Alessandro Costa Ribeiro

MODELO DE RECONHECIMENTO DE PADRÕES EM IDEIAS USANDO TÉCNICAS DE DESCOBERTA DE CONHECIMENTO EM TEXTOS

Dissertação submetida ao Programa de Engenharia e Gestão do Conhecimento da Universidade Federal de Santa Catarina para a obtenção do Grau de Mestre em Engenharia do Conhecimento.
Orientadora: Prof.ª Gertrudes Aparecida Dandolini, Dr.ª
Coorientador: Prof. João Artur de Souza, Dr.

Florianópolis
2018

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

Costa Ribeiro, Alessandro
Modelo de Reconhecimento de Padrões em Ideias Usando Técnicas de Descoberta de Conhecimento em Textos / Alessandro Costa Ribeiro ; orientador, Gertrudes Aparecida Dandolini, coorientador, João Artur de Souza, 2018.
172 p.

Dissertação (mestrado) - Universidade Federal de Santa Catarina, Centro Tecnológico, Programa de Pós Graduação em Engenharia e Gestão do Conhecimento, Florianópolis, 2018.

Inclui referências.

1. Engenharia e Gestão do Conhecimento. 2. Gestão de Ideias. 3. Descoberta de Conhecimento em Textos . 4. KDT. 5. Reconhecimento de Padrões. I. Aparecida Dandolini, Gertrudes . II. Artur de Souza, João . III. Universidade Federal de Santa Catarina. Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento. IV. Título.

Alessandro Costa Ribeiro

MODELO DE RECONHECIMENTO DE PADRÕES EM IDEIAS USANDO TÉCNICAS DE DESCOBERTA DE CONHECIMENTO EM TEXTOS

Esta Dissertação foi julgada adequada para obtenção do Título de "Mestre em Engenharia do Conhecimento", e aprovada em sua forma final pelo Programa Engenharia e Gestão do Conhecimento da Universidade Federal de Santa Catarina.

Florianópolis, 10 maio de 2018.

_____
Prof.ª Gertrudes Aparecida Dandolini, Dr.ª
Coordenadora do Programa

Banca Examinadora:

_____
Prof.ª Gertrudes Aparecida Dandolini, Dr.ª
Orientadora
Universidade Federal de Santa Catarina

_____
Prof. Roberto Raitz, Dr.
Universidade Federal do Paraná

_____
Prof. João Bosco da Mota Alves, Dr.
Universidade Federal de Santa Catarina

_____
Prof.ª Jandira Genka Palma, Dr.ª
Universidade Federal de Santa Catarina

AGRADECIMENTOS

Gostaria de agradecer a Deus por me guiar, iluminar e me dar tranquilidade para seguir em frente com os meus objetivos e não desanimar com as dificuldades.

Agradeço a minha família por acreditarem nos meus sonhos e darem o suporte necessário, especialmente a minha esposa Yohani Domink pessoa com quem partilho a vida. Obrigado pelo carinho, a paciência e por sua capacidade de me trazer paz na correria que é a nossa vida.

Agradeço a Universidade do Estado de Mato Grosso (UNEMAT) pelo apoio imprescindível para esta capacitação.

Agradeço aos orientadores Prof.ª Gertrudes Aparecida Dandolini e Prof. João Artur de Souza, pela dedicação aos seus alunos, pelo compartilhamento de conhecimento, incentivo e orientação ao longo desta pesquisa. Obrigado pela confiança depositada em mim.

A todos os mestres, colegas e amigos queridos, especialmente ao grupo Núcleo de Estudos em Inteligência, Gestão e Tecnologias para Inovação (IGTI) pela acolhida e apoio para construção desta jornada acadêmica.

"*Todo o conhecimento genuíno tem origem na experiência direta*."

(Mao Tse Tung)

RESUMO

O processo de inovação impulsiona as organizações a se desenvolverem rapidamente e/ou sobreviverem no mercado altamente competitivo. Como primeira etapa deste processo tem-se o *Front End* da Inovação (FEI) que compreende a criação de ideias, identificação de oportunidades, seleção e análise destas. Trata-se de uma etapa importante no processo como um todo, de forma que pode representar o sucesso ou fracasso das organizações. Para apoiar a gestão de ideias no *Front End*, há uma crescente utilização de Sistemas de Gestão de Ideias, os quais buscam, organizar, coletar, enriquecer, avaliar e selecionar ideias. Contudo, ao considerar as incertezas que circundam essa etapa e a quantidade de informações não estruturadas, são indispensáveis métodos, técnicas e ferramentas para os Sistemas de Gestão de Ideias no auxílio ao ciclo de vida das ideias dentro das organizações. Desta maneira, esta dissertação possui como objetivo propor um modelo de reconhecimento de padrões em ideias amparado por técnicas de descoberta de conhecimento em texto. Para demonstração de viabilidade do modelo proposto, foi desenvolvido um protótipo para apoiar as fases de criação, enriquecimento, seleção e avaliação das ideias, e este protótipo foi aplicado no cenário da iniciativa do Senado Federal chamada de Ideia Legislativa. A partir da aplicação do modelo, identificou-se como resultado por meio da métrica do cosseno, que há um grande de número de ideias semelhantes concorrendo entre si; já por meio da classificação das ideias por temáticas pré-estabelecidas com o algoritmo de Naive Bayes, evidenciou-se que esta técnica probabilística auxilia na classificação de ideias que podem pertencer a mais de uma classe. De modo que reconhecer padrões em ideias, dados não estruturados, em busca de gerar clusters auxilia no processo de gestão desta etapa tão importante e ao incorporar as atividades do modelo no ciclo de vida das ideias, visa-se criar ideias mais robustas com a formação de redes entre colaboradores e também facilitar o trabalho dos especialistas de domínio quanto a aprovação e classificação destas ideias.

