UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO DE CIÊNCIAS DA EDUCAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO

Daniel San Martin Pascal Filho

UM MODELO PARA VIGILÂNCIA TECNOLÓGICA AUTOMATIZADA DE PORTAIS WEB E REDES SOCIAIS

Florianópolis (SC)
2020

Daniel San Martin Pascal Filho

UM MODELO PARA VIGILÂNCIA TECNOLÓGICA AUTOMATIZADA DE PORTAIS WEB E REDES SOCIAIS

Dissertação submetida ao Programa de Pós-Graduação em Ciência da Informação da Universidade Federal de Santa Catarina para a obtenção do título de mestre em Ciência da Informação.
Orientador: Prof. Douglas Dyllon Jeronimo De Macedo, Dr.

Florianópolis (SC)
2020

Ficha de identificação da obra elaborada pelo autor,
através do Programa de Geração Automática da Biblioteca Universitária da UFSC.

San Martin Pascal Filho, Daniel
Um modelo para vigilância tecnológica automatizada de portais web e redes sociais / Daniel San Martin Pascal Filho ; orientador, Douglas Dyllon Jeronimo De Macedo, 2020.
116 p.

Dissertação (mestrado) - Universidade Federal de Santa Catarina, Centro de Ciências da Educação, Programa de Pós Graduação em Ciência da Informação, Florianópolis, 2020.

Inclui referências.

1. Ciência da Informação. 2. Vigilância tecnológica. 3. Monitoramento tecnológico. 4. Mineração de textos. I. Dyllon Jeronimo De Macedo, Douglas. II. Universidade Federal de Santa Catarina. Programa de Pós-Graduação em Ciência da Informação. III. Título.

Daniel San Martin Pascal Filho

UM MODELO PARA VIGILÂNCIA TECNOLÓGICA AUTOMATIZADA DE PORTAIS WEB E REDES SOCIAIS

O presente trabalho em nível de mestrado foi avaliado e aprovado por banca examinadora composta pelos seguintes membros:

Prof. Moisés Lima Dutra, Dr.
Universidade Federal de Santa Catarina

Juliano Anderson Pacheco, Dr.
Federação das Indústrias de Santa Catarina

Prof. José Eduardo Santerém Segundo, Dr.
Universidade de São Paulo

Certificamos que esta é a **versão original e final** do trabalho de conclusão que foi julgado adequado para obtenção do título de mestre em Ciência da Informação.

Documento assinado digitalmente
Adilson Luiz Pinto
Data: 29/12/2020 18:14:48-0300
CPF: 184.218.848-83

Prof. Adilson Luiz Pinto, Dr.
Coordenação do Programa de Pós-Graduação

Documento assinado digitalmente
Douglas Dyllon Jeronimo de Macedo
Data: 29/12/2020 15:37:23-0300
CPF: 633.987.372-34

Prof. Douglas Dyllon Jeronimo De Macedo, Dr.
Orientador

Florianópolis (SC), 2020.

Este trabalho é dedicado a todos aqueles que entenderam
minha ausência e não mediram esforços para que eu pudesse
chegar a esta etapa da minha vida.

AGRADECIMENTOS

Registro meus sinceros agradecimentos a toda equipe do Observatório FIESC pelo incentivo, colaboração e apoio oferecidos durante o tempo em que esta pesquisa foi desenvolvida, em especial ao Dr. Eng. Juliano Pacheco, quem viabilizou sua execução, contribuindo com ideias e soluções para os problemas encontrados, ao Sr. Vanderson Sampaio pelo suporte e atenção dedicada e ao Sr. Dérick Pereira Costa por sua parceria e orientações.

Agradeço ao Prof. Dr. Prof. Moisés Lima Dutra por contribuir diretamente com minha formação acadêmica por meio de suas aulas, observações e gentileza com que sempre me recebeu.

Agradeço a Universidade Federal de Santa Catarina por me acolher enquanto seu aluno desde a graduação até o mestrado e por contribuir em minha formação acadêmica e social, ao Programa de Pós Graduação da Ciência da Informação por me proporcionar um novo horizonte de conhecimentos e aos membros da Banca por sua disponibilidade, avaliação e contribuições para o trabalho.

Agradeço ao meu pais, Daniel e Verônica, e aos meus irmãos, Liber e Renato, por serem minhas referências como ser humano, por sempre acreditarem em mim e me apoiarem nas minhas decisões. Agradeço à minha amada companheira Franciele e à minha filha Sofia por estarem ao meu lado em todos os momentos e por fazem a minha vida tão bonita.

Por fim, deixo meu agradecimento mais do que especial ao meu orientador Prof. Dr. Douglas Dyllon Jeronimo de Macedo por compartilhar sua experiência e conhecimentos comigo, por me apoiar durante toda esta evolução acadêmica e por um ser humano diferenciado cujo profissionalismo e generosidade tive o privilégio de conhecer.

“A dúvida é o princípio da sabedoria.”
(Aristóteles)

RESUMO

A tecnologia é uma forte aliada da indústria. É capaz de promover a redução de custos operacionais, controle de qualidade e ganho de escala. Serve, ainda, como barreira de entrada e diferencial competitivo para concorrentes. Sua importância motivou organizações públicas e privadas a investirem esforços para construir e manter panoramas tecnológicos atualizados, dando origem a sistemas de Vigilância Tecnológica. Contudo, o aumento dos acervos digitais e dos canais por onde circulam as informações trouxeram complexidade e aumento de custos para as atividades de monitoramento tecnológico, principalmente àquelas cujos processos são executados manualmente, cenário que se agrava na medida em que os cenários de Big Data se tornam cada vez mais comuns. Assim, novos métodos que auxiliem as organizações em suas atividades de vigilância são úteis e necessários. Este trabalho apresenta um modelo de vigilância tecnológica automatizada de tecnologias-chaves a partir de portais web e redes sociais em cenários de Big Data. Para dar sustentação ao modelo, construiu-se um referencial teórico sobre os fundamentos da vigilância e assuntos correlatos além de uma revisão sistemática da literatura com um período de recorte entre os anos de 2013 a 2020. O modelo foi generalizado a partir de métodos existentes e pela análise comparativa de softwares especializados. Ele foi dividido em quatro módulos principais (coleta, preparação, análise e difusão) e dois auxiliares (parametrização e persistência), os quais viabilizam a geração de produtos da vigilância tecnológica sem a necessidade de intervenção humana. Por fim, foi desenvolvido um estudo de caso onde se implementou um sistema para validar o modelo. Os resultados experimentais demonstraram a viabilidade da abordagem proposta.

Palavras-chave: vigilância tecnológica, monitoramento tecnológico, mineração de textos, ciências da informação.

ABSTRACT

Technology is a strong ally of the industry. It is capable of promoting the reduction of operating costs, quality control, and gain of scale. It also serves as an entry barrier and competitive advantage against competitors. The private and public sectors are motivated by its importance in maintaining an updated overview of technological scenarios, giving rise to Technological Watch systems. However, the increase in digital databases of information and the channels through which they circulate brought complexity and increased costs to technological monitoring activities, especially to organizations that execute their processes manually or are in Big Data scenarios. Thus, new methods to do technology watch activities are useful and necessary. This work presents a conceptual model of automated Technology Watch of key technologies from web portals and social networks in Big Data scenarios. To support it, it was built a theoretical framework considering the fundamentals of surveillance and related issues. In addition, a systematic review of the literature with a clipping period between the years 2013 to 2020 was done. The model was generalized based on existing Technology Watch methods and specialized softwares analysis. It was divided into four main modules (collection, preparation, analysis, and dissemination) and two auxiliaries (parameterization and persistence), which enable the generation of technological surveillance products without the need for human intervention. Finally, a case study was developed where a system was implemented to validate the model. The experimental results demonstrated the feasibility of the proposed approach.

Keywords: technology watch, technology monitoring, technological surveillance, text mining, information science.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE ABREVIATURAS E SIGLAS

CIESC	Centro das Indústrias do Estado de Santa Catarina
CNM	Clauset-Newman-Moore
IC	Inteligência Competitiva
IEL	Instituto Euvaldo Lodi
LDA	Latent Dirichlet Allocation
PDIC	Programa de Desenvolvimento Industrial Catarinense
RSL	Revisão Sistemática da Literatura
SENAI	Serviço Nacional de Aprendizagem Industrial
SESI	Serviço Social da Indústria
TF-IDF	Termo–inverso da Frequência nos Documentos
VT	Vigilância Tecnológica

SUMÁRIO

1 INTRODUÇÃO

Há um século, as organizações mantinham suas informações em registros escritos e bem estruturados organizados em seus arquivos. As inovações tecnológicas eram conhecidas por meio da leitura de jornais ou revistas. Porém, com o advento da computação e da internet esse cenário mudou radicalmente. Os ciclos de inovação tornaram-se cada vez mais curtos e difíceis de serem percebidos através de canais de comunicação mais tradicionais; e para as empresas, a informação tornou-se um patrimônio de alto valor.

O progresso tecnológico alcançado transformou o fluxo da informação a qual passou a circular por uma infinidade de canais físicos e digitais, tornando seu acompanhamento e monitoramento atividades complexas e dependentes da informática. Com este tipo de problemática vigente, a Ciência da Informação ganhou espaço. Sendo uma ciência interdisciplinar, ela oferece suporte teórico e prático para lidar com os problemas informacionais, por exemplo, o monitoramento e a vigilância tecnológica. Este trabalho aponta o estado da arte da vigilância tecnológica e apresenta um modelo para lidar com ela nos atuais cenários de Big Data.

1.1 CONTEXTUALIZAÇÃO

O sucesso ou fracasso de uma organização são determinados, em grande parte, por sua capacidade de competir no mercado. As atividades envolvidas nessa competição incluem sua capacidade de se diferenciar, de inovar, de formar parcerias estratégicas, criar barreiras de entrada para os competidores ou reduzir custos. Como salienta Renata Peregrino de Brito e Luiz Artur Ledur Brito (2012), quando uma empresa apresenta um desempenho superior, é-lhe atribuída a existência de vantagens competitivas, ou seja, uma capacidade de criação de valor para seus clientes acima da média de seus concorrentes.

Em seu livro "*COMPETITIVE ADVANTAGE: Creating and Sustaining Superior Perfomance*", Porter (1985) define estratégia competitiva como a busca de uma posição competitiva favorável em um setor, ambiente onde a competição ocorre, com o objetivo de estabelecer uma posição lucrativa e sustentável com relação às forças que regem a concorrência no setor. Para ele, a vantagem competitiva aumenta quando a capacidade da empresa de criar valor para seus clientes excede o custo para criar esse valor. Neste contexto, o valor é o que os clientes já estariam dispostos a pagar, e o valor superior é definido como a capacidade de se oferecer o mesmo produto por um preço inferior ao do mercado com os mesmos benefícios ou oferecer benefícios superiores (premium) que compensariam inclusive um preço mais alto.

Estando em um ambiente competitivo por natureza, as indústrias e empresas precisam analisar seu posicionamento em relação à concorrência. Assim como no ambiente militar, ilustrado por Sun Tsu em seu livro "A Arte da Guerra" (TZU; PIN, 2015) onde ele apresenta a espionagem dos adversários como uma atividade fundamental para se ganhar uma batalha, as organizações corporativas passaram a desenvolver métodos específicos para esse propósito. O termo Inteligência Competitiva (IC) foi cunhado para expressar este tipo de atividade. Calof e

Sheila Wright (2008) destacam que a IC envolve a coleta de informações, internas, externas e dos concorrentes, como também informações de fornecedores, clientes, tecnologias e relações comerciais. Com isso, a IC pode ajudar na antecipação de mudanças de concorrentes, parceiros, clientes e governos. Como resume Canongia *et al.* (2004), a Inteligência Competitiva é geralmente utilizada como um instrumento por empresas para identificar, coletar, sistematizar e interpretar informações relevantes sobre seu setor de forma ética.

Um conceito fortemente associada a Inteligência é a Vigilância. Palop e Vicente (1999) definem vigilância como um esforço sistemático e organizado da empresa para observar, capturar, analisar, disseminar e recuperar informações de forma precisa sobre os fatos do ambiente econômico, tecnológico, social ou comercial, relevantes em seu contexto, em busca de se identificar oportunidades ou ameaças. Pere Escorsa (2001) sugere que o processo de vigilância de uma organização seja composto por quatro eixos:

- Vigilância Competitiva: responsável por colher informações sobre os concorrentes atuais e em potencial;
- Vigilância Comercial: avalia informações de mercado sobre clientes e fornecedores;
- Vigilância Tecnológica: busca informações sobre tecnologias disponíveis ou em desenvolvimento;
- Vigilância do Entorno: busca informações sobre fatos externos que possam afetar o futuro da organização.

Apesar dos conceitos de inteligência competitiva e vigilância tecnológica estarem associados e muitas vezes tratados como iguais, eles possuem diferenças sutis. Hidalgo *et al.* (2002) escrevem:

> "A inteligência difere da vigilância, pois constitui uma etapa adicional no processo de gerenciamento das informações obtidas. A vigilância busca obter as informações mais relevantes do ambiente para os interesses das organizações e sua análise, enquanto a inteligência coloca ênfase especial em outros aspectos, como a apresentação das informações em um formato adequado para a tomada de decisão e a análise da avaliação dos resultados obtidos com a sua utilização" (apud RAMÍREZ *et al.*, 2012, p. 245, tradução nossa).

Ainda que Pere Escorsa (2001) discorra timidamente sobre a Vigilância Tecnológica (VT), Palop e Vicente (1999), J Marcela Sánchez e Palop (2002) e Salgado Batista *et al.* (2003) trabalham o conceito com maior profundidade elevando sua importância no contexto empresarial. Salgado Batista *et al.* (2003) definem a VT como um sistema estruturado para coordenar as atividades de recuperação de informação, processamento, análise e disseminação da informação interna e do meio ambiente de acordo com um plano e uma estratégia organizacional. Como afirmam Ramírez *et al.* (2012), a Vigilância Tecnológica e a Inteligência Competitiva (IC) são duas ferramentas ou processos que se complementam e se tornam muito úteis quando o objetivo é se manter informado para antecipar um evento e melhorar a competitividade de uma organização.

Assim, sendo a tecnologia uma importante vantagem competitiva na atualidade, esta ferramenta vem ganhando cada dia mais espaço, sendo, inclusive, normatizada para tornar-se uma boa prática organizacional (ABNT, 2011), (AENOR, 2019).

A tecnologia da informação e a gestão estratégica da informação criaram a necessidade das organizações serem cada vez mais "Data Driven", ou seja, ter suas decisões balizadas por dados. Assim, elas precisam manter registros digitais sobre a situação tecnológica interna e externa a fim de realizar cruzamentos e detectar oportunidades no mercado, dado que uma interpretação precisa sobre o panorama tecnológico é um dos pré-requisitos na obtenção de vantagens competitivas em relação aos concorrentes (GOORHA; UNGAR, 2010).

Não obstante, com uma velocidade nunca antes vista, mercados inteiros estão desaparecendo e outros estão surgido. Tecnologias como o Blockchain [1] prometem eliminar a necessidade de cartórios e os Drones, veículos aéreos não tripulados e controlados remotamente por tecnologias sem fio, apontam para um futuro próximo onde as entregas já não serão feitas por humanos. Por sua vez, as tecnologias de digitalização de áudio, como o MP3, praticamente extinguiram as mídias físicas, como CDs e DVDs, e mudaram completamente o mercado fonográfico.

Um exemplo comum de aplicação da VT está documentado por Andrade Navia *et al.* (2018) em seu artigo entitulado de "Vigilancia tecnológica aplicada a la cadena productiva de cacao" em que por meio de um estudo sistemático os autores trazem um panorama sobre a cadeia produtiva de cacau, identificando os principais países com participação na produção científica, as mais destacadas áreas de investigação, possíveis relações de intercâmbio entre autores e países, as relações de intercâmbio entre instituições e autores, a produção tecnológica ao longo do tempo, instituições requerentes de patentes entre outros.

Padilla *et al.* (2018a), por sua vez, utilizaram um processos e vigilância tecnológica associado ao mapeamento do ciclo de vida tecnológico através da Curva "S" para identificar inovações e novos processos em carne como subproduto do curtimento, atuando sobre bases de patentes e artigos.

A VT também pode ser um ator em busca de economia em investimentos para determinados setores. Na Colômbia, o baixo domínio tecnológico constitui uma desvantagem para os fornecedores de peças automotivas os quais não tem capacidade financeira suficiente para orquestrar diferentes projetos de inovações. Para eles, a implementação de processos de inovação baseados em Vigilância Tecnológica e Inteligência Competitiva pode ajudar empresas, universidades e governo a encontrarem respostas em escala global, conforme descrito no trabalho de López C. e Zartha Sossa (2014). Eles apresentam uma aplicação da vigilância tecnológica em conjunto com técnicas de prospecção baseada no método Delphi para identificar aspectos como as principais características das tendências de pesquisa no desenvolvimento e implantação de

[1] Blockchain é uma tecnologia utilizada para registrar transações entre seus usuários de forma segura. Como resume bem (CROSBY *et al.*, 2016): ela é basicamente um banco de dados distribuído de registros de todas as transações ou eventos digitais que foram executados e compartilhados entre as partes participantes. A veracidade dos eventos é verificada por consenso entre a maioria dos participantes da blockchain. Outra característica importante, é que uma vez inserida, as informações sobre a transação nunca poderão ser removidas. A blockchain tem como principal exemplo de seu uso o Bitcoin, uma cripto-moeda digital descentralizada.

aços avançados utilizados na fabricação de peças para veículos.

Conforme os exemplos apresentados, percebe-se a necessidade que existe para a sobrevivência das organizações que as mesmas mantenham um monitoramento constante do ambiente tecnológico, sendo este muitas vezes um fator de vida ou morte delas. Porém, torna-se custoso e complexo manter este tipo de atividade uma vez que os ciclos de desenvolvimento tecnológicos e inovação são cada vez mais curtos e a informação é diversificada e abundante. De acordo com a IDC Research, em 2020 terão sido gerados 44 zettabytes de dados (IDCEMC2, 2014) e esse número crescerá rapidamente nos cinco anos subsequentes, atingindo 163 zettabytes em 2025 (DATAAGE2025, 2017). Parte desses dados é composta por publicações científicas, patentes, notícias e dados de redes sociais, os quais são matérias-primas fundamentais para que as organizações possam realizar suas análises e estar à frente do mercado, mantendo-se competitivas.

Encontrar e monitorar informações estratégicas nesse volume de dados em constante mudança requer o uso e desenvolvimento de novas tecnologias. São cenários de dados complexos conhecidos como cenários de Big Data (WHITE, 2015), onde os paradigmas utilizados em sistemas computacionais tradicionais não conseguem atender adequadamente. Entre outros motivos, isso acontece porque os cenários de Big Data demandam grande capacidade de armazenamento e processamento, comumente obtidos por meio de computação paralela e armazenamento distribuído. A análise também requer técnicas diferenciadas como mineração de textos (*text mining*), o uso de inteligência artificial e agentes de softwares capazes de monitorar e até interpretar informações e mudanças em fontes de interesse.

1.2 JUSTIFICATIVA E MOTIVAÇÃO

Com o propósito de identificar e monitorar tendências tecnológicas, têm sido realizados um grande número de estudos por uma série de entidades como a European Commission (2009), International Telecommunications Union (2014), Manchester Institute of Innovation Research (2014), empresas como a Shell (2007), IBM (2013), Microsoft-Fujitsu (2011) e consultorias como Gartner (2013), Lux Research (2014), Deloitte (2012) (ENA *et al.*, 2016). As metodologias utilizadas costumam envolver coleta e análise de dados de fontes estruturadas e não estruturadas em busca de padrões ocultos pelos quais são identificadas tendências.

Segundo Ena *et al.* (2016), a maior parte dos trabalhos sobre monitoramento de tendências tecnológicas recaem sobre patentes e publicações científicas. Contudo, essas informações não são suficientes para entender o ciclo completo do desenvolvimento tecnológico ou verificar seu possível impacto no mercado. Um sistema completo, deve englobar informações em seus diferentes estágios, como artigos de pesquisa básica nos quais ainda não há um produto concebido, de pesquisa aplicada onde já se vislumbra uma solução, patentes que refletem o início da proteção de um produto em vias de ir ao mercado e publicações em revistas, notícias e portais de eventos pelas quais já se pode avaliar um produto e seu impacto no mercado.

Esta variedade de canais e tipos de informações multiplicam o volume de dados que preci-

sam ser capturados e analisados pelos especialistas. Nestes cenários, os métodos tradicionais ou manuais de vigilância tecnológica não são suficientes para as organizações por deixarem espaço para subjetividade na análise e interpretação das informações, capacidade limitada de coleta de dados em diferentes fontes, tempo proibitivo de processamento e dificuldade de armazenamento e recuperação. Na empresa Koniker S.Coop, segundo estudo de caso desenvolvido por A Perez *et al.* (2018), apenas 35% do tempo de sua equipe agregava valor às atividades de Vigilância Tecnológica. De forma geral, constatou-se que se investia um tempo considerável na leitura de grandes quantidades de documentos para categorizá-los e enviá-los manualmente para outros especialistas. Estas são atividades passíveis de automatização. Os autores utilizaram técnicas não convencionais para aprimorar o processo de vigilância tecnológica na Koniker filtrando e classificando as informações, como patentes, notícias, boletins oficiais do governo do sistema de VT automaticamente através de técnicas de aprendizagem de máquina, além de enriquecer os textos agregando anotações semânticas com a DBpedia Spotlight(DBPEDIA SPOTLIGHT, 2020).

No estado de Santa Catarina, Brasil, a Federação das Indústrias de Santa Catarina (FIESC) é uma peça chave para a saúde industrial. Ela possui mais de 140 sindicatos filiados, 16 regionais e apoio das câmaras especializadas, tem o objetivo de promover um ambiente favorável aos negócios e buscar a redução dos custos logísticos, melhorando a competitividade das indústrias locais (FIESC, 2019). Dentre suas ações estão, por exemplo, monitorar as questões relevantes para os setores produtivos e sugerir alternativas (FIESC, 2014).

Recentemente, visando ampliar a competitividade da indústria catarinense, a FIESC construiu o PDIC 2022 (Programa de Desenvolvimento Industrial 2022), que envolveu e conectou empresas, governo, terceiro setor e instituições de ensino em múltiplas iniciativas, para que as oportunidades geradas sejam absorvidas pelas indústrias e permitam fortalecer o posicionamento do setor industrial catarinense em âmbito Nacional e Internacional.

O PDIC 2022 busca, entre outras ações, identificar os setores indutores de desenvolvimento e elaborar visões de futuro para cada um. Nele, foram realizados painéis com especialistas, os quais apontaram dezesseis "Setores Portadores de Futuro para a Indústria Catarinense" e identificaram as tecnologias-chave para a indústria (FIESC, 2013), cujo processo contou com a realização de vários painéis envolvendo dezenas de especialistas durante meses de trabalho. Vencida a etapa de identificação das tecnologias, surgiu a necessidade de manter os dados atualizados, atividade que requer um esforço similar ou ainda maior. Desta maneira, uma estratégia de vigilância tecnológica que proporcione um panorama atualizado de forma rápida e automatizada a um custo adequado para organizações como a FIESC sobre o cenário tecnológico interno e externo é um desafio e ao mesmo tempo uma oportunidade de pesquisa.

1.3 PROBLEMA DE PESQUISA

Os estudos sobre as tecnologias-chaves para setores específicos da indústria costumam ser conduzidos por especialistas da área, cujo processo pode envolver entrevistas, revisões

bibliográficas e reuniões. Entre outras atividades, é necessário produzir e manter registros sobre cada etapa executada para que seja possível produzir relatórios gerenciais a fim de nortear decisões estratégicas para as organizações envolvidas. Este processo caracteriza um sistema de vigilância tecnológica tradicional, o qual demanda tempo e custos significativos.

Um fator decisivo nos ambientes permeados pela competitividade tecnológica, no entanto, são a velocidade e precisão com que uma organização percebe as evoluções tecnológicas no seu entorno que podem representar ameaças ou oportunidades. Esta percepção está em grande parte associada a um trabalho constante de avaliação do desempenho de cada tecnologia em relação as outras considerando diferentes fontes de dados, cujos formatos, como visto, podem variar.

O cenário descrito reflete o *problema* enfrentado pelas grandes indústrias e organizações que utilizam informações de inteligência competitiva de cunho tecnológico e que precisam manter todo ciclo de vigilância funcional sob custos e tempo aceitáveis.

1.3.1 Pergunta de pesquisa

Em um contexto onde é necessário manter um panorama permanentemente atualizado sobre o cenário tecnológico de interesse para as organizações, este trabalho busca responder a seguinte questão de pesquisa:

- Como identificar e monitorar tecnologias de interesse de forma automatizada e constante a partir de múltiplas fontes cujos dados e comportamentos que caracterizem cenários de Big Data?

1.4 OBJETIVOS

Nesta seção são detalhados os principais objetivos deste trabalho.

1.4.1 Objetivo Geral

Propor um modelo para a realização de vigilância tecnológica automatizada em fontes disponíveis eletronicamente como artigos de portais web ou rede sociais.

1.4.2 Objetivos Específicos

Com o propósito de alcançar o objetivo geral foram definidos os seguintes objetivos específicos:

- Levantar o estado da arte sobre vigilância tecnológica em busca de insumos para este trabalho por meio da revisão da literatura e análise de softwares especializados;
- Elaborar um modelo conceitual que permita a realização de vigilância tecnológica de maneira automatizada em publicações disponíveis na internet considerando cenários de Big Data;

- Desenvolver um estudo de caso prático para aplicar o modelo.
- Avaliar a abordagem proposta junto à especialistas.

1.5 DELIMITAÇÃO DE PESQUISA

Esta pesquisa está delimitada na linha de pesquisa de informação, gestão e tecnologia, com foco no eixo informação e tecnologia na área das Ciências da Informação. Considera a vigilância tecnológica aplicada exclusivamente em publicações disponíveis eletronicamente na internet e redes sociais com conteúdos em inglês e português no formato HTML, excluindo-se documentos em PDF, vídeos ou áudio. Não obstante, o modelo proposto pode ser estendido para outros tipos de publicações como patentes e artigos científicos, efetuando-se os ajustes necessários.

Serão utilizadas como referência apenas as tecnologias-chaves definidas pelos especialistas que integraram o projeto Programa de Desenvolvimento Industrial Catarinense (PDIC) (FIESC, 2014) da FIESC e as destacaram como fundamentais para os Setores Portadores do Futuro para o Estado de Santa Catarina. Por fim, esta pesquisa procura trazer uma contribuição à Vigilância Tecnológica pela proposição de um modelo que seja totalmente automatizável, assim ela considera apenas trabalhos que permitam um fluxo de informação integrado desde a captura até a divulgação dos resultados encontrados.

O estudo de caso está limitado aos recursos disponíveis no Observatório da Indústria (FIESC, O., 2019), onde ele será implantado e avaliado. Como produtos do estudo de caso, objetiva-se a construção de painéis de vigilância tecnológica para comunicação automatizada dos monitoramentos, um sistema de processamento para automatizar as análises e a organização da informação.

1.6 ALINHAMENTO DO TEMA À CIÊNCIA DA INFORMAÇÃO

A ciência da informação (CI) nasceu para lidar com os problemas relacionados ao aumento da complexidade dos acervos de informação, da necessidade de adaptação das metodologias e do surgimento de tecnologias para esses fins. Nas últimas décadas, a rápida evolução das tecnologias computacionais de criação, representação, armazenamento, organização, disseminação e consumo, fizeram com que as características que motivaram o surgimento desta ciência fossem sentidas de maneira exponencial (ROCHA SOUZA *et al.*, 2015). Ainda, segundo Borko (1968), ela se apresenta como uma disciplina que investiga as propriedades e o comportamento da informação, as forças que governam seu fluxo e os meios para processá-la. Consequentemente, suas características interdisciplinar, dinâmica e ubíqua emprestam à maioria das pesquisas atuais alguns de seus conceitos e ferramentas.

A pesquisa em vigilância tecnológica (VT) tem a agregação de informações, suas transformações e interpretações como principais elementos de estudo (SÁNCHEZ, J Marcela; PALOP, 2002), dado que ela pode ser entendida como um processo estruturado para visualizar

e comparar os cenários tecnológicos internos e externos às organizações através da análise de objetos informacionais, principalmente aqueles disponíveis em formato digital. Quando a relacionamos com a definição de Borko (1968) sobre os aspectos tratados pela Ciência da Informação, fica evidente que este ramo de pesquisa pode naturalmente se tornar parte do escopo de investigação de um profissional da Ciência da Informação.

1.7 ESTRUTURA DA DISSERTAÇÃO

Este trabalho está estruturado em seis seções, sendo esta a Introdução. Na seção 2, é levantado o referencial teórico sobre vigilância tecnológica, incluindo seus benefícios, normatizações, principais metodologias e plataformas integrais web. Disserta-se, também, sobre conceitos de Big Data, análise de domínio, ontologias, agentes de softwares e mineração de textos. Na seção 3, são apresentados os aspectos metodológicos que direcionam esta pesquisa. Na seção 4, são apresentados o modelo conceitual para vigilância tecnológica automatizada de tecnologias-chaves a partir de publicações web e redes sociais, uma arquitetura de referência baseada no modelo e seu *workflow*. A discussão sobre os resultados experimentais onde se apresenta a implementação de um estudo de caso para validar o modelo proposto acontece na seção 5. Na seção 6, são trazidos para a discussão os resultados experimentais desta pesquisa. Finalmente, na seção 7 são feitas as considerações finais e proposições de trabalhos futuros.

2 REFERENCIAL TEÓRICO

2.1 VIGILÂNCIA TECNOLÓGICA

Antes da globalização e da aceleração das mudanças tecnológicas, estar atualizado em relação à evolução e aos resultados dos esforços tecnológicos era relativamente mais simples (SÁNCHEZ, J Marcela; PALOP, 2002). A comunidade científica e tecnológica era menor em número e em países, os principais trabalhos científicos apareciam em um volume "administrável" de publicações, as sobreposições entre comunidades de pesquisa não eram habituais e a disseminação dos avanços era feita de forma mais pessoal. A velocidade com que novidades surgiam era mais lenta do que hoje e, frequentemente, as mudanças de estado da arte coincidiam com o ciclo de vida de um profissional de uma empresa.

Com o crescimento da industrialização, importantes mudanças de paradigmas foram implementadas no mercado, resultando nas "revoluções industriais" (LASI *et al.*, 2014). A partir delas, o uso intensivo de tecnologias se consolidou como um diferencial competitivo para as organizações e acelerando o desenvolvimento tecnológico. Ao longo das últimas décadas, a juntada de informações sobre competidores ou tecnologias como forma de apoio à tomada de decisão ganhou importância e passou a exigir formas mais organizadas na gestão de informações.

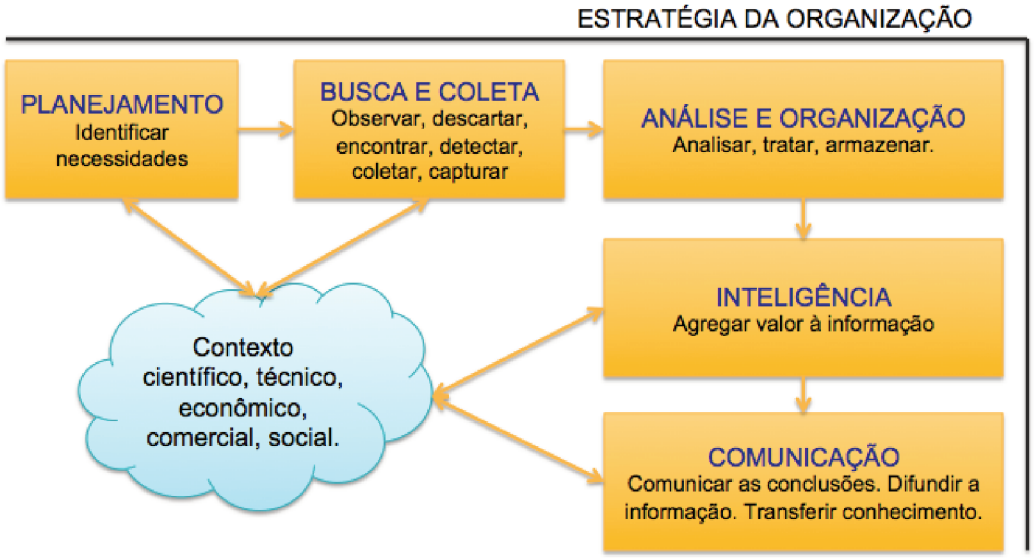
As tentativas de se criar uma forma estruturada de se obter e analisar informações tecnológicas competitivas deu origem ao termo **Vigilância Tecnológica (VT)**. Em linhas gerais, pode-se conceituar a VT como um processo estruturado para monitorar e avaliar fontes de informações formais e informais, eletronicamente disponíveis ou não, em busca de evidências que sinalizem alterações no cenário tecnológico no qual uma organização está inserida. O quadro 1 apresenta cinco definições elaboradas por autores de referência na área pelas quais se podem observar pontos convergentes.

Dentre os benefícios buscados pelas organizações com a VT estão a necessidade de antecipação de mudanças para evitar desvantagens competitivas, a redução de custos, os progressos em relação ao mercado, a necessidade de inovar e a identificação de novos parceiros. Por outro lado, a não adoção de um sistema estruturado pode acarretar em perdas de competitividade devido a entrada de um parceiro tecnologicamente mais avançado. Um exemplo desta situação vem acontecendo com as empresas de distribuição fonográfica que precisaram se adaptar à diversas tecnologias físicas como discos de vinil, fitas cassetes e CDs, mas não conseguiram reagir na mesma proporção ao surgimento das plataformas digitais como o iTunes e Spotify[1].

De forma geral, os métodos de vigilância tecnológica possuem etapas similares, dividindo-se em coleta de informações, análise do material coletado e comunicação dos resultados aos interessados. A Figura 1 ilustra as etapas propostas por Palop e Vicente (1999), em que os autores sugerem as etapas de **planejamento** na qual se busca identificar as necessidades informacionais

[1] O iTunes e Spotify permitem aos usuários comprarem faixas de músicas desde smartphones, tablets e computadores ou consumirem conteúdo por assinatura alterando a forma como as pessoa se relacionavam com o conteúdo de seus artistas(GOMES, C. *et al.*, 2015).

Figura 1 – Metodologia de J Marcela Sánchez e Palop (2002).



Fonte: traduzida pelo autor.

da organização, a etapa de **busca e coleta** em que as fontes de informação são observadas e têm sua informação capturada, a etapa de **análise e organização** onde se analisam os dados coletados, tratando-os e armazenando o que for relevante, a etapa de **inteligência** a partir da qual são feitas as análises e cruzamentos de informação e com o propósito de agregar valor à informação bruta capturada gerando os chamados *produtos da vigilância tecnológica* e, por último, a etapa de **comunicação** em que se encaminham as análises produzidas às pessoas interessadas a fim fundamentar suas tomadas de decisões.

O escopo exato de cada etapa e como elas devem de fato ser implementadas depende, dentre outras coisas, das necessidades informacionais ou do volume de canais e informações a serem coletadas, processadas e distribuídas. Um sistema de vigilância pode ser constituído por rotinas de pesquisas patentarias realizadas por um colaborador de uma organização ou até por um complexo processo contendo as etapas de coleta, análise, inteligência e comunicação de forma recorrentes, atuando sobre enormes volumes de informações e entregando resultados em curtos períodos de tempo, o que demanda o uso de tecnologias de apoio. Os exemplos anteriores ilustram duas classificações básicas em que estes tipos de sistemas podem ser divididos segundo o uso das tecnologias utilizadas, sendo eles:

- **Vigilância tecnológica:** este trabalho define um sistema de VT tradicional (não automatizado) como aquele que tem a maior parte de suas atividades realizadas de forma manual. Estão inclusos nesta categoria os estudos técnicos, as pesquisas e compilações de patentes executados preponderantemente de forma manual, incluindo aquelas em que são utilizadas ferramentas como bases especializadas para consultas e planilhas eletrônicas para compilação de dados durante as atividades.

Quadro 1 – Definição de vigilância tecnológica.

