transformações de dados e interações do usuário no processo de construção e apreensão das informações por um usuário humano.

Foi destacada também a relevância da Visualização da Informação nas práticas da Ciência da Informação, em especial nas áreas da informetria, cientometria, bibliometria, altmetria e outros métodos quantitativos ligados à interpretação de um grande volume de documentos. A intersecção dessas áreas com funções de recuperação da informação se mostrara promissora na composição de interfaces.

De forma a concretizar os estudos, foi feito um levantamento de exemplos de sistemas de recuperação da informação que utilizam visualizações em suas interfaces. Identificamos uma grande variedade de aplicações, observadas em aplicações no contexto do Jornalismo de Dados, a exploração de bases de produção cinematográfica e as navegações analíticas oferecidas nas plataformas de pesquisa *Web Of Science* (WoS) e *Scopus*.

Uma constatação da pesquisa foi que boa parte dos sistemas documentados em artigos científicos foram experimentos temporários, e as plataformas, na maior parte das vezes, não estão mais disponíveis digitalmente.

A seção 2.7 traz um levantamento prático de recursos de visualização para interfaces de SRI. Identificamos tipos de representações e seus casos de uso de acordo com os modelos de organização utilizados. Foram elencados exemplos de representação de hierarquias (árvore, invólucros, sunburst), visualização de grupos e relações (grafos e gráficos de dispersão); novas possibilidades de uso de representações em 3D e plataformas imersivas (realidade virtual); e o desafio imposto pela variedade de dispositivos em que as visualizações, e mesmo os SRI, deverão funcionar de forma coesa.

Dentre os desafios identificados no trabalho, podemos destacar atributos físicos, cognitivos e sociais. Do ponto de vista do paradigma físico, o desafio está nos métodos utilizados para estruturação e organização dos documentos que devem servir aos dois objetivos – visualização e recuperação – simultaneamente.

A dificuldade está em realizar as transformações de dados necessárias no processo da visualização da informação, mantendo íntegros os metadados, em grande parte categóricos, dos documentos. Essa tarefa é, muitas vezes, comprometida pela capacidade de armazenamento e processamento de informações pelo sistema, mas o contexto Big Data, apresenta evolução na infraestrutura e nas metodologias, fato que permite suportar os requisitos necessários.

Do ponto de vista cognitivo, observou-se que certas visualizações ou interfaces não são acessíveis a qualquer público, o que restringe a utilidade de ferramentas a usuários mais experientes, que possuam competências para a compreensão de determinado gráfico.

Pelo viés social, vale destacar o impacto que a visualização tem na percepção real de valor por usuários e instituições. Uma interpretação errada ou uma estruturação enviesada pode ter efeitos diretos no sucesso do uso de sistemas de recuperação da informação.

Ao final do referencial teórico, na seção 2.8., foi contextualizada a proposta de dois critérios de avaliação do uso de visualizações em SRIs, a extroversão, definida como a capacidade e o preparo do sistema para gerar visualizações de dados; e descoberta, que se refere a capacidade de otimização do sistema em facilitar a recuperação de informações de acordo com a necessidade de um usuário. Ainda que de forma introdutória, a seção pontua questões-chave que ajudam na compreensão dos objetivos e resultados desejados, e analisa os exemplos apresentados ao longo do referencial teórico.

O Estudo de Caso apresentado na terceira seção buscou identificar os atributos e conceitos trabalhados ao longo da pesquisa. O sistema apresentado é uma evidência empírica dos aprendizados que reforça os desafios que devem ser considerados no projeto e gestão de SRI que utilizem visualizações como recursos interativos de interface. Nele, foram identificados cada um dos temas trabalhados, de forma

empírica, e os resultados são apresentados mostrando as potencialidades e limitações da aplicação.

Foram exploradas as visualizações hierárquicas, gráficos de área e de linha e tabelas e gráficos de dispersão, disponibilizados em painéis de controle. A interatividade foi um fator largamente explorado, identificados nos recursos listados na seção 2.7.4 – realce e foco, detalhamento e hiperlinks, panorama e contexto, parâmetros editáveis, troca de representações e fusão temporal.

Pudemos reforçar, por fim, as duas funções da visualização da informação aplicada em SRI, e aferidas pelos critérios de extroversão e descoberta. A extroversão, pois os gráficos e outros recursos visuais, extrovertem as características completas de uma base de documentos; e a descoberta, ao ser incorporada (a visualização) a uma interface utilizada por um usuário para recuperar novas informações que sejam relevantes a ele. O estudo de caso deixa clara a complexidade da avaliação, pois além dos aspectos visíveis na interface e o sucesso em sua utilização, é necessário levar em conta questões institucionais, limitações técnicas e o conhecimento do usuário e sua necessidade de informação.

4.1. Desdobramentos e futuros trabalhos

Espera-se que a pesquisa colabore ativamente na instrumentalização de pesquisadores da Ciência da Informação sobre os principais conceitos, formatos e aplicações da Visualização da Informação.

Nesse sentido, são diversos os caminhos e desdobramentos de pesquisa possíveis, se considerados os três paradigmas da informação – físico, cognitivo e social. Podem ser exploradas novas técnicas e metodologias, estudos sobre usuários, o impacto nas corporações, por exemplo.

No entanto, a aplicação da visualização da informação em interfaces de SRI mostra ter um caráter experimental e empírico. Se considerarmos o contexto Big Data, largamente referenciado no trabalho, nunca tivemos tantos dados disponíveis e acessíveis na nuvem, passíveis de serem explorados, coletados, organizados,

visualizados e recuperados. Dentre esses dados estão os acervos de bibliotecas digitais e conjuntos diversos de documentos não estruturados.

A hipótese introduzida neste trabalho, que define os critérios de extroversão e descoberta para avaliação de SRIs que utilizam recursos de visualização de dados, carece claramente de maior aprofundamento e casos de aplicação.

Certamente é um solo fértil para a execução de experimentos e testes dirigidos para muitas áreas do conhecimento, o que envolve a escolha de modelos de visualização para cada tipo de finalidade. O aprofundamento de conhecimentos sobre as possibilidades de uso de cada tipo de visualização, assim como seus aspectos positivos e negativos no processo de recuperação da informação é outra perspectiva a ser explorada em novas pesquisas.