Palavras-chave: Gestão de Ideias; Descoberta de Conhecimento em Textos; KDT; Reconhecimento de Padrões; RP.

ABSTRACT

The innovation process drives organizations to develop rapidly and / or survive in the highly competitive marketplace. The first step in this process is the Front End of Innovation (FEI), which includes the creation of ideas, identification of opportunities, selection and analysis of these. This is an important step in the process as a whole, so it can represent the success or failure of organizations. To support the management of ideas in the Front End, there is a growing use of Idea Management Systems, which seek, organize, collect, enrich, evaluate and select ideas. However, considering the uncertainties surrounding this stage and the amount of unstructured information, methods, techniques and tools for Idea Management Systems are indispensable in helping the life cycle of ideas within organizations. In this way, this dissertation aims to propose a model of recognition of patterns in ideas supported by techniques of discovery of knowledge in text. To demonstrate the feasibility of the proposed model, a prototype was developed to support the creation, enrichment, selection and evaluation phases of the ideas, and this prototype was applied in the scenario of the initiative of the Federal Senate called the Legislative Idea. From the application of the model, it was identified as a result by means of the metric of the cosine, that there is a great number of similar ideas competing with each other, already by means of the classification of the ideas by pre-established thematic ones with the algorithm of Naive Bayes , it was evidenced that this probabilistic technique assists in the classification of ideas that can belong to more than one class. Thus, recognizing patterns in ideas, unstructured data, seeking to generate clusters assists in the management process of this very important stage, and by incorporating the activities of the model in the life cycle of ideas, it is aimed at creating more robust ideas with the formation of networks between collaborators and also facilitate the work of the domain experts regarding the approval and classification of these ideas.

Keywords: Idea Management; Knowledge Discovery in Texts; KDT; Pattern Recognition; PR.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ABNT	Associação Brasileira de Normas Técnicas
CAE	Comissão de Assuntos Econômicos
CAS	Comissão de Assuntos Sociais
CCJ	Comissão de Constituição, Justiça e Cidadania
CT	Categorização de Textos
CCT	Comissão de Ciência, Tecnologia, Inovação, Comunicação e Informática
CE	Comissão de Educação, Cultura e Esporte
CDH	Comissão de Direitos Humanos
CDR	Comissão de Desenvolvimento Regional e Turismo
CI	Comissão de Serviços de Infraestrutura
CMA	Comissão de Meio Ambiente
CRE	Comissão de Relações Exteriores e Defesa Nacional
CTFC	Comissão de Transparência, Governança, Fiscalização e Controle e Defesa do Consumidor
CRA	Comissão de Agricultura e Reforma Agrária
CRM	*Customer Relationship Management*
CSF	Comissão Senado do Futuro
CESM	*Composition, Environment, Structure, Mechanism*
DSRM	*Science Research Methodology*
DSR	*Design Science Research*
EG	Engenharia do Conhecimento
ERP	Enterprise Resource Planning
FEI	*Front End* da Inovação
GI2MO	*Semantically Empowerede Idea Manegement*
IGTI	Núcleo de Estudos em Inteligência, Gestão e Tecnologias para Inovação
IMS	*System Management Idea*
IBM	*International Business Machines*
KDD	*Knowledge Discovery in Database*
KDT	*Knowledge Discovery in Text*
MDS	*Multidimensional Scaling*
NCD	Desenvolvimento de Novos Conceitos
NLTK	*Natural Language Toolkit*
NLP	*Natural Language Processing*
PLM	*Product Lifecycle Management*
PLN	Processamento de Linguagem Natural

PPGEGC	Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento
P&D	Planejamento e Desenvolvimento
PDMS	*Product Data Management Systems*
PLM	*Product Lifecycle Management*
PDM	Gerenciamento de Dados de Produto
RDF	*Resource Description Framework*
RP	Reconhecimento de Padrões
SCM	*Supply Chain Management*
SPARQL	*Structured Query Language*
UFSC	Universidade Federal de Santa Catarina
UNEMAT	Universidade do Estado de Mato Grosso

SUMÁRIO

1 INTRODUÇÃO

Neste primeiro capítulo apresentam-se informações referentes ao tema, à contextualização e a problematização, o objetivo geral e os objetivos específicos. Contém também a justificativa, as delimitações do estudo para a elaboração desta dissertação, sua aderência ao Programa de Pós-Graduação de Engenharia e Gestão do Conhecimento da Universidade Federal de Santa Catarina (PPGEGC/UFSC) e a estrutura do trabalho.