Definição	Fonte
Vigilância (Tecnológica) é o esforço sistemático e organizado empreendido pela empresa para observação, captura, análise, divulgação precisa e recuperação de informações sobre os fatos do ambiente econômico, tecnológico, social ou comercial, relevantes para que seja capaz de implicar uma oportunidade ou ameaça a este.	Palop e Vicente (1999)
Sistema estruturado para coordenar as atividades de recuperação de informação, processamento/análise e disseminação, na informação interna e do meio ambiente de acordo com um plano e uma estratégia organizacional.	Salgado Batista *et al.* (2003)
A vigilância tecnológica passa pelas etapas de diagnóstico, pesquisa e captura de informação, análise da informação, valorização da informação relevante, divulgação e comunicação, oferecendo orientação à tomada de decisões.	OVTT (2019a)
Um modelo de vigilância tecnológica é composto por um conjunto de processos: identificação das necessidades, definição das fontes e meios de acesso à informação; busca, tratamento e validação; valorização da informação, resultados, medição e melhoria (ALZATE et al., 2012).	UNE 166006 de 2006 (AENOR, 2019)
Um processo sistemático que objetiva identificar, organizar e correlacionar os resultados da prospecção tecnológica de forma a torná-los úteis às estratégias da organização.	Norma ABNT NBR 16501:2011 (ABNT, 2011)

Fonte: o autor.

- **Vigilância tecnológica automatizada:** sistemas de VT automatizados são úteis quando os decisores precisam analisar grandes volumes de informações em um curto período de tempo. Eles possuem a maior partes de seus processos sistematizados com o auxílio de tecnologias computacionais e seu fluxo de informação transita com baixas taxas de intervenção humana.

Adicionalmente, os processos de vigilância podem ser categorizados de acordo com a necessidade informacional de quem os demanda. Segundo OVTT (2019a), os tipos mais comuns são a vigilância ativa ou monitoramento, a vigilância passiva, inteligência competitiva (IC), e previsão tecnológica, os quais estão detalhados a seguir.

- **Vigilância Ativa ou Monitoramento:** é um tipo de vigilância pela qual se buscam informações ou dados em fontes específicas para suprir uma necessidade de informação previamente estabelecida pelas organizações. Sua execução pode acontecer pontualmente. As pesquisas patentárias são um exemplo deste tipo de vigilância, pois exigem que se coletem patentes de bases especializadas para posterior análise (GUZMÁN SÁNCHEZ; SOTOLONGO AGUILAR, 2002).

- **Vigilância Passiva:** é um tipo de vigilância que consiste no recebimento de informações de interesse por uma organização. Tem sua utilização difundida no monitoramento em saúde pública em que órgãos ou institutos disponibilizam canais de atendimento para receber notificações sobre incidentes ou registro de substâncias.

- **Inteligência Competitiva:** segundo Ruthes (2007) este é um processo ético baseado em informações legalmente disponíveis sobre tendências, eventos e atores externos às fronteiras de uma organização que objetiva subsidiar a tomada de decisão de seus gestores e contribuir para que suas metas sejam atingidas. Ainda, segundo o autor, a IC contribui para o entendimento das estratégias e à operação dos concorrentes-chaves. Como principais etapas de um processo de IC, tem-se: a coleta, a análise e a disseminação das informações consolidadas para os usuários.

- **Previsão Tecnológica:** refere-se aos estudos técnicos de projeção da situação tecnológica atual e passada para um plano futuro estritamente científico-tecnológico. A previsão ou extrapolação de dados, conforme as ideias de Garcez e James Terence Coulter Wright (2010), ao se basear na utilização de dados, informações e eventos já ocorridos para prever o futuro, torna-se válida desde que o futuro não represente uma ruptura radical com o passado, pelo menos em certos aspectos técnicos ou mercadológicos e por um período limitado. Neste tipo de vigilância as amostras devem permitir fazer uma comparação histórica e quando disponível em grande quantidade contribui na redução da incerteza e melhora a validação estatística.

Outro importante ponto a ser considerado são as ferramentas ou softwares de apoio à vigilância. Elas são essenciais para que se obtenha maior sucesso e precisão, especialmente quando se lida com volumes significativos de fonte de informações, especialistas ou interessados. A escolha das ferramentas deve estar alinhada ao objeto estratégico da organização e aos resultados esperados do monitoramento. Alguns exemplos que estão disponíveis atualmente são:

- **Alertas:** são serviços especializados em notificar aos interessados sobre novidades de um setor ou assunto escolhido, como alterações em legislação, eventos, publicação de patentes e licitações. Um mecanismo de alerta disponível é o Google Alerts [2] pelo qual o usuário pode cadastrar termos que deseja acompanhar em páginas e portais monitorados pela ferramenta e o sistema passa a enviar boletins periódicos com as publicações nas quais eles apareçam.

- **Buscadores especializados:** são buscadores desenvolvidos para recuperar informações de um determinado tipo de fonte, como patentes, ou uma área em particular, como medicina.

- **Bases especializadas:** são muito utilizadas nas atividades de vigilância pois concentram uma densa quantidade de documentos sobre uma área temática. Nesta categoria, temos as bases de patentes, de artigos científicos e de teses.

- **Meta-pesquisadores:** também são conhecidos como agregadores. Permitem realizar consultas em múltiplas fontes de dados por meio de uma interface única de maneira organizada

[2] https://www.google.com.br/alerts

e utilizando os motores de buscas das fontes pesquisadas. Exemplos comuns são as plataformas de pesquisas de voos ou de preços em sites de e-commerce.

- **Marketplace:** são ambientes onde se divulgam necessidades tecnológicas, oportunidades para colaboradores, parceiros ou sócios e ofertas de tecnologias.
- **Softwares de vigilância tecnológica:** são softwares para gestão completa do processo de vigilância tecnológica. Devido à sua importância, estas ferramentas serão melhor abordadas na próxima seção.

Como foi visto acima, as ferramentas têm a característica de ampliar a capacidade de coletar, analisar e distribuir as informações. Algumas, inclusive, estão disponíveis de forma gratuita, como o Alerts do Google. Utilizar as bases e os buscadores especializados tende a trazer mais qualidade à entrada de informações, o que é fundamental para que se produzam boas análises.

2.1.1 Plataformas de Vigilância Tecnológicas

Na literatura, é comum encontrar a vigilância tecnológica aplicada à áreas ou contextos específicos (ANDRADE NAVIA *et al.*, 2018), (MARULANDA *et al.*, 2016), (KARVONEN *et al.*, 2016), (PADILLA *et al.*, 2018b), onde sua execução é realizada de forma pontual com o apoio de algumas ferramentas. Contudo a implantação de sistemas de vigilância tecnológica em organizações envolve desafios financeiros, humanos e culturais, e exige formalizações e padronizações de processos (PEREZ, L. G. *et al.*, 2017) para sustentar as atividades necessárias à execução e manutenção do mesmo (VILLARROELG *et al.*, 2015), principalmente quando existe a necessidade de um monitoramento constante. Nesse sentido, a norma europeia UNE 166006 de 2006 (AENOR, 2019) e a brasileira ABNT NBR 16501:2011 (ABNT, 2011) proporcionam diretrizes para sua implantação.

Uma opção para conduzir a vigilância de um modo mais sistemático e se obter automatização nos processos de busca e análise da informação é o uso de plataformas integrais web, softwares disponíveis na internet ou intranet que podem ser acessados por navegadores como Google Chrome ou Firefox. São exemplos de recursos oferecidos por estas plataformas (BERGES-GARCIA *et al.*, 2016):

- Sistematização, automação e centralização de processos de VT / IC;
- Monitoramento de fontes de informação como notícias, patentes e artigos científicos;
- Filtragem de informações;
- Análise visual da informação;
- Gerenciamento de conteúdo;

- Exportação de informações em vários formatos;
- Gerenciamento de usuários;
- Envio de boletins, newsletters, etc; e
- Gerenciamento de alertas.

Atualmente, estão disponíveis diversas opções de plataformas integrais web para dar suporte às atividades de vigilância tecnológica. Complementarmente, existe a possibilidade de se utilizar softwares para problemas específicos, como a captura de documentos ou mineração de textos (SÁNCHEZ, Jenny Marcela; PALOP, 2002). Neste trabalho, foram analisadas as plataformas que oferecem a maior cobertura possível para as atividades de VT. O processo de escolha das plataformas considerou aquelas apontadas pelos trabalhos de Berges-Garcia *et al.* (2016) e de Martínez Rivero e Maynegra Díaz (2014), além da seleção feita pelo Observatório Virtual de Transferência de Tecnologia da Universidade de Alicante (OVTT, 2019a). Como resultado, foram analisadas oito plataformas: SoftVT, Vicubo Cloud, Vigiale, Innguma, MUSSOL, Minera, Intelligent Watcher e Hontza. A seguir, são discutidas e apresentadas as características de cada uma.

SoftVT

A SoftVT (AIMPLAS, 2020) é uma plataforma para a automação dos principais processos de Vigilância Tecnológica tais como captura, gerenciamento e disseminação de informações. É desenvolvida pela empresa espanhola de mesmo nome que o produto e tem como seus principais clientes organizações de médio e de grande porte como institutos e federações do mesmo país que desejam um sistema completo e centralizado para monitoramento e gestão de informações. Está composta por quatro módulos.

O módulo de recuperação busca informações diária e automaticamente, realiza o controle de publicações duplicadas, controle de registros duplicados, indexação e classificação automática, além de permitir a adição de novas fontes. O módulo de gestão trabalha com o controle de usuários, permissões de acesso, parâmetros de busca, gestão de índice temático e gestão de boletins eletrônicos. O terceiro módulo, chamado de módulo de oportunidades, oferece a possibilidade de vinculação de informações do sistema a propostas de clientes, a gestão de grupos de trabalhos e sistemas de avaliação de oportunidades. O módulo de análise estratégica é utilizado para o acompanhamento da concorrência tecnológica e de mercado, criação de mapas tecnológicos, análise de informação bibliométrica e estatística e tendências tecnológicas.

Vicubo Cloud

Vicubo Cloud (E-INTELLIGENT, 2020) é uma plataforma que permite extrair informações estratégicas do ambiente no qual uma organização inserida, favorecendo a inteligência

competitiva na tomada de decisões e buscando vantagens frente à concorrência. É bastante completa em termos de funcionalidades sendo capaz de monitorar diferentes tipos de fontes de dados e produzir gráficos para facilitar a compreensão de seus usuários.

Mais especificamente, a plataforma Vicubo Cloud oferece a possibilidade de monitorar marcas, concorrentes, tecnologias, redes sociais, sites e blogs. Permite, ainda, analisar, normalizar, filtrar e rotular o conteúdo capturado, otimizando o trabalho do usuário final nas atividades de consulta e análise. Segundo informações do seu site, ela é capaz de controlar de páginas mal formatadas em XML, links quebrados, edições peculiares e interrupções de serviços. Também é possível criar relatórios personalizados, visualizar de forma gráfica seus dados através de indicadores e enviar boletins informativos em Word, Excel, CSV e PDF.

Assim como a plataforma SoftVT, a Vicubo Cloud auxilia na centralização e na sistematização do processo de vigilância tecnológica e inteligência competitiva, contribuindo para que as organizações que a utilizem possam se manter em conformidade com a norma europeia UNE 166.006 (AENOR, 2019).

Vigiale

A Vigiale (TECNOLOGÍA, 2020) consiste em uma Plataforma Web para o gerenciamento de Vigilância Tecnológica que permite o rastreamento de fontes de informação selecionadas e a notificação de alterações detectadas aos seus usuários. Ela também permite gerir diferentes fontes de informação de forma integrada, ordená-las, classificá-las e fazer sua atualização, por meio de tecnologias de captura, categorização, indexação e filtros customizáveis de acordo com as necessidades e requisitos de cada empresa ou organização.

Dentre as principais funcionalidades da plataforma estão o monitoramento de fontes de informações, detecção de mudanças em publicações monitoradas, a disponibilização de painéis para análise visual da informação que contabiliza os volumes de dados recolhidos periodicamente classificados por tipo e categoria de informação e os termos mais frequentes divididos por grupos temáticos. Complementarmente oferece a possibilidade de enviar alertas por e-mail contendo as últimas notícias geradas e coletadas. Também é possível gerar boletins em PDF automaticamente com os itens coletados, segmentando-os por tópicos.

Innguma

Innguma (IDEKO, 2020) é um software para inteligência competitiva e vigilância tecnológica que permite aos usuários capturarem dados de fontes escolhidas, filtrar conteúdos, organizar a informação, fazer pesquisas nos documentos coletados e difundir as informações para dar apoio na tomada de decisões. Dentre as principais atividades que ele permite, estão:

- Possibilidade de adicionar fontes de informação a serem monitoradas, no padrão RSS[3] e

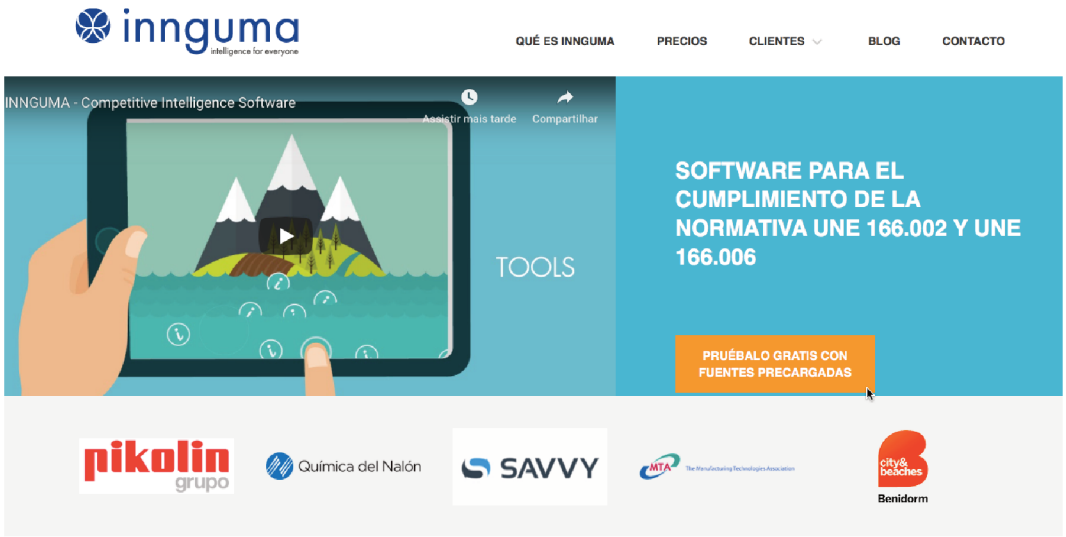
[3] RSS (do inglês para Rich Site Summary ou Really Simple Syndication) é um formato utilizados pelos sites para disponibilizar um resumo de seus conteúdos. Tecnicamente, eles o fazem por meio de um arquivo no formato

redes sociais;

- Mecanismos de pesquisas tradicional, semântica e filtragem de publicações;
- Tradução automática de publicações;
- Estatísticas das fontes de informação monitoradas;
- Rotulagem de informação e repositório categorizado;
- Divulgação das análises e conteúdos capturados por meio de boletins informativos;
- Painéis (dashboards) dinâmicos.

O software é usado por empresas de médio e grande porte para garantir sua atualização tecnológica. Tem como casos de sucesso a empresa Savvy Data Systems, desenvolvedora de soluções para monitoramento e análise de dados para fabricantes de máquinas e ferramentas que por meio da plataforma manteve o controle sobre notícias e atualizações sobre cibersegurança. Outro "case" de sucesso foi feito em conjunto com a MONDRAGON Corporation, grupo de identidade cooperativa composto por mais de 100 cooperativas com presença global, que atua nos setores industrial, financeiro e de logística, em que a esta plataforma foi utilizada para oferecer informações estratégicas e transversais para suas cooperativas. Uma tela do website da ferramenta pode ser vista na Figura 2, onde também é possível encontrar outros estudos de casos.

Figura 2 – Website do software de vigilância tecnológica Innguma.



Fonte: Innguma. Acessado em 05 de julho de 2019.

O conjunto de funcionalidades apresentado mostra seu potencial para aqueles usuários que sabem exatamente quais fontes precisam monitorar. De todo modo, as estatísticas sobre as fontes que a ferramenta disponibiliza permite que seja feita constantemente uma curadoria,

XML o qual costuma trazer um conjunto de itens com título, resumo, data e hora das publicações do site.

melhorando progressivamente a qualidade das mesmas. Assim como as anteriores, a Innguma sintetiza o conteúdo de forma visual, em boletins e e-mails para que seus usuários possam consumir as informações de um modo mais amigável.

MUSSOL

MUSSOL (S.L., A. information technology, 2020) é uma solução para auxílio de atividades de inteligência competitiva nas organizações. A plataforma é comercializada no formato Saas (software como serviço) e tem como especialidade coletar informação relevante e analisar de forma automática, para apoiar a alta gestão na tomada de decisão.

A solução permite coletar informações de sites (completos ou parciais), feeds RSS de notícias, newsletters ou boletins recebidos por e-mail, bases de dados (e.g. patentes ou artigos científicos), repositórios corporativos de documentos, redes sociais e também armazenar de informações geradas pela própria organização. Ela utiliza inteligência artificial para auxiliar na seleção das publicações mais relevantes. Segundo análise feita pelo OVTT (OVTT, 2019a), as principais funcionalidades da ferramenta são:

- Disparar alertas sobre fatos que possam impactar no negócio, por exemplo oportunidades de negócio e obstáculos;
- Monitorizar competidores;
- Apoiar processos de exportação e de internacionalizar mercados;
- Implementar mecanismos de monitoramento ativa sobre produtos e organização;
- Apoiar a gestão de ideias e inovação na organização.

Intelligent Watcher

Intelligent Watcher (IW) (WATCHER, 2020) é um software gratuito cujo foco é a extração de informações de sites, blogs, redes sociais e bases de dados online de forma simples e automatizada, segundo o site dos desenvolvedores. O software permite enviar boletins sobre o conteúdo capturado e visualizar gráficos básicos sobre o volume de dados capturados. Ele foi projetado para auxiliar as organizações a cumprir a norma UNE 166006 de Gestão do conhecimento e vigilância tecnológica.

Minera

A plataforma Minera (S.L., M., 2020) contribui para as atividades de monitoramento, armazenamento, análise e visualização das informações. Ela permite trabalhar com o monitoramento automatizado de patentes, literatura científica, notícias, páginas, blogs, twitter e tweets. As informações podem ser capturadas manual e automaticamente. As informações capturadas

podem ser categorizadas com rótulos que definem diferentes áreas. São oferecidos gráficos e "Geomapas", para que o analista explore cada tópico de interesse. O sistema ainda permite realizar análises de tendências por meio de mineração de texto e disponibiliza um fórum de discussão para questões interesse dos usuários.

Hontza

A plataforma Hontza (CDE INTELIGENCIA COMPETITIVA, 2020) foi concebida para dar suporte ao ciclo de vigilância de tecnológica e inteligência competitiva de organizações públicas e privadas. Ela é distribuída como software livre, mas pode ser contratada coo serviço. Utiliza como base o sistema de gestão de conteúdo Drupal [4]. Hontza foi criada pelo Centro de Vigilância de Normas e Patentes da Espanha (CDE)[5], como uma evolução do sistema de conteúdo Drupal.

Segundo a empresa, Hontza foi projetada para organizações que tenha definido um ambiente competitivo e uma equipe de pessoas que monitorem as oportunidades para se antecipar. Algumas das principais vantagens do Hontza sobre as outras plataformas avaliadas por Martínez Rivero e Maynegra Díaz (2014) em conjunto com especialistas são:

- É uma plataforma colaborativa;
- Oferece suporte para o ciclo completo da TV;
- É grátis;
- O acesso ao Hontza através de uma intranet corporativa pode oferecer mais segurança;
- Respeita a norma espanhola AENOR UNE166006: 2011;
- Oferece módulos de análise e semântica que podem ser acoplados ao Drupal para aprimorar os recursos de análise do Hontza.

A plataforma conta com recursos para: automação e integração com fontes públicas ou privadas para coleta de dados nos formatos RSS, JSON, Google Sheets e através de Scripting + Scraping por meio de uma solução chamada Hound[6]; realizar vigilância de conteúdo coletado via RSS ou inserida pelos usuários, sendo possível categorizar, comentar e incluir elas em boletins; gerar alertas personalizados e boletins personalizados para divulgação e marketing; criação de debates vinculados à notícias ou eventos; criação de documentos editados pela equipe de notícias como relatórios.

[4] https://www.drupal.org/
[5] http://www.cde.es/es/index.html
[6] http://www.hontza.es/hound/

2.1.1.1 Análise comparativa das plataformas

Aurelio Berges-Garcia *et al.* (2016) em seu trabalho "Methodology for evaluating functions and products for technology watch and competitive intelligence (TW/CI) and their implementation through web" apresentam um conjunto de indicadores para que as organizações possam avaliar as plataformas web voltadas à VT de acordo com suas necessidades e circunstâncias concretas. Os indicadores ou critérios foram divididos em dois domínios, no domínio número 1 estão as funções associadas aos estágios do ciclo global de VT e no domínio número 2, as funções mais tecnológicas e horizontais. Para este trabalho, os softwares foram observados do ponto de vista do primeiro domínio, considerando cinco critérios relativos ao ciclo de VT:

1. Busca e extração de informação;
2. Filtragem e valorização da informação;
3. Análise de informação;
4. Inteligência estratégica;
5. Difusão.

O Quadro 2 compara os softwares por meio de suas principais funcionalidades e benefícios. As informações foram inferidas a partir das peças de divulgação oferecidas pelas empresas, como sites e encartes. Estão presentes no quadro apenas as funcionalidades constantes em mais de uma das plataformas. Além disso, elas foram agrupadas de acordo com os indicadores citados. Abaixo de cada no sistema quadro anotou-se "S" quando a funcionalidade estava descrita no material da empresa e "N" quando não se pôde confirmar sua presença.

Quadro 2 – Comparativo das plataformas de vigilância tecnológica integrais.

Critério	Funcionalidade	SoftVT	Vicubo	Vigiale	Innguma	MUSSOL	IW	Minera	Hontza
1	Adição de novas fontes	S	S	S	N	N	S	S	S
1	Captura de patentes e legislações	S	S	S	S	S	S	S	N
1	Captura de RSS	S	S	S	S	S	S	S	S
1	Captura de Redes Sociais	S	S	S	S	S	S	S	N
1	Detecção de alteração de conteúdo monitorado	N	N	S	N	S	N	N	N
3	Controle de categorias do conteúdo	S	S	N	S	S	S	S	S
3	Rotulagem de conteúdo	N	S	N	S	N	N	S	S
3	Ferramentas automatizadas de apoio à análise	S	S	S	N	N	S	S	N
4	Gráficos com indicadores	S	S	S	S	N	S	S	N
5	Geração de boletins personalizados	S	S	N	S	S	S	N	S
5	Geração de RSS	S	N	N	S	S	S	N	S
5	Envio de alertas por e-mail	S	N	S	N	N	S	N	S

Fonte: elaborado pelo autor.

Dentre os sistemas avaliados, cinco deles oferecem algum suporte para as atividades de inteligência. O SoftVT traz um módulo de análise estratégica que ajuda a traçar um panorama tecnológico sobre a concorrência. O Vicubo Cloud traz ferramentas para filtrar e normalizar as informações para o analista. O Vigiale disponibiliza gráficos para visualizar estatísticas sobre os dados coletados incluindo categorias e tipos de informação. O Minera é permite realizar análises de tendências por meio da mineração de texto. Estas são funcionalidades que enriquecem conteúdo coletado e agregam valor aos olhos de seus consumidores. Em sua maioria, elas são comercializadas por meio do modelo de assinatura (software como serviço), cobradas para serem implantadas no cliente ou distribuídas livremente.

Apesar de não constar no Quadro 2, outras funcionalidades que facilitam o processo de vigilância mas estavam presentes de forma individualizada nas plataformas estudadas são: inibição de duplicidade em documentos, indexação automática, classificação automática de conteúdo, tradução automática de conteúdo, controle do período para cada coleta, controle dos parâmetros de busca, ambiente para criação de mapas tecnológicos, análise bibliométrica automatizada, visualização facilitada para identificar tendências tecnológicas, parametrização para normalização do conteúdo coletado, aplicativo móvel e pesquisa semântica.

Em resumo, a análise realizada aponta que tende a existir um ganho de produtividade e de assertividade com o uso das ferramentas disponíveis. No entanto o gargalo que surge na atualidade é a capacidade de se lidar com incontáveis fontes e formatos de dados que se multiplicam de modo exponencial. Gerenciar as informações em constante crescimento demanda muito mais do que capacidade analítica, sendo necessária uma enorme capacidade de armazenamento e de processamento. O cenário atual em grande parte é caracterizado pela era do Big Data, que é abordada em maior detalhes na seção 2.2 e para a qual estas ferramentas analisadas se apresentam como apropriadas.

2.2 CENÁRIOS DE BIG DATA

No ano de 1880, o excesso de informação ficou evidente quando se tornou um dos principais gargalos no censo dos Estados Unidos. Eram necessários oito anos para que ele fosse calculado. Posteriormente, graças à máquina tabuladora de Hollerith (HOLLERITH, 1894), que utilizava cartões perfurados, o censo pôde ser calculado em pouco mais de um ano de trabalho. Os resultados positivos e a crescente demanda por armazenamento e poder computacional para um volume cada vez maior de dados tornou Hermann Hollerith fundador, em 1924, da International Business Machines Corporation (IBM).

A partir do ano de 1941, o crescimento exponencial de informação tomou ainda mais espaço nas organizações e nos governos e passou a ser conhecida por especialistas como "explosão informacional", definição consolidada por um artigo publicado no jornal "The Lawton Constitution". Em 1944, Fremont Rider, um bibliotecário americano da Wesleyan University, levantou uma bandeira vermelha sobre o problema de armazenamento de informações ao prever que a cada dezesseis anos as bibliotecas das Universidades Americanas tendiam a dobrar de

tamanho. Com essa taxa de crescimento persistindo, Rider estimou que a biblioteca de Yale no ano de 2040 tenderia a acumular 200.000.000 de volumes (SANTHIYA, 2018).

A partir dos anos 2000, com a popularização da internet e mais recentemente com o uso intensivo de tecnologias como smartphones, computação em nuvem e internet das coisas, houve um aumento na geração e armazenamento de dados em proporções nunca antes vistas e de fontes totalmente heterogêneas (redes sociais, sensores, marketing, finanças, governo, saúde, comércio eletrônico, entre outras). Até o ano de 2003, haviam sido criados pelo menos 5 Exabytes (10^{18} bytes) em todo o mundo. Já em 2012, eram gerados aproximadamente 2,5 exabytes por dia, o que significa que seriam necessários 20 bilhões de computadores pessoais com capacidade de 500 gigabytes (10^9) para armazená-los (SAGIROGLU; SINANC, 2013). Percebeu-se, também, que o volume estaria dobrando a cada 40 meses aproximadamente. Em 2013, a International Data Corporation (IDC) estimou que tenham sido gerados, replicados e consumidos cerca de 4,4 Zettabytes (ZB) de dados e que este volume estaria dobrando a cada dois anos. Em 2015, os dados gerados já estariam em 8 ZB.

O crescimento vertiginoso e a necessidade de alto poder de processamento ficam claros quando se olha para as projeções futuras. A rede de supermercados Wallmart, por exemplo, gera mais de 2,5 PB de dados sobre as transações efetuadas por seus consumidores a cada hora (OUSSOUS *et al.*, 2018). Em maior escala, estima-se que volume de dados gerados no mundo poderá saltar de 45 Zettabytes em 2019 para 175 Zettabytes em 2025 (REINSEL *et al.*, 2018), sendo que a estimativa é de que quase 30% destes dados precisarão de processamento em tempo real.

Para descrever esses cenários foi cunhado o termo "Big Data". O termo teve sua formalização nos meios científicos atribuída aos pesquisadores da NASA Michael Cox e David Ellsworth (COX; ELLSWORTH, 1997), mas foi larga e rapidamente adotado por toda a indústria de tecnologia. O termo *Big Data* refere-se a cenários onde transitam ou são armazenados crescentes volumes e variedades de dados, com ou sem estruturas previamente definidas, difíceis de serem suportados pelas tecnologias de armazenamento ou processamento tradicionais.

O sistemas de gerenciamento de bancos de dados tradicionais (SGBDs) são baseados na lógica de predicados e na teoria dos conjuntos. Eles representam os dados em banco de dados como tabelas, também chamadas de relações, como no SGBD PostgreSQL [7]. Neste paradigma, é possível oferecer maior capacidade de armazenamento por meio de clusters de banco de dados. Para Macedo *et al.* (2008) os clusters oferecem preocupações como escalabilidade, autonomia e replicação dos dados, sendo que ao crescer o número de nós o sistema deve manter o mesmo desempenho.

No entanto, os bancos de dados tradicionais não são suficientes para lidar com os cenários de Big Data já que seus dados costumam ser compostos por dados estruturados, semi-estruturados e não estruturados, como textos, vídeos e imagens. Sendo assim, o Big Data demanda

[7] O PostgreSQL (www.postgresql.com) é um sistema de gerenciamento de banco de dados de código aberto com suporte ao modelo híbrido objeto-relacional contemplando recursos como chaves estrangeiras, indexação textual, store procedures em várias linguagens, integridade transacional.

operações complexas como mineração de textos e processamento de imagens. De todo modo, o particionamento dos dados em diversos computadores, ou nós, tende a ser necessário para suportar o grande volume e permitir que sejam processados e analisados em um tempo e em um custo aceitável.

Para Sagiroglu e Sinanc (2013), o termo Big Data é utilizado para definir conjuntos de dados massivos com estruturas grandes, variadas e complexas, e que oferecem dificuldades adicionais de armazenamento, análise e visualização para seus processos e resultados. De Mauro *et al.* (2016) realizaram uma pesquisa sobre as principais características presentes em cenários de Big Data e elaboraram uma definição levando-as em consideração. Em sua pesquisa eles concluíram que "Volume", "Velocidade" e "Variedade" descrevem as características da Informação. "Tecnologia" e "Métodos analíticos" definem os requisitos necessários para que seja possível utilizar de forma adequada essas informações; e "Valor" descreve a transformação de informações brutas em informações úteis capazes de gerar vantagens econômicas para empresas e sociedade. Em uma tradução livre, eles escreveram: "Big Data é um ativo de informação caracterizado por volume, velocidade e variedade tão altos que exige tecnologia e métodos analíticos específicos para sua transformação em valor".

O processo de análise em cenários de Big Data é conhecido como "Big Data Analytics". Suthaharan (2014) entende esta atividade como o processo de analisar e compreender as características de conjuntos de dados massivos e extrair padrões geométricos e estatísticos. Para Sagiroglu e Sinanc (2013) Big Data Analytics consiste em um processo de pesquisa em grandes volumes de dados em busca de padrões ocultos ou correlações entre os mesmos. Estes autores prosseguem destacando que as informações obtidas desta atividade são úteis para empresas e organizações uma vez que ajudam a obter "insights" ricos e aprofundados gerando vantagem competitiva sobre a concorrência. Como observam De Mauro *et al.* (2016), no mercado competitivo atual ser capaz de explorar dados para segmentar de clientes e entender seus comportamentos, oferecer serviços personalizados e obter informações sobre os dados fornecidos por várias fontes são a chaves para a vantagem competitiva.

É possível encontrar aplicações em cenários de Big Data nas mais variadas áreas, como exemplo em aplicações governamentais, indústrias, varejo, diagnóstico médico por imagem, seguros de saúde, internet das coisas, segurança computacional, prevenção de desastres naturais (BIG DATA, 2019) e cidades inteligentes (GOMES, E. H. *et al.*, 2018).

A consultoria McKinsey (MANYIKA *et al.*, 2011) publicou uma pesquisa apresentando o potencial da geração de valor do Big Data para cinco grandes áreas: saúde nos Estados Unidos da América, administração pública na União Europeia, varejo nos Estados Unidos, manufatura global e dados pessoais de localização, os quais representaram cerca de 40 por cento do produto interno bruto global. A pesquisa identificou 15 "alavancas" com potencial de melhorar a eficiência e a eficácia da área da saúde, explorando o grande volume de informações eletrônicas já disponíveis. O mesmo foi feito para as demais áreas. Um resumo dos principais itens considerados como alavancas na pesquisa citada é apresentado a seguir.

- **Saúde:** sistemas de apoio à decisão clínica. Transparência sobre dados médicos. Monitoramento remoto de paciente. Análise avançada aplicada aos perfis dos pacientes. Sistemas automatizados para detecção de fraudes em cobranças. Modelagem preditiva de novos medicamentos. Análise de dados clínicos de pacientes para identificar indicações adicionais e descobrir efeitos adversos de medicamentos. Medicina personalizada. Análise de padrões de doenças. Plataformas e comunidades online para coleta de dados valiosos. Aprimoramento da vigilância em saúde pública;

- **Administração Pública:** criação de transparência. Permitir a experimentação para descobrir necessidades, expor a variabilidade e melhorar o desempenho. Segmentar populações para personalizar ações públicas. Substituição ou suporte à tomada de decisão humana por algoritmos automatizados. Inovação de novos modelos, produtos e serviços de negócios com Big Data, fornecendo ferramentas e análises para que a iniciativa privada, organizações sem fins lucrativos e indivíduos criem valor para o setor público;

- **Varejo:** venda cruzada. Marketing baseado em localização. Análise de comportamento na loja. Micros-segmentação de clientes. Análise de sentimentos. Aprimorando a experiência do consumidor multicanal. Otimização de preços. Otimização de posicionamento e design. Gerenciamento de estoque. Distribuição e otimização logística. Informação de negociações com fornecedores, serviços de comparação de preços. Mercados baseados na Web;

- **Indústria:** gerenciamento do ciclo de vida do produto. Design para valorizar. Inovação aberta. Fábrica digital. Operações acionadas por sensor. Cadeia de suprimentos;

- **Dados pessoais de localização:** roteamento inteligente. Telemática automotiva. Serviços baseados em localização de celulares. Publicidade segmentada por área geográfica. Cobrança eletrônica de pedágio. Preços de seguro. Resposta de emergência. Planejamento urbano. Inteligência de negócios de varejo.

As aplicações que fazem uso intensivo de dados e processamento são necessárias nas mais variadas áreas, como aponta o estudo da McKinsey. O governo dos Estados Unidos lançou no ano de 2009 o "Data.gov" com o propósito de dar mais transparência e responsabilidade às ações dos governantes. O Data.gov é um data warehouse, ou repositório de dados, no modelo Open Data (BARBOSA *et al.*, 2016, p. 3), com 261,318 *datasets* (conjuntos de dados) contendo dados sobre agricultura, clima, consumo, educação, energia, finanças, saúde, governo, manufatura, segurança pública, ciência e tecnologia (DATA.GOV, 2020). A partir dos dados disponibilizados, empresas, ongs e cidadão têm desenvolvido aplicativos como o City-Data.com[8], um portal onde é possível pesquisar perfis detalhados e informativos para todas as cidades dos Estados Unidos, incluindo taxas de criminalidade e padrões climáticos. O Fooducate[9] permite que os compradores de supermercados façam escolhas saudáveis sobre seus alimentos. Com ele é possível fotografar

[8] http://www.city-data.com
[9] http://www.fooducate.com/

o código de barras das embalagens e receber informações nutricionais dos mesmos. O "Where are the Jobs" [10] permite que os usuários explorem interativamente salários e estatísticas sobre trabalhos para várias profissões em níveis estadual e nacional dos Estados unidos.

O governo do Brasil também mantém uma iniciativa similar conhecida como Portal Brasileiro de Dados Abertos, construído para atender a Lei de Acesso a Informação Pública (Lei 12.527/2011) que regula o acesso aos dados e informações do governo brasileiro. No portal são disponibilizados 7.233 conjuntos de dados sobre saúde, educação, meio ambiente, entre outros (DADOS.GOV.BR, 2020).

Outros exemplos do uso de Big Data podem ser encontrados na biomedicina. Análises em Big Data já estão influenciando as decisões de saúde e atendimento ao paciente e mudando a forma como se trabalha com os dados em virtude da preocupação com a privacidade de segurança dos dados pessoais. A área de medicamentos personalizados vê com bons olhos estas novas tecnologias. Segundo Costa (2014), o Big Data em biomedicina é impulsionado pela importante preocupação de se ter programas de medicina personalizados que melhorarão significativamente o atendimento ao paciente. O autor prossegue destacando que os avanços no entendimento de diferentes informações ômicas possibilitam maior entendimento sobre os fatores genéticos causais que podem ajudar a gerenciar o triângulo de ouro do tratamento: o alvo certo, o medicamento e o paciente certo. Viceconti *et al.* (2015) apresentam um estudo de caso onde se utilizaram tecnologias preditivas de Big Data para prever o risco de ocorrer uma fratura óssea em um paciente acometido por osteoporose (uma doença que provoca redução da massa óssea e fragilização da microarquitetura do tecido ósseo). O setor de transportes também tem se beneficiado das tecnologias de Big Data e Analytics. Na Índia, por exemplo, são reservados cerca de 250.000 assentos todos os dias, sendo que estas reservas podem ser feitas com dois meses de antecedência. O país possui uma das maiores malhas ferroviárias do mundo, por isso fazer previsões de demanda é uma tarefa complicada uma vez que precisa levar em conta variáveis como feriados, finais de semanas, viagens noturnas e paradas intermediárias. Utilizando-se algorítimos de aprendizagem de máquina nestes cenários de Big Data, torna-se possível minerar os dados e aplicar "advanved analytics" nas coleções de dados sobre o passado assim como nas novas. Desta forma, pode-se garantir uma boa precisão nos resultados das previsões (OUSSOUS *et al.*, 2018).