REFERÊNCIAS

- AGOSTI, M.; CRESTANI, F.; PASI, G. (Eds.). **Lectures on Information Retrieval**. Third European Summer-School, ESSIR 2000 Varenna, Italy, September 11-15, 2000 Revised Lectures. Berlin: Springer, 2001.
- AHLBERG, C.; SHNEIDERMAN, B. Visual Information Seeking: tight coupling of dynamic query filters with starfield displays (1993). In CARD et al. **Readings in information visualization: using vision to think**. San Francisco, CA: Morgan Kaufmann Publishers Inc., 1999.
- ALSALLAKH, B.; MICALLEF, L.; AIGNER, W. et al. The state-of-the-art of set visualization. **Computer Graphics Forum**, v. 35, n. 1, 2016, p. 234-260.
- ALLEN, M. **The hype over hyperbolic browsers**. 2002. Disponível em: <<http://www.onlinemag.net/may02/allen.htm>>. Acesso em: 16 jun. 2019.
- ALMEIDA, M. Revisiting Ontologies: A Necessary Clarification. *Journal of the american Society for information science and technology*, v. 64, n. 8, 2013. p.1682-1693
- ALVARENGA, L. Representação do conhecimento na perspectiva da Ciência da Informação em tempo e espaço digitais. *Encontros Bibli: Revista Eletrônica de Biblioteconomia e Ciência da Informação*, Florianópolis, n. 15, p. 1-23, jan/jun 2003, Disponível em http://www.encontros-bibli.ufcs.br/Edicao_15.html Acesso em 28 de Mai. 2016.
- ANZANELLO, C. A. **OLAP: conceitos e utilização**. Instituto de Informática – Universidade Federal do Rio Grande do Sul, 2007.
- BAEZA-YATES, R.; RIBEIRO-NETO, B. **Modern information retrieval**. New York: Acm Press, 1999.
- BARROS, M. Qual é a diferença entre bibliometria, cientometria, infometria, webmetria e altmetria?. **Bibliotecários sem Fronteiras**, 24 mar. 2015. Disponível em: <<https://bsf.org.br/2015/03/24/diferenca-definicao-conceito-bibliometria-cientometria-infometria-altmetrics/>>. Acesso em: ago. 2019.
- BELKIN, N. Interaction with texts: information retrieval as information-seeking behavior. *Proceedings of the Fifth International Symposium for Information*

Science, 1993. Disponível em: <<http://citeseer.ist.psu.edu/belkin93interaction.html>>. Acesso em: 12 dez. 2018.

BELKIN, N.; ODDY, R.; BROOKS, H. M. ASK for information retrieval: part I. background and theory. **Journal of Documentation**, v. 38, n. 2, 1982, p. 61-71.

BEZERRA, E. et al. Regime de informação: abordagens conceituais e aplicações práticas. *Em Questão*, Porto Alegre, v. 22, n. 2, 2016, p. 60-86

BLUMENSTEIN, K.; NIEDERER, C.; WAGNER, M. et al. Evaluating information visualization on mobile devices: gaps and challenges in the empirical evaluation design space. Proceedings of the **Beyond Time and Errors on Novel Evaluation Methods for Visualization**. Austria: St. Poelten University of Applied Sciences, 2016, p. 125-132. Disponível em: <<https://dl.acm.org/citation.cfm?doid=2993901.2993906>>. Acesso em: 18 mar. 2018.

BORKO, H. Information Science: what is it?. **American Documentation**, v. 19, 1968, p. 3-5.

BRÄSCHER, M.; CAFÉ, L. Organização da informação ou organização do conhecimento? In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 9, 2008, São Paulo, **Anais**. São Paulo: ANCIB, 2008. Disponível em: <<http://www.enancib2008.com.br>>. Acesso em: 18 mar. 2008.

BUSH, V. As we may think. **The Atlantic**, 1945. Disponível em: <<https://www.theatlantic.com/magazine/archive/1945/07/as-we-may-think/303881/>>. Acesso em: ago. 2019.

BUCKLAND, M. K. Information as thing. **Journal of the American Society for Information Science**, v. 42, n. 5, p. 351-360, 1991.

BUCKLAND, M. K. What is a “document”? **Journal of the American Society of Information Science**, ed. 48, n. 9, 1997, p. 804-809.

BURCH, M. et al. IMDb explorer: exploration of a movie database. Proceedings of the **11th International Symposium on Visual Information Communication and Interaction**, 2018, p. 88-91. Disponível em: <<http://dl.acm.org/citation.cfm?doid=3231622.3231633>>. Acesso em: 16 jun. 2019.

CAFÉ, L.; SALES, R. Organização da informação: conceitos básicos e breve fundamentação teórica. In: ROBREDO, Jaime; BRÄSCHER, Marisa. (Org.). *Passeios*

pelo bosque da informação: estudos sobre a representação e organização da informação e do conhecimento – eroic. Brasília DF: IBICT, 2010. p. 115-129.

CAPURRO, R. Epistemologia e Ciência da Informação. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 5, 2003, Belo Horizonte, **Anais**. Tradução de Ana Maria Rezende Cabral, Eduardo Wense Dias, Isis Paim, Ligia Maria Moreira Dumont, Marta Pinheiro Aun e Mônica Erichsen Nassif Borges. Belo Horizonte: ANCIB, 2003. Disponível em: <<http://www.enancib2003.com.br>>. Acesso em: 12 dez. 2018.

CAPURRO, R. Knowledge map of information science: Rafael Capurro's responses to Chaim Zins. In ZINS, C. Knowledge Map of Information Science; **Journal of the American Society for Information Science and Technology (JASIST)** 2007, Vol. 58, N. 4, p. 526-535. Disponível em: <http://www.capurro.de/zins.html>

CAPURRO, R., HJORLAND, B. O conceito de informação. **Perspectivas em Ciência da Informação**, v.12, n.1, p.148-207, 2007. Disponível em <http://portaldeperiodicos.eci.ufmg.br/index.php/pci/article/view/54/47>

CARD, S. K.; MACKINLAY, J. D.; SHNEIDERMAN, B. **Readings in information visualization: using vision to think**. San Francisco, CA: Morgan Kaufmann Publishers Inc., 1999.

CHEN, C. **Information visualization: beyond the horizon**. London: Springer, 2004.

CHITTARO, L. Visualizaing information on mobile devices. Computer v. 39, n. 3, 2006, p. 40-45

COLUMBIÉ, R. **Ciencia de la Información: su historia y epistemologia**. Bogotá: Rojas Eberhard Editores, 2005.

CORRÊA, C. A. **Indexação automática e visualização de informações: um estudo baseado em Lógica paraconsistente**. 2011. 152 fl. Tese (Doutorado em Ciência da Informação) Universidade de São Paulo, São Paulo, 2011.

CORREA, R.; VIEIRA, J. Representações visuais para recuperação de informação na BDTD-UFPE. **Perspectivas em Ciência da Informação**, v. 18, n. 4, 2014, p. 18-34.

DE MAURO, A.; GRECO, M.; GRIMALDI, M. A formal definition of Big Data based on its essential features. **Library Review**, vol. 65, iss. 3, 2016, p. 122-135.