1.1 CONTEXTUALIZAÇÃO E PROBLEMA DE PESQUISA

Com o advento da sociedade do conhecimento, a inovação é vista como um dos fatores fundamentais ao desenvolvimento das organizações e pode representar um diferencial para obtenção de vantagem competitiva (BAREGHEH; ROWLEY; SAMBROOK, 2009; GIBSON; SKARZYNSKI, 2008). Dada a sua relevância no cenário organizacional, sejam instituições públicas ou privadas, existem muitas pesquisas que compreendem a inovação como um processo que necessita ser gerenciado. Barecheh, Rowley e Sambrook (2009) definem inovação como um processo composto por várias etapas, onde as ideias são transformadas em novos [ou melhorados] produtos, serviços ou processos, para que as organizações consigam avançar, competir e se diferenciar no mercado de atuação.

Para Tidd e Bessant (2015) as organizações buscam estruturar este processo, respeitando as peculiaridades que envolvem a questão de flexibilidade, de maneira que promovam o processo criativo. Embora as pesquisas sobre inovação tenham iniciado na área de desenvolvimento de novos produtos, onde o foco principal era o desenvolvimento propriamente dito de algo novo e sua comercialização, na atual sociedade do conhecimento, surge a premissa de melhorar/criar os seus processos, em especial a etapa de pré-desenvolvimento. Etapa essa que, se bem gerenciada, possibilita uma maior agilidade no processo de inovar.

O resultado deste processo, antes focado em produto, pressupõe agora inovações em outras áreas, como: serviços, processos, marketing ou organizacional (OECD, 2005), na busca de ideias potenciais e diferenciadas. Para Murah *et al.* (2013) esse grande volume de ideias advindas dessas áreas, geram um desafio para a gestão organizacional, visto que dificultam o tratamento de todos esses dados e conteúdo.

Para Bessant *et al.* (2005) e Löwer e Heller (2014) a capacidade das organizações para inovar está diretamente relacionada com a busca por novas ideias, em prol de aprimorar seus produtos, serviços e processos. A fim de tornar essa busca mais consistente as organizações estabelecem processos de inovação (HORTON; GOERS, 2014). Contudo, esses processos são complexos, visto que os dados gerados a partir deles nem sempre estão estruturados (MURAH *et al.*, 2013).

Os autores Smith e Reinertsen (1991) apresentam modelos de inovação que estruturam seu processo em três grandes etapas: *front end* da inovação (pré-desenvolvimento), desenvolvimento e implementação. O processo da inovação, segundo Koen *et al.* (2002) pode ser dividido também em três grandes subprocessos, similares aos de Smith e Reinertsen, sendo esses: 1) *Front End* da inovação (FEI) 2) desenvolvimento de produtos e 3) comercialização. Para estes autores, o FEI é reconhecido como a primeira etapa do processo e corresponde às atividades como: identificação de oportunidades, geração de ideias e concepção de um novo conceito; a segunda etapa, o desenvolvimento refere-se às atividades executadas no sentido de especificar e detalhar o conceito de forma a tornar possível sua implementação, como prototipagem, testes e detalhamentos de projeto (TEZA, 2018). E a terceira e última etapa, a implementação, envolve as atividades como produção e introdução no mercado (SMITH; REINERTSEN, 1991; KOEN *et al.*, 2001; HERSTATT *et al.*, 2006).

Cooper e Edgett (2008), Koen *et al.* (2002) e Kempe *et al.* (2011) definem que são as fases iniciais da inovação que ditam o sucesso desta, pois é neste momento que as ideias são criadas para posteriormente serem desenvolvidas e comercializadas.

Neste contexto, ideias se integram ao processo de inovação como um ponto inicial no processo de desenvolvimento de novos produtos. (BJÖRK; BOCCARDELLI; MAGNUSSON, 2010). Além disso, tornam-se um elemento essencial para o sucesso deste processo inicial que, na maioria das vezes, não linear. É no FEI, que as ideias são geradas, enriquecidas, reconstruídas até serem classificadas e selecionadas (Koen *et al.* 2002). Neste sentido, as ideias passam por diversas atividades e processos, até serem classificadas e selecionadas. Estas atividades devem ser organizadas e gerenciadas, de forma a desafiar a gestão quanto à seleção dos conteúdos gerados.