2.2.0.1 Caracterização do Big Data

Apesar dos cenários de Big Datas já serem uma realidade comum na atualidade, sua definição ainda não encontrou um consenso. Uma das definições mais populares para o termo Big Data busca delimitar o conceito por meio de algumas de suas principais características e é popularmente conhecido como três V's, sendo eles: *variedade, volume e velocidade* (IBM *et al.*, 2011), (LAURILA *et al.*, 2012), (SAGIROGLU; SINANC, 2013) e (OUSSOUS *et al.*, 2018), os quais são discutidos a seguir.

[10] http://www.where-are-the-jobs.com/app.php

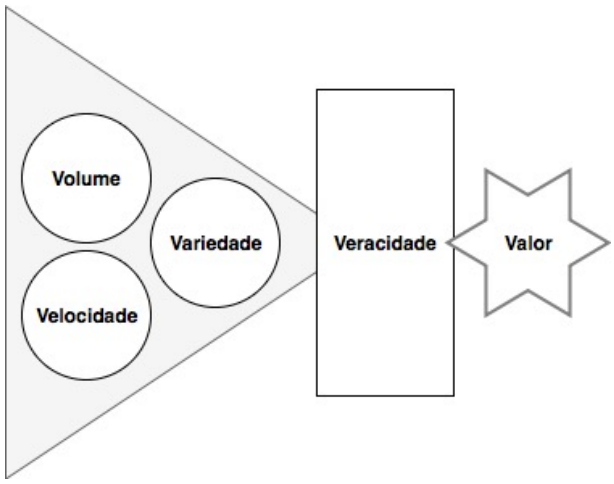
- Volume: os cenários de Big Data costumam ser compostos por grande quantidade de dados, podendo chegar facilmente a petabytes. O Google, por exemplo, monitora 7,2 bilhões de páginas por dia e processa 20 petabytes por dia (SAGIROGLU; SINANC, 2013). Isto torna inviável a utilização apenas de bancos de dados tradicionais neste tipo de cenário.

- Variedade: esta é uma das características marcantes dos cenários de Big Data pois dificulta seu armazenamento e análise. Os dados nestes cenários costumam vir de uma grande variedades de fontes. Quanto a suas estruturas, podem ser compostos por dados estruturados, como tabelas, semi-estruturados como arquivos no formato XML ou JSON e não estruturados como áudio, vídeo ou textos. Para Suthaharan (2014) a variedade representa o crescimento da variedade de dados no conjunto.

- Velocidade: é a celeridade com que os dados chegam ou devem ser processados. Segundo Oussous *et al.* (2018), os dados envolvidos nos cenários de Big Data são gerados de maneira rápida por suas fontes e devem ser processados rapidamente para que se possa extrair informações úteis e insights relevantes aos interessados.

Possuir ou não todas as características descritas anteriormente não faz de um cenário automaticamente classificado como Big Data. É preciso entender o problema em sua integralidade e nesse sentido outas definições foram propostas. Com base em suas análises e trabalhos correlatos, Demchenko *et al.* (2013) propuseram uma definição mais ampla de Big Data estendendo de 3 para 5 Vs: Volume, Velocidade, Variedade, Valor e Veracidade, mostrado na Figura 3. Esta definição foi bem aceita no meio acadêmico e de negócios (WHITE, 2015). O valor está ligado à serventia que os dados coletados são capazes de trazer ao processo, atividade ou predição buscada. Também impacta o resultado final almejado. Para os autores, o valor está intimamente relacionado ao volume e variedade de dados. O valor dos dados dependerá dos processos ou eventos que eles representam.

Figura 3 – Os 5Vs do Big Data.



Fonte: Elaborado pelo autor.

A veracidade é composta por dois aspectos: a consistência dos dados (ou certeza) e a confiabilidade dos dados, que pode ser dada por vários fatores como métodos de coleta e processamento, origem dos dados, entre outros. Em um cenário de Big Data, a veracidade deve garantir que os dados utilizados sejam autênticos, confiáveis e protegidos contra acesso e modificação não autorizados.

É cada vez mais comum se deparar com cenários de Big Data durante a estruturação de um processo de vigilância tecnológica, pelo qual se monitoram inúmeras fontes, capturando os mais variados tipos de dados. Os dados são oriundos majoritariamente de **fontes formais de informação** como bases de patentes, artigos científicos e publicações de portais especializados, conforme visto na seção 2.1. Um dado que não ofereça valor e que não tenha sua veracidade confirmada, dificilmente será útil ao tomador de decisão, mesmo após sofrer diversas análises e transformações. Desse modo, este trabalho adota a definição dos 5 Vs para caracterizar cenários de Big Data.

2.3 ANÁLISE DE DOMÍNIO

A **análise de domínio** (AD) é uma abordagem filosófico-realista, que busca encontrar a base para a Ciência da Informação em fatores externos às percepções individualistas-subjetivas dos usuários, em oposição, por exemplo, aos paradigmas comportamentais e cognitivos. Segundo a ela, quando se deseja desenvolver um software para a Geografia Brasileira, por exemplo, não se deve focar somente em seus usuários, mas sim convidar um geógrafo especializado em Geografia brasileira para contribuir com a solução.

Para este trabalho, o conceito "domínio de conhecimento" se torna relevante uma vez que representa "uma demarcação de determinado conhecimento, seja ele fixado num contexto profissional ou não" (THELLEFSEN, T.; THELLEFSEN, M., 2005, p 179). O paradigma de domínio analítico da Ciência da Informação afirma que a melhor maneira de compreender a informação é estudar os domínios do conhecimento como comunidades de pensamento ou discursivas[11], que são partes da divisão do trabalho da sociedade (HJØRLAND; ALBRECHTSEN, 1995).

O termo "domínio" é visto na CI como uma estrutura meta teórica útil para entender a informação dentro de assuntos específicos e para grupos específicos. O conceito foi inserido no contexto desta ciência por Albrechtsen e Hjørland (HJØRLAND; ALBRECHTSEN, 1995), que, à época, não definiram claramente o que entendiam como domínio. Teve sua aplicação trazida da Ciência da Computação e foi largamente difundida em meados da década de 80 (NEIGHBORS, 1980). Uma aplicação prática da análise de domínio pode ser vista no estudo de caso aplicado na área industrial no trabalho de Bastarrica *et al.* (2006) em que foi desenvolvida abordagem baseada em um conjunto de diretrizes, métricas, funções, entradas e saídas para sistemas de repositório com foco em reuso de software.

[11] Comunidades discursivas são definidas como distintos grupos sociais sincronizados em pensamento, linguagem e conhecimento, constituintes da sociedade moderna (HJØRLAND; ALBRECHTSEN, 1995).

Outro exemplo é encontrado no trabalho de Koziolek *et al.* (2013). Os autores apresentam uma abordagem de análise de domínio estruturada em sete etapas para avaliar o potencial das linhas de produtos de software (SPL). A abordagem tenta lidar com a identificação e análise de uma infinidade de fontes de informação, entrevistando os principais interessados nos recursos dos produtos e analisando a arquitetura, bem como com avaliação da perspectiva de reutilização e a criação de casos de negócios nas SPL. Os autores realizaram um estudo de caso sobre reuso de software para uma das maiores empresas de engenharia do mundo na área de tecnologia de energia e automação, a ABB.

Segundo Shapere (SHAPERE, 1977), domínio é o corpo total de informações para o qual, idealmente, uma resposta para um problema deva ser considerada. Em particular, se o problema requer uma teoria como resposta, o domínio constitui o corpo total de informações que, idealmente, devem ser levadas em conta por uma teoria que resolva esse problema. Já na visão de Smiraglia (P. SMIRAGLIA, 2014, p. 85), um domínio pode ser definido como um grupo que compartilha uma *ontologia*, realiza uma pesquisa ou um trabalho em comum e também se envolve em discurso ou comunicação, formal ou informalmente. Como exemplo de domínio pode-se citar Comunidade Budista, Psicologia, Hinduísmo e Arquitetura de Computadores.

As ontologias podem ser entendidas como um ramo da filosofia preocupado com o estudo do que existe. De uma perspectiva computacional, a formalização de um domínio do conhecimento por meio de ontologias tem possibilitado o desenvolvimento de algoritmos que suportam a geração de inferências a partir de dado conjunto de fatos sobre o mundo, sendo útil para a gestão do conhecimento, especialmente quando se lida com grandes quantidades de conhecimento. Supondo que se deseje modelar um conjunto de conhecimentos relacionados à assistência médica, então *paciente, doença, sintoma, diagnóstico e tratamento* podem estar entre os conceitos-chaves que descrevam um domínio. Esses conceitos e seus significados juntos definem uma ontologia para à assistência médica. (JURISICA *et al.*, 1999).

Santoso *et al.* (2015) trazem um exemplo de utilização prática de ontologias no ambiente computacional. Os autores utilizaram um extenso conjunto de 2000 artigos extraídos da Wikipedia sobre plantas e animais, organizando-os, e classificando-os em seus respectivos domínios. Em seguida, os autores aplicaram seu método de criação automática de ontologias para os domínios por meio de técnicas de mineração de textos. Por fim, utilizando as ontologias criadas, os autores executaram seu classificador construído no modelo Map Reduce com o algoritmo Naive Bayes para classificar os artigos em um dos domínios, obtendo uma precisão de aproximadamente 98.5%.

Na World Wide Web há disponível uma linguagem criada especialmente para se definir uma ontologia chamada de Ontology Web Language (OWL). Ela possibilita o registro de descrições de classes, propriedades e relacionamentos. A OWL foi projetada para que fosse possível às aplicações processarem conteúdos informacionais com um vocabulário estruturado e uma semântica formal. Esta é uma tecnologia de grande relevância nos estudos de Web semântica. Atualmente, é possível criar e editar ontologias no formato OWL por meio de softwares de edição

como Protege, OntoEdit e ONTOLIS.

Com o suporte conceitual oferecido pela análise de domínio bem como pelas ontologias formais é possível modelar o conhecimento sobre os domínios ou setores tecnológicos em termo de tecnologias-chaves e suas categorias a fim de facilitar sua identificação e classificação em publicações presentes nas fontes monitoradas.

2.4 AGENTES DE SOFTWARE

Parte da enorme quantidade de informação disponível na internet pode ser acessada via navegadores web, como o Google Chrome. Contudo pode ser necessário consumir e analisar dados e recursos digitais que dispersos em diferentes locais. Essa foi uma das necessidades que motivou a criação conceito de agentes de softwares. Os agentes são programas que podem operar de forma autônoma executando tarefas sem a interferência direta de um ser humano. No contexto deste trabalho os agentes de software são utilizados para capturar dados de portais web e redes sociais.

Existem dois tipos básicos de agentes: *agentes* e *agentes inteligentes*. Os agentes não-inteligentes ou simplesmente agentes, são programados para trabalhar de forma específica trazendo algum tipo de informação ao seus usuários. Por sua vez, além de coletar informações, os agentes inteligentes costumam utilizar inteligência artificial para executar e aprimorar suas tarefas. Eles têm maior liberdade e autonomia em sua execução. Agentes inteligentes podem ser treinados para procurar notícias, poderiam, por exemplo, adaptar-se a trazer notícias de acordo com as preferências pessoais de seus usuários (HEATON, 2002). Entender o conceito de agente é importante para compreender os dois próximos conceitos: *web crawlers* e *data scraping* presentes na construção de coletores de dados da web.

2.4.1 Data Scraping

Data scraping (em português raspagem de dados) é uma técnica computacional utilizada para extrair de páginas web dados legíveis para os seres humanos, serviços ou aplicativos, sendo o resultado final gerado em formato JSON, CSV e XML. Alguns, ainda, guardam a saída diretamente em bancos de dados.

O que diferencia o data scraping de uma atividade de "parsing" comum é que o primeiro coleta e extrai o dado para ser consumido por um usuário humano, descartando dados binários ou marcações não úteis à compreensão da informação. Os dois tipos mais comuns de *data scraping* são *screen scraping* e *web scraping*.

Um software de **web scraping** navega pelos elementos pelos quais as páginas web são construídas como marcações HTML e XHTML para extrair os elementos textuais relevantes aos usuários finais. As técnicas de extração podem envolver processamento em linguagem natural, visão computacional e técnicas como variação de IPs [12] para não ser bloqueados pelos servidores

[12] Endereço de Protocolo da Internet (Endereço IP) é um identificador numérico atribuído a dispositivos como

de hospedagem dos sites-alvos.

O **screen scraping** é um tipo de técnica que tenta extrair informações de maneira visual. É comumente usado para extrair textos de imagens quando por exemplo se deseja quebrar um "captcha" ou entender elementos de imagens em uma publicação noticiosa. É comum que softwares de *screen scraping* façam uso de visão computacional, inteligência artificial e processamento de imagens sendo software complexos e sofisticados por natureza.

2.4.2 Web Crawlers

Os rastreadores de rede, em inglês "Web Crawler", por sua vez, são softwares que navegam pela internet para capturar informações de forma automatizada. Diferente das técnica de *data scraping* os *web crawlers* tem maior liberdade para navegar pela web em busca de informações relevantes. Estas ferramentas são utilizadas para indexação de páginas web, como o buscador da Google, para comparação de preços entre diversos sites de comércio eletrônico e atualização de valores da bolsa por exemplo. Os Web Crawlers são também conhecidos como *web spiders, web robot ou bots*.

Em geral, os rastreadores possuem uma lista de endereço de partida conhecida como *seeds* a partir da qual iniciam o mapeamento e extração das informações de interesse. Os Web Crawlers mais avançados utilizam recursos de inteligência artificial para identificar os conteúdos, páginas alvos e a frequência com que devem fazer as coletas.

Em suma, a construção de ferramentas de "data scraping" e "web crawling" tem o propósito de otimizar as atividades de coleta de informação. Estes agentes automatizam essas atividades e entregam os conteúdos minimamente organizados. Na vigilância tecnológica, as informações coletadas estão em sua maioria disponíveis em formato de textos, principalmente no monitoramento de notícias, patentes e redes sociais. Dessa forma, para que seja possível extrair valor, são necessários métodos adequados para minerar o conteúdo e identificar os termos de interesse. Outra necessidade comum é a de se verificar similaridades entre documentos, extrair suas características ou detectar tópicos. Na próxima seção, a mineração de textos, atividade fundamental para se conseguir os resultados comentados, é tratada com mais clareza.

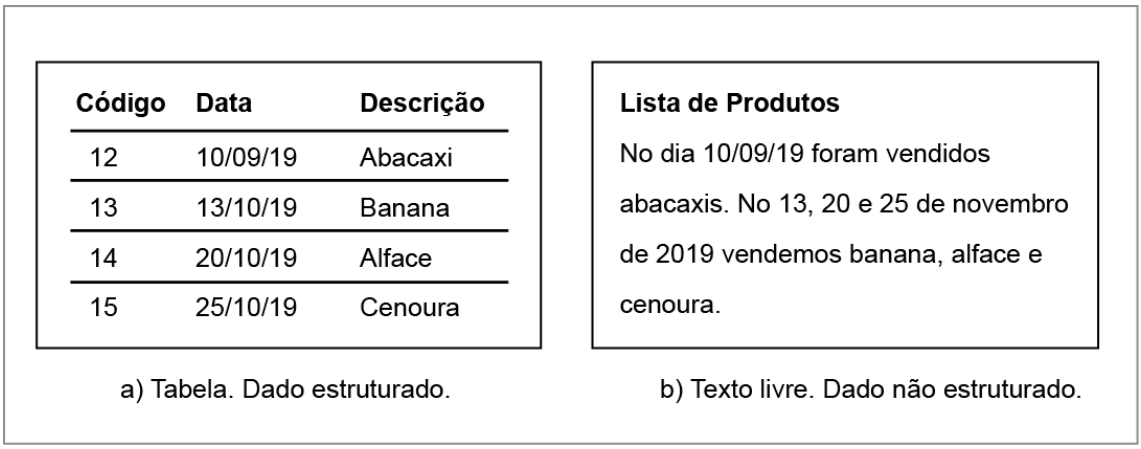
2.5 MINERAÇÃO DE TEXTOS

Há pouco mais de duas décadas, os computadores eram desenvolvidos com foco em ambientes corporativos. Com o passar do tempo, porém, eles tomaram os lares e passaram a assumir diferentes funções que vão desde o uso para o trabalho até o entretenimento, e formas como notebooks, tablets e smartphones. Houve, assim, uma revolução no setor que se intensificou com a chegada da internet, impulsionando ainda mais seu uso.

computadores, smartphones e até impressoras conectados a uma rede de computadores que utiliza o Protocolo de Internet para comunicação(WIKIPEDIA, 2019a). As técnicas de variação de endereços IP buscam contornar o problema de ser bloqueado pelos servidores das páginas webs ou recursos que sofrem "scraping" ao serem identificados como tentativa de ataque ou com oum tráfego não usual.

Os softwares corporativos costumam ter seus dados organizados em tabelas de banco de dados, onde cada coluna mantém tipos de dados específicos, como números ou datas. Este modelo confere às informações um aspecto estruturado e quais são facilmente processadas e analisadas por computadores. Com a utilização massiva das redes sociais, dos registros médicos digitais e dos portais de notícias, houve um crescimento dramático na produção de dados não estruturados, principalmente textos, por parte dos usuários. São dados nos quais não é possível perceber uma estrutura padrão para cada entrada de dados, como ilustrado na Figura 4. Mas, como afirmam Allahyari *et al.* (2017), este volume de textos gerado é uma fonte inestimável de informações e conhecimentos.

Figura 4 – Dados estruturados e não estruturados.



Fonte: Elaborado pelo autor.

Diferentemente de dados numéricos ou categóricos, os quais tendem a ter formatos claros e uma grande variedade de algorítimos e ferramentas para manipulação, os dados textuais não possuem uma estrutura clara. Eles podem serem vistos essencialmente como vetores de palavras em que a posição, o significado e a relação entre os termos podem ser relevantes dependendo do objetivo da análise.

Dessa forma, são caracterizados por serem esparsos e terem alta dimensionalidade o que prejudica, inclusive, sua análise através de bancos de dados relacionais tradicionais (AGGARWAL; ZHAI, 2012b). Estas particularidades motivaram a criação de algoritmos específicos para lidar com eles, fazendo com que a mineração de textos, também conhecida por *Text Mining*, passasse a receber cada vez mais atenção nos últimos anos (ALLAHYARI *et al.*, 2017). Nas próximas seções os principais técnicas utilizadas em *Text Mining* serão discutidas.

2.5.1 Pré-processamento de textos

O pré-processamento é uma das principais tarefas em algoritmos de mineração de texto. A estrutura de um algorítimo de categorização tradicional compreende, por exemplo, as etapas de pré-processamento, extração de características, seleção de características e classificação

(ALLAHYARI *et al.*, 2017). Entre as atividades de um pré-processamento de textos estão normalmente as seguintes:

- **Tokenização:** do inglês *tokenization*, é a tarefa de quebrar os textos, aqui considerados sequências de caracteres, em palavras ou frases (tokens). A lista de *tokens* resultantes serve de entrada para os próximos estágios de processamento;
- **Filtragem:** a filtragem é aplicada em documentos para remover informações irrelevantes para as análises. Um tipo de filtro comum é aquele utilizado para remover *stop-words*, que são palavras comuns nos textos mas que não agregam valor à informação. Exemplos de *stop-words* são as preposições e conjunções;
- **Lematização:** a lematizacão tenta fazer uma análise morfológica nas palavras, agrupando-as para que o texto possa ser analisado ignorando-se o tempo verbal, caso seja um verbo, o gênero, o plural ou singular das palavras. No geral, a lematização mapeia os tempos verbais para o infinitivo e os substantivos para o singular;
- **Stemização:** o primeiro algoritmo de *stemming* (termo em inglês) foi apresentado por Lovins (1968). Esta classe de algoritmo visa extrair a raiz das palavras, sendo muito dependente do idioma. Ao aplicar métodos de *stemming* no conjunto de palavras [fazendo, fazer, fazia] o resultado seria [faz, faz, faz].

Como benefícios obtidos por um pré-processamento estão a redução de dimensionalidade do conjunto de dados, entregando um conjunto de dados limpos e relevantes para as próximas etapas, a normalização dos dados, permitindo que sejam analisados em uma mesma escala e a transformação do dado de um domínio para outro, como a transformação de palavras para identificadores numéricos.

É comum durante o pré-processamento que os textos sejam codificados para se buscar uma simplificação e facilitar seu processamento. Um modelo muito comum é o *Bag-of-words* (BOW), em que o texto é simplesmente quebrado em um vetor de palavras. Este vetor serve de entrada para diversos algoritmos como o Latent Dirichlet Allocation (LDA) (BLEI *et al.*, 2003) e SVM (ALLAHYARI *et al.*, 2017), detalhados mais adiante. Ele é usado como entrada em algoritmos de busca por similaridade como o Termo–inverso da Frequência nos Documentos (TF-IDF), o qual é um tipo de representação do documento. O termo–inverso da frequência nos documentos (TF-IDF) é uma medida estatística que visa mensurar a relevância de uma palavra presente em um documento em relação a uma coleção de documentos. O número de ocorrências de uma palavra em um documento causa um aumento proporcional no em seu TF-IDF, contudo esse aumento é normalizado pela frequência da palavra na coleção. Na seção 2.5.3 sobre redução de dimensionalidade e modelagem de tópicos algumas aplicações e transformações como *Bag-of-words* serão discutidas.

2.5.2 Clusterização

A clusterização pode ser definida como a tarefa de se encontrar grupos de dados similares dentro de um determinado conjunto. A semelhança entre os dados é medida por meio de uma função de cálculo de similaridade a qual leva em contra, entre outras coisas, suas características. Em um conjunto de pontos cartesianos, por exemplo, para medir a similaridade de pontos poderia ser utilizada uma função que calcule a distância euclidiana entre os mesmos.

A mineração de textos se beneficia desta classe de algoritmos pois ela agrupa os textos em diferentes granularidades como frases, parágrafos ou textos inteiros e ao dividir os textos de acordo com suas características, o processo de recuperação e navegação pelos dados categorizados torna-se mais prático (AGGARWAL; ZHAI, 2012a). Segundo Shah e Mahajan (2012), as principais aplicações da clusterização em documentos são:

- Localizando documentos semelhantes: o usuário pode recuperar documentos semelhantes ao documento de interesse. Este tipo de pesquisa difere de uma pesquisa simples de palavras em documentos a qual traria apenas uma lista de documentos contendo a palavra pesquisada ao invés de trazer documentos conceitualmente semelhantes;
- Organização de grandes coleções de documentos: a clusterização agrupa os documentos conceitualmente, categorizando-os em taxonomias similares às que um ser humano criaria;
- Detecção de conteúdo duplicado: a clusterização é empregada para detecção de plágio e agrupamento de patentes;
- Sistema de recomendação: ao se armazenar artigos que já tenham sido lidos por um usuário, é possível traçar seu perfil e oferecer novos conteúdos aderentes as suas preferencias;
- Otimização de pesquisa: o tempo de pesquisa de documentos pode ser reduzido ao se comparar a consulta com os agrupamentos ao invés de se comparar diretamente com os documentos.

McCowan *et al.* (2006) utilizaram clusterização de textos para classificar o estágio de câncer de pulmão de pacientes com base na análise de seus relatórios médicos. Eles treinaram um modelo baseado no algoritmo *Support Vector Machines* (máquina de vetores de suporte) para cada estágio com base nas ocorrências de palavras dos relatórios histológicos dos pacientes com as patologias. Também utilizaram processamento de linguagem natural para transformar o texto do relatório. Uma visão mais ampla sobre clusterização de textos pode ser encontrada em Berry e Castellanos (2004) e Allahyari *et al.* (2017).

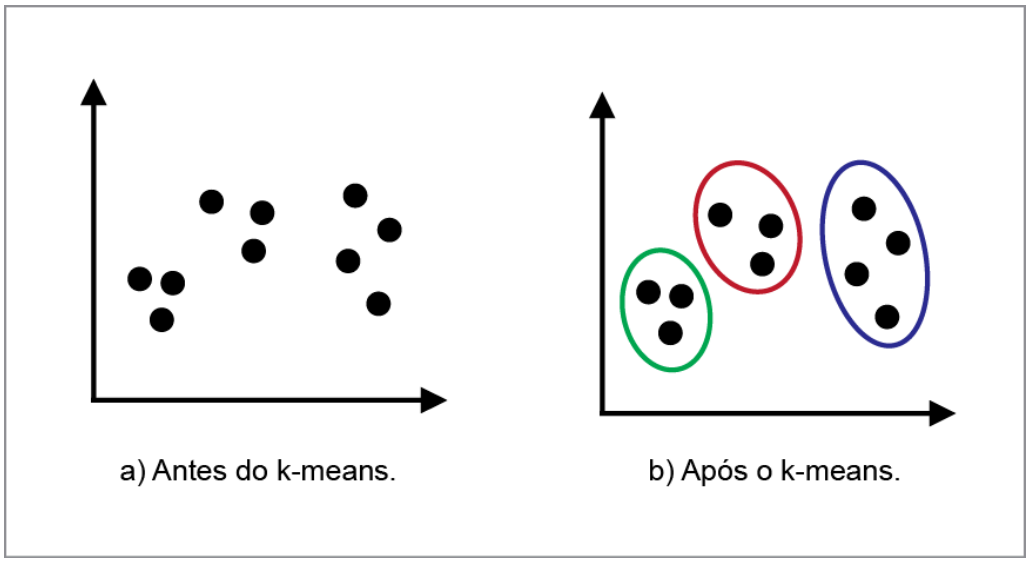
2.5.2.1 Tipos e classificações dos métodos de clusterização

De acordo com o nível de restrição durante a criação dos *clusters* (agrupamentos), os métodos de clusterização são classificados como rígidos, quando cada documento pode pertencer

a apenas um grupo, ou flexível, quando um documento pode pertencer a mais de um grupo (SHAH; MAHAJAN, 2012). Os algoritmos de clusterização, por sua parte, podem ser classificados de maneira geral como hierárquicos e de partição (SHAH; MAHAJAN, 2012), (NAGPAL *et al.*, 2013). Os algoritmos hierárquicos geram grupos de *clusters* como uma hierarquia construída de maneira descendente (divisiva) ou ascendente (aglomerativa). Estes algoritmos são baseados em distância, aplicando-se uma função de similaridade para medir a proximidade entre documentos (ALLAHYARI *et al.*, 2017). É possível representar os resultados da clusterização hierárquica no formato de dendrogramas, cujos nós ou folhas, representam o subconjunto de uma coleção de documentos. Os algoritmos HAC, do inglês Hierarchical Agglomerative Clustering (YAARI, 1997) e BIRCH, do inglês Balanced Iterative Reducing and Clustering using Hierarchies (ZHANG *et al.*, 1996) pertencem a esta categoria.

Os algoritmos de partição tentam encontrar todos os *clusters* simultaneamente por meio do particionamento dos dados, gerando uma ou mais partições. Eles alocam os documentos em um número fixo de *clusters* não vazios. Os método de particionamento mais conhecido é o **K-means**. O funcionalmento K-means é simples. Inicialmente ele aloca os dados em vários *clusters* de forma aleatória.

Figura 5 – Exemplo de aplicação do K-means a um conjunto de dados.



Fonte: elaborado pelo autor.

A cada iteração ele calcula a média de cada *clusters* e redistribui os dados de acordo com a média mais próxima a dele. As iterações chegam ao fim quando não houver alterações nos *clusters*. Por exemplo, a figura 5 ilustra de forma simplificada o que aconteceria se aplicarmos o k-means a um conjunto de dados definindo o número de *clusters* desejados como três. É possível perceber que após sua aplicação os dados foram divididos em três grupos com base na distância das médias dos dados à média do grupo.

2.5.3 Redução de dimensionalidade e modelagem de tópicos

Embora abundantes e muito comuns nos sistemas atuais, os dados textuais ainda oferecerem importantes desafios à maioria dos algoritmos de *text mining*. Um mesmo conceito pode ser expresso por diferentes termos (sinonímia) e, inversamente, um mesmo termo pode ter significados muito diferentes em contextos diferentes (polissemia). A palavra "Apple" que pode significar uma fruta ou o nome de uma empresa. Estas características contribuem com a alta dimensionalidade dos dados do tipo texto o que prejudica o desempenho da maioria dos algoritmos de *text mining*, notadamente os de clusterização (PHIL; COLLEGE, 2011), tornando seu processamento custoso e pouco eficaz.

Para ter sucesso, a mineração de textos automatizada precisa compreender o contexto e identificar maneiras diferentes de expressar o mesmo conceito para ser capaz desambiguar termos polissêmicos (AGGARWAL; ZHAI, 2012a, p. 130). Se isso não for feito, os algoritmos estarão sendo sobrecarregados com informações irrelevantes ou fora do contexto. Um método muito comum para codificar textos é o *bag-of-words*, todavia o produto de sua aplicação apresenta alta dimensionalidade porque cada dimensão corresponde a um termo da linguagem.

O desafio da redução de dimensionalidade é transformar dados de alta dimensionalidade em uma representação de dimensão reduzida, mas significativa e relevante. No contexto dos dados textuais, a aplicação de métodos de pré-processamento traz contribuições diretas, principalmente as técnicas para redução de sinônimos, filtragem de *stop-words* e stemização. Conforme Van Der Maaten *et al.* (2009) a nova representação deve corresponder à dimensionalidade intrínseca dos dados, que é a quantidade mínima de características, ou *features* como é mais conhecida no meio da ciência de dados, necessárias para explicar as propriedades observadas originalmente. Como resultados positivos da redução da dimensionalidade, continua o autor, estão, entre outros, a facilitação da classificação, da visualização e da compactação de dados de alta dimensão.

Os algoritmos de redução de dimensionalidade podem ser divididos em duas categorias: de extração e de seleção de características. Na extração de características, novas características são combinadas a partir das originais por meio de transformações. De acordo com Shah e Mahajan (2012) embora eficazes, esses algoritmos apresentam grande sobrecarga computacional, o que dificulta sua aplicação em dados textuais.

Na seleção de características, selecionam-se características diretamente. Devido a sua eficiência, esta classe de algoritmos é muito utilizada. Para contornar as limitações de ambas as técnicas, Yan *et al.* (2011) propõe um método combinando ambas chamado de TOFA, que pode encontrar os subconjuntos de *features* ideais de acordo com uma função objetivo (função que se deseja maximizar ou minimizar de acordo com o objetivo da otimização).

Mesmo após a simplificação feita por um pré-processamento eficiente, pode ser ideal trabalhar com um espaço semântico ainda menor, mapeando cada dimensão para um tópico. Dessa forma, a redução da dimensionalidade pode ser aplicada para encontrar um espaço semântico adequado em relação a sua representação BOW (AGGARWAL; ZHAI, 2012b, p. 131). Ao final, a nova representação em tópicos pode ser até mais clara do que a representação original.

De acordo com Alghamdi e Alfalqi (2015), a modelagem em tópicos tem como principal objetivo a descoberta de padrões no uso de palavras bem como conectar documentos que compartilham os mesmos padrões. Assim, a ideia por trás da modelagem é que os documentos são uma mistura de tópicos, nos quais um tópico é visto como uma distribuição de probabilidade de palavras, ou seja, um tópico é um modelo generativo para documentos.

Topic modeling ou modelagem de tópicos são algoritmos de agrupamento probabilístico populares. O objetivo principal da modelagem de tópicos é criar um modelo generativo probabilístico para um corpus de documentos de texto. Na modelagem de tópicos, os documentos são uma conjunto de tópicos, em que um tópico é uma distribuição de probabilidade por palavras (ALLAHYARI *et al.*, 2017).

Um dos algoritmos mais populares para modelagem de tópicos para corporas é o **Latent Dirichlet Allocation** (LDA), ou modelo de alocação de Dirichlet latente, um modelo probabilístico generativo que utiliza técnica não supervisionada (BLEI *et al.*, 2003). O LDA busca fornecer um modelo para a distribuição de entradas e saídas baseado nas variáveis latentes e tenta descrever como os documentos foram gerados a partir de um conjunto de tópicos distribuídos obedecendo a distribuição de Dirichlet. Neste modelo generativo, é utilizado o resultado da amostragem de Dirichlet para alocar as palavras dos diferentes tópicos que formarão os documentos. No LDA o texto é codificado no modelo *bag-of-words*, no qual a ordem das palavras não é relevante. Assim, os documentos são gerados palavra por palavra. Outro ponto importante, é que os parâmetros das distribuições, ou hiper-parâmetros, são passados a priori no modelo. Para um entendimento mais aprofundado sobre este algoritmo consulte Blei *et al.* (2003).

2.5.4 Classificação

A classificação de textos tem o objetivo de atribuir a probabilidade de um texto pertencer a uma ou mais classes. Esta é uma técnica muito popular utilizada em problemas reais. Como exemplos de aplicações tem-se classificadores criados para tentar dizer se um e-mail é ou não um *Spam* e verificar sobre qual assunto uma notícia pode estar falando (esporte, política, etc).

O *Support Vector Machine* (SVM) é um algoritmo popular em problemas de classificação de textos. Este é um algoritmo de aprendizado supervisionado de classificação linear cujos modelos após treinados tomam decisões baseados nas combinações lineares das características dos documentos. Um SVM simples busca separar o conjunto de características em duas classes, traçando um hiperplano entre elas com a maior distância possível. Tal divisão também é chamada de vetores de suporte (ALLAHYARI *et al.*, 2017).

2.6 CONSIDERAÇÕES FINAIS DO CAPÍTULO

Este capítulo trouxe fundamentação teórica por meio da revisão bibliográfica de temas relacionados a esta dissertação. O conceitos discutidos servem para embasar o modelo proposto.

O capítulo inicia-se discorrendo sobre o assunto principal deste trabalho, a vigilância tecnológica. Traz as principais definições, tipos e categorias de ferramentas especializadas. Em seguida, foram analisadas e discutidas oito plataformas web para vigilância tecnológica e se enumeraram as principais funcionalidades e características comuns entre elas, o que auxiliou no processo de abstração do modelo. As plataformas analisadas foram a SoftVT, Vicubo Cloud, Vigiale, Innguma, MUSSOL, Minera, Intelligent Watcher e a Hontza. A seleção das plataformas deu-se por meio de um levantamento feito em trabalhos e sites de referência.

Na sequência, foi dissertado sobre cenários de Big Data em que se trouxe um resgate histórico acerca do surgimento e consolidação do termo. Falou-se, ainda, sobre as projeções que preveem que estes cenários se tornem cada vez mais comuns. Em 2025, por exemplo, é projetado que tenha sido gerados pelo menos 175 zettabytes de dados no mundo dos quais 30% necessitem de processamento em tempo real. Ademais, foram enumerados exemplos de cenários de Big Data nas áreas da indústria, administração pública e varejo. Esta seção trouxe à luz um importante desafio que os sistemas de vigilância tecnológica automatizados precisam enfrentar, uma vez que o monitoramento e o processamento de diferentes fontes e formatos de dados em curto espaço de tempo, faz-se necessários para gerar resultados rápidos aos tomadores de decisão nas organizações.

Em seguida, foi abordada a análise de domínio, uma importante ferramenta para a delimitação do conhecimento sobre determinado escopo tecnológico que se deseje monitorar. Também foram discutidos o conceito de ontologias e como elas podem representar um conjunto de tecnologias e suas relações.

Mais adiante, tratou-se dos "agentes de softwares", que podem ser construídos no formato de robôs especializados para capturar informações automaticamente em fontes digitais. No contexto deste trabalho, os agentes de softwares para Web Crawling e Data Scrapping são peças fundamentais porque contribuem na automatização de sistemas de vigilância, atuando diretamente na busca e no monitoramento de informações relevantes.

Por último, foram discutidos conceitos e técnicas relacionadas à *mineração de textos*, cuja aplicação vem crescendo nos últimos anos devido ao aumento de dados não estruturados, principalmente no formato de publicações de redes sociais. A mineração de texto traz um importante conjunto de ferramentas para os sistemas de vigilância que precisam analisar documentos como patentes, artigos científicos e notícias de forma automatizada. Por meio do pré-processamento é possível eliminar textos duplicados, filtrar dados relevantes, identificar tecnologias e até classificar ou agrupar documentos, facilitando a compreensão e análise de seus conteúdos.