DE LIRA, J.; ASSIS, F. A contribuição da organização e da visualização da informação para os sistemas de recuperação de informação. *Inf. Inf.*, Londrina, v. 20, n. 1, 2015 p. 110 - 136

DIAS, M. **A contribuição da visualização da informação para a Ciência da Informação**. 2007. 126 fl. Dissertação (Mestrado em Ciência da Informação) Pontifícia Universidade Católica de Campinas, Campinas, 2007.

DIAS, M.; CARVALHO, J. O. A Visualização da Informação e a sua contribuição para a Ciência da Informação. **DataGramaZero – Revista de Ciência da Informação**, v. 8, n. 5, 2010. Disponível em: <http://www.dgz.org.br/out07/Art_02.htm>. Acesso em: 6 jun. 2019.

DIX, A. Introduction to information visualization. In: AGOSTI, M.; FERRO, N.; PAMELA, F. et al. (Eds.). **Information visualization meets information retrieval**. Switzerland: Promise Winter School, 2012, p. 1-27.

ELMQVIST, N.; MOERE, A.; JETTER, H. et al. Fluid interaction for information visualization. **Information Visualization**, v. 10, n. 4, 2011, p. 327-340.

ENTENDA as irregularidades envolvendo uso do WhatsApp na eleição. **Folha de S.Paulo**, São Paulo, 22 out. 2018. Disponível em: <<https://www1.folha.uol.com.br/poder/2018/10/entenda-as-irregularidades-envolvendo-uso-do-whatsapp-na-eleicao.shtml>>. Acesso em: 29 nov. 2018.

FACELI, K.; LORENA, A. C.; GAMA J.; CARVALHO, A. C. P. L. F. **Inteligência Artificial: uma abordagem de Aprendizado de Máquina**. Rio de Janeiro: LTC Editora, 2011.

FERNEDA, E. **Recuperação de informação: análise sobre a contribuição da Ciência da Computação para Ciência da Informação**. 2003. 147p. Tese (Doutorado em Ciência da Informação) Universidade de São Paulo, São Paulo, 2003.

FLORES, P. O que a Cambridge Analytica, que ajudou a eleger Trump, quer fazer no Brasil. **Nexo Jornal**, São Paulo, 8 dez. 2017. Disponível em: <<https://www.nexojornal.com.br/expresso/2017/12/08/O-que-a-Cambridge-Analytica-que-ajudou-a-eleger-Trump-quer-fazer-no-Brasil>>. Acesso em: 29 nov. 2018.

FRIENDLY, M.; DENIS, D. The early origins and development of the scatterplot. **Journal of the History of the Behavioral Sciences**, v. 41, n. 2, 2005, p. 103-130.

FROHMANN, B. Revisiting “what is a document?”. **Journal of Documentation**, v. 65, n. 2, 2009, p. 291-303.

FROHMANN, B. O caráter social, material e público da informação. In: FUJITA, M.S.L.; MARTELETO, R.M.; LARA, M.L.G. (org). A dimensão epistemológica da Ciência da Informação e suas interfaces técnicas, políticas e institucionais nos processos de produção, acesso e disseminação da informação. São Paulo: Cultura Acadêmica/Marília: FUNDEPE, 2008. p.13-34.

GARCIA GUTIÉRREZ, A. Cientificamente favelados: uma visão crítica do conhecimento a partir da epistemografiaTransinformação, Campinas, v.18, n.2, p.103-112, 2006. Disponível em: <http://revistas.puc-campinas.edu.br/transinfo/viewarticle.php?id=151>. Acesso em: 16 agosto 2016.

GIL, A. C. **Métodos e técnicas de pesquisa social**. 4. ed. São Paulo: Atlas, 1994.

GIL, A. C. **Como elaborar projetos de pesquisa**. 4. ed. São Paulo: Atlas, 2007.

GILLILAND, A. J. Setting the stage. In: BACA, M. (ed.). **Introduction to metadata**. 2nd ed. Los Angeles: The Getty Research Institute, 2008, p. 1-19.

GOOGLE TRENDS, **Volume de buscas pelo termo “Big Data”**, de 2004 a 2018. Disponível em: <<https://trends.google.com/trends/explore?date=all&q=big%20data>> Acesso em: 8 jun. 2019.

GUSMÃO, F.; DELAZERI, B.; MATOS, S. et al. Hierarchical visualization techniques: a case study in the domain of meta-analysis. **International Journal of Engineering Research & Science (IJOER)**, v. 2, n. 5, 2016, p. 168-180.

HABIB, M. Sneak Preview Scopus Analyze Results: Overview and use case. **Michael Habib**, 8 dez. 2011. Disponível em: <<http://www.mchabib.com/2011/12/08/sneak-preview-scopus-analyze-results-overview-and-use-case/>>. Acesso em: 22 maio 2019.

HEER, J.; BOSTOCK, M.; OGIEVETSKY, V. **A Tour through the Visualization Zoo**. Communications of the ACM, v.53 n.6, 2010. Disponível em: <<https://dl.acm.org/citation.cfm?id=1743567>>. Acesso em: agosto 2019.

INGWERSEN, P. **Information retrieval interaction**. London: Taylor Graham, 1992.

JACOB, E. K.; SHAW, D. Sociocognitive perspectives on representation. **Annual Review of Information Science and Techonolgy (ARIST)**, v. 33, p. 131-185, 1998.

JOHNSON, B.; SHNEIDERMAN, B. Tree-maps: a space-filling approach to the visualization of hierarchical information structures. **IEEE**, 1991, p. 284-291.

JUNG, C. **Tipos psicológicos**. 3. ed. Petrópolis, RJ: Vozes, 2009.

KLEINER PERKINS. **Internet Trends**, 2018. Disponível em: <<https://www.kleinerperkins.com/perspectives/internet-trends-report-2018/>>. Acesso: 20 janeiro 2019.

KOBASHI, N. Y.; FERNANDES, J. C. Pragmática Lingüística e organização da informação. In: Encontro nacional de pesquisa e pós-graduação em Ciência da Informação, v. 9., 2008, São Paulo. Anais. São Paulo: Ancib, 2008.

LAMPING, J.; RAO, R. The hyperbolic browser: a focus + contexto technique for visualizing large hierarchies. **Journal of Visual Languages and Computing**, v. 7, n. 1, 1995, p. 33-55.

LATOUR, B. Redes que a razão desconhece: laboratórios, bibliotecas, coleções. In PARENTE, A. (ed.). **Tramas da Rede Sulina**. Porto Alegre, 2004, p. 39-63.

LE COADIC, Y. F. **A Ciência da Informação**. Brasília: Briquet de Lemos, 1996.

LIM, Y. H. What movies have we produced?. **Tableau Public**, 17 mar. 2014. Disponível em: <<https://public.tableau.com/en-us/s/gallery/imdb-movies-visualized>>. Acesso em: 27 jan. 2019.