Mediante a necessidade de gerenciar estas etapas do processo, surge o conceito de gestão de ideias. A gestão de ideias tem como um dos objetivos, facilitar a organização das ideias e possibilitar a implementação

destas de forma mais eficiente e eficaz. Como forma de melhorar o processo, cada vez mais os gestores das organizações necessitam de conhecimentos de domínios específicos, bem como a necessidade de investimento em criação de sistemas computacionais para facilitar a gestão do conteúdo com intuito de analisá-lo e processá-lo e classificá-lo (MURAH *et al.*, 2013). A interação coletiva, além dos conhecimentos dos gestores e dos sistemas é outra característica fundamental para os sistemas de gestão de ideias (PEREZ; LARRINAGA; CURRY, 2014).

Para Westerski, Iglesias e Garcia (2012) a necessidade de gerenciar ideias sempre existiu, inicialmente como simples "caixas de sugestões" nas organizações. No entanto, ao passar do tempo, sua complexidade foi aumentando e novas alternativas foram criadas, de modo que atualmente o conceito de gestão de ideias está fortemente relacionado ao conceito de Sistemas de Gestão de Ideias (Idea Management System-IMS) (MIKELSONE; LIELA, 2015).

Os sistemas de gestão de ideias são plataformas que proporcionam ferramentas para criar, armazenar, procurar, editar, comentar e votar ideias (PEREZ; LARRINAGA; CURRY, 2014) com o intuito de auxiliar na administração, geração, avaliação e seleção de ideias inovadoras (WESTERSKI; DALAMAGAS; IGLESIAS, 2013; LI; LI; CHEN, 2014). Estes sistemas são desenvolvidos embasados em tecnologias da informação e são considerados um ramo com futuro promissor (WESTERSKI; IGLESIAS, 2011). Entretanto, sistemas de gestão de ideias deparam-se com desafios, tais como: sobrecarga de informações, devido aos picos de ideias triviais e redundantes, e esforço humano despendido com o processo de avaliação e seleção de ideias, quando realizado de forma manual (WESTERSKI; DALAMAGAS; IGLESIAS, 2013).

Assim, grande quantidade de ideias coletadas na fase inicial do processo de inovação pode tornar o processo de gestão de ideias uma atividade complexa, não trivial (KAMPA; CZIULIK, 2016; LUNING; PENGZHU, 2009; WESTERSKI; IGLESIAS; RICO, 2010; WESTERSKI; IGLESIAS, 2011; JANSEN, 2012), sendo necessário o uso de técnicas para analisá-las. Neste sentido as organizações começam a se preocupar em como gerir esta etapa inicial da inovação e descobrir ideias com potencial em base de dados de grande volume e não estruturadas. À vista disso para reduzir esta complexidade pode-se buscar agrupar ideias semelhantes ou classificá-las de acordo com critérios pré-definidos (WESTERSKI; DALAMAGAS; IGLESIAS, 2013), facilitada pelas tecnologias da informação.

Poveda, Westerski e Iglesias (2012) evidenciam que ao usar técnicas para clusterizar/classificar ideias pode favorecer o trabalho dos especialistas de domínio no processo de examinar e avaliar as ideias coletadas a partir de comunidades online. Neste sentido ao se classificar/clusterizar usando aprendizado supervisionado e não supervisionado em busca de reconhecer padrões em ideias pode auxiliar a gestão de ideias em suas atividades, de modo que o valor de tais métodos está relacionado ao fato de que estes possuem uma forma de operar imparcial e também a capacidade de trabalhar com grandes volumes de dados. (GRIMMER *et al*., 2009).

Para Magnusson, Netz e Wästlund (2014) um fator relevante em sistemas de gestão de ideias é a ocorrência de uma ideia surgir inúmeras vezes na base de ideias, ainda que com algumas características diferentes, porém fortemente relacionadas. De forma isolada estas ideias podem não serem interessantes, mas a potencialidade destas ideias cresce quando agrupadas. Os autores ainda destacam que se uma ideia tem um grande número de ocorrências e repetições pode assinalar para uma possível necessidade ou demanda da comunidade geradora. Sendo assim, importante o seu agrupamento e/ou classificação, para que apontem uma oportunidade diferenciada.

Quando se classifica ideias alinhadas as temáticas específicas das organizações, com o objetivo de dar suporte às decisões, evidenciam-se conhecimentos armazenados, porém até então não utilizados por estes especialistas de domínio (KAMPA; CZIULIK, 2016; PEREZ; LARRINAGA; CURRY, 2014; MURAH *et al.*, 2013). Jansen (2012) apresenta que as pesquisas acerca de gestão de ideias cresceram nos últimos anos, devido à dificuldade de identificar as ideias com potencial quando estas estão em um banco de dados de ideias não estruturado.