3 ASPECTOS METODOLÓGICOS

Neste capítulo são abordados os aspectos metodológicos desta pesquisa. São feitas sua caracterização e explicados os procedimentos adotados para o seu desenvolvimento.

3.1 CARACTERIZAÇÃO DA PESQUISA

A realização de uma pesquisa científica tem um sucesso maior quando se delimitam os critérios a serem adotados antes de sua execução. Entender estes critérios contribui para sua classificação e aderência aos objetivos.

A caracterização quanto à abordagem, natureza, objetivos e procedimentos é feita nesta seção. Os pontos estão resumidos na Quadro 3 e detalhados a seguir.

1. **Quanto aos Objetivos:** com o propósito de compreender as relações entre causa e efeitos do objeto de estudo esta pesquisa se caracteriza como explicativa e, de forma estruturada, baseia-se em métodos experimentais para atingir seu objetivo;

2. **Quanto à Natureza:** busca-se com esta pesquisa obter um modelo para vigilância tecnológica automatizada para cenários de Big Data. Deste modo, ao produzir conhecimentos para resolver um problema específico, sua natureza é aplicada;

3. **Quanto à Abordagem:** ao tentar desenvolver e validar um modelo junto à especialistas esta pesquisa faz uso de abordagens qualitativas (FLICK, 2009). Também serão aplicadas abordagens quantitativas durante os experimentos. A combinação das abordagens permite obter resultados mais consistentes do que utilizar elas de forma isolada (FONSECA, 2002);

4. **Quanto aos Procedimentos:** ao longo deste trabalho são utilizadas pesquisas bibliográficas, principalmente na fundamentação teórica onde são estruturadas as bases para o restante do trabalho além de proporcionar novos conhecimentos para o autor. Este tipo pesquisa se caracteriza por explicar um problema a partir de referências publicadas em documentos como livros e artigos científicos, sem a necessidade de elaboração de hipóteses pelo autor. Devido à natureza tecnológica e aplicada deste trabalho, ele usa procedimentos experimentais por meio dos quais, segundo Gil (2002), deve-se apontar um objeto de estudo, selecionar as variáveis que o influenciam e se definir formas de controle e de observação dos efeitos que tais variáveis produzem no objeto.

3.2 PROCEDIMENTOS METODOLÓGICOS

Foram utilizados como procedimentos metodológicos na construção da abordagem proposta uma revisão da literatura pela qual se levantou o estado da arte atual na área de vigilância tecnológica em bases conceituadas para se compreender os procedimentos e métodos

Quadro 3 – Aspectos metodológicos da pesquisa.

Critérios de caracterização metodológica	**Características desta Pesquisa**
Objetivo	Experimental
Natureza	Aplicada
Abordagem	Qualitativa e quantitativa
Procedimentos	Bibliográficos e experimentais
Instrumentos	Estudo de caso

Fonte: elaborado pelo autor.

utilizados pelos autores. Analisaram-se, ainda, oito plataformas integrais web para VT das quais se extraíram e se compararam suas principais funcionalidades a fim de traçar um paralelo sobre quais componentes já consolidados no mercado deveriam ser levados ao modelo. Também, estudou-se o *Big Data Analytics*, conhecido como o processo de se analisar informações em cenários de Big Data. Adicionalmente, houveram entrevistas abertas com especialistas para entender seu processo de trabalho. Como produtos dessas atividades, generalizou-se o modelo conceitual buscado.

Para construir uma visão aplicada do modelo foi desenvolvida uma arquitetura computacional onde seus componentes se alinharam aos do modelo, sendo projetados cada qual com suas funções e responsabilidades. Com base na arquitetura foi desenvolvido um protótipo e avaliado por meio de um estudo de caso real em uma organização, sendo possível mensurar a efetividade do modelo. Assim, as etapas necessárias para o desenvolvimento deste trabalho podem ser enumeradas como se segue:

1. Revisão bibliográfica para fundamentar este trabalho, incluindo temas como vigilância tecnológica, Big Data, análise de domínio, mineração de textos e agentes de software;
2. Execução de uma revisão sistemática da literatura sobre vigilância tecnológica para verificar o seu estado da arte;
3. Análise dos principais softwares comerciais para vigilância e tecnológica;
4. Proposta de um modelo para sustentar um sistema de vigilância automatizada em portais web e redes sociais;
5. Desenvolvimento de um estudo de caso;
6. Validação do modelo proposto.

3.3 REVISÃO SISTEMÁTICA DA LITERATURA

Este trabalho visa a concepção de um modelo para a realização de vigilância tecnológica automatizada em fontes disponíveis eletronicamente como artigos de portais web ou rede sociais. Logo, compreender o atual estado da arte e as diferentes abordagens utilizadas na resolução

de problemas similares são fundamentais para reaproveitar métodos, conceitos e buscar novas contribuições ao tema.

Pelo caráter científico, esta pesquisa prima pela consulta às fontes formais de informação, sendo a revisão da literatura a opção mais adequada para compreender o estado da arte. Os dois principais métodos de revisões utilizados atualmente são os integrativos e sistemáticos. As revisões integrativas buscam agrupar os resultados obtidos em pesquisas sobre um tema de forma ampla, porém sistemática e ordenadamente. Elas permitem uma variedade de composição, combinando dados de literatura teórica e empírica, tendo assim um caráter mais generalista.

Por outro lado, a revisão sistemática, diferentemente da revisão integrativa, tem um caráter mais especializado, sendo utilizada para responder questões de um problema específico. Para Keele (2007) a revisão sistemática da literatura é um meio de identificar, avaliar e interpretar toda a pesquisa disponível relevante para uma questão de pesquisa específica, área de tópico ou fenômeno de interesse. Por ter seu problema de pesquisa e objetivos claros, este trabalho optou por utilizar este segundo tipo de revisão por estar mais aderente às suas necessidades informacionais.

Dessa forma, realizou-se uma Revisão Sistemática da Literatura (RSL) entre os meses de janeiro e fevereiro do ano de 2020, pela qual se buscou traçar um panorama sobre as publicações feitas entre o ano de 2013 a 2020 com o objetivo de responder a questão de pesquisa sobre quais métodos foram propostos no sentido de conceber ou apoiar um sistema de vigilância tecnológica automatizado para monitoramento de tecnologias de interesse em fontes eletronicamente disponíveis.

A RSL deste trabalho foi elaborada com base na metodologia proposta por Kitchenham (2004), com sólida aceitação no meio acadêmico e cuja aplicação foi novamente demonstrada em conjunto com outros autores posteriormente (KITCHENHAM *et al.*, 2009). A condução da RSL apoiou-se no método proposto por Galvão e Pereira (2014) pelo qual a construção de revisões sistemáticas deve prever a elaboração de uma pergunta de pesquisa delimitando o escopo do trabalho, a busca na literatura em bases de qualidade, a seleção de artigos nas bases, a extração de dados dos artigos, a avaliação da qualidade das evidências informacionais encontradas nos trabalhos e por fim a divulgação dos resultados obtidos.

A escolha das bases a serem utilizadas nesta pesquisa foi feita com base em três critérios: (1) maturidade, levando-se em conta há quanto tempo ela existe e a qualidade de sua tecnologia de busca, (2) o reconhecimento no meio científico e acadêmico e (3) o volume de trabalhos científicos existentes para a área pesquisada. Foram escolhidas cinco bases de dados para compor esta revisão sistemática:

- **SCOPUS**: uma base multidisciplinar de resumos e citações de artigos de mais de 5.000 editoras internacionais que cobre periódicos desde 1960;
- **Library and Information Science Abstracts (LISA)**: uma ferramenta internacional de resumos e citações projetada para profissionais de informações que abstrai mais de 440

periódicos de mais de 68 países;

- **IEEE Xplore**: biblioteca digital publicada pela IEEE (Institute of Electrical and Electronics Engineers) que viabiliza o acesso a mais de cinco milhões de documentos em texto completo, dentre as quais estão as publicações mais citadas do mundo em áreas como engenharia elétrica, ciência da computação e eletrônica;
- **ACM Digital Library**: é uma das bases mais abrangentes do mundo que engloba literatura completa e bibliográfica das áreas de computação e tecnologia da informação. É mantida pela ACM (Association for Computing Machinery), instituição fundada em 1947 como a primeira sociedade científica e educacional voltada à computação. A base contém quase 3 milhões de publicações e mais de 16 milhões de citações, incluindo artigos datados desde de 1936;
- **Web of Science**: o portal Web of Science (WOS) foi criado pelo Institute for Scientific Information e atualmente é mantida pela Clarivate Analytics. É uma base que oferece acesso a aproximadamente 12 milhões de periódicos das mais diversas áreas de conhecimentos, contando ainda com ferramentas para análise de citações e análises bibliométricas.

Nas bases de dados supracitadas, procedeu-se as buscas de trabalhos relevantes para esta pesquisa em conteúdos de *journals*, periódicos acadêmicos e artigos. Realizaram-se buscas utilizando principalmente as opções de pesquisa avançadas das ferramentas com os termos *technology watch*, *technology surveillance*, *technology monitoring*, *technological surveillance* e *vigilancia tecnologica* nos campos *título* e *abstract (resumo)*. Delimitou-se nas buscas o período de 2013 a 2020. A quantidade de trabalhos encontrados, os tipos de documentos pesquisados em cada base e as strings de buscas estão detalhadas no Quadro 4.

3.3.1 Critérios de inclusão e exclusão

Após a etapa de pesquisas nas cinco bases de dados, obtiveram-se **263 publicações**. As publicações duplicadas foram removidas com a ajuda do software Mendeley[1], restando ainda **197 publicações**. Sobre estas publicações resultantes foram lidos seus títulos e resumos e aplicados critérios de inclusão e exclusão com o propósito de identificar somente trabalhos com importante relevância para esta pesquisa. Os critérios de **exclusão** definidos e aplicados com base na pergunta de pesquisa foram os que seguem:

1. Artigos que não contenham um método para vigilância tecnológica definida;
2. Artigos que utilizem apenas dados não disponíveis eletronicamente;
3. Artigos que trabalhem apenas com fontes informais de informação;
4. Trabalhos sem propostas de arquitetura, protótipo ou de cunho apenas teórico.

[1] https://www.mendeley.com/

Quadro 4 – Aspectos metodológicos da pesquisa.

Base / Tipo	**String de Busca**	**Total**
IEEE — Journals e Magazines	("Publication Title": "technology watch"OR "Abstract": "technology watch"OR "Publication Title": "technology surveillance"OR "Abstract": "technology surveillance"OR "Publication Title": "technological surveillance"OR "Abstract": "technological surveillance"OR "Publication Title": "technology monitoring"OR "Abstract": "technology monitoring"OR "Publication Title": "vigilancia tecnologica"OR "Abstract": "vigilancia tecnologica")	18
ACM Digital Library — ACM Full-Text Collection	[Publication Title: "vigilancia tecnologica"] OR [Publication Title: "technology surveillance"] OR [Publication Title: "technological surveillance"] OR [Publication Title: "technology monitoring"] OR [Abstract: "vigilancia tecnologica"] OR [Abstract: "technology surveillance"] OR [Abstract: "technological surveillance"] OR [Abstract: "technology monitoring"] AND [Publication Date: (01/01/2013 TO 12/31/2020)]	3
SCOPUS — Articles	TITLE-ABS-KEY ("technology watch"OR "technology surveillance"OR "technological surveillance"OR "technology monitoring"OR "vigilancia tecnologica") AND PUBYEAR > 2013 AND (LIMIT-TO (SRCTYPE , "j")) AND (LIMIT-TO (DOCTYPE , "ar")) AND (LIMIT-TO (LANGUAGE , "English") OR LIMIT-TO (LANGUAGE, "Spanish") OR LIMIT-TO (LANGUAGE , "Portuguese"))	138
LISA(Proquest) — Periódicos acadêmicos	ab("technology watch") OR ti("technology watch") OR ab("technology surveillance") OR ti("technology surveillance") OR ti("technological surveillance") OR ab("technological surveillance") OR ti("technology monitoring") OR ab("technology monitoring") OR ti("vigilancia tecnologica") OR ab("vigilancia tecnologica")	7
Web of Science — Artigos	TITLE: ("vigilancia tecnologica"OR "technology watch"OR "technology surveillance"OR "technological surveillance"OR "technology monitoring") OR TOPIC: ("vigilancia tecnologica"OR "technology watch"OR "technology surveillance"OR "technological surveillance"OR "technology monitoring")	97
Total de artigos únicos encontrados		263

Fonte: elaborado pelo autor.

Os trabalhos que se enquadraram em pelo menos um dos quatro critérios de exclusão previamente definidos foram desconsiderados. Em seguida, sobre as publicações restantes foram aplicados os critérios de **inclusão** definidos abaixo.

1. Artigos que contenham um método, *framework*, metodologia, arquitetura ou modelo para Vigilância Tecnológica;

2. Artigos que contenham métodos parcial ou totalmente automatizados para monitoramento tecnológico em artigos, patentes, portais, notícias, redes sociais ou outros documentos eletronicamente disponíveis.

Após a aplicação dos critérios, obtiveram-se **20 publicações** cuja a análise foi realizada na etapa seguinte da revisão sistemática, envolvendo suas leituras integrais. No ANEXO I estão listadas todas as publicações avaliadas, assim como o(s) critério(s) de exclusão e inclusão aplicados em cada uma delas.

3.3.2 Análise dos trabalhos relacionados

A lista de trabalhos relacionados é fruto da aplicação dos critérios de exclusão e inclusão supracitados e está sintetizada no Quadro 5. Os trabalhos foram lidos em sua íntegra, avaliando-se técnicas e métodos propostos ou utilizados pelos autores para oferecer atividades de vigilância tecnológica com um significativo grau de automatização em suas etapas.

Quadro 5 – Resumo dos artigos analisados.

Nº	**Título do Artigo**	**Autor(es)**	**Base(s)**
1	Comparing data sources for identifying technology trends	Mikova e Sokolova (2019)	SCOPUS, WOS
2	Trends in Logistics in the Last Five Years - A Review Through Technological Surveillance	Moreno C e Díaz (2019)	IEEE
3	A case study on the use of machine learning techniques for supporting technology watch	A Perez *et al.* (2018)	SCOPUS, WOS
4	A dynamic forward-citation full path model for technology monitoring: An empirical study from shale gas industry	Wei *et al.* (2017)	SCOPUS, WOS
5	A proposal for a technological surveillance unit aimed at regional competitiveness	L G Perez *et al.* (2017)	SCOPUS
6	Monitoring Newly Adopted Technologies Using Keyword Based Analysis of Cited Patents	Nam e Kwangsoo Kim (2017)	SCOPUS, WOS
7	Patent bibliometrics and its use for technology watch	Jürgens e Herrero-Solana (2017)	SCOPUS, WOS

Quadro 5 – Resumo dos artigos analisados.

Nº	**Título do Artigo**	**Autor(es)**	**Base(s)**
8	Technological surveillance of s curves and cycle life of technology	Jiménez González *et al.* (2017)	SCOPUS
9	The empirical research on patent-based models of technology entropy: A case of carbon capture technology	Hou (2017)	SCOPUS
10	Technological surveillance and technology life cycle analysis: Usability assessment techniques, metrics and tools in the ICT sector	Tobón Clavijo *et al.* (2017)	SCOPUS
11	Technology surveillance and curves in 'S': Environmental technologies in Tourism, Quindio Innova project	Grajales López *et al.* (2017)	SCOPUS
12	Developing a Mobile Application for Technological Alerts	Marulanda Echeverry *et al.* (2016)	WOS
13	Generating patent development maps for technology monitoring using semantic patent-topic analysis	M Kim *et al.* (2016)	SCOPUS, WOS
14	Identification and monitoring of possible disruptive technologies by patent-development paths and topic modeling	Momeni e Rost (2016)	SCOPUS; WOS
15	Using a semantic wiki for technology forecast and technology monitoring	Färber (2016)	LISA; SCOPUS; WOS
16	Trends in 3-D printing from a patent information analysis (APA)	Henri e Clerc (2015)	SCOPUS
17	Estudo de caso utilizando mapeamento de prospecção tecnológica como principal ferramenta de busca científica	Dos Santos Amparo *et al.* (2014)	SCOPUS
18	Technology mapping as a tool for technology analysis in foresight studies	Gudanowska (2014)	IEEE
19	Patterns of technological innovation and evolution in the energy sector: A patent-based approach	Kyungpyo Lee e Sungjoo Lee (2013)	WOS
20	Identifying technological opportunities using the novelty detection technique: a case of laser technology in semiconductor manufacturing	Geum *et al.* (2013)	WOS

Fonte: elaborado pelo autor.

Trabalho: Comparing data sources for identifying technology trends (MIKOVA; SOKOLOVA, 2019)

O trabalho "Comparing data sources for identifying technology trend" (MIKOVA; SOKOLOVA, 2019) avalia estratégias para se identificar tendências tecnológicas a partir de fontes de dados como publicações científicas, patentes, mídia, projetos prospectivos, conferências, projetos internacionais, dissertações e apresentações. Os autores apresentam um quadro comparativo com as vantagens e desvantagens sobre cinco estratégias comuns de buscas de tecnologias em bases de dados. Para a área de inteligência competitiva e monitoramento tecnológico os autores apontam como positivas as estratégias de pesquisas que compilem um conjunto de documentos com base em uma consulta nominal, como artigos de periódicos especializados, ou que recuperem publicações mais citadas.

Um estudo de caso prático foi feito para a área de energia verde no qual foram realizadas pesquisas por tópicos e palavras-chave ("keywords") fornecidos por especialistas em 8 diferentes fontes de dados, o que resultou em um conjunto de 35.986 documentos com diferentes formatos de arquivos (* .txt, * .html, * .doc, * .pdf, * .ppt). Afim de padronizar os resultados dessa pesquisas, os dados dos documentos foram normalizados utilizando o formato XML. Dessa forma, os dados puderam ser processados por meio do software Vantage Point. Por meio dele, os metadados dos documentos foram clusterizados. Finalmente, os autores identificaram três tipos de clusters: os clusters diretamente relacionados à área, os adjacentes à área e outros, sendo os resultados submetidos à especialistas para uma filtragem final. Como principal contribuição do trabalho está o método de quatro estágios desenvolvido que pode guiar novos trabalhos que busquem identificar tendências tecnológicas, cujas etapas são: (1) Formação da estratégia de busca, (2) Coleta de dados para identificar tendências tecnológicas (3) Processamento dos dados, e (4) Compilação uma lista final de tendências.

Trabalho: Trends in Logistics in the Last Five Years - A Review Through Technological Surveillance (MORENO C; DÍAZ, 2019)

No trabalho "Trends in Logistics in the Last Five Years - A Review Through Technological Surveillance", Moreno C e Díaz (2019) realizaram um exercício de vigilância tecnológica para o setor de Logística utilizando o software Vantage Point version 11 para processar informações capturadas a partir de artigos filtrados na base Scopus. Como filtro foi utilizada uma equação genérica que resultou em mais de mil artigos. Em seguida foi aplicado um novo filtro que reduziu o número de artigos, melhorando a relevância dos documentos resultantes. Os autores executaram sua vigilância em 4 etapas: (1) Definição da equação de pesquisa genérica, (2) Validação de palavras-chave, (3) Definição de equação de pesquisa específica e (4) Processamento e análise de software. O artigo não traz contribuições significativas para este trabalho por não apresentar um protótipo, arquitetura ou método automatizado.

Trabalho: A case study on the use of machine learning techniques for supporting technology watch; de (PEREZ, A. *et al.*, 2018)

O trabalho de A Perez *et al.* (2018) apresenta um estudo de caso conduzido na empresa Koniker S.Coop que utilizou técnicas de aprendizagem de máquina para agilizar processos de vigilância tecnológica, com ênfase na etapa de classificação de documentos. Segundo o trabalho, os analistas de vigilância da empresa precisam ler muitos documentos para categorizá-los e enviá-los manualmente à especialistas que utilizam esses documentos selecionados como matéria prima em suas análises. Segundo estimativas feitas pela Koniker, apenas 35% do tempo dos especialistas é gasto com contribuições de valor agregado ao processo de vigilância, sendo que 65% do tempo é passível de automatização. Com esta problemática em vista, os autores focaram na redução do tempo gasto pelo primeiro grupo de analistas e desenvolveram técnicas para filtrar e classificar as informações do sistema de vigilância tecnológica automaticamente.

No estudo de caso, foram utilizados 7379 documentos entre patentes, notícias, boletins oficiais do governo e documentação de competência, coletados em processos de vigilâncias anteriores, todos contendo título e textos. Os textos foram filtrados por meio de técnicas de tokenizing, stemming e remoção de "stop words". Eles foram classificados utilizando três algoritmos de classificação largamente adotados na indústria: Support Vector Machine (SVM), Decision tree (J48) e Naive Bayes. A qualidade dos resultados foi avaliada por dois experimentos, onde no primeiro usaram-se apenas os textos e no segundo os textos foram enriquecidos com anotações semânticas a partir da DBpedia Spotlight API.

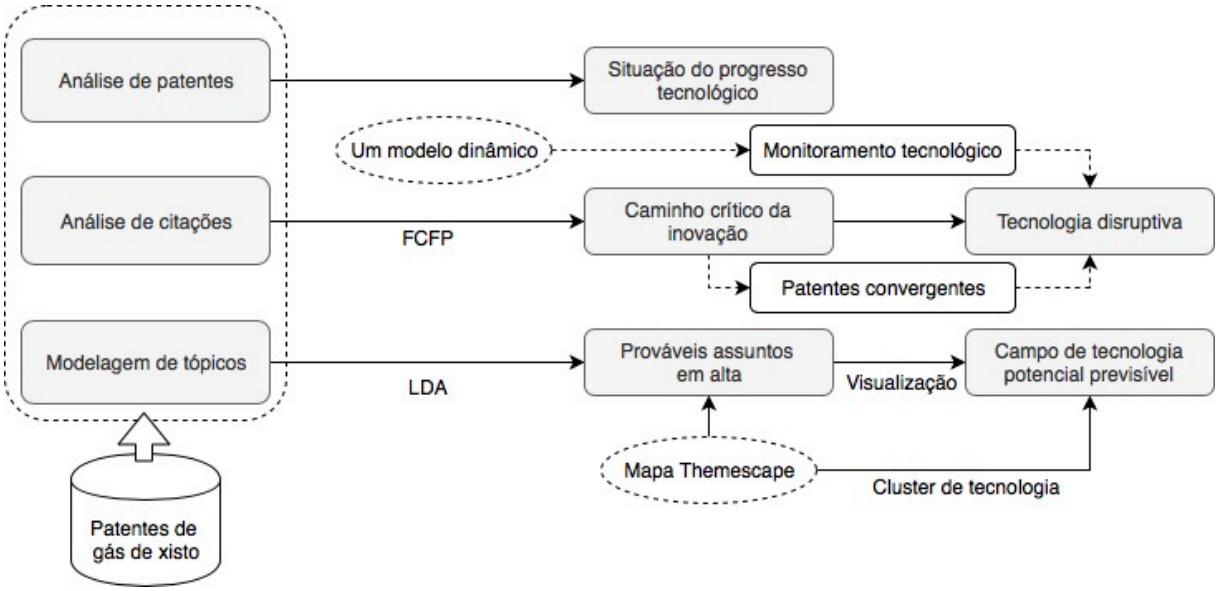
Os autores constataram que as anotações não melhoraram os resultados e que o algoritmo Decision tree (J48) apresentou o melhor resultado entre aqueles avaliados. Segundo o artigo, foi possível reduzir a quantidade de leituras realizadas por agentes humanos. O J48 e SVM reduziram quantidade de leitura em todos os casos (redução de 34,55 a 24,89% com J48 e 13,22 a 5,26% com SVM).

Trabalho: A dynamic forward-citation full path model for technology monitoring: An empirical study from shale gas industry (WEI *et al.*, 2017)

O artigo "A dynamic forward-citation full path model for technology monitoring: An empirical study from shale gas industry" (WEI *et al.*, 2017) prevê quais tecnologias em gás xisto são promissoras por meio do monitoramento tecnológico e mineração de texto, analisando quantitativamente as tendências de desenvolvimento e status quo da inovação tecnológica do gás de xisto combinado com mapas de patentes. A Figura 6 ilustra o método utilizado pelos autores em alto nível. Na parte mais alta da figura pode-se ver como a situação do progresso tecnológico é proveniente de uma análise de patentes. Também, vê-se que o caminho crítico da inovação é fruto de uma análise de citações e junto com o monitoramento tecnológico e o conjunto de patentes convergentes de acordo com os critérios estabelecidos, produzem uma visão sobre possíveis tecnologias disruptivas. A parte inferior da figura, mostra como o processo

de modelagem de tópicos usando o algoritmo LDA geram um conjunto de possíveis assuntos ou categorias de patentes permitindo sua visualização através de um software especial.

Figura 6 – Método proposto por Wei *et al.* (2017).



Fonte: Wei *et al.* (2017). Traduzido pelo autor.

Tecnicamente, os autores apresentam um algoritmo denominado *Forward-Citation Full Path* para identificar os principais caminhos de desenvolvimento tecnológico dentro de cada "cluster" de tecnologia de gás de xisto e, assim, monitorar as possíveis tecnologias presentes nesses caminhos. Eles utilizaram modelagem de tópicos por meio de LDA e mineração semântica de textos em combinação com o mapa de patentes do software ThemeScape da empresa Clarivate(CLARIVATE, 2019) para visualizar a distribuição de tópicos dos "clusters" de tecnologia para prever as tecnologias de gás de xisto mais promissoras.

Trabalho: A proposal for a technological surveillance unit aimed at regional competitiveness (PEREZ, L. G. *et al.*, 2017)

L G Perez *et al.* (2017) propuseram o design de uma unidade de vigilância tecnológica para o setor de materiais para construção civil em Sucre, Colômbia. A pesquisa foi conduzida metodologicamente através do desenvolvimento de um *framework* teórico para revisão literária, diagnósticos e identificação de "gaps" tecnológicos em setores e subsetores por meio de entrevistas, observação direta e análise de artigos. Os autores realizaram uma revisão literária nas bases Scopus, Science Direct, Web of Knowledge, Redalyc, Scielo, Proquest e Google Scholar. Especificamente para as bases Scopus, Science Direct e Web of Science, foi construída uma visualização com a ferramenta VOSViewer, a qual foi utilizada para o processo de análise e descoberta de "gaps" entre a dados de entrevistas em 21 empresas do setor de materiais de

construção local e o que está sendo adotado na literatura.

Trabalho: Monitoring Newly Adopted Technologies Using Keyword Based Analysis of Cited Patents (NAM; KIM, K., 2017)

Nam e Kwangsoo Kim (2017) propõem um método baseado em frequência do termo–inverso da frequência nos documentos (TF-IDF) e agrupamento K-means[2] para monitorar a adoção de determinadas tecnologias com base em citações de patentes. O método considera a citações entre patentes e se mostrou útil mesmo para um pequeno número de documentos. Para testar a validade do seu método, os autores fizeram um estudo no qual buscaram identificar quais tecnologias foram adotadas para permitir que os tratores agrícolas alcançassem uma direção automatizada.

Trabalho: Technological surveillance of s curves and cycle life of technology (JIMÉNEZ GONZÁLEZ *et al.*, 2017)

Os trabalho "Technological surveillance of s curves and cycle life of technology" (JIMÉNEZ GONZÁLEZ *et al.*, 2017) aplica uma metodologia para encontrar o ponto de inflexão da curva S para a identificação e análise dos ciclos de vida das tecnologias nos âmbitos da inovação, maturidade e declínio das mesmas, segundo as mudanças em modelos de difusão ao longo do tempo, levando em conta a observação dos pontos de inflexão nas mudanças de fases. Foram utilizadas análise estatística, especialmente modelos de regressão não lineres (Sigmoidal, Logístico e Gompertz) do volume de citações da tecnologia alvo em patentes e artigos científicos. Eles fizeram um estudo de caso sobre duas tecnologias: LCD em computadores e carros elétricos. No trabalho, foi utilizada uma conjunto de artigos e patentes coletados por 30 anos de bases de dados como Scopus y Free Patents Online respectivamente. A análise foi feita utilizando o software Sigmaplot pelo qual se avaliaram o ajuste da curva em 13 modelos estatísticos oferecidos pela ferramenta. Ao analisar a correlação entre as curvas obtidas nos artigos e nas patentes, os autores destacam que se pode entender melhor o comportamento das tecnologias e identificar com mais clareza os pontos de inflexão das curvas, ou seja, onde o produto deixa de ser uma novidade e passa a entrar em um período de adoção estável ou em declínio. O fluxo de dados entre as etapas aconteceu de forma manual pelos especialistas, assim como a escolha final dos modelos.

Trabalho: Technological surveillance and technology life cycle analysis: Usability assessment techniques, metrics and tools in the ICT sector (TOBÓN CLAVIJO *et al.*, 2017)

A aplicação da vigilância tecnológica para identificar tecnologias de avaliação da usabilidade foi abordada no trabalho "Technological surveillance and technology life cycle analysis:

[2] K-means é um método de aprendizagem de máquina não supervisionado de clusterização que busca dividir n observações dentre k grupos, onde cada observação pertence ao grupo mais próximo de sua média e que utiliza as centróides dos clusters para modelar seus dados (WIKIPEDIA, 2019b)

Usability assessment techniques, metrics and tools in the ICT sector" de Tobón Clavijo *et al.* (2017), limitado ao setor de TICs de Quindío (Colombia). Os autores basearam-se na metodologia de vigilância tecnológica proposta por Palop e Vicente (1999). A fase de planejamento foi realizada junto à especialistas e empresários da região, pela qual se determinou os fatores críticos da vigilância (FCV). Na etapa de busca e captura foram coletados artigos e patentes das bases de dados Scopus y Free Patents Online.

Por recomendação de especialistas, foram escolhidas 7 tecnologias para avaliar seus graus de maturidade utilizando suas curvas S. Foram contabilizados os volumes de artigos que citavam cada uma ao longo do tempo e aplicados modelos de regressão não linear para descrever e ajustar as curvas de crescimento, principalmente logísticos e sigmoidais, facilitar a visualização e encontrar os pontos de inflexão (vendo quando a tecnologia deixa de ser uma novidade e entra em sua fase de maturidade).

Por fim, para aprofundar a análise e encontrar comportamentos similares entre as tecnologias, foi realizada clusterização com o método *Nearst Neighbor* e distância euclidiana quadrática para se criar um diagrama de dendograma. Além do exercício da vigilância tecnológica aplicada, os autores salientam a importância de sua região empreender esforços na adoção de tecnologias com Eye tracking, Métricas y análise de Emoções para se tornar ainda mais competitiva.

Trabalho: Technology surveillance and curves in "S": Environmental technologies in Tourism, Quindio Innova project (GRAJALES LÓPEZ *et al.*, 2017)

No trabalho de Grajales López *et al.* (2017) realizou-se um exercício de vigilância tecnológica sobre tecnologias ambientais no turismo para o projeto Quindío Innova, seguindo uma metodologia similar a Jiménez González *et al.* (2017) e Tobón Clavijo *et al.* (2017). Foram coletados artigos e patentes. Os artigos foram coletados da base Scopus. As patentes foram obtidas por meio dos softwares Matheo patent e AcclaimIP. Em seguida, os autores extraíram as séries de dados acumulados no tempo sobre as publicações e patentes. A análise da curva S sobre os dados foi feita por meio de equações estatísticas, aplicando-se 13 modelos (Sigmoidal 3, Sigmoidal 4, Sigmoidal 5, Logístico 3, Logístico 4, Weibull 4, Weibull 5, Gompertz 3, Gompertz 4, Gompertz 5, Hill 3, Hill 4, Chapman 3 y Chapman 4 parameter) com o propósito de identificar quais oferecem melhores precisões para o ponto de inflexão da curva, facilitando a visualização das etapas do ciclo de vida (emergente, entrante, chave, madura ou em declínio) das tecnologias analisadas. Com isso, o trabalho identificou para o setor e para as tecnologias pesquisadas os principais líderes tecnológicos, os mercados com maior proteção de tecnologias, o alvo dos estudos científicos, as tendências mais marcantes do desenvolvimento tecnológico, os níveis de maturidade das tecnologias e as organizações líderes mundiais em pesquisa. Este trabalho não automatizou o fluxo de informações entre as etapas da vigilância tecnológica, tendo grande parte dos processos executados de forma manual, mas com o auxílio de algumas ferramentas.

Trabalho: Developing a Mobile Application for Technological Alerts (MARULANDA

ECHEVERRY *et al.*, 2016)

Marulanda Echeverry *et al.* (2016) descrevem como desenvolveram um protótipo de aplicativo de Vigilância Tecnológica para dispositivos móveis que dispara alertas sobre tecnologias relevantes para áreas de interesses previamente selecionadas pelos usuários interessados. Com a ferramenta, os usuários podem receber informações na forma de alertas quando uma nova fonte de conhecimento é adicionada ou sobre avanços tecnológicos sob diferentes óticas da comunidade acadêmica. Apesar de mencionar o fato de utilizar a inteligência coletiva e redes sociais, os autores não deixam claro como o aplicativo seria alimentado e nem se deveriam ser utilizadas apenas fontes formais de informação.

Trabalho: Generating patent development maps for technology monitoring using semantic patent-topic analysis (KIM, M. *et al.*, 2016)

M Kim *et al.* (2016) propõem um método composto por quatro etapas para construção de mapas de desenvolvimento de patentes (PDM) baseado em técnica de análise semântica de tópicos. Na primeira etapa, as patentes são coletadas e pré-processadas. São definidas strings de buscas em bases de patentes e os resultados são armazenados em planilhas. Na segunda etapa, os termos mais relevantes das patentes são extraídos por meio da técnica TF-IDF (frequência do termo–inverso da frequência nos documentos). Na terceira etapa são identificadas as taxonomias tecnológicas das patentes coletadas pela aplicação do algoritmo *Latent Dirichlet allocation* (LDA) sobre os dados gerados na etapa anterior. Assim, é gerado um modelo probabilístico conhecido como generativo, muito usado para *corpora* de documentos (coleções de dados discretos). Na quarta e última etapa é gerada uma visualização dos caminhos de desenvolvimento de patentes por meio da análise de sensibilidade baseada nas similaridades semânticas das patentes e citações. O protótipo que ilustrou o método utilizou patentes relacionadas à impressão 3D. O método é útil já pode ser aplicado sobre grandes volumes de patentes exigindo pouca intervenção humana.

Trabalho: Identification and monitoring of possible disruptive technologies by patent-development paths and topic modeling (MOMENI; ROST, 2016)

Momeni e Rost (2016) propõem um método baseado nas técnicas "Patent-Development Paths" (ou em português, caminhos de desenvolvimento de patentes), na análise "k-core" (k-núcleo) e em modelagem de tópicos para analisar a complexa relação entre patentes altamente citadas e a disrupção tecnológica. Conforme descrito no trabalho, a técnica *Patent-Development Paths* avalia as citações entre patentes para identificar os relacionamentos complexos entre elas e sua importância relativa para fornecer informações sobre o estágio atual e o histórico de uma determinada tecnologia.

Já a análise *k-core* permite distinguir entre as tecnologias nos caminhos principais do

grafo de citações, ou seja, permite identificar grupos coesos de tecnologias a partir de subgrupos de nós dentro de uma rede. Um *k-core* é um subgrafo conectado dentro de um grafo *P*, no qual todos os vértices têm um grau de pelo menos *k*. É formado pela exclusão repetida de todos os vértices de graus menores do que *k*. Essa abordagem remove, assim, os "corenesses" mais baixos da rede e tenta dividi-los em subgrupos que nos ajudam a detectar subconjuntos específicos de nós na rede. Consequentemente, pode-se analisar o maior componente da rede. O método foi aplicado para no setor da indústria fotovoltaica. Seus resultados permitem aos gerentes preverem sua própria posição na indústria e tomarem decisões mais assertivas com base nas tendências tecnológicas identificadas.