LIU, S.; CUI, W.; WU, Y. et al. Survey on information visualization: recent advances and challenges. **Vis Comput**, v. 30, 2014, p. 1.373-1.393.

LIU, Z.; JIANG, B.; HEER, J. imMens: Real-time Visual Querying of Big Data. **Eurographics Conference on Visualization (EuroVis)**, v. 32, n. 3, 2013, p. 421-430.

LUND, N.W. Document theory. *Annual Review of Information Science and Technology*, v.43, n.1, p.1-55, 2009. Disponível em: <http://onlinelibrary.wiley.com/doi/10.1002/aris.2009.1440430116/full>. Acesso em: 13 jul. 2016.

MACEDO, R. N.; KOBASHI, N. Y. Estudos bibliométricos avançados como dispositivos indicadores de tendências nas atividades de ciência, tecnologia e inovação. **Estudos Universitários Revista de Cultura**. Pernambuco: Universidade Federal de Pernambuco, 2012, p. 57-71.

MACEDO, R. N.; KOBASHI, N. Y. Bibliometria, cientometria infometria: conceitos e aplicações. **Pesquisa Brasileira em Ciência da Informação e Biblioteconomia**, Brasília, v. 2, n. 1, 2009, p. 155-172.

MACKINLAY, J. Automating the design of graphical Presentations of relational information; **ACM Transactions on Graphics**, V. 5, 1986, p. 110-141.

MALINI, F. Um método perspectivista de análise de redes sociais: cartografando topologias e temporalidades em rede. **XXV Encontro Anual da Compós**, Universidade Federal de Goiás, Goiânia, 2016.

MARCONDES, C. H.; SAYÃO, L. F. Documentos digitais e novas formas de cooperação entre sistemas de informação em C&T. **Ciência da Informação**, Brasília, v. 31, n. 3, 2002, p. 42-54.

MARCONDES, C. H.; SAYÃO, L. F. À guisa de introdução: repositórios institucionais e livre acesso. In SAYÃO, L. F. et al. (org.). **Implantação e gestão de repositórios institucionais**: políticas, memória, livre acesso e preservação. Salvador (BA): EDUFBA, 2009, p. 9-21.

MAYER-SCHÖNBERGER, V.; CUKIER, K. **Big Data**, a revolution that will transform how we live, work, and think. New York: Houghton Mifflin Harcourt Publishing Company, 2013.

MAZZOCCHI, F. Images of thought and their relation to classification: the tree and the net, **Knowledge Organization**, 40, n. 6, 2013, p. 366-374. Disponível em: <<https://doi.org/10.5771/0943-7444-2013-6-366>>. Acesso em: 16 jun. 2019.

MARKS, L.; HUSSEL, J.; MCMAHON, T. et al. ActiveGraph: a digital library visualization tool. **International Journal on Digital Libraries**, v. 5, n. 5, 2005, p. 57-69.

MANYIKA, J. et al. **Big Data**: the next frontier for innovation, competition, and productivity. Nova York: McKinsey Global Institute, 2011. Disponível em: <<https://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation>>. Acesso em: 13 jul. 2019.

MOOERS, C. (1951). Zatocoding applied to mechanical organization of knowledge. **American Documentation**, v. 2, n. 1, p.20-32.

MORUJÃO, Alexandre F. Prefácio da edição portuguesa. In: KANT, Immanuel. **Crítica da razão pura**. Lisboa: Fundação Calouste Gulbenkian, 2013.

NASCIMENTO, H.; FERREIRA, C. **Visualização de informações: uma abordagem prática**; XXV JAI; p. 1262-1312

NEWBY, G. B. Empirical study of a 3D visualization for information retrieval tasks. **Journal of Intelligent Information Systems**, n. 18, 2002, p. 31-53.

OLSON, H. Classification and universality: application and construction. **Semiotica**, v. 139, 2002, p. 377-391.

PAJIC, D. Browse to search, visualize to explore: Who needs an alternative information retrieving model?. **Computers in Human Behavior**, n. 39, 2014, p. 145-153.

PARISER, E. **O filtro invisível**: o que a internet está escondendo de você. Rio de Janeiro: Zahar, 2012.

PARK, D.; KIM, S. H.; ELMQVIST, N. Gatherplots: generalized scatterplots for nominal data. **arXiv:1708.08033**, v. 1, 2017. Disponível em: <<http://arxiv.org/abs/1708.08033>>. Acesso em: 16 jun. 2019.

POMBO, O. Da classificação dos seres à classificação dos saberes. **Leituras**. Revista da Biblioteca Nacional de Lisboa, n. 2, 1998, p. 19-33. Disponível em: <<http://cfcul.fc.ul.pt/textos/OP%20-%20Da%20Classificacao%20dos%20Seres%20a%20Classificacao%20dos%20Saberes.pdf>>. Acesso em: 16 jun. 2019.

PONJUAN, D.; DANTE, G. Influencias ergnómicas em la visualización de la información. *Inf. Inf.*, Londrina, v. 15, n. 1, 2010, p. 17 - 36

PRIEM, J. Altmetrics. In CRONIN, B.; SUGIMOTO, C. (Eds.), **Beyond bibliometrics**: harnessing multidimensional indicators of scholarly impact. London: MIT Press, 2014.

PRITCHARD, A. Statistical bibliography or bibliometrics?. **Journal of Documentation**, v. 25, n. 4, 1969, p. 348-349.

RAMOS, D. **A expansão do jornalismo para o ambiente numérico**. Curitiba: Appris, 2016.

RAYWARD, W.B. The origins of information science and the International Institute of Bibliography/International Federation for Information and Documentation (FID).

Journal of the American Society for Information Science, v. 48, n. 4, 1997, p. 289-300. Disponível em:

<<https://www.marilia.unesp.br/Home/Instituicao/Docentes/EdbertoFerneda/MRI%201%20-%20Rayward,%20WB%20-%201997.pdf>>. Acesso em: 16 jun. 2019.

RENDÓN-ROJAS, M. Epistemologia da Ciência da Informação: objeto de estudo e principais categorias. **InCID**, Revista de Ciência da Informação e Documentação, v. 3, 2012, p. 3-14.

SALTON, G. e MCGILL, M. J. (1983) Introduction to Modern Information Retrieval. McGraw Hill.

SANDERSON, M.; CROFT, W. B. The History of Information Retrieval Research. **IEEE**, v. 100, Special Centennial Issue, 2012, p. 1.444-1.451. Disponível em: <<https://ieeexplore.ieee.org/document/6182576>>. Acesso em: 22 jun. 2019.

SANTOS, R. N.; KOBASHI, N. Y. Bibliometria, cientometria, infometria: conceitos e aplicações. **Pesquisa Brasileira em Ciência da Informação e Biblioteconomia**, v. 2, n. 1., 2009, p. 155-172.