Assim, as tecnologias da informação são ferramentas empregadas pelos sistemas de gestão de ideias (FENN; LEHONG, 2011) e possibilitam a administração, avaliação e seleção de ideias para serem utilizadas pelas organizações (WESTERSKI; DALAMAGAS; IGLESIAS, 2013; LI; LI; CHEN, 2014). Conforme observado na revisão da literatura, tecnologias da informação, com foco em sistema de gestão de ideias estão sendo mencionadas como um promissor ramo da indústria de software (FENN; LEHONG, 2011).

Assim, diante do contexto apresentado surge a seguinte questão de pesquisa: Como reconhecer padrões em uma base de ideias, de modo a melhorar o processo de gestão de ideias?

O contexto refere-se a bancos de ideias não estruturados, no qual estes podem estar alocados em tabelas, documentos textuais, base de dados, entre outras formas de armazenamento, e que por meio de alguma semelhança entre estas ideias, seja possível que as técnicas de descoberta de conhecimento possam criar clusters de ideias para auxiliar os especialistas de domínio e usuários a tomar uma decisão nesta fase tão complexa do processo de inovação.

Para tanto foi realizado uma busca sistemática da literatura descrevendo os constructos que dão base para esta dissertação. Os principais temas são: *Front End* da Inovação e gestão de ideias além das abordagens para descoberta de conhecimento em textos.

1.2 OBJETIVOS

Considerando a problemática discutida, são descritos os objetivos da pesquisa a seguir.

1.2.1 Objetivo Geral

Propor um modelo de reconhecimento de padrões em ideias amparado por técnicas de descoberta de conhecimento em texto.

1.2.2 Objetivos Específicos

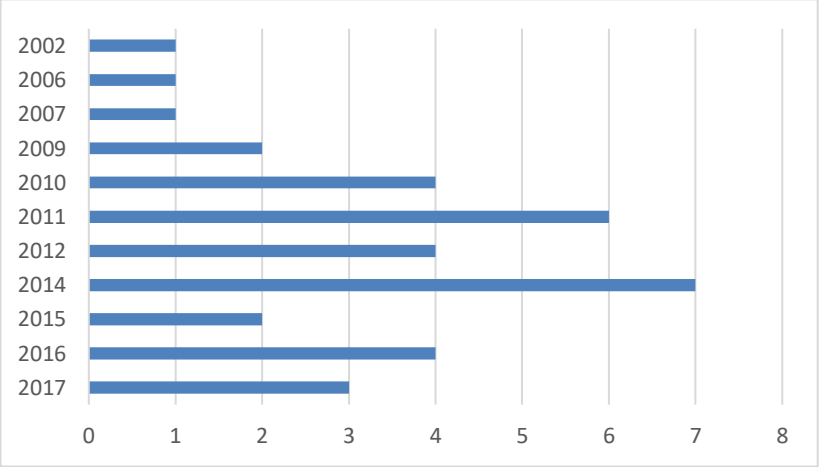
- Analisar métodos, técnicas e ferramentas utilizadas para tratamento de dados textuais na gestão de ideias;
- Criar protótipos para reconhecimento de padrões com base nas técnicas de KDT evidenciadas no modelo;
- Verificar a viabilidade do modelo proposto a partir de uma aplicação em um cenário.

1.3 JUSTIFICATIVA

Na conjuntura atual apresentada na problematização evidencia-se a necessidade estudos desta temática, pois, corroborando com Westerski e Iglesias (2011), na fase inicial do processo, grande quantidade de ideias são coletadas e este fato pode tornar o processo de gestão um desafio para as organizações. Assim são necessários modelos ou ferramentas que auxiliam na sua seleção e até mesmo na classificação dessas ideias.

Na busca sistemática e análise bibliométrica, realizada para esta dissertação (detalhes no apêndice A) constatou-se que as pesquisas sobre a descoberta de conhecimento em textos aplicados a base de ideias, vêm crescendo desde 2009, de forma que se pode associar este aumento pela grande recessão mundial entre 2008 a 2009, que desencadeou nas organizações uma busca por meios de se manterem competitivas. Nota-se também que os ápices das publicações de estudos foram nos anos de 2011 e 2014. Porém, não foram encontradas explicações para a queda em 2012 e de 2015 adiante. A Figura 1 ilustra esta análise.

Figura 1 - Evolução das publicações ao longo dos anos



Fonte: do autor.

Do ponto de vista profissional e de aplicabilidade, esta pesquisa contribui principalmente para o cenário organizacional, uma vez que a competitividade tem levado as organizações a investir em novos produtos/serviços e estratégias de atuação. Em se tratando de competitividade, a inovação torna-se fator essencial para manter as organizações com vantagens competitivas (GIBSON; SKARZYNSKI, 2008).