Trabalho: Using a semantic wiki for technology forecast and technology monitoring (FÄRBER, 2016)

Färber (2016) propõe uma ferramenta para armazenar, organizar e permitir a visualização de radares e portfólios tecnológicos, as quais são ferramentas extremamente difundidas na área de monitoramento tecnológico. Seu trabalho se diferencia ao propor o uso da Semantic Media Wiki como ferramenta de coleta e organização enriquecida pela adição de dois *plug-ins*[3] criados pelo autor, um para portfólios e outro para radares tecnológicos. Segundo o trabalho, o uso da Semantic Media Wiki[4] traz vantagens sobre wiki tradicionais ou ferramentas como o Microsoft Office por permitir criar relacionamentos entre informações por meio de atributos, semelhante aos conceitos das ontologias, facilitando o reuso, a pesquisa e a visualização. Adicionalmente, o autor desenvolveu um *plugin* para aumentar a segurança no acesso da informação. O trabalho foi validado por meio de questionários junto à usuários que deram excelentes notas com relação a eficiência e eficácia de seu uso. O trabalho dedica-se a apresentar um configuração da Semantic Wiki com as extensões necessárias para armazenar as informações e preparar seus resultados para a comunicação com os usuários. Como crítica ao trabalho, cabe destacar que este não oferece uma cobertura para o processo integral da vigilância ou monitoramento tecnológico ou algum grau significativo de automatização entre suas etapas.

Trabalho: Trends in 3-D printing from a patent information analysis (APA) (HENRI; CLERC, 2015)

A análise das informações contidas em patentes é uma importante atividade em projetos de pesquisa e desenvolvimento. Por meio delas, é possível entender determinada área sob a ótica de suas tendências tecnológicas e domínios de aplicação. No trabalho "Estudo de caso utilizando mapeamento de prospecção tecnológica como principal", Henri e Clerc (2015) aplicaram o método de APA (do inglês, Automatic Patent Analysis) para um estudo de caso de monitoramento tecnológico em patentes relacionadas à Impressão 3D. Os documentos foram

[3] Plugin é um software desenvolvido para adicionar funcionalidades complementar a outros software maiores.
[4] `https://www.semantic-mediawiki.org/wiki/Semantic_MediaWiki`

coletados da Worldpatent database a partir da European Patent Office (EPO) e salvos localmente em um computador. Para realizar suas análises os autores utilizaram o Matheo Software, pelo qual foi possível filtrar as patentes e montar diversos gráficos. No trabalho, não é apresentada uma metodologia especifica nem tampouco algum framework. Como contribuição, os autores destacam como a APA se mostra uma técnica útil para se obter em um curto espaço de tempo uma boa visão de um assunto de interesse.

Trabalho: Patterns of technological innovation and evolution in the energy sector: A patent-based approach (LEE, K.; LEE, S., 2013)

O trabalho "Patterns of technological innovation and evolution in the energy sector: A patent-based approac" de Kyungpyo Lee e Sungjoo Lee (2013) explora padrões de inovação e evolução em tecnologias energéticas. Utilizando patentes extraídas da *United States Patents and Trademark Office* (USPTO), os autores extraíram características dos documentos para realizar usas análises, incluindo mapas de patentes e clusterização hierárquicas para investigar padrões de inovações em grupos tecnológicos.

Trabalho: Identifying technological opportunities using the novelty detection technique: a case of laser technology in semiconductor manufacturing (GEUM *et al.*, 2013)

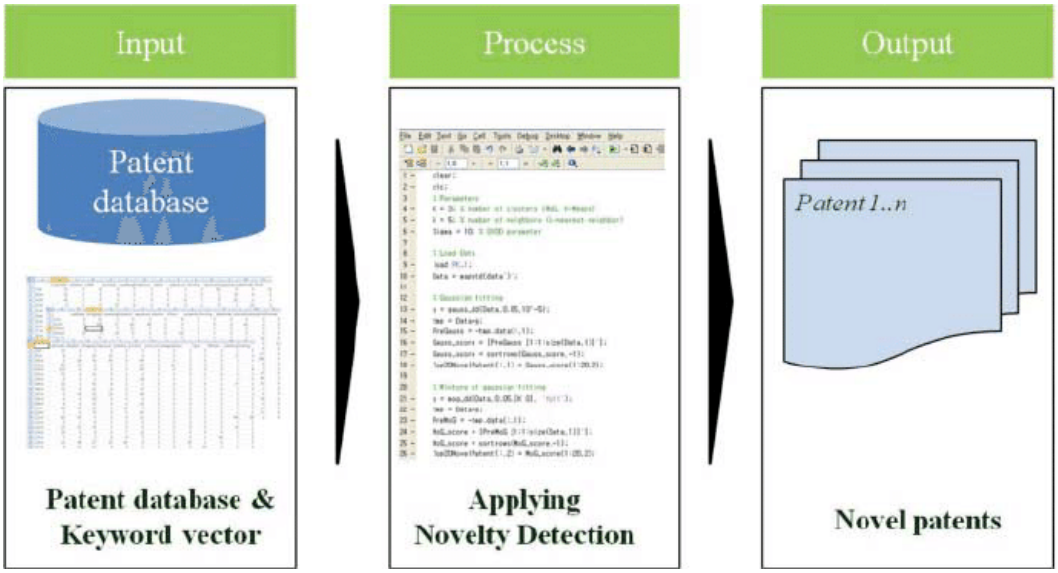
Em seu artigo, Geum *et al.* (2013) desenvolveram um *framework* sistemático para identificar oportunidades tecnológicas pela análise de bases de patentes. Foi empregada uma técnica de detecção de novidades cujo principal objetivo é detectar um novo padrão dentro dos textos de patentes. Como oportunidades tecnológicas os autores consideraram as "tecnologias com novidades ou tecnologias com potencial".

Foi realizado um estudo de caso utilizando a base de patentes *US Patent and Trademark Office* (USPTO). Cada patente foi pré-processada para encontrar as palavras-chaves mais relevantes para sua representação. Na fase de pré-processamento, após a extração, foi necessária a intervenção de especialistas para eliminar palavras irrelevantes. Por fim, as palavras foram derivadas para englobar abreviações, sinônimos, suas versões no singular e plural. Assim, um método próprio de *Detecção da Novidades* foi aplicado, recebendo como entrada os vetores que representam os documentos. As macro etapas descritas estão ilustradas na Figura 7.

O *framework* proposto pelos autores é composto por seis métodos com abordagens estatísticas para Detecção de Novidades: *Gaussian Fitting, Mixture of Gaussian Fitting, Parzen Window Density Estimation Fitting, k-means clustering fitting e Support Vector Data Description (SVDD)*, todos desenvolvidos em MATLAB[5]. No *framework*, os conjuntos de patentes identificados por cada algoritmo recebem pontuações para que possam ser caracterizados em uma escala de novidade pelos critérios de inovação, impacto tecnológico, escopo de aplicação, potencial de

[5] O MATLAB (MATrix LABoratory) é um software interativo de alta performance voltado para o cálculo numérico. Ele permite utilizar diversas bibliotecas que trazem facilidades na construção de equações matemáticas, aprendizagem de máquina, entre outros, sendo necessário codificar as operações em linguagem de programação.

Figura 7 – Entrada, processamento e saída da *detecção de novidades*.



Fonte: Geum *et al.* (2013).

mercado, intensidade competitiva e possibilidade de sucesso comercial e devem ser avaliadas pelos especialistas.

Um estudo de caso sobre a tecnologia de laser em litografia foi desenvolvido pelos autores para dar fundamentação à proposta. Nele, 384 patentes foram analisadas e pontuadas previamente por especialistas com mais de 10 anos de experiência na indústria de semicondutores. Em seguida, o *framework* proposto foi aplicado. Feito isso, os especialistas avaliaram os resultados obtidos pelo método, resultando em 0.9427 de acurácia e 0.8730 de precisão na indicação de quais patentes representariam uma novidade/inovação tecnológica. O resultado da avaliação mostrou um alto nível de precisão e acuracidade, justificando a validade da técnica. O ponto negativo da proposta é que o método de Detecção de Novidade é aplicado nas palavras-chaves extraídas das patentes, tornando a detecção dependente do grau de atualidade das palavras-chaves presentes nos documentos.

Trabalho (extra): Extraction of multi-dimensional research knowledge model from scientific articles for technology monitoring (HAKIM; DJATNA, 2016)

O trabalho de Hakim e Djatna (2016) foi localizado por meio de uma busca complementar à revisão sistemática, mas foi avaliado devido à sua relevância para este trabalho. Os autores propõem um método para extração de conhecimento de periódicos científicos(journals) para o desenvolvimento de mapas de pesquisas para o monitoramento e a avaliação de desenvolvimento tecnológico. Pelo método proposto, o conhecimento é representado pelos vocabulários presentes nos artigos dos *journals*. Em seu método, as palavras chaves são extraídas e tratadas, removendo-se *stopwords*, convertendo-se as palavras em seus radicais por *stemming*, entre outras técnicas. Para extrair-se as palavras mais significativas foi aplicado o método TF-IDF. Entre

as principais etapas do método está o cálculo da associação entre pesquisas utilizando sua frequência de palavras para gerar o grafo de pesquisas. O grafo gerado foi clusterizado por meio do método *Clauset-Newman-Moore (CNM)* para facilitar sua visualização. Métodos de clusterização populares como *Hierarchical Clustering*, *K-means* e *Support Vector Machine* (SVM) são usados para *clusterizar* documentos que não tenha sido modelados como grafos. O CNM, mais especificamente, é um algoritmo de aglomeração hierárquica para detectar a estruturas de comunidade (*community structure*)[6] e eficiente na clusterização de grandes grafos.(CLAUSET *et al.*, 2004). Como estudo de caso, os autores utilizaram artigos publicados entre 2004 e 2013 no periódico científico "Jurna Teknologi Tndustri Pertanian". Como conclusão, os autores obtiveram o desenvolvimento um mapa de tecnologia a partir das publicações coletadas, permitindo monitorar o desenvolvimento de tecnologia e possibilitar mais análises de rede.

3.3.3 Discussão sobre a análise dos trabalhos relacionados

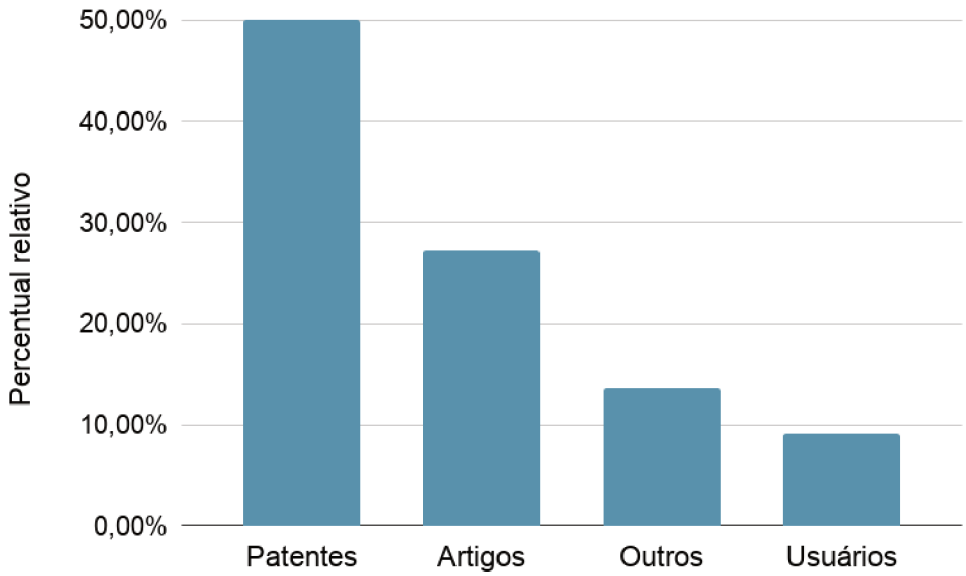
Com a realização da análise dos trabalhos relacionados foi possível comparar os métodos utilizados pelos autores nas etapas de inteligência e análise, suas fontes de dados, seus entendimentos sobre o posicionamento de suas propostas frente à cenários de Big Data e o nível de automatização existente no fluxo de informações entre as etapas que compreendem a coleta de informação, sua análise e comunicação aos interessados. O quadro 6 apresenta uma comparação entre os trabalhos com relação às características extraídas. Os trabalhos de Gudanowska (2014), Dos Santos Amparo *et al.* (2014) e Jürgens e Herrero-Solana (2017) foram desconsiderados após sua leitura por não oferecem contribuições relevantes para esta pesquisa, tendo pouco ou nenhum grau de automatização nos processos apresentados ou alguma arquitetura ou *framework* útil. O texto integral de Hou (2017) não foi encontrado.

De forma geral, os trabalhos utilizaram fontes formais de informação, com ênfase nas patentes, como pode ver visto no gráfico da figura 8. Outras características que levaram à sua escolha são características como confiabilidade, estruturação devido ao sistemas de classificação de patentes e a facilidade de coleta em bases conhecidas, como a *United States Patents and Trademark Office*. Com características semelhantes, o segundo tipo de documento mais utilizado foram artigos científicos. Ao monitorar este tipo de documento, por exemplo, é possível perceber comportamento de empresas cujos produtos ainda estão em estágio de pesquisa e desenvolvimento.

A aplicação da inteligência para a agregação de valor à informação foi a atividade com maior grau de automatização nos processos de vigilância e monitoramento tecnológicos. Nela, foram utilizadas principalmente técnicas de mineração de textos e clusterização, dentre as quais se destacaram o TF-IDF para extração de termos relevantes e *K-means* ou *Latent Dirichlet Allocation* (LDA) para geração e modelagem tópicos e identificação tecnologias de interesse (novas ou predominantes).

[6] Uma rede ou grafo tem uma estrutura de comunidade se for possível agrupar facilmente seus nós em subconjuntos cujos nós estejam densamente conectados entre si.

Figura 8 – Percentual relativo dos tipos de fontes utilizadas nos trabalhos analisados



Fonte: o autor.

As outras atividades do processo de vigilância variaram quanto ao grau de automatização. Os estudos de casos desenvolvidos, modelos e arquiteturas apresentados não evidenciam automatização no fluxo de informação entre suas atividades ou etapas sem a intervenção humana, tornando os processos caros e complexos para serem mantidos e executados de forma permanente e periódica. Também, não foi encontrado um modelo ou método que suporte trabalhar com cenários de Big Data, conforme definido na seção 2.2. Observou-se que as atividades de extração de dados das bases foram feitas de forma manual ou semi-automatizada e o bom funcionamento ou avaliação dos resultados foram dependentes da intervenção humana. Mesmo sem a afirmação explicita dos autores, pressupõe-se que a abordagem de A Perez *et al.* (2018) poderia ser utilizada em cenários de Big Data se observado o aspecto da variedade de documentos, porém seria necessário avaliar se as tecnologias utilizadas seriam capazes de lidar com grandes volumes de dados e oferecer resposta em um tempo aceitável.

Com isso, a revisão sistemática da literatura traçou um panorama preciso sobre os métodos de Vigilância Tecnológica com algum grau de automatização propostos em artigos publicados entre 2013 e 2020 nas principais bases científicas disponíveis, permitindo comparar métodos, estratégias utilizadas pelos autores em suas arquiteturas ou estudos de casos, fontes de dados mais utilizadas e o estado da arte sobre o tema, oferendo bases sólidas para que esta pesquisa pudesse ser desenvolvida aproveitando o conhecimento científico já produzido.

Quadro 6 – Análise comparativa dos trabalhos relacionados.

Autor(es)	**Fonte**	**Métodos**
Mikova e Sokolova (2019)	Publicações científicas, patentes, mídia, conferências, dissertações, apresentações	LDA
Moreno C e Díaz (2019)	Artigos	Clusterizações
A Perez *et al.* (2018)	Patentes, notícias, boletins oficiais do governo e documentação de competência	J48, SVM e Naive Bayes
Wei *et al.* (2017)	Patentes	LDA
L G Perez *et al.* (2017)	Artigos e entrevistas	Revisão da literatura
Nam e Kwangsoo Kim (2017)	Patentes	K-means
Jiménez González *et al.* (2017)	Artigos e patentes	Regressões não lineares
Tobón Clavijo *et al.* (2017)	Artigos e patentes	Regressões não lineares, clusterização (nearest neighbor)
Grajales López *et al.* (2017)	Artigos e patentes	Regressões não lineares
Marulanda Echeverry *et al.* (2016)	Inserção manual	Aplicativo
M Kim *et al.* (2016)	Patentes	LDA
Momeni e Rost (2016)	Patentes	LDA
Färber (2016)	Inserção manual	Semantic Wiki Plugin
Henri e Clerc (2015)	Patentes	Automatic Patent Analysis (APA)
Kyungpyo Lee e Sungjoo Lee (2013)	Patentes	Clusterização hieráquica
Geum *et al.* (2013)	Patentes	Novelty detection techniques (Gaussian, mixture of Gaussian, Parzen window density, k-means, k-nearest neighbour e SVDD)

Fonte: elaborado pelo autor.

4 PROPOSTA

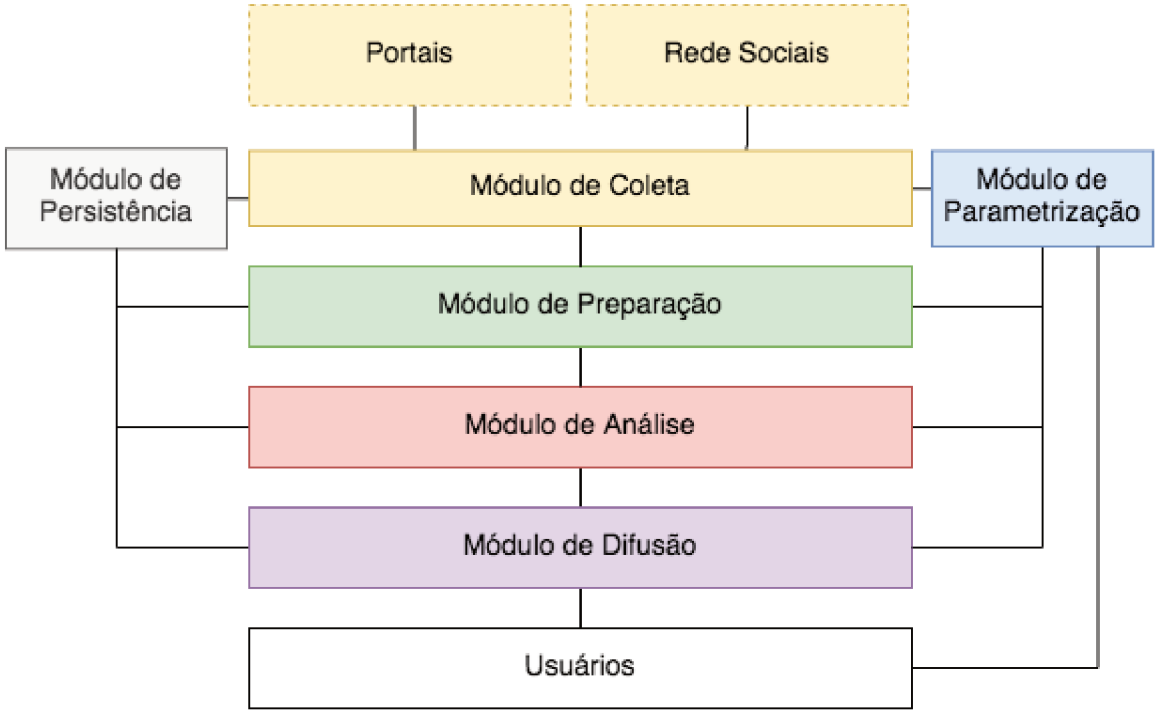
Este capítulo apresenta um modelo a partir do qual se pode implementar sistemas de vigilância tecnológica automatizada. Sua construção e generalização foram feitas a partir da revisão da literatura, onde foram estudados métodos e conceitos análogos e a partir da análise de softwares disponíveis no mercado apresentados na seção 2.1.1. Neste capítulo, também são apresentadas uma arquitetura de referência com base no modelo para viabilizar a implementação de um sistema real e seu *workflow* que detalha o fluxo da informação. Por fim, são feitas as considerações finais do capítulo.

4.1 MODELO CONCEITUAL

Um modelo conceitual é uma composição de conceitos utilizados para facilitar a compreensão ou permitir simulações sobre determinados assuntos. Ele abstrai informações estruturais, funcionais, de interações e interfaces com usuários ou outros sistemas, possuindo a vantagem de descrever em alto nível de abstração a semântica de um sistema. Neste trabalho, o modelo foi projetado depois de um processo de generalização baseado em uma revisão da literatura pela qual foram estudados conceitos relacionados ao tema, em metodologias propostas por trabalhos relevantes na área e na análise das plataformas de software apresentadas na seção 2.1.1. Como resultado, gerou-se um modelo conceitual contendo quatro módulo principais (coleta, preparação, análise e difusão) e dois módulos auxiliares (parametrização e persistência), que serão detalhados ao longo deste capítulo. Na Figura 9 pode-se ter uma visão de seus módulos e conexões.

J Marcela Sánchez e Palop (2002) propõe um ciclo de Vigilância Tecnológica e Inteligencia Competitiva composto por cinco etapas, em uma tradução livre, são: planejamento, busca e captura, análise e organização, inteligência e comunicação. Esse é um trabalho de relevância na área, sendo amplamente citado. Por essa razão, foi utilizado como referência para generalização do modelo. Nele, a etapa de *planejamento* engloba as atividades de identificação das necessidades informacionais da organização. Por envolver atividades dependentes de interações manuais e discussões entre especialistas ela não foi transportada ao modelo. Alternativamente, incluiu-se o módulo de *parametrização* que mantém algumas das definições decorrentes do planejamento, como as especificações das tecnologias de interesse a serem "vigiadas" ou os dicionários de termos a serem identificados nas publicações. A etapa de *busca e captura* se preocupa em coletar os documentos e as informações que alimentam o sistema e influenciou o *módulo coleta*. A etapa de *análise e organização* se dedica a analisar, tratar e armazenar as informações, sendo inspiração para o módulo *módulo de preparação*. Contudo, este módulo carrega conceitos mais próximos do *Big Data Analytics* como será apresentado mais adiante. A etapa de *inteligência* tem o objetivo de agregar valor às informações e produzir os resultados esperados pela organização, tendo uma representação similar no modelo pelo *módulo de análise*. Por fim, a etapa de *comunicação* objetiva comunicar os resultados à toda a organização e lhe transferir os conhecimentos gerados. No modelo, está última etapa esta relacionada ao *módulo*

Figura 9 – Modelo de Vigilância Tecnológica Ativa Automatizada proposto.



Fonte: elaborado pelo autor.

de difusão.

Na revisão da literatura realizada, construiu-se um panorama sobre o estado da arte em Vigilância Tecnológica por meio da seleção de trabalhos que apresentavam propostas de modelos ou arquiteturas de bases de dados científicas nacionais e internacionais conceituadas, excluindo-se aqueles de cunho apenas teórico. Dos 263 trabalhos encontrados e avaliados, nenhum considerou cenários de Big Data explicitamente conforme definido na seção 2.2. Esta limitação reforça a necessidade da construção de um modelo que considere estes cenários uma vez que grande parte das organizações estão imersas neles.

Como resposta, o modelo conceitual considera os cenários de Big Data caracterizados pela definição dos 5 vs (volume, variedade, velocidade, veracidade e valor), apresentada na seção 2.2. Todavia, a variação para mais ou para menos de um dos cinco Vs não descaracteriza necessariamente um cenário de Big Data. Isto pode ser melhor compreendido pelo exemplo a seguir. Um sistema de Vigilância Tecnológica poderia ter que lidar com um número significativo de documentos coletados de forma periódica (volume), 10000 documentos por hora por exemplo, com formatos não estruturados (variedade) compostos por textos, planilhas e diagramas, provenientes de bases de patentes e de sites especializados (veracidade) que precisam ser coletados, processados e analisados rapidamente (velocidade) a fim de se gerar um panorama tecnológico confiável sobre um determinado setor industrial (valor). No exemplo apresentado, a característica de "volume" não é proeminente. Mesmo assim é evidente o cenário de Big Data no qual o sistema precisa operar. Sem as automatizações e ferramentas computacionais adequadas, provavelmente

os resultados produzidos pelo processo de vigilância não atenderão aos seus usuários.

Os módulos do modelo conceitual devem operar sobre diversos tipos de informações sendo capaz de escalar horizontalmente quando for preciso aumentar a capacidade de armazenamento e oferecer processamento em paralelo para atender as demandas em um tempo aceitável. Cada um deles é discutido a seguir.

- Módulo de Parametrização: a delimitação das fontes de dados e de um escopo a ser monitorado levando em conta os interesses das organizações é fundamental para qualidade do resultado das atividades de vigilância. Transpor o conhecimento gerado durante o planejamento de uma equipe de especialistas para um sistema computacional pode ser um grande desafio. Uma alternativa interessante apresentada na seção 2.3 é a *análise de domínio*. Após o processo de análise é possível codificar o conhecimento sobre determinado escopo tecnológico no formato de *ontologias* e assim alimentar o sistema. Embora as atividades de planejamento sejam importantes, é o seu resultado que se procura inserir no sistema. Esta delimitação sobre o escopo da vigilância deve estar presente em um modelo que deseja trazer respostas precisas aos tomadores de decisões e este é o papel do módulo de parametrização. Este componente representa todas as configurações necessárias para o funcionamento do sistema. Visa adequá-lo antes ou durante suas execuções. O módulo deve possuir as fontes de informações e as tecnologias a serem avaliadas, as categorias de informações para eventuais categorizações, além de quaisquer parâmetros necessários para os demais módulos do sistema. A parametrização permite ao sistema se adequar às expectativas informacionais de seus usuários. Este componente deve permitir configurações periódicas para melhorar suas entradas com base no desempenho geral do sistema.

- Módulo de Coleta: um aspecto que deve ser incluído em um modelo para vigilância tecnológica automatizada são os coletores de informação, uma vez que para que o processo seja estruturado e capaz de ser executado com o mínimo de intervenção humana em um tempo aceitável, as informações devem ser buscadas ou recebidas por agentes de softwares de softwares e não apenas por seus usuários (seção 2.4), o que daria muito trabalho. A coleta de informações foi uma limitação verificada nos trabalhos analisados na revisão sistemática da literatura (seção 3.3), cujas extrações de informações das bases de dados foram realizadas de forma manual ou semi-automatizada, indo de encontro às necessidades atuais. Para que haja um ganho de produtividade, principalmente devido à periodicidade nas atividades de busca de informações nas diferentes fontes é preciso que o sistema de vigilância permita a automatização destas atividades. Assim, este módulo representa a entrada de dados no sistema. Ele pode atuar de forma ativa, buscando as informações ou de forma passiva, recebendo os dados de usuários ou sistemas. Ele deve ser capaz de coletar todas as informações necessárias para subsidiar a vigilância tecnológica. Por meio deste módulo deve ser possível capturar publicações de diversos portais e redes sociais quando

necessário.

- Módulo de Persistência: é encarregado de persistir as informações do sistema, como seus parâmetros, dados coletados e análises, ou interações dos usuários. Deve abstrair a indexação e distribuição dos documentos em unidades de armazenamentos adequadas para os modelos de dados escolhidos, devendo ser possível manter os dados de forma centralizada ou distribuída, dependendo da infraestrutura ou tecnologias escolhidas. Este módulo é um dos mais afetados em cenários de Big Data e pode envolver uma variedade de modelos de armazenamento. Todavia, não se deve restringir o armazenamento de dados exclusivamente em bancos de dados não relacionais comuns em ambientes de Big Data. As ontologias podem ser modeladas no formato de arquivo OWL e os dicionários de dados ou tesauros podem estar em formato de arquivos de textos. Dessa forma, uma grande quantidade de arquivos pode exigir um sistema de arquivos distribuídos, o qual distribui os arquivos ou suas cópias em clusters de computadores permitindo que se armazenem e se acessem arquivos remotos como se fossem locais. Alguns exemplos são HDFS, CEPH, FhGFS, PVFS e Lustre (MACEDO *et al.*, 2015), sendo o HDFS[1] o mais popular atualmente.

- Módulo de Preparação: visa preparar e organizar a informação a para sua análise e recuperação eficientes. A preparação dos dados é um processo consolidado nas atividades de *Big Data Analytics* (SIRIWEERA *et al.*, 2017), (STOREY; SONG, 2017). Podem ser empregadas técnicas de mineração de textos, conforme apresentado na seção 2.5, para reduzir a dimensionalidade dos documentos não estruturados como patentes ou artigos científicos coletados ou prover a possibilidade de filtragem, limpeza e extração de dados de interesse. A preparação é um ponto crítico no sistema já que sua baixa eficiência pode comprometer o tempo geral de entrega de resultado aos usuários.

- Módulo de Análise: tendo-se inserido as informações no sistema de vigilância, sua manipulação como o enriquecimento das informações ou cálculos também devem funcionar de maneira independente da intervenção humana. Os trabalhos científicos analisados na RSL utilizaram diferentes técnicas e algoritmos para processar e agregar valor à informação. Os autores adotaram principalmente técnicas como extração de termos de documentos por meio de técnicas como TF-IDF (frequência do termo–inverso da frequência nos documentos) (HAKIM; DJATNA, 2016), do algoritmo Latent Dirichlet Allocation (LDA) para modelagem de tópicos (NAM; KIM, K., 2017), (WEI *et al.*, 2017), (MOMENI; ROST, 2016), (MIKOVA; SOKOLOVA, 2019) e regressões não lineares (JIMÉNEZ GONZÁLEZ *et al.*, 2017), (TOBÓN CLAVIJO *et al.*, 2017), (GRAJALES LÓPEZ *et al.*, 2016). Outra técnica utilizada foi a clusterização dos documentos, destacando-se para esta finalidade o

[1] O HDFS é um sistema de arquivos distribuído inspirado no MapReduce e no GoogleFS que possibilita o armazenamento de dados em clusters de computadores comuns, sob demanda, possibilitando o processamento de enormes volumes de dados, oferecendo tolerância a falhas. Mais informações podem ser obtidas no site `https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html`

algoritmo de k-means (NAM; KIM, K., 2017), (GEUM *et al.*, 2013), o qual viabiliza a detecção de tecnologias proeminentes em determinados contextos. Nos trabalhos, o bom funcionamento ou avaliação dos resultados, contudo, foram dependentes de intervenções humanas, principalmente para buscar os documentos em bases especializadas e depois inserir eles em softwares de análise. O que se observou, é que a maioria dos trabalhos destaca uma etapa de análise. Assim, o módulo de análise deste modelo mantém as atividades de inteligência mais profundas de um sistema de vigilância tenológica. Como entrada, ele recebe os dados pré-processados e limpos pelo *módulo de preparação* e os analisa para produzir como saída as informações necessárias ao *módulo de difusão*. No módulo de análise podem ser executadas atividades como cálculos, predições e agrupamentos. Também é possível aplicar as técnicas de mineração de textos já discutidas como a modelagem de tópicos ou classificação de documentos. Como principal responsabilidade para este módulo tem-se a identificação dos termos chaves, como as tecnologias a serem monitoradas, dentro dos artigos, patentes, redes sociais ou qualquer documento coletado;

- Módulo de Difusão: tem por objetivo facilitar a hermenêutica dos dados e resolver as deficiências informacionais das organizações, sistemas e indivíduos. Enquanto J Marcela Sánchez e Palop (2002) nomeia sua etapa correlata de "comunicação", neste trabalho ela foi chamada de difusão, seguindo a nomenclatura proposta pelo Observatorio de Virtual de Transferencia de Tecnologia - OVTT, da Universidade de Alicante, Espanha (OVTT, 2019b). Entre as atividades viabilizadas por este módulo estão o envio de mensagens eletrônicas e a disponibilização de painéis com os dados levantados. Por fim, este módulo atua na geração de conhecimento e valorização das informações processadas.

Na Figura 9, os retângulos coloridos representam os módulos, exceto os de bordas pontilhadas (Portais e Redes Sociais). As linhas conectando os retângulos representam a conexão entre eles, abstraindo o fluxo de informações. Os módulos de coleta, preparação, análise e difusão estão conectados ao módulo de persistência devido à necessidade de se armazenar e recuperar dados. Também, estão conectados ao módulo de parametrização para se ter acesso direto às configurações necessárias. Além disso, vê-se uma representação para os usuários que interagem com o módulo de difusão, consumindo as informações resultantes do processo de vigilância.

Na seção 2.1, onde a vigilância tecnológica foi discutida, enumeraram-se os principais tipos de vigilância (vigilância ativa, vigilância passiva, inteligência competitiva e previsão tecnológica). O modelo proposto visa servir de referência para àqueles que desejem implementar sistemas de vigilância tecnológica **ativa** sobre informações **formais**. A importância do uso de fontes formais foi sustentada nos trabalhos analisados durante a revisão sistemática da literatura. Wei *et al.* (2017) e Momeni e Rost (2016) desenvolveram métodos para monitoramento em patentes. Geum *et al.* (2013) e Nam e Kwangsoo Kim (2017) também, mas se detiveram a patentes presentes da US Patent and Trademark Office - USTPO. Abe e Tsumoto (2009) e Hakim e Djatna (2016) utilizaram técnicas envolvendo artigos científicos enquanto Shiryaev *et al.* (2017)

atuaram sobre publicações disponíveis em portais da internet. Como foi destacado anteriormente, os métodos e fontes utilizadas não foram caracterizados como sendo pertencentes ou possíveis de operar em cenários de Big Data nem tampouco tiveram suas atividades automatizadas na maior parte das etapas ou processos, sendo necessária a constante intervenção dos usuários.

O grande volume de informações que podem ser coletadas em um processo de vigilância tecnológica exige que qualquer sistema que implemente este modelo não precise seguir passos lineares, ou seja, muitas vezes todos os componentes trabalharão em paralelo. Por exemplo, ao mesmo tempo em que um sistema pode capturar novas informações ele pode limpar ou filtrar outras, realizando a atualização de painéis (dashboards) em tempo real.

4.2 ARQUITETURA PROPOSTA

O modelo apresentado na seção 4.1 oferece uma visão macro dos componentes que devem compor um sistema de vigilância tecnológica automatizado em um nível conceitual. Nesta seção, é apresentada uma arquitetura para suportar o modelo. Aqui, exploram-se aspectos mais técnicos e detalhados, sendo útil para fundamentar a implementação do modelo em um cenário real. A arquitetura tem seus componentes agrupados na mesma estrutura do modelo conceitual: coleta, preparação, análise, difusão, persistência e parametrização. No componente de parametrização são definidas as limitações e escopo de atuação do sistema. Nesta arquitetura, os parâmetros são as *ontologias* das tecnologias de interesse, as fontes de informação e dicionários de dados contendo regiões geográficas. As informações dos dados são gerenciadas e armazenadas pelo módulo de armazenamento. As ontologias são armazenadas em arquivos OWL (do inglês Web Ontology Language ou linguagem de ontologia web em uma tradução livre), onde cada arquivo representa um domínio ou área da indústria. Pode haver, por exemplo, um arquivo que contém todas as tecnologias-chave do setor têxtil. Os dicionários de dados podem ser armazenados em formato texto ou mesmo em bancos relacionais uma vez que são dados mais bem estruturados. Um exemplo de dicionário seria um conjunto de nomes de países ou de regiões geográficas que se deseje extrair dos documentos coletados. A arquitetura prevê um componente de coleta chamado de "DataImporter" que se conecta a diferentes robôs de coleta de dados (webcrawlers) ou diretamente a bases de dados para receber as informações capturadas. As informações para este tipo de arquitetura estão no formato de artigos, notícias, publicações de redes sociais como youtube, facebook, twitter e linkedin, doravante chamadas de "publicações", comumente utilizados pelas organizações. O módulo de coleta é parametrizado podendo ter um canal ou uma fonte de informação adicionada ou removida.

As informações coletadas (publicações, artigos e notícias) são enviadas ao módulo de persistência para serem armazenadas em um Banco de Dados orientados a Documentos e adequado para Big Data. É importante que o banco escolhido possua escalabilidade horizontal que permite aumentar a capacidade do sistema adicionando-se mais servidores de maneira simples e sem impacto na escrita ou leitura dos dados. É importante, ainda, que o banco esteja preparado para manipular grandes volumes de dados de maneira praticamente transparente para

o usuário.