SANTOS, R. N.; KOBASHI, N. Y. Estudos bibliométricos avançados como dispositivos indicadores de tendências nas atividades de ciência, tecnologia e inovação. **Estudos Universitários**, Revista de Cultura da Universidade Federal de Pernambuco, v. 31, n. 12, 2012, p. 57-71.

SARACEVIC, T. Relevance: a review of the literature and framework for thinking on the notion in Information Science. Part II. Advances in Librarianship. **Journal of the American Society for Information Science and Technology**, v. 58, n. 13, 2007, p. 1.915-1.933.

SAYÃO, L. F. Bibliotecas Digitais e suas utopias. **PontodeAcesso**, Salvador, v. 2, n. 2, 2008, p. 2-36.

STAKE, R. E. The case study method in social inquiry. **Educational Researcher**, v. 7, n. 2, p. 5-8, 1978.

SVENONIUS, E. **The intellectual foundation of information organization**. Cambridge: MIT Press, 2000.

SHNEIDERMAN, B. et al. Visualizing Digital Library Search results with categorical and hierarchical axes. Proceedings of the **Fifth ACM conference on Digital Libraries**,

Maryland, Human-Computer Interaction Laboratory, 2000. Disponível em: <<https://dl.acm.org/citation.cfm?id=336637>>. Acesso em: 16 jun. 2019.

SILVA, N. N.; SÁNCHEZ, C.; PROAL, C. et al. Visual exploration of large collections in digital libraries. Proceedings of the **Latin American Conference on Human-Computer Interaction**, 2003, p. 147-157. Disponível em: <<https://dl.acm.org/citation.cfm?id=944535>>. Acesso em: 22 jun. 2019.

SMIT, J. A informação na Ciência da Informação. InCID: R. Ci. Inf. e Doc., Ribeirão Preto, v. 3, n. 2, p. 84-101, jul./dez. 2012.

SOUZA, R.; ALMEIDA, M.; BARACHO, R. Ciência da informação em transformação: Big Data, nuvens, redes sociais e web semântica. **Ciência da Informação**, IBICT, v. 42; n. 2, 2013, p. 159-173.

TAYLOR, A. G. **The organization of information**. Englewood: Libraries Unlimited Inc., 1999.

TECHNISCHE UNIVERSITÄT GRAZ. Hyperbolic Browser: 205 nodes. **FluidDiagrams**, 2014a. Disponível em: <<https://projects.isds.tugraz.at/fluididiagrams/fd/examples/Hyperbolic/Hyperbolic-small.html>>. Acesso em: 24 maio 2019.

TECHNISCHE UNIVERSITÄT GRAZ. Hyperbolic Browser: 3.418 nodes. **FluidDiagrams**, 2014b. Disponível em: <<https://projects.isds.tugraz.at/fluididiagrams/fd/examples/Hyperbolic/Hyperbolic.html>>. Acesso em: 24 maio 2019.

TRUCOLO, C. C.; DIGIAMPIETRI, L. A. Análise de tendências da produção científica nacional na área de Ciência da Informação: estudo exploratório de mineração de textos. **AtoZ**, 2014. Disponível em: <<http://www.atoz.ufpr.br/index.php/atoz/article/view/79>>. Acesso em: 15 dez. 2018.

TURTLE, H. R.; CROFT, W. B. Uncertainty in information retrieval system. In MOTRO, A.; SMETS, P. (Eds.). **Uncertainty Management in Information Systems**: from needs to solutions. Boston/London/Dordrecht: Kluwer Academic Publisher, 1997.

TUFTE, E. **The visual display of quantitative information**, 2nd ed. Cheshire, Connecticut: Graphics Press, 1983.

VAKKARI, P. Task Complexity, problem structure and information actions. Integrating studies of information seeking and retrieval. **Information Processing and Management**, v. 35, 1999, p. 819-837.

VAN RIJSBERGEN, C. J. **Information retrieval**. London: Butterworths, 1979.

VIEIRA, J.; CORREA, R. Visualização da informação na construção de interfaces amigáveis para sistemas de recuperação de informação. **Encontros Bibli**, Revista eletrônica de biblioteconomia e ciência da informação, v. 16, n. 32, 2011, p. 73-93.

Disponível em:

<http://www.brapci.inf.br/_repositorio/2011/11/pdf_6d28abda1b_0018918.pdf>.

Acesso em: 12 dez. 2018.

VIEIRA, J.; PINHO, F. A contribuição da organização e da visualização da informação para os sistemas de recuperação de informação. **Revista Informação**, Londrina, v. 20, n. 1, 2015, p. 110-136.

YIN, R. K. **Estudo de caso: planejamento e métodos**. 2. ed. 1. reimp. Porto Alegre: Bookman, 2003

WAGNER FILHO, J. A. **Evaluating immersive approaches to multidimensional information visualization**. 2018. Dissertação (Mestrado em Ciência da Computação), Universidade Federal do Rio Grande do Sul. Programa de Pós-graduação em Computação, Porto Alegre, 2018. Disponível em: <<https://www.lume.ufrgs.br/handle/10183/175082>>. Acesso em: 16 jun. 2019.

WE ARE SOCIAL; HOOTSUITE. **Digital 2019**: Essential insights into how people around the world use the internet, mobile devices, social media, and e-commerce. New York, 2019. Disponível em: <<https://wearesocial.com/global-digital-report-2019>>. Acesso em: ago. 2019.

WU, X. et al. Data Mining with Big Data. **IEEE**, v. 26, n. 1, 2013, p. 97-107. Disponível em: <<https://dl.acm.org/citation.cfm?id=2553838>>. Acesso em: 13 jul. 2019.

ANEXO 1: Análise de interdisciplinaridade a partir da recorrência de palavras-chave no histórico de projetos selecionados na FEBRACE (Feira Brasileira de Ciências e Engenharia)

Análise de interdisciplinaridade a partir da recorrência de palavras-chave no histórico de projetos selecionados na FEBRACE (Feira Brasileira de Ciências e Engenharia)

Alexandre Antonino Gonçalves Martinazzo
Nº USP 3726627
Doutorando do Programa de Pós Graduação em
Engenharia Elétrica (PPGEE - EPUSP)
Orientadora: Roseli de Deus Lopes
alexandremartinazzo@gmail.com

Eduardo Caetano Barbosa
Nº USP 2860824
Mestrando do Programa de Pós Graduação em
Ciência da Informação (PPGCI - ECA)
Orientadora: Nair Kobashi
educaeba@gmail.com

RESUMO - Este artigo é o resultado de conclusão da disciplina PCS 5031 - Introdução à Ciência de Dados, ministrada pelos professores doutores Pedro Corrêa e Fernando Corrêa, no segundo semestre de 2016, e tem como finalidade aplicar conceitos e aprendizados ao longo do curso. Para isso, propomos uma avaliação de recorrência de palavras-chave no histórico de projetos selecionados na FEBRACE (Feira Brasileira de Ciências e Engenharia). A metodologia utilizada inclui técnicas de *Text Data Mining*, descrição e pré-processamento de bases de dados, programação em Python e o software GEPHI para gerar grafos.