Neste contexto, estudar e aprender sobre as técnicas da engenharia do conhecimento aplicadas na gestão das ideias compõem a matéria-prima essencial para o processo de inovação e cooperam para identificar oportunidades (BJÖRK; BOCCARDELLI; MAGNUSSON, 2010; BOTHOS; APOSTOLOU; MENTZAS, 2012). Assim, necessitam ser gerenciadas de modo a estarem disponíveis quando necessário. A gestão

de ideias está se tornando uma ferramenta relevante ao incremento da produtividade das organizações, pois agiliza o desenvolvimento de novos produtos/serviços ou ainda melhora alguns processos da organização, acarretando competitividade (XIE; ZHANG, 2010).

Para Kampa e Cziulik (2016) o processo de ideação para novos produtos amparado no *crowdsourcing*[1] pode gerar um grande número de ideias o que dificulta a classificação. Complementando essa visão da necessidade de classificar as ideias, Poveda, Westerski e Iglesias (2012) salientam a importância do uso de técnicas e ferramentas que facilitam esse trabalho.

Murah *et al.* (2013) afirmam que esse elevado volume de conteúdo aponta para um desafio à gestão. Neste sentido, salientam que o processo se torna dependente de gestores com conhecimento específico. Os autores apontam ainda como alternativa o foco na criação de sistemas computacionais, com objetivo de facilitar a gestão do conteúdo, sendo mais rápida sua análise, classificação e agrupamento, para que estejam disponíveis no momento certo (MURAH *et al.*, 2013).

Fenn e Lehong (2011) também destacam para o uso de tecnologias da informação como ferramentas empregadas pelos sistemas de gestão de ideias e Li, Li, e Chen *et al.* (2014) afirmam que estas possibilitam a administração, avaliação e seleção de ideias para serem empregadas pelas organizações. E estão sendo mencionadas como um promissor ramo da indústria de software (FENN; LEHONG, 2011; WESTERSKI; DALAMAGAS; IGLESIAS, 2013, LI; LI; CHEN, 2014).

Assim, com foco nos estudos apresentados, esta dissertação se justifica, visto que apresenta um modelo suportado por ferramentas para um tema em expansão e que demanda de soluções mais efetivas.

Além disso, este estudo faz parte das pesquisas do Núcleo de Estudos em Inteligência, Gestão e Tecnologias para Inovação - IGTI, em relação ao *Front End* da Inovação, entre os temas a Gestão de Ideias.

1.4 DELIMITAÇÃO DA PESQUISA

Esta pesquisa tem como foco a etapa inicial do processo de inovação, definido na literatura como *Front End* da Inovação (FEI). O FEI é composto por 3 elementos distintos: Oportunidades, Ideias e Conceito. Esses elementos compreendem cinco atividades: identificação

[1] *Crowdsourcing*: o termo refere-se à colaboração coletiva como processo de obtenção serviços conteúdos e ideias de um grande número de pessoas, propiciado também pelas comunidades online.

de oportunidade, análise de oportunidades, geração e enriquecimento de ideias e seleção de ideias e geração de conceito segundo Koen *et al.* (2001).

Esta dissertação limita-se a estudar o elemento ideia, com foco em reconhecer padrões em ideias coletadas e armazenadas (de forma não estruturada) em um banco de ideias, de modo a auxiliar o processo de gestão de ideias. Este estudo é parte de pesquisas mais amplas do IGTI, a dimensão se delimita pelo objetivo da dissertação em propor um modelo de reconhecimento de padrões em ideias amparado por técnicas de descoberta de conhecimento em texto. Outras questões relacionadas aos critérios de criação, avaliação e seleção de ideias adotadas pelas organizações não fazem parte do escopo. São questões fundamentais a área de estudo, porém apontadas como estudos futuros.

1.5 ADERÊNCIA AO PROGRAMA DE PÓS-GRADUAÇÃO

O conhecimento é o principal objeto de pesquisa tratado no Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento da Universidade Federal de Santa Catarina (PPGEGC/UFSC), este é considerado nesta dissertação um recurso fundamental para que os processos de inovação ocorram. Assim, considera-se que o conhecimento é "conteúdo ou processo efetivado por agentes humanos ou artificiais em atividades de geração de valor científico, tecnológico, econômico, social ou cultural" (PACHECO, 2014). Bettoni *et al.* (2010) afirmam que a gestão de ideias, um dos constructos basilares que norteiam esta dissertação, é uma ferramenta de grande valia para a gestão de conhecimento, visto que o conhecimento pode surgir por meio do processo de geração de ideias de modo que se torne uma das peças fundamentais para gestão do conhecimento dentro das organizações.