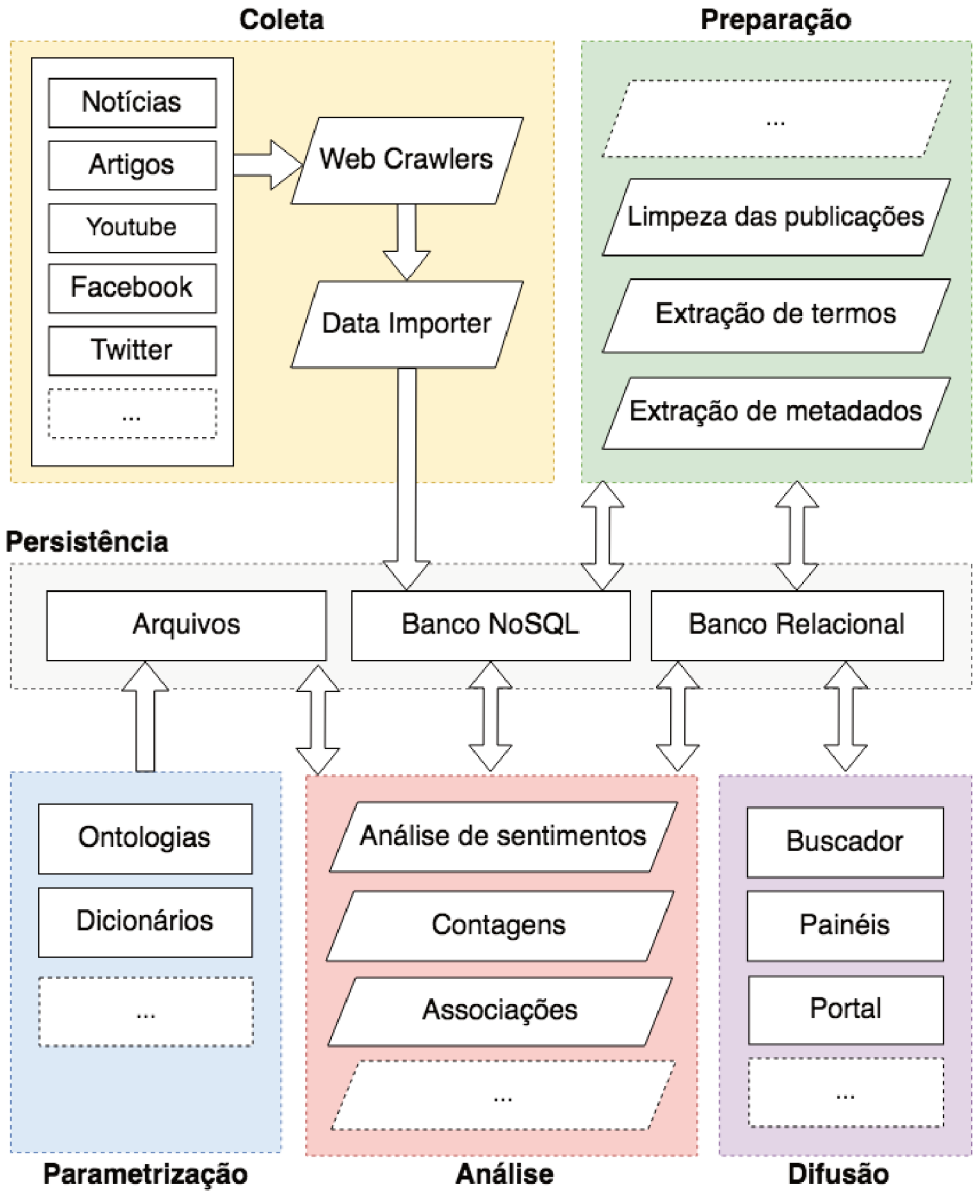
Dentre as alternativas de bancos existentes, o MongoDB[2] se mostra como uma ótima opção por oferecer os benefícios descritos e ainda ser de código aberto, o que traz segurança e reduz custos. Este tipo de banco de dados tem a vantagem de ser escalável horizontalmente, ou seja, de permitir a distribuição dos dados em diversos computadores ou locais. Bancos desta natureza costumam ter um desempenho melhor do que os bancos relacionais para este tipo de dado (CHICKERUR *et al.*, 2015), (JUNG *et al.*, 2015). Sugere-se uma organização de dados no formato de *coleções* específicas para cada tipo de fonte a qual é suportado pela maioria de bancos de dados orientados a documentos. Pode haver uma *coleção* de documentos para publicações do Facebook, para artigos de portais, para publicações do Twitter, e assim por diante. O módulo de preparação consome os dados armazenados no banco de dados NoSql para efetuar suas operações. Ele responsabiliza-se por eliminar documentos duplicados, algo comum já que um artigo pode estar presente em mais de um portal ou base de dados. Responsabiliza-se, também, pela identificação e extração das tecnologias-chaves das publicações e pela extração de dados como países citados além de metadados da publicação como data, hora, título, categoria, entre outros. Neste módulo, ocorre, ainda, a indexação das publicações em tecnologias específicas para recuperação de documentos textuais, conhecidos como *Search Engines*, dentre as quais se destacam o Apache Solr[3] e o ElasticSeach[4]. Mediante configuração específica, o banco MongoDB pode funcionar como indexador próprio para ferramentas de buscas *full text search*. Esta arquitetura não prevê a extração, armazenamento e análise de vídeos, imagens e áudios, os quais devem ser tratados com tecnologias específicas.

A saída deste módulo são dados no formato de documentos (semi-estruturados) e dados estruturados para um banco de dados relacional com informações e campos úteis, que alimentam o módulo de análise. Se na entrada deste módulo temos, por exemplo, um texto completo, na saída teremos uma lista de tecnologias-chaves identificadas e países mencionados nos documentos.

O módulo de análise se beneficia diretamente dos resultados do módulo de preparação, o qual tem o potencial de lhe entregar dados filtrados e estruturados. A análise deve prover inteligência necessária ao sistema para agregar valor às informações que serão entregues ao tomador de decisão. Aqui são feitas contagens e agrupamentos de tecnologias-chaves para permitir comparações, são feitas associações entre tecnologias e países mencionados em um mesmo documento e ainda são feitas análises de sentimentos para entender a percepção em uma região em relação a uma determinada tecnologia. A este módulo podem ser acoplados inúmeros outros cálculos ou operações que precisem ser executados para produzir as avaliações necessárias. Os dados processados nesta etapa permanecem em bancos de dados relacionais no formato estruturado. O módulo de análise concentra a maior parte da "inteligência" do sistema, mas para que os usuários percebam o valor gerado eles precisam ser capazes de interpretar os resultados. Esta é a responsabilidade do módulo de difusão. A difusão visa entregar ao usuário

[2] https://www.mongodb.com/
[3] https://lucene.apache.org/solr/
[4] https://www.elastic.co/pt/

Figura 10 – Arquitetura de Vigilância Tecnológica Ativa Automatizada para o modelo conceitual.



Fonte: elaborado pelo autor.

as informações necessárias para sua tomada de decisão. Esta arquitetura prevê como interfaces deste módulo:

- Painéis ou Dashboards: são componentes gráficos interativos que sintetizam as informações processadas e armazenadas nos bancos relacionais. A vantagem destes painéis em relação aos gráficos estáticos é que permitem uma navegação mais natural pelos dados possibilitando que as filtragens sejam feitas ao se clicar sobre um item de um gráfico relacionado. Os *dashboards* podem ser construídos para contabilizar as tecnologias mencionadas, agrupando-as por categorias, apresentando correlações entre tecnologias-chaves

e países. Estes painéis podem ser construídos com tecnologias como Microsoft PowerBI[5] e QlikView[6];

- Portal: as publicações coletadas são disponibilizadas em um portal para o usuário. Neste portal é possível filtrar as publicações por tipo de fonte, por domínio ou setor. Também é possível disponibilizar marcações sobre as publicações destacando as tecnologias-chaves que se encontram nas mesmas e ordená-las por data. O portal funciona como um verdadeiro agregador de publicações permitindo que o usuário possa acessar somente publicações relevantes para seu domínio ao invés de precisar acessar dezenas de portais ou bases;

- Buscador: permite ao usuário recuperar ou pesquisar as publicações coletadas no formato *full text search*, tipo de pesquisa que considera trechos parciais dos termos pesquisados mesmo quando existem variações ortográficas como acontece na ferramenta de busca Google. O buscador traz agilidade aos usuários quando estes precisam pesquisar dados específicos de um segmento.

Este módulo pode ser expandido, ainda, para enviar alertas aos usuários quando uma determinada tecnologia for identificada em uma publicação ou para oferecer uma ferramenta pela qual seja possível elaborar e enviar boletins agrupando publicações de interesse para um determinado segmento de usuários. Por fim, há módulo de persistência que se conecta a todos os outros módulos que o utilizam para armazenar, recuperar ou pesquisar dados, ou ainda, usar suas estruturas como um estagiamento em seus processos. A arquitetura prevê que ele deve ser capaz de manipular arquivos de ontologias, disponibilizar bancos relacionais como PostgreSQL ou Microsoft SQLServer, bancos NoSQL como MongoDB ou dicionário de dados em arquivos de texto.

4.3 WORKFLOW DA ARQUITETURA PROPOSTA

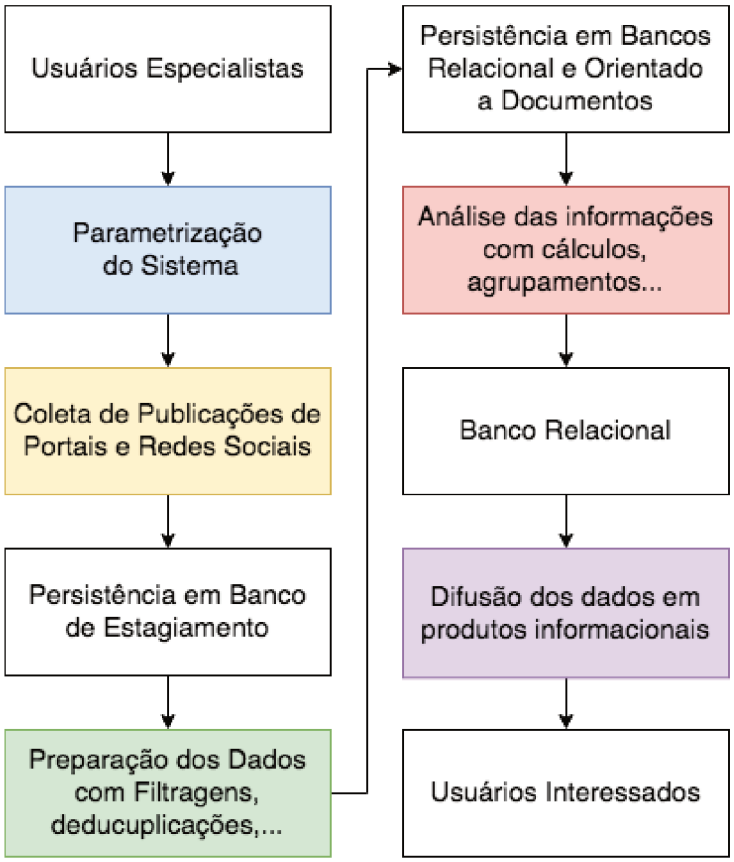
Workflow (em português, fluxo de trabalho) pode ser entendido como um conjunto de passos sequenciais para se automatizar processos de negócio, delimitados por um conjunto de regras definidas, auxiliando que o fluxo seja comunicado de uma pessoa para outra (WIKIPEDIA CONTRIBUTORS, 2020). Cichocki *et al.* (2012) entendem workflows como a execução coordenada de múltiplas tarefas desempenhadas por diferentes entidades de processamento. Para os autores, uma tarefa define algum trabalho a ser feito e pode ser especificado de diferentes maneiras, como textos, diagramas ou programas de computador. Já, as tarefas de um workflow são normalmente executadas por um sistema automatizado, segundo o trabalho, mas também podem ser executadas manualmente se desejado.

Um exemplo de workflow representando processos no contexto da vigilância tecnológica pode ser visto na Figura 1 que representa os passos da metodologia proposta por J Marcela Sánchez

[5] https://powerbi.microsoft.com/pt-br/
[6] https://www.qlik.com/pt-br/products/qlikview

e Palop (2002). Em seu trabalho, os autores dividem o ciclo de VT em cinco processos básicos: planejamento, busca e coleta, análise e organização, inteligência e comunicação) conforme detalhado na seção 2.1.

Figura 11 – Fluxo da arquitetura de Vigilância Tecnológica Ativa Automatizada para o modelo conceitual proposto.



Fonte: elaborado pelo autor.

Na Figura 11 é apresentado o workflow da arquitetura proposta, trazendo uma visão completa e de alto nível do encadeamento sequencial dos processos. Inicialmente, os usuários especialistas de cada domínio ou setor precisam parametrizar o sistema inserindo as ontologias com as tecnologias de interesse, além de delimitar quais portais e redes sociais que devem ser coletadas. Os sistema, então, inicia a coleta de forma automática, monitorando as fontes e buscando novos dados sempre que disponibilizados. As informações coletadas são então persistidas em um banco que serve como uma forma de estagiamento provisório, pois como os mesmos artigos ou publicações são normalmente reproduzidas em diversas fontes, ainda se faz necessária uma limpeza para evitar duplicações e extrair seus dados.

Em seguida, todos os dados coletados são pré-processados pela etapa de preparação. Ali, eles são removidas duplicidades mantendo apenas um registro, são feitas as extrações de metadados e identificadas as tenologias-chaves como descrito no modelo e na arquitetura. Após essa fase as publicações são armazenadas em um banco orientado a documentos com

sua estrutura já separada, como título, corpo do texto, data de publicação, etc; as tecnologias identificadas e os metadados das publicações são persistidos em um banco relacional.

Com todos os dados organizados e extraídos, a fase de análise terá insumos para sua execução. Ela concentra a maior parte da inteligência do sistema de vigilância tecnológica. Nela são feitos os cálculos, as associações e as demais operações necessárias para se construírem as informações com significado para os usuários e, por fim, são disponibilizadas em uma banco de dados. Logo após, inicia-se a última etapa: a difusão. Esta etapa consome as informações persistidas nos bancos de dados alimentando os *dashboards*, portal e buscador. Com os resultados da análise gerados, poder-se-ia gerar e enviar de boletins ou alertas para usuários previamente cadastrados.

4.4 CONSIDERAÇÕES FINAIS DO CAPÍTULO

O modelo para vigilância tecnológica automatizada em fontes de dados disponíveis na internet proposto foi apresentado neste capítulo. Ele oferece uma abstração em alto nível de tecnologias, informações estruturais, funcionais e de interações. Seus quatro módulo principais, divididos em coleta, preparação, análise e difusão, e seus dois módulos auxiliares, divididos em parametrização e persistência, foram detalhados e discutidos. Foi explicada a aderência do modelo à cenários de Big Data.

Em seguida, foi apresentada uma arquitetura de referência para implementar o modelo proposto, explorando com mais profundidade as questões tecnológicas associadas, como a sugestão de bancos de dados orientados à documentos para armazenar as informações de publicações coletadas, bancos relacionais para armazenar os dados gerados pelas análises automatizadas e arquivos para persistir dados das ontologias contendo os termos dos domínios tecnológicos a se monitorar.

Para a coleta automatizada das informações em suas fontes, foi proposto uso de agentes de softwares conhecidos como *webcrawlers*. Exemplificaram-se atividades para a preparação das informações como a limpeza dos documentos capturados e a extração de seus termos. Foram apresentados alguns tipos de análises para o módulo de análise, incluindo a análise de sentimentos sobre determinado termo, contagens do número de vezes em que um termo é mencionado nos documentos. Ao tratar da arquitetura, também se sugeriu a disponibilização de painéis gráficos (dashboards), buscadores especializados e portais agregadores de conteúdos coletados, alimentados automaticamente pelas atividades dos outros módulos como produtos da vigilância tecnológica.

Finalmente, um *workflow* foi detalhado a fim de facilitar o entendimento sobre como os dados transitam e são processados na arquitetura, partindo-se do momento em que usuários especialistas definem o escopo das atividades de vigilância tecnológica até o momento em que os interessados têm acessos as informações produzidas pelo sistema.

5 RESULTADOS EXPERIMENTAIS

Como resultados experimentais para esta pesquisa, tem-se um estudo de caso e as avaliações sobre o mesmo. Um detalhamento sobre o assunto é feito a seguir.

5.1 ESTUDO DE CASO

A avaliação do modelo proposto envolveu a realização de um estudo de caso para o qual foi desenvolvido um protótipo de sistema computacional de vigilância tecnológica automatizada que implementou os componentes e conceitos do modelo e da arquitetura de referência apresentados no capítulo 4.

Baseado no modelo, elaborou-se uma arquitetura de referência discutida na seção 4.2 e ilustrada na Figura 10. Ela viabiliza a automatização do fluxo de processamento e de dados. Arquitetura adota dois tipos de banco de dados: um banco relacional (ex: PostgreSQL) para armazenar metadados e os termos extraídos ou identificados das publicações e outro NoSQL próprio para lidar com Big Data, como MongoDB ou Apache Hive.

Este estudo de caso buscou atender às necessidades informacionais de vigilância tecnológica do Observatório da Indústria Catarinense - FIESC (Observatório FIESC, 2019). O observatório é mantido pela Federação das Indústrias de Santa Catarina sendo uma área voltada ao planejamento e desenvolvimento estratégico da indústria de Santa Catarina. Por meio dele são monitorados os principais fatores que afetam a competitividade industrial no Estado, utilizando-se da análise de desempenho econômico e das tendências tecnológicas dos setores estratégicos e, assim, disponibilizando as informações produzidas para indústrias e parceiros. O Observatório disponibiliza, também, o Portal Setorial da FIESC com informações relevantes aos empresários que buscam dados para tomadas de decisões relacionada à sua empresa. Neste portal, estão os resultados do PDIC (Programa de Desenvolvimento Industrial Catarinense).

Com o PDIC a FIESC busca identificar os setores indutores de desenvolvimento, as visões de futuro para cada setor, traçar o caminho mais provável para atingi-la e promover a articulação de todas as partes interessadas (FIESC, 2014). Os dois objetivos principais do programa são proporcionar uma dinâmica de prosperidade industrial de longo prazo e posicionar a indústria regional como protagonista do desenvolvimento do Estado.

Como forma de alcançar estes objetivos, a FIESC dividiu o programa em três grandes projetos, assim denominados:

1. Setores Portadores de Futuro para a Indústria Catarinense;
2. Rotas Estratégicas Setoriais;
3. Masterplan.

Este estudo de caso envolveu, como explicado anteriormente, o desenvolvimento de um protótipo de sistema computacional para vigilância tecnológica automatizada para atender

ao primeiro objetivo ao contribuir para o monitoramento das tecnologias-chaves dos setores portadores dos futuro gerando informações sobre inteligência competitiva para que os analistas da FIESC possam repassar às indústrias interessadas.

O público-alvo que participou deste estudo de caso e interagiu com o sistema de vigilância desenvolvido é formado por analistas de inteligência industrial e consultores de inovação do Sistema FIESC que desenvolvem suas atividades de atendimento institucional da FIESC, do Serviço Social da Indústria (SESI)/SC, do Serviço Nacional de Aprendizagem Industrial (SENAI)/SC, do Instituto Euvaldo Lodi (IEL)/SC e do Centro das Indústrias do Estado de Santa Catarina (CIESC) e das indústrias catarinenses, e cujas atividades são suportadas pelo monitoramento socioeconômico e de tendências tecnológicas e de mercado.

O desenvolvimento e a implantação do sistema construído neste estudo de caso em um cenário real trouxe experiência prática e conhecimentos tácitos que somados à revisão da literatura e ao referencial teórico aprimoraram o modelo proposto neste trabalho durante seu desenvolvimento. No sistema, foram utilizadas diferentes tecnologias como a linguagem de programação Python, devido a sua grande variedade de bibliotecas úteis na área de ciência de dados, o HTML e a linguagem Javascrpt para montar telas do sistema.

Para persistência dos dados utilizaram-se os bancos de dados relacionais PostgreSQL e Microsft SQL Server, e os bancos de dados NoSQL MongoDB e Hive. Os códigos foram hospedados em uma máquina virual (VM) na nuvem da Microsoft Azure. A configuração da VM possuía 7 GB de memória RAM, processador Intel(R) Xeon(R) com duas CPUs de 2,3GHz, sistema operacional Windows Server 2016 Data Center e 140Gb de disco. Outros softwares e tecnologias auxiliares também foram necessárias e serão detalhados nos próximos parágrafos.

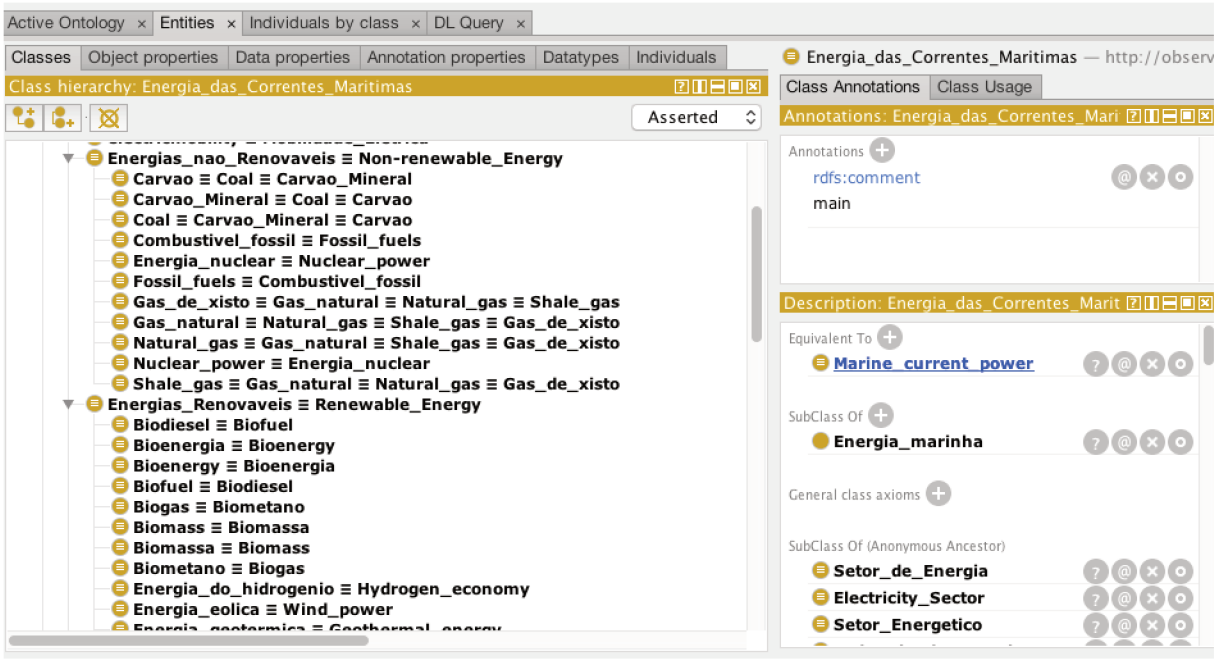
A fase de *planejamento* conforme recomendada por (SÁNCHEZ, J Marcela; PALOP, 2002) não está disponível no sistema já que este visa implementar o modelo proposto e parte do pressuposto de que as necessidades de informações dos usuários estão previamente definidas e modeladas. Sendo assim, cada um dos componentes do modelo foram implementados conforme detalhado a seguir.

A ***parametrização*** foi realizada com o auxílio de especialistas. Os domínios de conhecimento do estudo de caso foram estruturados em ontologias com base nos interesses dos Setores Portadores de Futuro para o Estado de Santa Catarina presentes nos cadernos disponibilizados pelo programa PDIC 2022. As *ontologias* contém as tecnologias-chaves e suas classes para cada setor, tendo sua validade atestada por especialistas. A hierarquia proposta para a ontologia do setor de Energia, por exemplo, contém a tecnologia de "Painel solar" pertencendo à classe de "Energias renováveis" e na classe "Energias não renováveis" está contida a tecnologia de "Energia nuclear". Para modelar as ontologias utilizou-se o editor Protegé [1] o qual gerou arquivos OWL. A figura 12 ilustra uma captura de tela do software com a ontologia de energia em edição.

Ao total foram modeladas 15 ontologias representando 15 setores industrias diferentes (Agroalimentar, Bens de Capital, Celulose e Papel, Cerâmica, Construção Civil, Economia do

[1] O software Protegé está disponível em https://protege.stanford.edu

Figura 12 – Tela do software Protegé com exemplo de ontologia de energia.



Fonte: tela capturada pelo autor.

Mar, Energia, Indústrias Emergentes, Meio Ambiente, Metal, Móveis e Madeira, Químicos e Plásticos, Saúde, Têxtil e Tic).

As fontes de informação, por sua vez, foram delimitadas em um arquivo indicando de onde a informação deveria ser buscada. Em conjunto, as fontes foram categorizadas de acordo com o tipo de fonte (rede social ou notícia) e de acordo com a fonte de informação (nome do Portal ou usuário monitorado).

Na implementação do ***módulo de coleta*** a captura das publicações se deu por meio do sistema Intellitotum da empresa Dígitro que funcionou como um agente WebCrawler buscando as informações das fontes (portais e redes sociais) de interesse periodicamente e que disponibilizou todas as publicações em uma única tabela de passagem em um banco de dados Microsoft SQL Server. Até 9 de maio do ano de 2019 haviam sido capturadas 878.375 publicações entre notícias de portais especializados e conteúdos das redes sociais Twitter, Facebook e Youtube.

O conceito do ***módulo de preparação*** foi implementado da seguinte forma. Nos conteúdos disponibilizados pela ferramenta de captura haviam muitos documentos duplicados. As datas não estavam em um formato adequado para manipulação, os nomes das fontes de notícias estavam concatenadas com os títulos entre outras características que precisaram ser normalizadas para a utilização deles. Para isso, foi construído um *parser* (conversor) que transformou o conteúdo em um formato mais amigável nesta fase de preparação.

No módulo de preparação, a informação capturada disponível na tabela de passagem foi importada por meio de scripts em Python denominados “Data Importer”. Ele foi responsável por remover conteúdos duplicados, identificar o tipo de cada publicação e importar para uma coleção

adequada dentro de um banco de dados não relacional MongoDB. Os conteúdos capturados do Twitter, por exemplo, foram importados para uma coleção chamada "twitter". Um exemplo de como o documento foi armazenado no formato JSON por ser visto no Código 5.1.

O MongoDB foi escolhido pela facilidade em se armazenar documentos no formato JSON, por ser escalável horizontalmente e ser um dos principais bancos utilizados para lidar com cenários de Big Data. No MongoDB criaram-se índices padrões sobre campos de IDs e URLs para facilitar a recuperação dos documentos e um índice especial que permite pesquisa textual sobre os campos de Título e Textos.

Código 5.1 – Amostra de documento da coleção twitter.

```
{
    "_id" : ObjectId("5b1d280f2e77caa637fef37a"),
    "ref_date" : ISODate("2017-10-23T19:59:07.000Z"),
    "old_id" : 1089331,
    "imported_at" : ISODate("2018-11-10T10:30:24.878Z"),
    "url" : "http://twitter.com/statuses/922582908371550208",
    "font" : "FoodInsight International Food Information",
    "tweet" : "Claims about #superfoods are just that, they are not proven
    scientific facts - http://undrarmr.co/2i2b3aO Via @MyFitnessPal #health
    ",
    "category_id" : "Agroalimentar",
    "category_description" : "Agroalimentar"
}
```

Fonte: elaborado pelo autor.

O módulo de preparação teve, também, como ator a classe Python “Metadata Extractor” que extraiu as tecnologias-chaves e os nomes de países de cada um dos documentos armazenados, gerando milhares de metadados. O “Metadata Extractor” minerou o corpo e o título de cada documento em busca dos termos das ontologias contida nos arquivos OWL de cada setor. A extração recebeu como entrada dados não estruturados (textos) e gerou uma coleção de dados estruturados (metadados). Seu resultado foi armazenado e indexado em um banco relacional (PostgreSQL).

Para implementar o conceito do ***módulo de análise*** foi construído um conjunto de scripts (programas de computador) na linguagem Python que compuseram um componente intitulado “Science Builder”, o qual executa consultas para identificar o número de menções sobre as tecnologias-chaves e países nas publicações agrupando ambas as análises em semanas, gerando séries temporais. Com isso, foi possível oferecer repostas para perguntas sobre quais países são mais citados quando um documento fala sobre determinada tecnologia. Também, foi construído um componente chamado de “Sentiment Extractor” utilizando como base a biblioteca de processamento de linguagem natural em Python chamada TextBlob para extrair os sentimentos das publicações associando-os com as tecnologias mencionadas.

O ***módulo de difusão***, segundo o modelo proposto, é o responsável por entregar os produtos da vigilância tecnológica. Para validá-lo, foram construídos quatro tipos de itens:

- Dashboards no Microsoft PowerBI [2];
- Boletins tecnológicos;
- Portal de Tendências;
- Buscador especializado de publicações.

Os painéis interativos ou *dashboards* feitos no Microsoft PowerBI têm seus dados atualizados automaticamente toda vez que uma nova análise é realizada e seus resultados são disponibilizados no banco de dados. Assim, periodicamente os usuários terão experiências e informações diferentes ao interagirem com os painéis. No total foram disponibilizados quatro painéis especializados e automatizados.

O *Painel de Monitoramento Tecnológico Geral* (Figura 13) é generalista e apresenta um panorama sobre as publicações capturadas e processadas. Ele traz gráficos dos termos mais mencionados no formato de nuvem de palavras onde aqueles mais citados são destacadas com uma fonte maior e os menos citados com fontes menores. Apresenta ainda informações sobre o volume de documentos capturados, a distribuição dos termos mencionados e a conexão entre conhecimento e tecnologia-chave.

Por sua vez, o comportamento das tecnologias-chaves ao longo do tempo pode ser visto no Painel de Monitoramento Tecnológico Temporal (Figura14), onde também são apresentadas as "top"cinco áreas conhecimentos e tecnologias chaves mais presentes nas publicações.

No *Painel de Monitoramento Tecnológico Geográfico* (Figura15) são exibidas aos usuários as conexões entre tecnologias-chaves e países e ainda os países mencionados quando se clica sobre uma determinada tecnologia. Neste painel, ao se clicar sobre um país as conexões entre as tecnologias são destacadas também.

Por fim, como resultado do processamento em linguagem natural feito sobre as publicações, é disponibilizado o *Painel de Monitoramento Tecnológico com Sentimentos* (Figura16). Este traz três indicadores de sentimentos (positivo, neutro e negativo) sobre cada setor com base na análise de sentimentos. Os resultados podem ser filtrados por países mostrando como é avaliação sobre as tecnologias em cada país.

Conforme citado, foi desenvolvido um portal que agrega todas as publicações, com exceção das redes sociais, capturadas e inclui marcações com as tecnologias-chaves monitoradas e presentes nelas. Ele foi chamado de *Portal de Tendências* e está disponível no Portal Setorial da FIESC. Na Figura 17 é possível ter uma visão do mesmo. No portal é possível filtrar as publicações de acordo com o setor industrial, como pelo setor Agroalimentar ou Cerâmica, clicando sobre o menu esquerdo conforme destacado na figura. As publicações estão dispostas

[2] Microsoft PowerBI é um sistema de business intelligence da Microsoft o qual é possível conectar a fontes de dados e visualizá-los de forma interativa.

Figura 13 – Monitoramento Tecnológico - Geral.

Fonte: FIESC. Capturada em maio de 2019.

Figura 14 – Monitoramento Tecnológico - Temporal.

Fonte: FIESC. Capturada em maio de 2019.

em *cards* (quadrados) com o título, data de publicação, fonte, data e horário da publicação, e ainda as categorias as quais pertencem, que neste caso são as classes das tecnologias de interesse

Figura 15 – Monitoramento Tecnológico - Geográfico.

Fonte: FIESC. Capturada em maio de 2019.

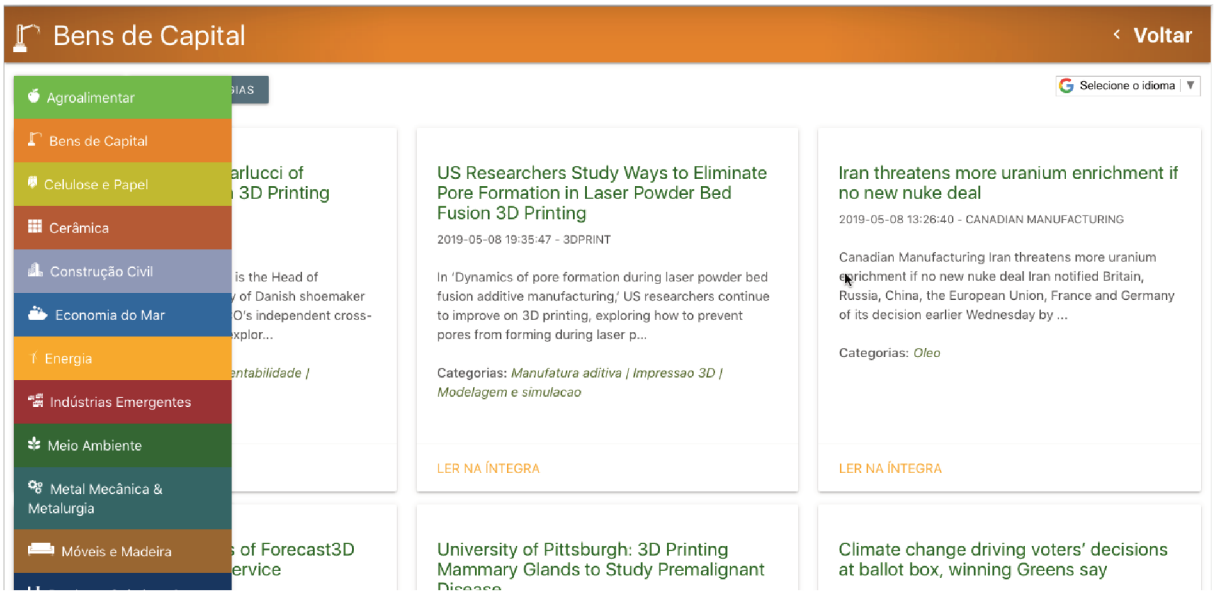
Figura 16 – Monitoramento Tecnológico - Sentimentos.

Fonte: FIESC. Capturada em maio de 2019.

presentes nas ontologias estruturadas pelos especialistas.

Uma ação desenvolvida pelo Observatório são os boletins digitais ou alertas nos quais se

Figura 17 – Portal de Tendências.



Fonte: FIESC. Capturada em maio de 2019.

resumem as publicações mais relevantes e se enviam aos interessados quinzenalmente. Eles são chamados de Monitores e não estão no escopo do protótipo desenvolvido por este trabalho, sendo por enquanto uma atividade semi-automatizada. As publicações mais relevantes são sugeridas pelo sistema implantado por meio da ferramenta *TechMonitor* por onde os especialistas mineram as publicações para escolher aquelas que mais se considera relevantes aos setores industriais.

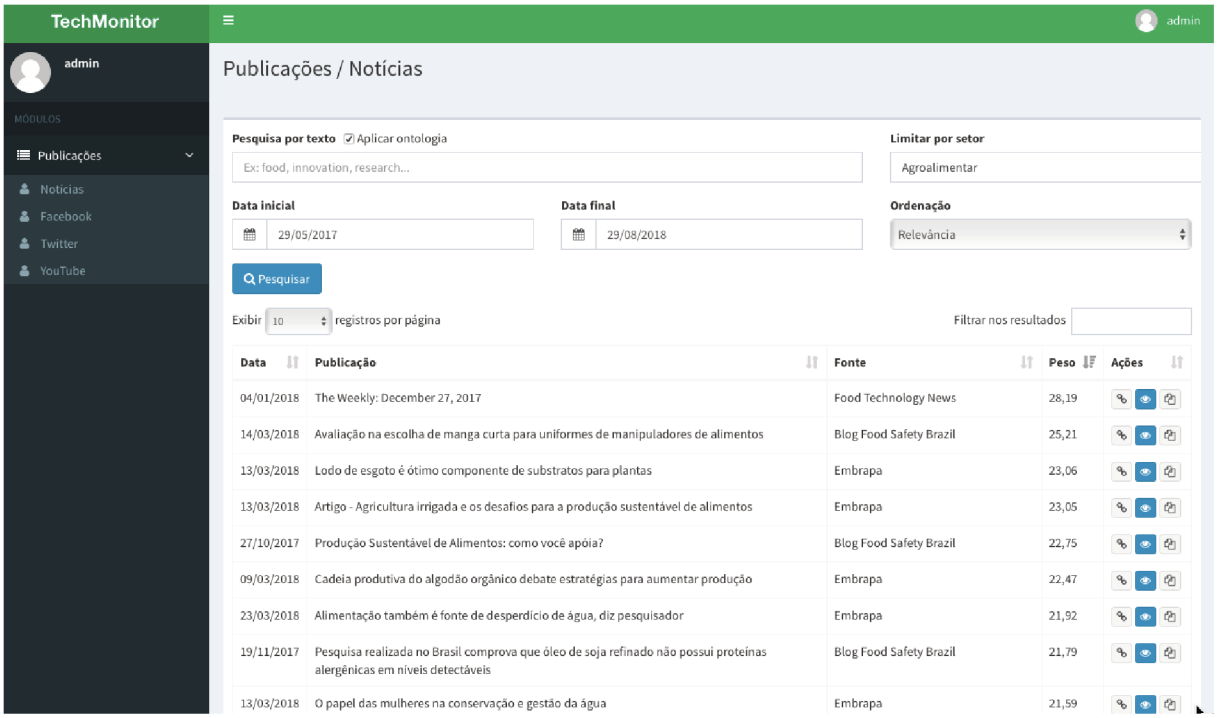
O *TechMonitor* foi desenvolvido como um sistema de recuperação de informação (RI) no formato de buscador especializado que permite localizar textos dentro das publicações coletadas ou recuperar as publicações de interesse. Ele possui funcionalidades importantes como a recuperação de informação baseada em relevância (calculada pela presença de termos das ontologias nas publicações). Segundo as ideias de Souza Inácio *et al.* (2014, p. 501, tradução nossa):

> "Uma das etapas mais importantes durante o processo de pesquisa em sistemas de RI é a previsão de quais documentos devem ser considerados relevantes para alguma tarefa de pesquisa e quais devem ser descartados. Essa tarefa de seleção é realizada por um algoritmo que, com base em heurísticas previamente definidas, decide quais documentos devem ser considerados relevantes para recuperação e os classifica de acordo com os critérios de relevância estabelecidos por essas heurísticas." .