Palavras-Chave - interdisciplinaridade; mineração de texto; heatmaps; grafos;

I. INTRODUÇÃO

“Por que o cientista fez sua aparição no começo do século XIX? Por que, pouco depois, surgiu um grupo sempre crescente de disciplinas especializadas? Ou, dito de outro modo, por que os estudiosos passaram a desconfiar dos polímatas? Uma resposta óbvia é que a explosão de conhecimento obrigou a maioria deles a restringir suas ambições intelectuais e, assim, a desconfiar dos poucos que se recusavam a limitá-las” [1, 2].

Na sequência desta passagem, o historiador Peter Burke levanta outras abordagens que levaram à especialização da ciência, respostas sociológicas que indicavam a profissionalização do cientista e a fundação de sociedades que promoviam a ciência. Não é, afinal, um processo isolado.

A ideia de uma explosão de conhecimento não é nova, ela é percebida com nitidez pela primeira vez, em suas devidas proporções e disponibilidade de acesso, na invenção da imprensa de Gutemberg no século XV. O invento, de um momento para o outro, tornou acessível um volume de conhecimento ilimitado para a apreensão de um estudioso da época. Este mesmo fenômeno é observado em outros conhecidos momentos de ruptura,

sempre - ou quase sempre - relacionados à uma evolução tecnológica. O surgimento das bibliotecas, do telégrafo, o rádio, a televisão, o computador, a internet e o Big Data, entre tantos outros exemplos, são tecnologias que transformaram, cada um a seu tempo, o acesso à informação e a capacidade de geração de conhecimento.

Em cada um desses ciclos, a comunidade científica se reorganiza para lidar com esta enxurrada de informações que não para de crescer exponencialmente.

Métodos, processos e tecnologias (em especial as tecnologias de informação e comunicação) evoluíram ao longo da história, e novos olhares sobre a organização de disciplinas se tornaram indispensáveis. A compreensão sobre especialização, como consolidada a partir do século XIX e não raramente presente nas instituições de hoje, não dá mais conta de enfrentar os desafios e oportunidades que surgem nestes momentos de ruptura.

Uma resposta óbvia é a colaboração entre especialidades e pessoas, organizadas em novas disciplinas, não mais isoladas, e com o poder de criar relações de interdisciplinaridade.

A Ciência dos Dados, assim com a Ciência da Informação, são exemplos deste contexto. São disciplinas que nascem da interação entre especialidades e não existem por si só, elas se aplicam a partir da relação com outras áreas e na solução de problemas de outras disciplinas. Conforme mostra a figura 1, a Ciência da Informação é intrinsecamente interdisciplinar, agregando áreas como Educação, Filosofia, Comunicação, Biblioteconomia, e outras.

Fig. 1: Conhecimentos/disciplinas envolvidas na formação das Ciências da Informação¹

Fig. 2: Conhecimentos/disciplinas envolvidas na Ciência dos Dados²

Analogamente, a Ciência dos Dados está na interseção entre Ciência da Computação e Matemática e Estatística com o domínio de aplicação do conhecimento. Existem diversas disciplinas aplicadas no domínio de conhecimento, como Ciências Sociais, Medicina e Economia. O panorama da interação entre diversas especialidades na perspectiva de Ciência dos Dados é ilustrado na figura 2.

Este artigo pretende dar vistas ao tema da interdisciplinaridade e à importância em estimular um ambiente propício para que aconteça.

¹ Imagem disponibilizada por CARVALHO, J. em Slides de aula sobre Conhecimento e Ciência da Informação, 2009. disponível em <http://professorjonathascarvalho.blogspot.com.br/2009/09/slides-de-aula-sobre-conhecimento-e.html> (acesso em 8/12/2016)

² Imagem disponibilizada por DUARTE, A.N, UFPB, em Slides de aula Introdução à análise de dados. disponível em <http://pt.slideshare.net/alexandrend/introducao-analise-de-dados-aula-01> (acesso em 8/12/2016)

O exercício proposto é entender se, e como, a interdisciplinaridade é trabalhada no ambiente dos Ensinos Fundamental e Médio, a partir da base de dados de trabalhos inscritos e selecionados para a Feira Brasileira de Ciências e Engenharia. Este repositório é oportuno pois tem uma representatividade nacional, e conta com a participação de instituições públicas e privadas, regulares e técnicas. Esta é nossa amostragem.

Pretende-se observar, aplicando conhecimento, técnicas e metodologias da Ciência dos Dados, a relação entre as categorias e subcategorias dos projetos inscritos na feira segundo classificação da FAPESP, e a presença de projetos potencialmente interdisciplinares que podem ser avaliados por diferentes abordagens.

II. CONSTRUÇÃO DO CORPUS

O corpus usado neste trabalho foi construído a partir da base de trabalhos inscritos e selecionados como finalistas FEBRACE³, na última edição. São aproximadamente 340 documentos, caracterizados principalmente por campos de texto descritivos que especificam:

- Título
- Estudantes (nomes separados por vírgula)
- Orientadores(as) (nomes separados por vírgula)
- Escola
- Rede da Escola
- Resumo
- Categoria
- Subcategoria
- UF
- Palavras-chave (separadas por vírgula)

Sabe-se que textos descritivos “puros” são dados não-estruturados, ou seja, não podem ser aplicados diretamente a modelos de análise estatística ou aprendizado de máquina. Para isso, é preciso criar uma representação numérica dos atributos do texto. Neste trabalho, adotou-se a representação conhecida como “*bag-of-words*” de unigramas, que usa as palavras únicas presentes no corpus como atributos [3]. A pontuação dos textos foi ignorada neste trabalho, bem como a diferenciação de letras maiúsculas e minúsculas. Ao final, a representação “*bag-of-words*” permite obter uma matriz com a frequência de cada uma das palavras (nas colunas) para cada um dos documentos disponíveis (linhas).

Os dados estão disponíveis no portal institucional da FEBRACE e, como já mencionado, são compostos por dados públicos (em sua maior parte composto por textos) e sigilosos (dados sobre avaliadores, suas notas e média

³ <http://febrace.org.br> (acessado em 4/12/2016)

final de cada trabalho). A aquisição dos dados consistiu em extrair uma versão da base apenas com textos publicados nos anais do evento, que possam ser disponibilizados em repositórios científicos de livre acesso.