Bettoni *et al.* (2010) ressaltam ainda os processos de aprendizagem e compartilhamento do conhecimento que ocorrem para gestão de ideias, tornam as atividades intensivas em conhecimento. Repko (2011) apresenta a interdisciplinaridade como um processo de responder a um tema muito amplo ou complexo forma adequada por diversas disciplinas ou profissão. Então pode-se compreender que a interdisciplinaridade tem seu foco em resolver um problema complexo e exige conhecimento de diversas áreas e profissionais. Pode-se considerar então que se baseia em perspectivas disciplinares e integra os seus conhecimentos para produzir uma perspectiva mais abrangente. Neste sentido, o foco do estudo dessa dissertação são as técnicas de classificação

e agrupamento de ideias dentro do processo de inovação, que envolve além dos conhecimentos em tecnologias da informação, também a parte de construção de processos de gestão de ideias, temas devidamente tratados na Administração e áreas correlatas.

Tomando-se ainda como disciplinas bases: a gestão do conhecimento, na análise do contexto e critérios de avaliação e seleção e a engenharia do conhecimento, nas técnicas extração e representação do conhecimento, a pesquisa se integra a concepção do programa na linha de pesquisa "Engenharia do Conhecimento aplicada às organizações". A relevância da gestão do conhecimento nos sistemas de gestão de ideias é identificada em 10 dos artigos analisados (PEREZ; LARRINAGA; CURRY, 2014; BETTONI; BERNHARD; BITTEL 2013; WESTERKI *et al.*, 2013; POVEDA; WESTERSKI; IGLESIAS, 2012; WESTERKI *et al.*, 2012; WESTERKI; IGLESIAS, 2011; WESTERKI *et al.*, 2010; BETTONI *et al.*, 2010; BAILEY; HORVITZ, 2010; HRASTINSKI *et al.*, 2010).

Além dos pontos acima descritos, o tema ideias já se apresenta consolidados como fonte de pesquisa nos trabalhos já realizados no PPGEGC/UFSC conforme descritos no Quadro 1.

Quadro 1 - Dissertações Realizadas no PPGEGC/UFSC

TÍTULO	**ANO**	**AUTOR**	**ORIENTADOR**	**NÍVEL**
Processo de Seleção de Ideias em Empresas Inovadoras.	2017	VALDATI, Aline de Brittos	Prof. João Artur de Souza, Dr.	M
Um Modelo Baseado em Ontologia e Análise de Agrupamento para Suporte à Gestão de Ideias.	2016	SÉRGIO, Marina Carradore	Prof. Alexandre Leopoldo Gonçalves, Dr.	M
Identificação de Critérios para Avaliação de Ideias: Um Método utilizando Folksonomias.	2016	ROCHADE L, Willian	Prof. João Artur de Souza, Dr.	M
Inteligência Competitiva na Web: Um Framework Conceitual para Aquisição de Ativos de conhecimento no	2013	SCHMITT, Maurílio Tiago Brüning	Prof. João Artur de Souza, Dr.	M

contexto do *Front End* da inovação.				
O Processo de Geração de Ideias para Inovação: Estudo de Caso em uma Empresa Náutica.	2013	DOROW, Patrícia Fernanda	Prof. João Artur de Souza, Dr.	M
Front end da Inovação: proposta de um modelo conceitual	2012	TEZA, P.	Prof.ª Aline França de Abreu, Ph.D.	M
Uma Abordagem de Geração de Ideias para o Processo de Inovação.	2012	MIGUEZ, Viviane Brandão	Prof. Rogério Cid Bastos, Dr.	M
Uma Arquitetura de *Business Intelligence* para Processamento Analítico Baseado em Tecnologias Semânticas e em Linguagem Natural.	2011	SILVA, Dhiogo Cardoso da	Prof. Denilson Sell, Dr.	M
Proposta de Modelo para o Gerenciamento de Portfólio de Inovação: Modelagem do Conhecimento na Geração de Ideias.	2009	PRADA, Charles A.	Prof.ª Aline França de Abreu, Ph.D.	M

Fonte: do autor, baseado na base de dados EGC.

Por fim, cabe ressaltar ainda, que o estudo é continuidade de pesquisas já realizadas no âmbito do PPGEGC/UFSC e um dos campos de estudo do grupo de pesquisa IGTI, como: Teza (2012), Miguez (2012), Dorow (2013), Schmitt (2013). Além disso, complementa a pesquisa de Valdati (2017) fornecendo ferramentas em prol de facilitar e agilizar o processo de avaliação e seleção de ideias dentro das organizações. E avança nas pesquisas de Rochadel (2016) e Sérgio (2016) trazendo uma outra abordagem com técnicas ainda não aplicadas neste contexto para classificação das ideias guiadas por temáticas específicas das organizações.

1.6 ESTRUTURA DO TRABALHO

Este trabalho é constituído por cinco capítulos. Além deste capitulo introdutório responsável por apresentar introduzir o tema desta

dissertação, asseverar o problema de pesquisa e descrever o objetivo geral e os específicos, a justificativa, a aderência do tema ao PPGEGC, bem como as delimitações e a estrutura do trabalho. O capítulo 2 demonstra a revisão da literatura, descrevendo o estado da arte sobre os constructos que dão base para esta dissertação. Os principais temas são: Inovação, Gestão de Ideias, Descoberta de Conhecimento em Textos.