Na Figura 18 tem-se uma visão do TechMonitor. Ela mostra o campo de pesquisa onde se pode digitar o texto a ser pesquisado. Adicionalmente, pode-se optar por considerar as ontologias das tecnologias-chaves nas pesquisas. Quando se seleciona o tipo de ordenação por "Relevância" conforme mostrado na figura, os documentos recuperados são ordenados por um peso relativo que indica o volume de menções de tecnologias de interesse nos documentos. As principais funcionalidades disponíveis no sistema são:

- Pesquisa por publicações contendo um ou mais termos.

Figura 18 – Captura de tela do TechMonitor.



Fonte: FIESC. Capturada em maio de 2019.

- Pesquisa por publicações contendo itens das ontologias dos setores.
- Recuperação de publicações por data ou relevância.
- Marcação das publicações com as tecnologias-chaves.
- Tradução automática dos textos.
- Organização das publicações por tipo e fonte.

O sistema desenvolvido neste estudo de caso está em uso pela equipe do Observatório e ganhou importância na instituição por facilitar o trabalho de sua equipe de especialistas e oferecer informações extremamente atualizadas para os seus usuários. Seu desenvolvimento foi baseado em uma versão inicial do modelo proposto e permitiu a seus conceitos abstratos evoluírem para uma ferramenta prática.

6 ANÁLISE E DISCUSSÃO DO EXPERIMENTO

Este capítulo descreve a análise realizada sobre o experimento elaborado para avaliar o modelo proposto. Com base nos resultados obtidos pela avaliação são feitas comparações com cenários anteriores sobre como o modelo proposto impactou o processo de vigilância tecnológica na instituição onde os experimentos ocorreram. Na primeira parte deste capítulo são apresentados o instrumento de avaliação, seu público-alvo e os aspectos avaliados. Na segunda parte, são discutidos os resultados obtidos e feitas as comparações relevantes.

6.1 INSTRUMENTO DE AVALIAÇÃO DO EXPERIMENTO

No capítulo anterior, detalhou-se o estudo de caso realizado Observatório FIESC que contemplou o desenvolvido e a implantação de um sistema de vigilância tecnológica automatizada baseado no modelo conceitual elaborado. O sistema teve como público-alvo analistas de inteligência industrial e consultores de inovação do Sistema FIESC que atuam no atendimento institucional das indústrias catarinenses por meio da FIESC, SESI/SC, SENAI/SC, IEL/SC e CIESC.

Antes do estudo de caso e da disponibilização do protótipo de sistema de vigilância tecnológica automatizada, o Observatório dispunha da ferramenta Intellitotum da empresa Dígitro para capturar e visualizar publicações de portais web e redes sociais. Quando se desejava construir uma análise mais sofisticada era necessário exportar os dados manualmente e os importar em alguma ferramenta especializada para criação de gráficos ou execução dos cálculos. A leitura das publicações tomava muito tempo dos especialistas porque o sistema de ordenação dos textos, apesar de ordenar por um nível básico de relevância, exigia uma complexa configuração e a inclusão ou alteração de domínios do conhecimento exigia uma complexa atividade de configuração. Pode dizer, que não era fácil recuperar as publicações coletadas de acordo com uma relevância relativa baseada na incidência de tecnologias-chaves no corpo do texto para um setor específico da indústria. Além disso, era necessário resumir e traduzir todos os textos de forma manual.

Para avaliar o experimento realizado, foi elaborado um questionário estruturado para extrair as percepções dos especialistas sobre os ganhos obtidos com a implantação do sistema de vigilância automatizada baseada no modelo proposto no Capítulo 4 em relação ao processo de vigilância que se utilizava na organização antes do estudo de caso. Sua aplicação teve como público-alvo os especialistas que participaram do estudo, ou seja, analistas de inteligência industrial e consultores de inovação do Sistema FIESC e que preferencialmente participaram do PDIC 2022 (Programa de Desenvolvimento Industrial 2022) onde contribuíram no mapeamento dos dezesseis "Setores Portadores de Futuro para a Indústria Catarinense"e pelo qual identificaram as tecnologias-chave para os diferentes setores industriais.

O conteúdo do questionário foi dividido em sete dimensões e 18 questões, sendo cinco dimensões relacionadas à avaliação dos produtos do ciclo de vigilância tecnológica e duas

relacionadas à pergunta de pesquisa e ao objetivo deste trabalho. Ele foi construído em duas etapas. Na primeira etapa, foram elaboradas questões para verificar se a pergunta de pesquisa foi respondida com o modelo e se o objetivo do trabalho foi atingido do ponto de vista dos usuários. Na segunda etapa incluíram-se questões para avaliação das funções e produtos de Vigilância Tecnológica e Inteligência Competitiva e sua implementação através de plataformas web por meio dos critérios propostos por Berges-Garcia *et al.* (2016).

As dezoito questões elaboradas e presentes no **questionário final** enviado aos especialistas estão listadas no Quadro 8. Elas possuem respostas no formato de múltipla escolha conforme o Quadro 7, sendo solicitado que se assinale apenas uma.

Quadro 7 – Opções de respostas do questionário.

Código	**Resposta**
R1	Discordo totalmente
R2	Discordo
R3	Neutro
R4	Concordo
R5	Concordo totalmente

Fonte – elaborado pelo autor.

6.1.1 Questões relacionadas às funções e produtos de Vigilância Tecnológica

O modelo e a arquitetura propostos são divididos em componentes ou módulos, os quais se responsabilizam por automatizar atividades dentro do ciclo ou processo de vigilância tecnológica. No modelo foram propostos quatro módulos principais (Coleta, Preparação, Análise e Difusão) e dois módulos auxiliares (Parametrização e Armazenamento) que juntos oferecem o fluxo necessário para se automatizar o processo de vigilância. As questões a seguir, buscam coletar as percepções e opiniões dos especialistas em relação aos impactos do modelo em seu trabalho tendo em vista que o mesmo foi implementado no protótipo utilizado por eles. Os critérios avaliados foram baseados no trabalho de Berges-Garcia *et al.* (2016) conforme mencionado anteriormente. Os autores propõe um conjunto de indicadores para se avaliar plataformas web de VT / CI sob a ótica do conjunto geral de funções do ciclo de VT / CI oferecidas, permitindo que seja escolhida a solução mais apropriada de acordo com suas necessidades e circunstâncias específicas.

O questionário ateve-se aos cinco critérios obre as funções associadas as etapas do ciclo global de VT/IC e propostos pelos autores: (critério 1) identificação de necessidades, busca e extração de informação; (critério 2) filtragem e avaliação de informações; (critério 3) análise de informações; e (critério 5) disseminação. As questões estão agrupadas e detalhadas abaixo.

- **Dimensão 1 (D1)**: Questões elaboradas para avaliar o modelo quanto à identificação das necessidades, busca e extração de informação:

  - O sistema impactou positivamente no processo coleta e organização das publicações?
  - A inclusão de novas tecnologias-chaves foi facilitado com o modelo implementado?

- **Dimensão 2 (D2)**: Questões elaboradas considerando o critério quanto à filtragem e valorização da informação:
  - O sistema extraiu as tecnologias de interesse e metadados (título, texto, data, etc) das publicações satisfatoriamente?
  - A interpretação do conjunto de publicações foi facilitada pelo novo sistema com a manipulação e extração de dados como tags, categorias e tradução automática?

- **Dimensão 3 (D3)**: Questões elaboradas considerando quanto à análise da informação:
  - O sistema facilitou a análise gráfica das informações coletadas?
  - O sistema melhorou as discussões sobre o monitoramento tecnológico entre os analistas e interessados?

- **Dimensão 4 (D4)**: Questões elaboradas considerando os critérios quanto à inteligência estratégica:
  - O sistema aumentou a capacidade de analisar as publicações e informações?
  - As análises produzidas pelo sistema (correlações, contagens, análise de sentimentos) agregaram valor às informações?

- **Dimensão 5 (D5)**: Questões elaboradas para avaliar o modelo quanto aos critérios de difusão, ou seja, quanto a sua eficácia em entregar ou garantir o acesso à informação aos interessados:
  - O sistema reduziu a carga de trabalho de leitura das publicações coletadas?
  - O sistema facilitou a visualização gráfica das informações coletadas?
  - O sistema simplificou a busca pelas informações/publicações de interesse?

6.1.2 Questões relacionadas à pergunta de pesquisa e objetivo geral

A pergunta de pesquisa que este trabalho busca responder é "Como identificar e monitorar de forma automatizada e constante tecnologias de interesse de segmentos industriais catarinenses a partir de múltiplas fontes cujos dados e comportamentos caracterizem cenários de Big Data?", a qual surgiu em um cenário onde é preciso manter um panorama atualizado sobre o comportamento das tecnologias de interesse para o setor industrial do Estado de Santa Catarina, facilitando a identificação de ameaças ou oportunidades.

Para avaliar se o modelo, materializado através do protótipo, trouxe respostas à pergunta de pesquisa, foram elaboradas as seguintes questões da **Dimensão 6 (D6)**:

- A identificação das tecnologias-chaves nas publicações e portais de interesse foi facilitada com o novo sistema?
- O monitoramento do comportamento das tecnologias-chaves nas fontes foi facilitado com o sistema?
- O sistema facilitou a entrega de dados para as demandas informacionais dos segmentos industriais catarinenses descritos no PDIC2022?
- O sistema ajudou a lidar melhor com o monitoramento em cenários de Big Data?

A avaliação do objetivo deste trabalho, "Propor um modelo para a realização de vigilância tecnológica automatizada em fontes disponíveis eletronicamente como artigos de portais web ou rede sociais", foi avaliado na **Dimensão 7 (D7)**, sendo composto por três questões. As duas primeiras avaliam se a abordagem trouxe melhorias na automatização do processo de VT e a terceira busca entender se ela poderia ser utilizada por outras organizações. As questões estão listadas a seguir.

- O sistema melhorou a automatização da vigilância tecnológica com artigos disponíveis em portais web?
- O sistema melhorou a automatização da vigilância tecnológica com redes sociais?
- Você acredita que o modelo implementado possa ser implementado em outras organizações?

6.2 ANÁLISE DAS AVALIAÇÕES COLETADAS

O questionário, composto com as perguntas de múltipla escolha, foi construído com a ferramenta on-line de elaboração de formulários Google Forms [1], a qual oferece ótima usabilidade e simplicidade para quem responde as questões, além de ficar disponível em um site web 24 horas por dia. O link para acesso ao formulário on-line foi compartilhado de forma privada e enviado por e-mail para um grupo de dez especialistas aos quais se garantiu o anonimato das respostas. Os especialistas selecionados para receber o questionário foram aqueles que tiveram contato direto com os produtos ofertados pelo protótipo desenvolvido no estudo de caso e que atuaram com monitoramento tecnológico no Observatório FIESC.

As respostas foram enviadas automaticamente a uma planilha digital (Google Spreadsheet) onde os dados foram consolidados. Segundo Filippo *et al.* (2011), quando um questionário é enviado por e-mail, o retorno médio esperado é de 30%. Quando é acessível via um site, a taxa de retorno média é de 40%. Contudo, estudos apontam que a taxa de resposta de um questionário em ambiente acadêmico pode ter taxas de respostas menores que 20% (VAN MOL, 2017). Neste

[1] https://forms.google.com

Quadro 8 – Questões do questionário.

Código	**Questão**
Q1	O sistema impactou positivamente no processo coleta e organização das publicações?
Q2	A inclusão de novas tecnologias-chaves foi facilitado com o modelo implementado?
Q3	O sistema extraiu as tecnologias de interesse e metadados (título, texto, data, etc) das publicações satisfatoriamente?
Q4	A interpretação do conjunto de publicações foi facilitada pelo novo sistema com a manipulação e extração de dados como tags, categorias e tradução automática?
Q5	O sistema facilitou a análise gráfica das informações coletadas?
Q6	O sistema melhorou as discussões sobre o monitoramento tecnológico entre os analistas e interessados?
Q7	O sistema aumentou a capacidade de analisar as publicações e informações?
Q8	As análises produzidas pelo sistema (correlações, contagens, análise de sentimentos) agregaram valor às informações?
Q9	O sistema reduziu a carga de trabalho de leitura das publicações coletadas?
Q10	O sistema facilitou a visualização gráfica das informações coletadas?
Q11	O sistema simplificou a busca pelas informações/publicações de interesse?
Q12	A identificação das tecnologias-chaves nas publicações e portais de interesse foi facilitada com o novo sistema?
Q13	O monitoramento do comportamento das tecnologias-chaves nas fontes foi facilitado com o sistema?
Q14	O sistema facilitou a entrega de dados para as demandas informacionais dos segmentos industriais catarinenses descritos no PDIC2022?
Q15	O sistema ajudou a lidar melhor com o monitoramento em cenários de Big Data?
Q16	O sistema melhorou a automatização da vigilância tecnológica com artigos disponíveis em portais web?
Q17	O sistema melhorou a automatização da vigilância tecnológica com redes sociais?
Q18	Você acredita que o modelo implementado possa ser implementado em outras organizações?

Fonte – elaborado pelo autor

experimento a taxa de resposta foi de 70%, possivelmente porque alguns dos especialistas já não faziam parte do quadro de colaboradores da empresa onde se empreendeu o estudo de caso.

Para fins de análise, as respostas estão dispostas nos quadros em percentuais relativos. A primeira coluna representa o código da questão e as demais os códigos das respostas. A avaliação do modelo quanto à identificação das necessidades, busca e extração de informação também trouxeram uma perspectiva positiva na qual concordam ou concordam totalmente que o sistema impactou positivamente no processo coleta e organização das publicações e na inclusão de novas tecnologias-chaves, conforme os dados coletados e expostos no Quadro 10.

Quadro 9 – Avaliação do modelo quanto à identificação das necessidades, busca e extração de informação.

Questão	**R1**	**R2**	**R3**	**R4**	**R5**
Q1	0	0	0	28,6%	71,4%
Q2	0	0	0	57,1%	42,9%

Fonte – elaborado pelo autor.

A filtragem e valorização da informação foram bem atendidas pelo modelo conforme a totalidade dos especialistas que responderam o questionário. O item mais bem avaliado, com 57,1% das respostas confirmando que concordam totalmente, foi a melhoria trazida para a interpretação do conjunto de publicações cuja facilitação foi promovida pelo protótipo disponibilizado no estudo de caso com a manipulação e extração de dados como tags, categorias e tradução automática.

Quadro 10 – Avaliação do modelo quanto à filtragem e valorização da informação.

Questão	**R1**	**R2**	**R3**	**R4**	**R5**
Q3	0	0	0	57,1%	42,9%
Q4	0	0	0	42,9%	57,1%

Fonte – elaborado pelo autor.

As estatísticas apontadas pelo Quadro 11 qualificam o modelo como uma alternativa viável para oferecer uma ferramenta útil para discussões sobre monitoramento tecnológico entre especialistas e ainda como um facilitador em suas análises gráficas principalmente, tendo recebido a respostas “concordo totalmente” por 71,4% dos respondentes.

Quadro 11 – Questões elaboradas considerando quanto à Análise da informação.

Questão	**R1**	**R2**	**R3**	**R4**	**R5**
Q5	0	0	14,4%	42,9%	42,9%
Q6	0	0	0	28,6%	71,4%

Fonte – elaborado pelo autor.

Conforme citado em capítulo anterior, o estudo de caso disponibilizou alguns produtos informacionais pelo módulo de difusão. Estes produtos, compostos por *dashboards*, portal especializado e ferramenta de busca nas publicações coletados. Segundo 57,1% dos especialistas o sistema aumentou suas capacidades de analisar as publicações e informações coletadas. Ainda, para 71,4%, as análises automatizadas executadas pelo sistema como correlações, contagens e análise de sentimentos agregaram valor às informações (Quadro 12).

Na avaliação do modelo quanto aos critérios de difusão, aproximadamente a metade dos respondentes concordam e a outra metade concorda totalmente que o sistema proporcionou um redução na carga de trabalho na leitura das publicações coletadas, facilitou a visualização gráfica das informações e ainda simplificou a busca pelas informações de interesse (Quadro 13).

Quadro 12 – Questões elaboradas considerando os critérios quanto à inteligência estratégica.

Questão	**R1**	**R2**	**R3**	**R4**	**R5**
Q7	0	0	0	57,1%	42,9%
Q8	0	0	0	28,6%	71,4%

Fonte – elaborado pelo autor.

Quadro 13 – Questões elaboradas para avaliar o modelo quanto aos critérios de difusão.

Questão	**R1**	**R2**	**R3**	**R4**	**R5**
Q9	0	0	0	57,1%	42,9%
Q10	0	0	0	42,9%	57,1%
Q11	0	0	0	57,1%	42,9%

Fonte – elaborado pelo autor.

No Quadro 14 estão listadas as respostas para as perguntas que avaliarão se o modelo foi capaz de solucionar a pergunta de pesquisa. Para 71,4% os especialistas, o protótipo baseado no modelo facilitou a identificação das tecnologias-chaves nas publicações e portais de interesse, o monitoramento do comportamento das tecnologias-chaves nas fontes, a entrega de dados para as demandas informacionais dos segmentos industriais catarinenses integrantes do projeto PDIC2022 e a operacionalizar o monitoramento no cenário de Big Data em que eles estão inseridos. Apenas 14,3% dos especialistas entendem que não houve ganhos significativos em relação a identificação das tecnologias-chaves e a operacionalização da vigilância nos cenários de Big Data.

Quadro 14 – Avaliação da Pergunta de Pesquisa.

Questão	**R1**	**R2**	**R3**	**R4**	**R5**
Q12	0	0	14,3%	14,3%	71,4%
Q13	0	0	0	28,6%	71,4%
Q14	0	0	0	28,6%	71,4%
Q15	0	0	14,3%	28,6%	57,1%

Fonte – elaborado pelo autor.

O Quadro 15 apresenta as respostas para as perguntas relacionadas ao objetivo deste trabalho. No geral, a avaliação dos especialista confirma que o objetivo foi atingido, sendo que a maioria das reposta se acumulam entre “concordo” e “concordo totalmente”. 14,3% dos especialistas não tem uma opinião formada sobre a possibilidade de se utilizar o mesmo modelo em outras organizações.

O gráfico da Figura 19 permite visualizar proporcionalmente a quantidade de opções escolhidas dentre respostas possíveis (R1, R2, R3, R4 e R5) dadas pelos especialistas para cada uma das dezoito questões do questionário. Por ele, infere-se que a abordagem proposta nesta dissertação foi entendida como superior àquela anteriormente utilizada por eles. Nota-se,

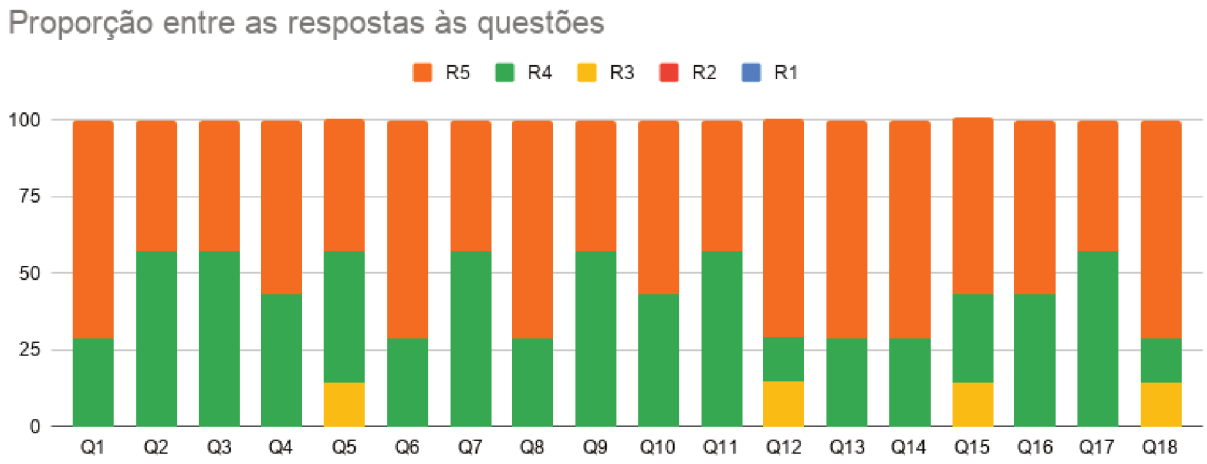
Quadro 15 – Avaliação do Objetivo do trabalho.

Questão	**R1**	**R2**	**R3**	**R4**	**R5**
Q16	0	0	0	42,9%	57,1%
Q17	0	0	0	57,1%	42,9%
Q18	0	0	14,3%	14,3%	71,4%

Fonte – elaborado pelo autor.

também, que a maior parte das respostas recaiu sobre a opção "concordo totalmente" reforçando a validação da abordagem. Não se encontraram respostas "discodo" ou "discordo totalmente", tendo havido apenas três perguntas onde 14,3% dos participantes se mostraram neutros em relação a elas (Q5, Q12, Q15 e Q18).

Figura 19 – Gráfico das proporções entre as respostas às questões do questionário.

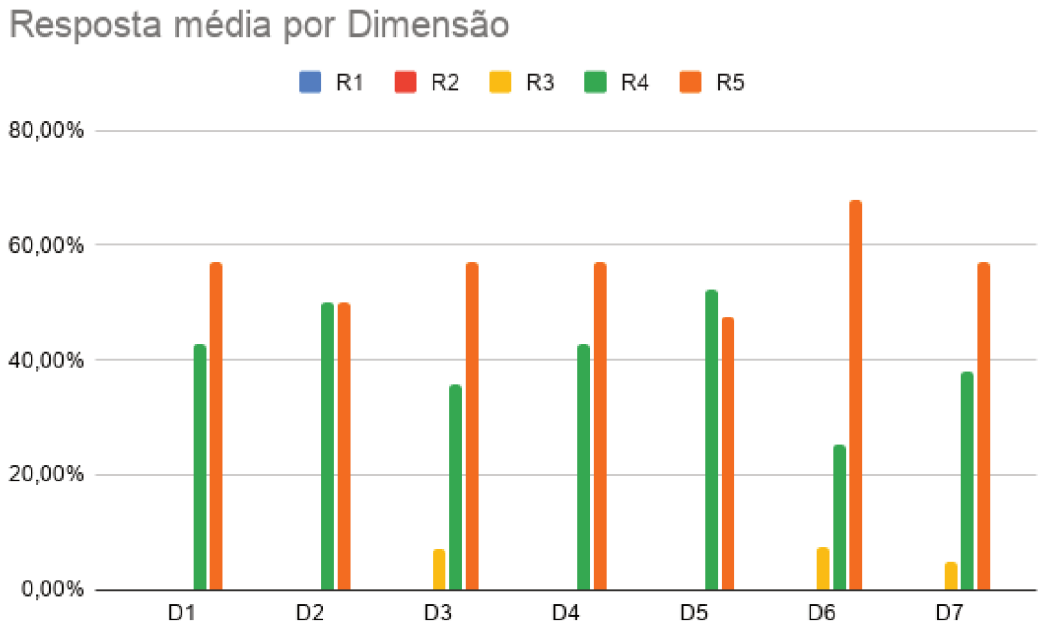


Fonte – elaborado pelo autor.

6.3 DISCUSSÕES FINAIS DOS EXPERIMENTOS

Os resultados atingidos na validação da abordagem proposta trouxeram maior clareza sobre os pontos fortes e fracos do modelo conceitual. A estruturação do questionário para avaliar pontos específicos facilitou sua discussão e trouxe maior clareza sobre a ótica dos especialistas. Dentre as sete dimensões analisadas, a dimensão seis (D6), que avaliou a capacidade da abordagem proposta responder a pergunta de pesquisa, foi a mais bem computada pelos especialistas, conforme pode ser visto no gráfico da Figura 20, onde as proporções das respostas foram agrupadas por dimensão e sua média foi extraída. Por outro lado, as questões que analisaram o modelo com base nos critérios de difusão (D5) obtiveram maior quantidade de respostas "concordo" relativamente.

Figura 20 – Gráfico com a proporção do somatório das respostas agrupadas por dimensões.



Fonte – elaborado pelo autor.

Para que um modelo tenha sucesso, pressupõe-se que seja possível reutilizá-lo em outras aplicações. Para 71,4% dos especialistas que participaram do estudo de caso e avaliaram a abordagem, esta características está presente na mesma. O resultado positivo das questões Q16 e Q17 que aferiram o potencial de automatização obtido corrobora para a eficácia da proposta. Não houve registros de repostas com as opções "discordo" ou "discordo totalmente", tendo apenas quatro perguntas (Q5, Q12, Q15, Q18) que receberam 1 (uma) resposta "neutro" de alguns especialistas.

6.4 CONSIDERAÇÕES FINAIS DO CAPÍTULO

Neste capítulo foi apresentado o instrumento de avaliação utilizado como parte do processo de validação da abordagem proposta. Ele foi estruturado no formato de questionário com 18 questões de múltipla-escolha divididas em 7 dimensões pelas quais se avaliaram aspectos específicos sobre as contribuições do modelo.

O questionário foi enviado e respondido por especialistas da FIESC que utilizaram o sistema desenvolvido com base no modelo conceitual e na arquitetura derivada. As respostas mostraram um alto grau de aprovação da abordagem, sendo ela avaliada como positiva em todas as dimensões analisadas. Assim, infere-se que o estudo de caso e sua avaliação demonstraram a validade da abordagem proposta. A escolha de implementar um protótipo de sistema completo de vigilância tecnológica incorporando os conceitos do modelo e da arquitetura, mesmo demandando

um esforço significativo, trouxe clareza sobre a aplicabilidade do modelo conceitual e serviu como uma referência objetiva para a discussão junto aos especialistas da organização participante.

7 CONCLUSÕES E TRABALHOS FUTUROS

Esta dissertação apresentou um modelo e uma arquitetura de referência para o criação de sistemas de vigilância tecnológica automatizada em fontes disponíveis eletronicamente considerando cenários de Big Data. Ambas tiveram seus conceitos e componentes implementados na forma de protótipo de sistema funcional de vigilância tecnológica automatizada que foi utilizado em um estudo de caso realizado no Observatório FIESC. Especialistas utilizaram o sistema, interagindo com seus componentes e discutindo seus conceitos. Ao final, avaliaram positivamente a abordagem proposta.

O modelo foi generalizado a partir de uma extensa revisão da literatura relacionada que avaliou processos, métodos e ferramentas, além da análise de plataformas integrais para VT disponíveis no mercado. Assim, o modelo, a arquitetura de referência, seu *workflow* e a revisão da literatura realizada são as principais contribuições deste trabalho. Estas e outras contribuições são melhor detalhadas nos próximos parágrafos.

O primeiro objetivo específico exigiu a revisão dos principais conceitos sobre vigilância tecnológica, os tipos existentes, suas definições e as principais ferramentas utilizadas, como disparadores de alertas, buscadores especializados e softwares para VT. Sobre estes últimos, foi feito um levantamento dos principais disponíveis no mercado e comparados por meio de critérios relativos ao ciclo da VT, proporcionando um panorama sobre as funcionalidades mais comuns entre entre eles. Por meio de uma revisão sistemática da literatura, verificou-se o estado da arte sobre vigilância tecnológica identificando os principais métodos utilizados atualmente, os tipos de documentos e de fontes de dados monitoradas e as propostas de arquiteturas e protótipos desenvolvidos pelos autores.

O levantamento bibliográfico em conjunto com a revisão sistemática da literatura estruturaram o arcabouço teórico necessário para instrumentar o trabalho para se atuar sobre a elaboração de um modelo conceitual (2° objetivo específico) para identificar e monitorar tecnologias em fontes disponíveis na internet cuja captura e processamento integrem um cenário de Big Data. A partir do modelo conceitual foi projetada uma arquitetura para dar suporte ao desenvolvimento de sistemas de VT automatizados contendo componentes similares ao modelo, porém com maiores detalhamentos técnicos e tecnológicos. Também, foi apresentado o *workflow* da arquitetura proposta.

No sentido de atender ao terceiro e ao quarto objetivos específicos desta dissertação que envolviam o desenvolvimento de um estudo de caso prático para aplicar o modelo e a avaliação da abordagem proposta junto à especialistas, foi conduzido um estudo de caso nas dependências da Federação das Indústrias de Santa Catarina (FIESC) para o qual foi desenvolvido e implantado um protótipo de sistema de vigilância tecnológica automatizada derivado da arquitetura e consequentemente do modelo conceitual propostos neste trabalho. O protótipo cobriu as etapas mais comuns de um sistema vigilância tecnológica, como a busca das informações, limpeza, organização e armazenamento, a identificação das tecnologias nos

documentos, análises envolvendo cálculos e cruzamentos dados, e a comunicação dos resultados aos interessados, automatizando-as por meio de seus componentes.

O protótipo coletou mais de 800 mil publicações de diferentes fontes de dados, incluindo portais e redes sociais. As publicações foram filtradas, organizadas e tiveram as tecnologias de interesse (tecnologias-chaves para os segmentos industriais catarinense) identificadas assim como outros metadados. Os resultados dos processamentos alimentaram de forma automatizada *dashboards* na tecnologias Microsoft PowerBI, um portal próprio (Portal de Tendências) e, ainda, foram disponibilizadas em uma ferramenta construída em conjunto com o protótipo chamada TechMonitor com a função de um buscador especializado permitindo que os especialistas recuperem publicações através de pesquisas que consideram a relevância de cada termo pesquisado ou ainda trazem resultados das pesquisas priorizando aqueles que possuam algumas das tecnologias chaves pré-definidas durante sua parametrização.

O segundo instrumento utilizado para avaliar o modelo foi a aplicação de um questionário contendo dezoito questões de múltipla escolha que cobriram sete dimensões ou aspectos que se queriam aferir, detalhadas no Capítulo 6, como o impacto do modelo no processo de monitoramento tecnológico ou ao ganho de capacidade dos especialistas em analisarem as informações coletadas.

Os dados obtidos pela aplicação do questionário demonstram que a abordagem proposta neste trabalho trouxe ganhos na totalidade dos processos de vigilância tecnológica. Os impactos positivos se estenderam nas sete dimensões avaliadas, tendo atendido à pergunta de pesquisa, o objetivo geral, como também às funções e produtos de Vigilância Tecnológica, ou seja, a identificação das necessidades, busca e extração de informação, a filtragem e valorização da informação, a análise da informação, a inteligência estratégica e sua difusão. Para todas as dimensões avaliadas, mais de 80% dos especialistas concordam ou concordam totalmente que a abordagem trouxe impactos positivos ao processo de vigilância tecnológica de sua organização. Tendo em vista os aspectos mencionados é possível inferir que o modelo proposto obteve sucesso ao sistematizar e permitir a construção de um sistema de VT automatizado no qual as informações foram coletadas, analisadas e difundidas sem a necessidade de intervenção humana.

Em síntese, esta dissertação partiu de um problema de pesquisa sobre processos de vigilância tecnológica que após seu estudo originou uma pergunta de pesquisa. Em seguida, foi realizada uma revisão da literatura para trazer a fundamentação teórica necessária à pesquisa e para colaborar com o entendimento dos principais conceitos utilizados na área. Em adição, foi realizada uma revisão sistemática da literatura com o recorte temporal entre 2013 a 2020 para se consolidar uma visão atualizada e abrangente sobre o tema vigilância tecnológica considerando principalmente trabalhos de aspectos práticos contendo métodos e propostas de arquitetura ou protótipos. Em conformidade com trabalhos análogos e o contexto da pergunta de pesquisa, foi elaborado um modelo conceitual e uma arquitetura derivada que foram materializados em um protótipo para que fosse possível a realização de um experimento no formato de estudo de caso e sua posterior avaliação por especialistas. Por fim, os resultados experimentais demonstraram a

viabilidade do modelo e sua eficácia na VT automatizada em cenários de Big Data.

7.1 TRABALHOS FUTUROS

O modelo proposto neste trabalho, conforme apresentado no Capítulo 4, é constituído de sete módulos que foram herdados pela arquitetura sugerida para a qual foi adicionalmente apresentado seu *workflow*. O sistema de VT automatizada implementado no protótipo automatiza todas as etapas da vigilância. Contudo, ainda é preciso a intervenção humana para avaliar a qualidade das fontes de informação ou incluir novas tecnologias-chaves a serem monitoradas.

Como sugestão de trabalho futuro, sugere-se que o modelo seja estendido para reduzir ainda mais o esforço humano na qualificação das fontes de dados e na identificação de novas tendências tecnológicas, isto é, não somente aqueles modelos utilizados nas ontologias e indicados por especialistas. Nesse sentido, sugere-se a inclusão de um novo módulo de *Aprendizagem* permitindo ao sistema se autoavaliar, automatizando processos como a qualificação das fontes de informação monitoradas com base no *feedback* e da interação dos usuários e sugerindo a descontinuidade das que não trouxerem informações úteis. Ademais, um novo módulo de *Descoberta* imbuído também com inteligência artificial ou aprendizagem de máquina que seja capaz de identificar novas tecnologias nas publicações reduzindo ainda mais a necessidade de intervenção humana é uma evolução factível para este trabalho.

Um tema não abordado neste trabalho, mas que pode ser crítico a depender da tipo de documento coletado e da eventual necessidade de auditoria no processo de vigilância é a proveniência dos dados (do inglês Data Provenance). Como afirma Sembay *et al.* (2020), “com o uso dos dados de proveniência é possível manter um registro completo de como um determinado cálculo ou processamento foi realizado.”. Um processo que considere a proveniência pode oferecer um registro histórico sobre os dados, permitindo saber por onde eles transitaram e quais foram suas origens. Logo, entende-se como desejável uma evolução do modelo para comportar este tipo de característica.

Por fim, considerando o aspecto multimídia cada vez mais presente nos documentos formais, como vídeos, imagens e áudios, encaminha-se como trabalho futuro a evolução do modelo para suportar a coleta, análise e visualização destes tipos de dados não estruturados ampliando a capacidade geral de monitoramento tecnológico.

REFERÊNCIAS

ABE, Hidenao; TSUMOTO, Shusaku. Detecting temporal trends of technical phrases by using importance indices and linear regression. **Foundations of Intelligent Systems**, p. 251–260, 2009.

ABNT/C-130. **ABNT NBR 16501:2011**. [*S.l.*], nov. 2011. Acessado em 10 de fevereiro de 2018. Disponível em: https://www.abntcatalogo.com.br/norma.aspx?ID=088796.

AGGARWAL, Charu C.; ZHAI, ChengXiang. A Survey of Text Clustering Algorithms. *In*: **Mining Text Data**. Edição: ChengXiang Aggarwal Charu C.and Zhai. Boston, MA: Springer US, 2012a. P. 77–128. ISBN 978-1-4614-3223-4. DOI: 10.1007/978-1-4614-3223-4_4.

AGGARWAL, Charu C.; ZHAI, ChengXiang. **Mining Text Data**. Edição: Charu C. Aggarwal e ChengXiang Zhai. Boston, MA: Springer US, 2012b. v. 53, p. 1–30. ISBN 978-1-4614-3222-7. DOI: 10.1007/978-1-4614-3223-4. arXiv: arXiv:1011.1669v3.

AIMPLAS. **Softvt**. [*S.l.*: *s.n.*], 2020. Acessado em 1 de novembro de 2020. Disponível em: https://www.softvt.com.

ALGHAMDI, Rubayyi; ALFALQI, Khalid. A Survey of Topic Modeling in Text Mining. **IJACSA) International Journal of Advanced Computer Science and Applications**, v. 6, n. 1, p. 147–153, 2015. ISSN 21565570, 2158107X. DOI: 10.14569/IJACSA.2015.060121.