Importante mencionar que a autoria dos títulos, resumos, assim como especificação de categoria, subcategoria e palavras-chave são de responsabilidade dos autores dos trabalhos inscritos anualmente na FEBRACE. As categorias e subcategorias propostas pela FEBRACE são um subconjunto da classificação de áreas do conhecimento conforme indicado na FAPESP⁴. Já as palavras-chave não obedecem nenhuma padronização ou modelo de referência, sendo sempre 3 para cada projeto inscrito e inserida pelos(as) autores(as).

III. PRÉ-PROCESSAMENTO DOS DADOS

Os documentos do corpus foram unificados na forma de um arquivo CSV (*comma-separated values*) com os campos disponíveis nos anais do evento: título, nomes de estudantes, nomes de orientadores, nome da escola, rede da escola, resumo, categoria, subcategoria, estado de origem (unidade federativa - UF), palavras-chave. Cada campo foi gravado como uma coluna no arquivo, sendo os projetos armazenados em linhas. Esse procedimento foi feito apenas para simplificar as etapas seguintes.

A. Seleção de campos

Nem todos os campos disponíveis nos documentos são necessários para a análise proposta neste trabalho. Assim, agregaram-se os campos título, resumo e palavras-chave como um único texto para cada um dos projetos, e foram descartados de dados de autoria (nomes de estudantes e orientadores) e de vínculo escolar. Só então foi construída a matriz de frequência de palavras (representação “*bag-of-words*”).

B. Redução de dimensionalidade

Palavras que ocorrem em apenas um projeto (independente de sua frequência num único documento) não estabelecem nenhuma relação entre trabalhos, e podem ser descartadas. Esta redução de dimensionalidade é desejável para que o volume de dados total permita mais agilidade nos processos computacionais seguintes. Também fez parte da redução de dimensionalidade o processo de remoção de sufixos (*stemming*), que permite

agregar palavras com radicais⁵ comuns [3]. Esse processo faz com que a frequência para as palavras “científico”, “científica” e “científicos” sejam agregadas por meio do radical “científic”, por exemplo.

Depois da remoção de sufixos, são integradas novamente à matriz de frequência de palavras, informações do estado de origem, categoria e subcategoria do conhecimento, e rede escolar, para que possam ser feitas as análises de interdisciplinaridade.

Para a manipulação dos textos, adotou-se a biblioteca Pandas [8]; para a contagem de palavras foi usada a biblioteca Scikit-learn [5], enquanto a biblioteca NLTK (Natural Language Toolkit) [7] permitiu implementar a remoção de palavras comuns da língua portuguesa [6] e a detecção de radicais comuns entre palavras. Todas as bibliotecas mencionadas estão disponíveis na linguagem Python. A redução de dimensionalidade permitiu sair de 10.886 termos e chegar-se a 2904.

IV. VISUALIZAÇÃO E ANÁLISES

A visualização dos dados partiu de três abordagens distintas e complementares. Cada uma delas permitiu construir e compreender diferentes perspectivas de interdisciplinaridade.

A. Análise sob o ponto de vista das categorias

A primeira análise faz uma seleção dos termos de maior ocorrência, verificando sua distribuição em relação às categorias do conhecimento (a saber Ciências Exatas e da Terra, EXA; Ciências Biológicas, BIO; Ciências da Saúde, SAU; Ciências Agrárias, AGR; Ciências Humanas, HUM; Ciências Sociais Aplicadas, SOC; Engenharias, ENG). Essa distribuição foi representada por um *heatmap*, mostrado na figura 3.

⁴ Referência de áreas do conhecimento assim determinas pela FAPESP no link http://www.bv.fapesp.br/pt/area_conhecimento/ (acesso em 4/12/2016)

⁵ Radical, neste contexto, não se refere ao sentido linguístico clássico de palavras com origem comum, e sim a palavras com sufixos removidos que podem ser agrupadas.

Fig. 3: 50 palavras de maior ocorrência distribuída por categorias

A gradação de cores indica a concentração de cada termo em relação à cada categoria.

Em uma leitura sobre cada coluna, é fácil identificarmos áreas de concentração específica de termos, como por exemplo a área de humanas, que concentra 3 dos termos mais utilizados.

Categorias mais densas, onde predominam tons mais escuros indicam áreas mais propensas ao uso de determinados termos de uso particular à ela, mostrando uma tendência de concentração que pode não ser indício de interdisciplinaridade.

Já sob o ponto de vista de análise das linhas, ou seja, a partir dos termos e não das categorias, podemos observar aquelas que tendem a ter incidência maior entre as áreas e aquelas concentradas em temas específicos.

Algumas relações parecem óbvias e já esperadas, como os termos derivados do radical “trabalh”, que aparece com uma frequência mais ou menos equilibrada entre as categorias.

Outros termos baseados em radicais como “ambient” e “plant” evidenciam concentração nas categorias Agrária e Biológicas. “Aliment”, por sua vez, além das áreas mencionadas mostra uma relação forte com a área de saúde.

A mesma técnica, isolando os 50 termos mais recorrentes de cada área e comparando sua incidência com as outras categorias, evidenciam melhor aquelas que tendem a usar o mesmo corpus de palavras ou não, como mostram as visualizações a seguir (figuras 4 a 10).

Fig. 4: 50 palavras de maior ocorrência na categoria AGRÁRIA

Fig. 5: 50 palavras de maior ocorrência na categoria BIOLÓGICAS

Fig. 6: 50 palavras de maior ocorrência na categoria ENGENHARIA

Fig. 7. 50 palavras de maior ocorrência na categoria EXATAS E DA TERRA

Fig. 10. 50 palavras de maior ocorrência na categoria SOCIAIS APLICADAS

Fig. 8. 50 palavras de maior ocorrência na categoria HUMANAS

Fig. 9. 50 palavras de maior ocorrência na categoria SAÚDE

A comparação entre as imagens sugere quais as categorias com maior potencial de relação com outras. Quando notamos uma concentração clara apenas na categoria em questão, a tendência é que ela seja considerada pouco (ou com menor potencial) interdisciplinar, ou com relações mais específicas. É o caso das Ciências Biológicas e das Engenharias, claramente concentradas em si mesmas.

As figuras também mostram que as relações são direcionais, ou seja, uma forte presença de palavras de uma categoria em outras. A categoria Exatas, por exemplo, mostra que seus termos mais comuns são observados com frequência em outras áreas. O contrário, porém, nem sempre é verdadeiro, como podemos ver na concentração dos termos em Humanas, que não se repetem na categoria Exatas.

Saúde tem uma relação relativamente forte com Sociais Aplicadas, mas a relação contrária, por outro lado, mostra uma relação mais fraca.

B. Análise sob o ponto de vista das subcategorias

Em uma outra abordagem procurou-se investigar a relação entre as subcategorias do conhecimento. Para isso foi explorada uma visualização em grafo não direcional.

A visualização foi gerada utilizando o software GEPHI e o plugin Circular Layouts. Os nós representam as subcategorias, e as arestas ligam os nós que possuem acima de 300 ocorrências de termos em comum. Quanto maior o número de ocorrências comuns, maior o peso da aresta. As cores indicam as categorias-pai.

Fig. 11. Grafo circular representando ligação entre subcategorias

Conseguimos inferir em conjunto com as visualizações em heatmap que a categoria EXATAS é a que mais tem conexões com outras, especialmente com ENGENHARIAS e HUMANAS, mas com muitas outras arestas menos pesadas distribuídas com nós de outras categorias.

Se destacam no grafo as subcategorias Ciência da Computação (EXA), Química (EXA), Eletrônica (ENG) e Educação (HUM), sucedidas pelas subcategorias Saúde Coletiva (SAU), Biologia Geral (BIO) e Ecologia (BIO).

O isolamento das categorias AGRÁRIAS e SOCIAIS APLICADAS ficam ainda mais evidentes, com pouca concentração apenas nas subcategorias Serviço Social (SOC).

C. Análise sob o ponto de vista dos projetos

A terceira abordagem utiliza como nós os trabalhos em si. O vetor de palavras por projeto foi utilizado para a criação de uma visualização da disparidade entre cada um deles. Para isso, adotou-se a técnica de escalonamento multidimensional (*multidimensional scaling* - MDS), que permite representar de forma geométrica a distância entre cada projeto [9, 10]. Tradicionalmente, adota-se a distância euclidiana como métrica de disparidade, mas esta não apresentou os resultados mais interessantes. Assim, a figura 11 mostra a dessemelhança entre os documentos usando como critério de distância a distância de cossenos.

A cor de cada nó indica sua categoria, e as ligações entre eles usam a semelhança de cossenos; sempre que a semelhança for maior que 0,4 entre dois documentos há uma aresta.

Fig. 11. Grafo de distância de projetos

Esta visão é importante para entendermos os resultados anteriores. Podemos identificar, por exemplo, que o isolamento da área de AGRÁRIAS e SOCIAIS APLICADAS (cores laranja e marrom na figura 11, respectivamente) se deve à pouca semelhança destes projetos com outros, já que há poucas arestas ligando nós destas categorias.

Notamos que muito projetos de engenharia estão ligados entre si. O grafo circular anterior não mostra arestas fortes entre as subcategorias de ENGENHARIA, o que nos leva a inferir que estas ligações se referem a projetos dentro das mesmas subcategorias.

Por fim, as conexões entre nós das categorias EXATAS e HUMANAS pode mais uma vez ser observada nessa visualização, ao identificarmos uma ocorrência maior de conexões entre nós das cores vermelho e azul. Essa constatação confirma a interdisciplinaridade observada no grafo das subcategorias e nos *heatmaps* das categorias.

V. PALAVRAS FINAIS

Ainda que os métodos e análises precisem de maiores apurações para chegar a conclusões mais contundentes, observa-se um potencial de estudo de descoberta de

conceitos ligados à mais de uma área do conhecimento que possam, sim, indicar novas fronteiras entre áreas, alertando também para, ao menos no que compete à Ciências da Informação, novas tendências de colaboração entre as disciplinas atuais.

A partir da classificação das categorias do conhecimento orientadas pela FAPESP, procurou-se investigar quais as palavras que representam conceitos intrínsecos em mais de uma disciplina. Um possível próximo passo deste experimento seria localizar as palavras que mais fazem sentido para as áreas conectadas e iniciar um processo, menos exato, de definição de conceitos e como são aplicados em cada campo do conhecimento.

Esta abordagem representa um esboço de metodologia que promova a interoperabilidade de repositórios de conteúdo de diferentes disciplinas e que possa se tornar uma ferramenta de descoberta de temas e projetos interdisciplinares que possam ajudar outras áreas de domínio.

A aplicação deste processo em um ambiente menos controlado, de engajamento de estudantes do fundamental e secundaristas, é também interessante por trabalhar a interdisciplinaridade num ambiente de educação ainda pautado pela visão clássica da especialização por disciplinas.

Se a organização das áreas de domínio das ciências estão em transformação, onde mais e mais áreas tem se beneficiado pela interação com outros campos do conhecimento, é urgente a preparação de estudantes para uma visão mais horizontal, mais inclusiva em termos de especialidades para que assim floresçam outras disciplinas e futuros pesquisadores que reúnam especialidades únicas baseadas na interdisciplinaridade.

O caso de aplicação deste projeto também deve ser entendido como um ensaio para uso em instâncias maiores. Uma abordagem comumente observada hoje nos uso de dados para a recuperação de informações em repositórios digitais focam em áreas de domínio específico, repositórios temáticos. Há uma lacuna para experimentos que pretendam estudar a integração de repositórios de áreas distintas, mas que tenham um potencial de conversa, onde procuramos semelhanças e diferenças entre áreas do conhecimento, e como pode haver colaboração entre elas.

Próximos passos deste estudo devem incluir aumentar o número projetos analisados, um maior corpus para mineração que forneça mais dados e inferências e, também relevante, experimentar a aplicação para repositórios de cunho científicos, com uma estrutura de metadados mais consistente para a análise.

VI.REFERÊNCIAS

- [1] BURKE, P. Uma História Social do Conhecimento: de Gutenberg a Diderot. Rio de Janeiro: Zahar Ed, 2003.
- [2] BURKE, P. Uma História Social do Conhecimento II: da Enciclopédia à Wikipedia. Rio de Janeiro: Zahar Ed, 2012.
- [3] S. Raschka, *Python Machine Learning*. Birmingham, UK: Packt Publishing, 2015.
- [4] S. Raschka, *Python Machine Learning*. Packt Publishing - ebooks Account, 2015.
- [5] F. Pedregosa *et al.*, “Scikit-learn: Machine Learning in Python”, *Journal of Machine Learning Research*, vol. 12, p. 2825–2830, out. 2011.
- [6] C. Huyck e V. Orenge, “A Stemming Algorithmm for the Portuguese Language”, in *String Processing and Information Retrieval, International Symposium on*, Los Alamitos, CA, USA, 2001, p. 0186.
- [7] S. Bird, E. Loper, e E. Klein, *Natural Language Processing with Python*. O’Reilly Media Inc., 2009.
- [8] W. McKinney, *Python for Data Analysis*. 2012.
- [9] J. B. Kruskal, “Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis”, *Psychometrika*, vol. 29, nº 1, p. 1–27, mar. 1964.
- [10] J. B. Kruskal, “Nonmetric multidimensional scaling: A numerical method”, *Psychometrika*, vol. 29, nº 2, p. 115–129, jun. 1964.