ALLAHYARI, Mehdi; POURIYEH, Seyedamin; ASSEFI, Mehdi; SAFAEI, Saied; TRIPPE, Elizabeth D.; GUTIERREZ, Juan B.; KOCHUT, Krys. **A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques**. [*S.l.*: *s.n.*], 2017. arXiv: 1707.02919 [cs.CL].

ANDRADE NAVIA, J M; RAMÍREZ PLAZAS, E; ORJUELA GARZÓN, A. Technological watch applied to the production chain of cocoa [Vigilancia tecnológica aplicada a la cadena productiva de cacao]. **Espacios**, v. 39, n. 9, 2018.

ASOCIACIÓN ESPAÑOLA DE NORMALIZACIÓN Y CERTIFICACIÓN AENOR. **Norma Española Experimental UNE 166006 Gestión de la I+D+i: Sistema de Vigilancia Tecnológica.** [*S.l.*: *s.n.*], 2019. Acessado em 10 de fevereiro de 2019. Disponível em: https://www.aenor.com/normas-y-libros/buscador-de-normas/une?c=N0059973.

BARBOSA, Sérgio AA; LEITE, George; OLIVEIRA, Andre S; JESUS, Telmo O de; MACEDO, Douglas DJ de; NASCIMENTO, Rogério PC do. An architecture proposal for the creation of a database to open data related to ITS in smart cities. *In*: IEEE. 2016 8th Euro American Conference on Telematics and Information Systems (EATIS). [*S.l.*: *s.n.*], 2016. P. 1–7.

BASTARRICA, María Cecilia; HITSCHFELD-KAHLER, Nancy; ROSSEL, Pedro O. The Domain Analysis Concept Revisited: A Practical Approach. **Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)**, v. 4039, June, p. 403–406, 2006. ISSN 03029743. DOI: 10.1007/11763864.

BERGES-GARCIA, A; MENESES-CHAUS, J M; MARTINEZ-ORTEGA, J F. Methodology for evaluating functions and products for technology watch and competitive intelligence (TW/CI) and their implementation through web [Metodologia para evaluar funciones y productos de vigilancia tecnológica e inteligencia competitiva (VT/IC) y. **Profesional de la Informacion**, v. 25, n. 1, p. 103–113, 2016. DOI: 10.3145/epi.2016.ene.10.

BERGES-GARCIA, Aurelio; MENESES-CHAUS, Juan M; MARTINEZ-ORTEGA, Jose F. Methodology for evaluating functions and products for technology watch and competitive intelligence (TW/CI) and their implementation through web. **PROFESIONAL DE LA INFORMACION**, v. 25, n. 1, p. 103–113, 2016. ISSN 1386-6710. DOI: 10.3145/epi.2016.ene.10.

BERRY, Michael W; CASTELLANOS, Malu. Survey of text mining. **Computing Reviews**, Springer, v. 45, n. 9, p. 548, 2004.

BIG DATA. **Big Data — Wikipedia, The Free Encyclopedia**. [*S.l.*: *s.n.*], 2019. Acessado em: 09 de janeiro de 2019. Disponível em: https://en.wikipedia.org/wiki/Big_data.

BLEI, David M; NG, Andrew Y; JORDAN, Michael I. Latent dirichlet allocation. **Journal of machine Learning research**, v. 3, Jan, p. 993–1022, 2003.

BORKO, Harold. Information science: what is it? **American documentation**, Wiley Online Library, v. 19, n. 1, p. 3–5, 1968.

BRITO, Renata Peregrino de; BRITO, Luiz Artur Ledur. Vantagem competitiva, criação de valor e seus efeitos sobre o desempenho. **RAE-Revista de Administração de Empresas**, Fundação Getulio Vargas, v. 52, n. 1, p. 70–84, 2012.

CALOF, Jonathan L; WRIGHT, Sheila. Competitive intelligence: A practitioner, academic and inter-disciplinary perspective. **European Journal of marketing**, Emerald Group Publishing Limited, v. 42, n. 7/8, p. 717–730, 2008.

CANONGIA, Claudia; SANTOS, Dalci M; SANTOS, Marcio M; ZACKIEWICZ, Mauro *et al.* Foresight, inteligência competitiva e gestão do conhecimento: instrumentos para a gestão da inovação. **Gestão & Produção**, SciELO Brasil, 2004.

CDE INTELIGENCIA COMPETITIVA, S.L. **Hontza**. [*S.l.*: *s.n.*], 2020. Acessado em 1 de novembro de 2020. Disponível em: `http://www.hontza.es`.

CHICKERUR, Satyadhyan; GOUDAR, Anoop; KINNERKAR, Ankita. Comparison of relational database with document-oriented database (mongodb) for big data applications. *In*: IEEE. 2015 8th International Conference on Advanced Software Engineering & Its Applications (ASEA). [*S.l.*: *s.n.*], 2015. P. 41–47.

CICHOCKI, Andrzej; ANSARI, Helal A; RUSINKIEWICZ, Marek; WOELK, Darrell. **Workflow and process automation: concepts and technology**. [*S.l.*]: Springer Science & Business Media, 2012. v. 432.

CLARIVATE. **Derwent Innovation: Explanation on ThemeScape**. [*S.l.*: *s.n.*], 2019. Acessado em: 01 setembro 2019. Disponível em: `https://support.clarivate.com/Patents/s/article/Derwent-Innovation-Explanation-on-ThemeScape?language=en%5C_US`. Acesso em: 1 set. 2019.

CLAUSET, Aaron; NEWMAN, M. E. J.; MOORE, Cristopher. Finding community structure in very large networks. **Physical Review E**, American Physical Society (APS), v. 70, n. 6, dez. 2004. ISSN 1550-2376. DOI: `10.1103/physreve.70.066111`.

COSTA, Fabricio F. Big data in biomedicine. **Drug Discovery Today**, v. 19, n. 4, p. 433–440, 2014. ISSN 1359-6446. DOI: `https://doi.org/10.1016/j.drudis.2013.10.012`.

COX, Michael; ELLSWORTH, David. Managing big data for scientific visualization, jan. 1997.

CROSBY, Michael; PATTANAYAK, Pradan; VERMA, Sanjeev; KALYANARAMAN, Vignesh *et al.* Blockchain technology: Beyond bitcoin. **Applied Innovation**, v. 2, n. 6-10, p. 71, 2016.

DADOS.GOV.BR. **Dados.gov.br**. [*S.l.*: *s.n.*], 2020. `http://dados.gov.br`. Acessado em 13 de janeiro de 2020.

DATA.GOV. **Data.gov**. [*S.l.*: *s.n.*], 2020. http://Data.gov. Acessado em 13 de janeiro de 2020.

DATAAGE2025. **Data Age 2025:The Evolution of Data to Life-Critical**. [*S.l.*: *s.n.*], 2017. Disponível em: https://www.seagate.com/www-content/our-story/trends/files/Seagate-WP-DataAge2025-March-2017.pdf. Acesso em:

DBPEDIA SPOTLIGHT. **DBpedia Spotlight API**. [*S.l.*: *s.n.*], 2020. https://www.dbpedia-spotlight.org/api. Acessado em 23 de setembro de 2020.

DE MAURO, Andrea; GRECO, Marco; GRIMALDI, Michele. A formal definition of Big Data based on its essential features. **Library Review**, v. 65, p. 122–135, mar. 2016. DOI: 10.1108/LR-06-2015-0061.

DEMCHENKO, Yuri; GROSSO, Paola; DE LAAT, Cees; MEMBREY, Peter. Addressing big data issues in Scientific Data Infrastructure. **Proceedings of the 2013 International Conference on Collaboration Technologies and Systems, CTS 2013**, May, p. 48–55, 2013. DOI: 10.1109/CTS.2013.6567203.

DOS SANTOS AMPARO, K K; DO RIBEIRO, M C O; GUARIEIRO, L L N. Case study using mapping technology foresight as the main tool of scientific research [Estudo de caso utilizando mapeamento de prospecção tecnológica como principal ferramenta de busca científica]. **Perspectivas em Ciencia da Informacao**, v. 17, n. 4, p. 195–209, 2014. DOI: 10.1590/S1413-99362012000400012.

E-INTELLIGENT. **vicubo Cloud**. [*S.l.*: *s.n.*], 2020. Acessado em 1 de novembro de 2020. Disponível em: https://www.vicubocloud.es.

ENA, Oleg; MIKOVA, Nadezhda; SARITAS, Ozcan; SOKOLOVA, Anna. A methodology for technology trend monitoring: the case of semantic technologies. **Scientometrics**, Springer Netherlands, v. 108, n. 3, p. 1013–1041, 2016. ISSN 15882861. DOI: 10.1007/s11192-016-2024-0.

FÄRBER, Michael. Using a semantic wiki for technology forecast and technology monitoring. English. **Program**, Emerald Group Publishing Limited, Bradford, v. 50, n. 2, p. 225–242, 2016. ISSN 00330337.

FIESC. **Sobre a FIESC**. [*S.l.*: *s.n.*], 2019. acessado em: 04 de dezembro de 2019. Disponível em: http://fiesc.com.br/sobre-fiesc. Acesso em: 4 dez. 2019.

FIESC. **Programa de Desenvolvimento da Indústria Catarinense 2022: competitividade com sustentabilidade.** [*S.l.*], 2014. acessado em: 09 de março de 2019. Disponível em: http://www4.fiescnet.com.br/homepdic. Acesso em: 10 mar. 2019.

FIESC. **Setores Portadores de Futuro para a Indústria Catarinense 2022**. [*S.l.*], 2013. acessado em: 09 de março de 2019. ISBN ISBN 978-85-66826-02-9. Disponível em: http://www4.fiescnet.com.br/images/banner-pedic/documento-oficial-setores.pdf.

FIESC, Observatório. **Observatório da Indústria Catarinense - FIESC**. [*S.l.*: *s.n.*], 2019. Acessado em 5 de maio de 2019. Disponível em: http://www.portalsetorialfiesc.com.br.

FILIPPO, Denise; PIMENTEL, Mariano; WAINER, Jacques. Metodologia de pesquisa cientıéfica em sistemas colaborativos. **Sistemas Colaborativos**, v. 1, p. 379–404, 2011.

FLICK, U. **Introdução à pesquisa qualitativa**. [*S.l.*]: Artmed, 2009. (Biblioteca Artmed : Métodos de pesquisa). ISBN 9788536317113. Disponível em: https://books.google.com.br/books?id=9O9oPgAACAAJ.

FONSECA, J.J.S. da. **Apostila de metodologia da pesquisa cientıéfica**. Fortaleza: João José Saraiva da Fonseca, 2002. Disponível em: https://books.google.com.br/books?id=oB5x2SChpSEC.

GALVÃO, Taıés Freire; PEREIRA, Mauricio Gomes. Revisões sistemáticas da literatura: passos para sua elaboração. **Epidemiologia e Serviços de Saúde**, SciELO Public Health, v. 23, p. 183–184, 2014.

GARCEZ, Marcos Paixão; WRIGHT, James Terence Coulter. Estudo de modelos de previsão tecnológica aplicados à substituição de embalagens de refrigerantes para o mercado brasileiro. **Revista de Administração**, v. 45, n. 3, p. 255–270, 2010. ISSN 0080-2107.

GEUM, Youngjung; JEON, Jeonghwan; SEOL, Hyeonju. Technology Analysis & Strategic Management Identifying technological opportunities using the novelty detection technique : a case of laser technology in semiconductor manufacturing. October 2016, 2013. DOI: 10.1080/09537325.2012.748892.

GIL, Antônio Carlos. **Como elaborar projetos de pesquisa**. São Paulo: EDITORA ATLAS S.A, 2002.

GOMES, Carolina; FRANÇA, Rosiane; BARROS, Taıés; RIOS, RIVERSON. Spotify: streaming e as novas formas de consumo na era digital. *In*: ANAIS do XVII Congresso de Ciências da Comunicação na Região Nordeste. Natal. [*S.l.*: *s.n.*], 2015. P. 1–11.

GOMES, Eliza HA; DANTAS, Mario AR; MACEDO, Douglas DJ De; ROLT, Carlos R De; DIAS, Julio; FOSCHINI, Luca. An infrastructure model for smart cities based on big data. **International Journal of Grid and Utility Computing**, Inderscience Publishers (IEL), v. 9, n. 4, p. 322–332, 2018.

GOORHA, Saurabh; UNGAR, Lyle. Discovery of significant emerging trends. **ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)**, p. 57, 2010. DOI: 10.1145/1835804.1835815.

GRAJALES LÓPEZ, C A; ZARTHA SOSSA, J W; HERNÁNDEZ ZARTA, R; ESTRADA REVEIZ, R; GUARNIZO GÓMEZ, C A; DÍAZ URIBE, J H; GÓMEZ GARCÉS, J. Technological surveillance and analysis of the life cycle of the technology: Review of tools for enterprise diagnosis and the application of the life cycle of the product in the tourism sector. **Espacios**, v. 37, n. 36, 2016.

GRAJALES LÓPEZ, C A; ZARTHA SOSSA, J W; HERNÁNDEZ ZARTA, R H; ESTRADA REVEIZ, R E; GUARNIZO GÓMEZ, C A; DÍAZ URIBE, J H; GÓMEZ GARCÉS, J G; VALENCIA GRISALES, L V. Technology surveillance and curves in 'S': Environmental technologies in Tourism, Quindio Innova project. **Espacios**, v. 38, n. 32, p. 78–86, 2017.

GUDANOWSKA, A E. Technology mapping as a tool for technology analysis in foresight studies. *In*: 2014 IEEE International Technology Management Conference. [*S.l.*: *s.n.*], 2014. P. 1–4. DOI: 10.1109/ITMC.2014.6918613.

GUZMÁN SÁNCHEZ, María V; SOTOLONGO AGUILAR, Gilberto. Mapas tecnológicos para la estrategia empresarial. Situación tecnológica de la neisseria meningitidis. es. **ACIMED**, scielocu, v. 10, p. 1–2, ago. 2002. ISSN 1024-9435.

HAKIM, Arif Rakhman; DJATNA, Taufik. Extraction of multi-dimensional research knowledge model from scientific articles for technology monitoring. **Proceedings - 2015 3rd International Conference on Adaptive and Intelligent Agroindustry, ICAIA 2015**, p. 300–305, 2016. DOI: 10.1109/ICAIA.2015.7506526.

HEATON, Jeff. **Programming Spiders, Bots, and Aggregators in Java**. 1st. Alameda, CA, USA: SYBEX Inc., 2002. ISBN 0782140408.

HENRI, D; CLERC, P. Trends in 3-D printing from a patent information analysis (APA). **International Journal of Technology Intelligence and Planning**, v. 10, n. 3-4, p. 354–372, 2015. DOI: 10.1504/IJTIP.2015.070854.

HIDALGO, Antonio; LEÓN, Gonzalo; PAVÓN, Julián. **La Gestión de la Innovación y la Tecnología en las Organizaciones**. [*S.l.*: *s.n.*], dez. 2002. ISBN 84-368-1702-8.

HJØRLAND, Birger; ALBRECHTSEN, Hanne. Toward a New Horizon in Information Science: Domain analysis. **Journal of the American Society for Information Science**, v. 46, n. 6, p. 400–425, 1995.

HOLLERITH, Herman. The electrical tabulating machine. **Journal of the Royal Statistical Society**, JSTOR, v. 57, n. 4, p. 678–689, 1894.

HOU, J. The empirical research on patent-based models of technology entropy: A case of carbon capture technology. **ICIC Express Letters, Part B: Applications**, v. 8, n. 6, p. 971–979, 2017.

IBM; ZIKOPOULOS, Paul; EATON, Chris. **Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data**. 1st. [*S.l.*]: McGraw-Hill Osborne Media, 2011. ISBN 0071790535, 9780071790536.

IDCEMC2. **The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things**. [*S.l.*: *s.n.*], 2014. Disponível em: https://www.emc.com/leadership/digital-universe/2014iview/executive-summary.htm. Acesso em:

IDEKO, Centro Tecnológico. **Innguma**. [*S.l.*: *s.n.*], 2020. Acessado em 1 de novembro de 2020. Disponível em: https://www.innguma.com.

JIMÉNEZ GONZÁLEZ, S; DÍEZ OCHOA, S; ARANGO ALZATE, B; HERNÁNDEZ ZARTA, R. Technological surveillance of s curves and cycle life of technology. **Espacios**, v. 38, n. 44, 2017.

JUNG, Min-Gyue; YOUN, Seon-A; BAE, Jayon; CHOI, Yong-Lak. A study on data input and output performance comparison of MongoDB and PostgreSQL in the big data environment. *In*: IEEE. 2015 8th International Conference on Database Theory and Application (DTA). [*S.l.*: *s.n.*], 2015. P. 14–17.

JÜRGENS, B; HERRERO-SOLANA, V. Patent bibliometrics and its use for technology watch. **Journal of Intelligence Studies in Business**, v. 7, n. 2, p. 17–26, 2017.

JURISICA, Igor; MYLOPOULOS, John; YU, Eric. Using ontologies for knowledge management: An information systems perspective. *In*: INFORMATION TODAY; 1998. PROCEEDINGS of the Annual Meeting-American Society For Information Science. [*S.l.*: *s.n.*], 1999. P. 482–496.

KARVONEN, Matti; KAPOOR, Rahul; UUSITALO, Antti; OJANEN, Ville. Technology competition in the internal combustion engine waste heat recovery: a patent landscape analysis. **JOURNAL OF CLEANER PRODUCTION**, v. 112, n. 5, p. 3735–3743, jan. 2016. ISSN 0959-6526. DOI: 10.1016/j.jclepro.2015.06.031.

KEELE, Staffs. Guidelines for performing systematic literature reviews in software engineering. **Technical report, Ver. 2.3 EBSE Technical Report. EBSE**, 2007.

KIM, M; PARK, Y; YOON, J. Generating patent development maps for technology monitoring using semantic patent-topic analysis. **Computers and Industrial Engineering**, v. 98, p. 289–299, 2016. DOI: 10.1016/j.cie.2016.06.006.

KITCHENHAM, Barbara. Procedures for performing systematic reviews. **Keele, UK, Keele University**, v. 33, n. 2004, p. 1–26, 2004.

KITCHENHAM, Barbara; BRERETON, O Pearl; BUDGEN, David; TURNER, Mark; BAILEY, John; LINKMAN, Stephen. Systematic literature reviews in software engineering–a systematic literature review. **Information and software technology**, Elsevier, v. 51, n. 1, p. 7–15, 2009.

KOZIOLEK, Heiko; GOLDSCHMIDT, Thomas; DE GOOIJER, Thijmen; DOMIS, Dominik; SEHESTEDT, Stephan. Experiences from identifying software reuse opportunities by domain analysis. **ACM International Conference Proceeding Series**, p. 208–217, 2013. DOI: 10.1145/2491627.2491641.

LASI, Heiner; FETTKE, Peter; KEMPER, Hans-Georg; FELD, Thomas; HOFFMANN, Michael. Industry 4.0. **Business & information systems engineering**, Springer, v. 6, n. 4, p. 239–242, 2014.

LAURILA, Juha; GATICA-PEREZ, Daniel; AAD, Imad; BLOM, Jan; BORNET, Olivier; DO, T.-M.-T; DOUSSE, Olivier; EBERLE, Julien; MIETTINEN, Markus. The mobile data challenge: Big data for mobile computing research. Nokia Research Center. *In*:

LEE, Kyungpyo; LEE, Sungjoo. Patterns of technological innovation and evolution in the energy sector: A patent-based approach. **Energy Policy**, Elsevier, v. 59, p. 415–432, 2013. ISSN 03014215. DOI: 10.1016/j.enpol.2013.03.054.

LÓPEZ C., C A; ZARTHA SOSSA, J W. Technological surveillance in advanced steel used in the automotive industry. **Espacios**, v. 35, n. 8, p. 1, 2014.

LOVINS, Julie B. **Development of a stemming algorithm. Electronic Systems Laboratory**. [*S.l.*]: MIT Information Processing Group, Cambridge, 1968.

MACEDO, Douglas DJ de; PERANTUNES, Hilton WG; MAIA, Luiz FJ; COMUNELLO, Eros; WANGENHEIM, Aldo von; DANTAS, Mario AR. An interoperability approach based on asynchronous replication among distributed internet databases. *In*: IEEE. 2008 IEEE Symposium on Computers and Communications. [*S.l.*: *s.n.*], 2008. P. 658–663.

MACEDO, Douglas DJ de; VON WANGENHEIM, Aldo; DANTAS, Mario AR. A data storage approach for large-scale distributed medical systems. *In*: IEEE. 2015 Ninth International Conference on Complex, Intelligent, and Software Intensive Systems. [*S.l.*: *s.n.*], 2015. P. 486–490.

MANYIKA, James; MICHAEL, Chui; BRAD, Brown; JACQUES, Bughin; RICHARD, Dobbs; CHARLES, Roxburgh; BYERS, Angela Hung. Big data: The next frontier for innovation, competition, and productivity. **McKinsey Global Institute**, June, p. 156, 2011. ISSN 14712970. DOI: 10.1080/01443610903114527.

MARTÍNEZ RIVERO, F; MAYNEGRA DÍAZ, E R. Evaluation of web platforms for their implementation in the system of Biomundi consulting technological surveillance [Evaluación de plataformas web para su implementación en el sistema de vigilancia tecnológica de la Consultoría Biomundi]. **Revista Cubana de Informacion en Ciencias de la Salud**, v. 25, n. 1, p. 99–109, 2014.

MARULANDA, C E; HERNÁNDEZ, A; LÓPEZ, M. Technology surveillance for university students. The case of the national university of Colombia, manizales campus [Vigilancia tecnológica para estudiantes universitarios. El caso de la universidad nacional de Colombia,

sede manizales]. **Formacion Universitaria**, v. 9, n. 2, p. 17–27, 2016. DOI: 10.4067/S0718-50062016000200003.

MARULANDA ECHEVERRY, Carlos Eduardo; LÓPEZ TRUJILLO, Marcelo; LÓPEZ VILLEGAS, Luis Ignacio. Desarrollo de una aplicación móvil para alerta tecnológica * Developing a Mobile Application for Technological Alerts. **REVISTA VIRTUAL UNIVERSIDAD CATOLICA DEL NORTE**, v. 48, p. 316–330, 2016. ISSN 0124-5821.

MCCOWAN, I.; MOORE, D.; FRY, M. Classification of Cancer Stage from Free-text Histology Reports. *In*: 2006 International Conference of the IEEE Engineering in Medicine and Biology Society. [*S.l.*: *s.n.*], ago. 2006. P. 5153–5156. DOI: 10.1109/IEMBS.2006.259563.

MIKOVA, N; SOKOLOVA, A. Comparing data sources for identifying technology trends. **Technology Analysis and Strategic Management**, v. 31, n. 11, p. 1353–1367, 2019. DOI: 10.1080/09537325.2019.1614157.

MOMENI, Abdolreza; ROST, Katja. Identification and monitoring of possible disruptive technologies by patent-development paths and topic modeling. **Technological Forecasting & Social Change**, Elsevier Inc., v. 104, p. 16–29, 2016. ISSN 0040-1625. DOI: 10.1016/j.techfore.2015.12.003.

MORENO C, J A; DÍAZ, D P. Trends in Logistics in the Last Five Years - A Review Through Technological Surveillance. *In*: 2019 Congreso Internacional de Innovación y Tendencias en Ingenieria (CONIITI). [*S.l.*: *s.n.*], 2019. P. 1–5. DOI: 10.1109/CONIITI48476.2019.8960691.

NAGPAL, Arpita; JATAIN, Arnan; GAUR, Deepti. Review based on data clustering algorithms. **2013 IEEE Conference on Information and Communication Technologies, ICT 2013**, Ict, p. 298–303, 2013. DOI: 10.1109/CICT.2013.6558109.

NAM, Sunghyun; KIM, Kwangsoo. Monitoring Newly Adopted Technologies Using Keyword Based Analysis of Cited Patents. **IEEE ACCESS**, v. 5, p. 23086–23091, 2017. ISSN 2169-3536. DOI: 10.1109/ACCESS.2017.2764478.

NEIGHBORS, James Milne. **Software Construction Using Components**. 1980. Tese (Doutorado). AAI8106784.

OBSERVATORIO DE VIRTUAL DE TRANSFERENCIA DE TECNOLOGIA. **CONCEITO DE VIGILÂNCIA TECNOLÓGICA**. [*S.l.*: *s.n.*], 2019a. Acessado em 20 de fevereiro de 2019. Disponível em: https://pt.ovtt.org/vigilancia-tecnologica-conceitos.

OBSERVATORIO DE VIRTUAL DE TRANSFERENCIA DE TECNOLOGIA - OVTT. **CICLO DA VIGILÂNCIA TECNOLÓGICA**. [*S.l.*: *s.n.*], 2019b. Acessado em 1 de maio de 2019. Disponível em: https://pt.ovtt.org/vigilancia-tecnologica-metodos-pt.

OUSSOUS, Ahmed; BENJELLOUN, Fatima Zahra; AIT LAHCEN, Ayoub; BELFKIH, Samir. Big Data technologies: A survey. **Journal of King Saud University - Computer and Information Sciences**, King Saud University, v. 30, n. 4, p. 431–448, 2018. ISSN 22131248. DOI: 10.1016/j.jksuci.2017.06.001.

P. SMIRAGLIA, Richard. **The Elements of Knowledge Organization**. [*S.l.*]: Springer, jul. 2014. P. 1–101. DOI: 10.1007/978-3-319-09357-4.

PADILLA, J B; ZARTHA, J W; ALVAREZ, V T; OROZCO, G L. Technological surveillance for the identification of innovations in leather tanning byproducts. **Informacion Tecnologica**, v. 29, n. 4, p. 127–141, 2018a. DOI: 10.4067/s0718-07642018000400127.

PADILLA, J B; ZARTHA, J W; ALVAREZ, V T; OROZCO, G L. Technological surveillance for the identification of innovations in leather tanning byproducts [Vigilancia tecnológica para la identificación de innovaciones en subproductos de la curtición]. **Informacion Tecnologica**, v. 29, n. 4, p. 127–141, 2018b. DOI: 10.4067/s0718-07642018000400127.

PALOP, Fernando; VICENTE, José M. **Vigilancia Tecnológica e Inteligencia Competitiva. Su potencial para la empresa española.** [*S.l.*], 1999.

PERE ESCORSA, Ramon Maspons. **De la vigilancia tecnologica a la inteligencia competitiva.** Madrid: Editorial Financial Times Prentice Hall., 2001.

PEREZ, A; BASAGOITI, R; CORTEZ, R A; LARRINAGA, F; BARRASA, E; URRUTIA, A. A case study on the use of machine learning techniques for supporting technology watch. **Data and Knowledge Engineering**, v. 117, p. 239–251, 2018. DOI: 10.1016/j.datak.2018.08.001.

PEREZ, L G; DOMINGUEZ, E R; OVALLOS-GAZABON, D. A proposal for a technological surveillance unit aimed at regional competitiveness. **Journal of Engineering and Applied Sciences**, v. 12, n. 21, p. 5566–5571, 2017. DOI: 10.3923/jeasci.2017.5566.5571.

PHIL, M; COLLEGE, P S G R Krishnammal. Survey on Feature Selection in Document Clustering. v. 3, n. 3, p. 1240–1244, 2011.

PORTER, Michael E. **Competitive advantage: Creating and sustaining superior performance**. New York e London: Free Press, 1985. P. xviii, 557.

RAMÍREZ, María Isabel; RUA, David Escobar; ALZATE, Sandra Bibiana Arango. Vigilancia tecnológica e inteligencia competitiva. **Gestión de las Personas y Tecnología**, Facultad Tecnológica, v. 4, n. 13, p. 149–153, 2012.

REINSEL, David; GANTZ, John; RYDNING, John. Data age 2025: the digitization of the world from edge to core. **IDC White Paper Doc# US44413318**, p. 1–29, 2018.

ROCHA SOUZA, Renato; ALMEIDA, Mauricio; BARACHO, Renata. Ciência da Informação em transformação: Big Data, Nuvens, Redes Sociais e Web Semântica. **Ciência da Informação**, v. 42, p. 159, ago. 2015.

RUTHES, S. **Inteligência competitiva para o desenvolvimento**. [*S.l.*]: PEIROPOLIS, 2007. (Série transportátil). ISBN 9788575961001. Disponível em: https://books.google.com.br/books?id=iDw92JLJ9LIC.

S.L., Antara information technology. **Antara Mussol**. [*S.l.*: *s.n.*], 2020. Acessado em 1 de novembro de 2020. Disponível em: https://www.antara.ws/es/soluciones-software/inteligencia-competitiva-semantica.

S.L., Miniera. **Miniera**. [*S.l.*: *s.n.*], 2020. Acessado em 1 de novembro de 2020. Disponível em: http://www.miniera.es/pt-br/plataforma-inteligencia-competitiva.

SAGIROGLU, Seref; SINANC, Duygu. Big data: A review. **Proceedings of the 2013 International Conference on Collaboration Technologies and Systems, CTS 2013**, IEEE, p. 42–47, 2013. DOI: 10.1109/CTS.2013.6567202.

SALGADO BATISTA, Darlin; GUZMÃ¡N SÃ¡NCHEZ, MarÃa Victoria; CARRILLO CALVET, Humberto. Establecimiento de un sistema de vigilancia científico-tecnológica. es. **ACIMED**, scielocu, v. 11, p. 0 - 0, dez. 2003. ISSN 1024-9435.

SÁNCHEZ, J Marcela; PALOP, Fernando. Herramientas de Software para la práctica de la Inteligencia Competitiva en la empresa. **Valencia: Triz XXI**, 2002.

SÁNCHEZ, Jenny Marcela; PALOP, Fernando. Herramientas de Software especializadas para Vigilancia Tecnológica e Inteligencia Competitiva, abr. 2002.

SANTHIYA, K. An Automated MapReduce Framework for Crime Classification of News Articles Using MongoDB. v. 13, n. 1, p. 131–136, 2018.

SANTOSO, Joan; YUNIARNO, Eko Mulyanto; HARIADI, Mochamad. Large scale text classification using map reduce and naive bayes algorithm for domain specified ontology building. **Proceedings - 2015 7th International Conference on Intelligent Human-Machine Systems and Cybernetics, IHMSC 2015**, v. 1, p. 428–432, 2015. DOI: 10.1109/IHMSC.2015.24.

SEMBAY, Márcio José; MACEDO, Douglas Dyllon Jeronimo de; DUTRA, Moisés Lima. A Proposed Approach for Provenance Data Gathering. **Mobile Networks and Applications**, Springer, p. 1–15, 2020.

SHAH, Neepa; MAHAJAN, Sunita. Document clustering: a detailed review. **International Journal of Applied Information Systems**, Citeseer, v. 4, n. 5, p. 30–38, 2012.

SHAPERE, Dudley. Scientific Theories and Their Domains, jan. 1977. DOI: 10.1007/978-94-010-9731-4_13.

SHIRYAEV, Alexey P.; DOROFEEV, Andrey V.; FEDOROV, Alexey R.; GAGARINA, Larisa G.; ZAYCEV, Vladimir V. LDA models for finding trends in technical knowledge domain. **Proceedings of the 2017 IEEE Russia Section Young Researchers in Electrical and Electronic Engineering Conference, ElConRus 2017**, v. 4, n. 2, p. 551–554, 2017. DOI: 10.1109/EIConRus.2017.7910614.

SIRIWEERA, T. H.Akila S.; PAIK, Incheon; KUMARA, Banage T.G.S. QoS and Customizable Transaction-Aware Selection for Big Data Analytics on Automatic Service Composition. **Proceedings - 2017 IEEE 14th International Conference on Services Computing, SCC 2017**, IEEE, p. 116–123, 2017. DOI: 10.1109/SCC.2017.22.

SOUZA INÁCIO, Andrei de; ANDRADE, Rafael; WANGENHEIM, Aldo von; MACEDO, Douglas DJ de. Designing an information retrieval system for the STT/SC. *In*: IEEE. 2014 IEEE 16th International Conference on e-Health Networking, Applications and Services (Healthcom). [*S.l.*: *s.n.*], 2014. P. 500–505.

STOREY, Veda C.; SONG, Il Yeol. Big data technologies and Management: What conceptual modeling can do. **Data and Knowledge Engineering**, Elsevier B.V., v. 108, February, p. 50–67, 2017. ISSN 0169023X. DOI: 10.1016/j.datak.2017.01.001.

SUTHAHARAN, Shan. Big data classification: Problems and challenges in network intrusion prediction with machine learning. **Performance Evaluation Review**, v. 41, n. 4, p. 70–73, 2014. ISSN 01635999. DOI: 10.1145/2627534.2627557.

TECNOLOGÍA, IALE. **Vigiale**. [*S.l.*: *s.n.*], 2020. Acessado em 1 de novembro de 2020. Disponível em: https://www.vigiale.com.

THELLEFSEN, Torkild; THELLEFSEN, Martin. Pragmatic Semiotics and Knowledge Organization. **Knowledge Organization**, v. 31, p. 177–187, jan. 2005.

TOBÓN CLAVIJO, M L; ZARTA, R H; ZARTHA SOSSA, J W; REVEIZ, R E; DÍAZ URIBE, J H; GÓMEZ GARCÉS, J G. Technological surveillance and technology life cycle analysis: Usability assessment techniques, metrics and tools in the ICT sector. **Espacios**, v. 38, n. 22, 2017.

TZU, Sun; PIN, Sun. **A arte da guerra**. [*S.l.*]: WWF Martins Fontes, 2015.

VAN DER MAATEN, Laurens; POSTMA, Eric; VAN DEN HERIK, Jaap. Dimensionality reduction: a comparative. **J Mach Learn Res**, v. 10, n. 66-71, p. 13, 2009.

VAN MOL, Christof. Improving web survey efficiency: the impact of an extra reminder and reminder content on web survey response. **International Journal of Social Research Methodology**, Taylor & Francis, v. 20, n. 4, p. 317–327, 2017.

VICECONTI, M.; HUNTER, P.; HOSE, R. Big Data, Big Knowledge: Big Data for Personalized Healthcare. **IEEE Journal of Biomedical and Health Informatics**, v. 19, n. 4, p. 1209–1215, jul. 2015. ISSN 2168-2208. DOI: 10.1109/JBHI.2015.2406883.

VILLARROELG, C; COMAI, A; KARMELICPAVLOV, V; FERNÁNDEZO, A; ARRIAGADAV, C. Design and implementation of a technological surveillance and competitive intelligence unit. **Interciencia**, v. 40, n. 11, p. 751–757, 2015.

WATCHER, Intelligent. **Intelligent Watcher**. [*S.l.*: *s.n.*], 2020. Acessado em 1 de novembro de 2020. Disponível em: https://intelligentwatcher.com.

WEI, Yi-Ming; KANG, Jia-Ning; YU, Bi-Ying; LIAO, Hua; DU, Yun-Fei. A dynamic forward-citation full path model for technology monitoring: An empirical study from shale gas industry. **APPLIED ENERGY**, v. 205, p. 769–780, nov. 2017. ISSN 0306-2619. DOI: 10.1016/j.apenergy.2017.08.121.

WHITE, Tom. **Hadoop: The Definitive Guide**. 4. ed. Beijing: O'Reilly, 2015. ISBN 978-1-4919-0163-2.

WIKIPEDIA. **Endereço IP — Wikipedia, The Free Encyclopedia**. [*S.l.*: *s.n.*], 2019a. [Online; acessado 04 de setembro 2019]. Disponível em: https://pt.wikipedia.org/wiki/Endere%C3%A7o_IP.

WIKIPEDIA. **K-means — Wikipedia, The Free Encyclopedia**. [*S.l.*: *s.n.*], 2019b. Acessado em: 01 setembro 2019. Disponível em: https://pt.wikipedia.org/wiki/K-means.

WIKIPEDIA CONTRIBUTORS. **Conceptual Model**. [*S.l.*]: Wikimedia, 2020. Disponível em: https://pt.wikipedia.org/wiki/Fluxo_de_trabalho.

YAARI, Yaakov. Segmentation of expository texts by hierarchical agglomerative clustering. **arXiv preprint cmp-lg/9709015**, 1997.

YAN, Jun; LIU, Ning; YAN, Shuicheng; YANG, Qiang; FAN, Weiguo; WEI, Wei; CHEN, Zheng. Trace-oriented feature analysis for large-scale text data dimension reduction. **IEEE Transactions on Knowledge and Data Engineering**, IEEE, v. 23, n. 7, p. 1103–1117, 2011. ISSN 10414347. DOI: 10.1109/TKDE.2010.34.

ZHANG, Tian; RAMAKRISHNAN, Raghu; LIVNY, Miron. BIRCH: an efficient data clustering method for very large databases. *In*: ACM, 2. ACM Sigmod Record. [*S.l.*: *s.n.*], 1996. P. 103–114.