No capítulo 3 são apresentados os procedimentos metodológicos que norteiam a pesquisa para construção deste modelo. No capítulo 4 é apresentado o modelo e os principais resultados e constatações encontrados mediante da aplicação deste modelo no case do Portal e-Cidadania. Por fim, o capítulo 5 é responsável pelas considerações finais da dissertação, contribuições, limitações e as recomendações para trabalhos futuros.

2 REVISÃO DA LITERATURA

O presente capítulo tem por objetivo explicitar conceitos basilares para o desenvolvimento, fundamentação e compreensão dessa pesquisa, possibilitando na chegada do resultado proposto. Buscou-se conceitos para o desenvolvimento da dissertação nos seguintes temas: Inovação, Gestão de Ideias, Descoberta de Conhecimento em Textos.

2.1 INOVAÇÃO

Tendo em vista a existência de uma competitividade, as organizações têm buscado investir em novos produtos e estratégias de atuação, em virtude disso, a inovação tornou-se algo crucial para manter a organização viva no mercado (GIBSON; SKARZYNSKI, 2008). A inovação é frequentemente associada às questões tecnológicas, no entanto, tem seu entendimento e aplicabilidade em várias áreas (BANERJEE, 2014).

Schumpeter foi um dos primeiros autores a tratar sobre o conceito de inovação, em seus trabalhos "*The theory of economic development"* (1912) e "*Capitalism, Socialism and Democracy"* (1942) discursou sobre uma força que causava transformação contínua das estruturas sociais, institucionais e econômicas (SHIMA, ESTEVÃO 2016).

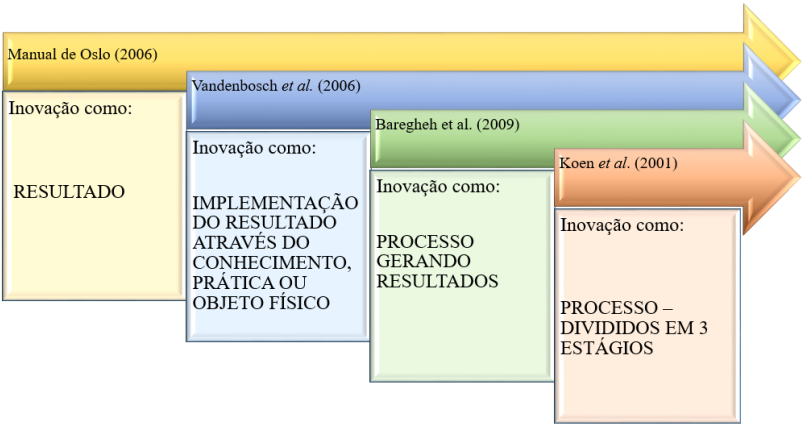
Como mencionado no primeiro parágrafo, a inovação tem sido um ponto forte para que as organizações se mantenham competitivas no mercado. Com a necessidade de possuir um melhor entendimento dos mecanismos que propiciam ou prejudicam o processo de inovação, a OECD criou o Manual de Oslo. Este documento apresenta propostas que são diretrizes para coleta e interpretação de dados sobre inovação, tratando-a de forma abrangente e multidimensional (OECD, 2005).

O Manual de Oslo define a inovação como resultado, seja em produto, processo, marketing ou método organizacional (CROSSAN; APAYDIN, 2010). Vandenbosch, Saatcioglu e Fay (2006) apresentam em seus estudos, que a inovação é a implementação de uma ideia criativa, que pode ser expressa na forma de conhecimento, de uma prática ou de um objeto. Já Baregheh, Rowley e Sambrook (2009) definem a inovação como um processo composto de várias etapas, das quais as organizações transformam as ideias em produtos novos ou melhorados, serviços ou processos, afim de buscar diferenciação positiva no mercado.

Corroborando com os conceitos acima apresentados, Quintane *et al.* (2011) coloca que o conhecimento possui concomitantemente as

características de ser duplicável, ser novo no contexto em que é introduzido e de demonstrar utilidade, sendo a inovação apresentada como resultado. Baragheh Rowley e Sambrook (2009) confirmam o conceito de inovação como processo de Kanter (1984), quando diz que inovação é o processo de trazer novas ideias de resolução de problemas que estão em uso. A Figura 2 ilustra a evolução dos conceitos de inovação.

Figura 2 - Complementaridade dos conceitos de Inovação



Fonte: do autor.

Para a presente dissertação, será utilizada a complementaridade dos conceitos apresentados, conforme Figura 2, pois a inovação será tratada como um processo que possui várias etapas, e tem por objetivo minimizar os riscos e aumentar as possibilidades de sucesso, gerando um resultado, sendo este implementado por meio do conhecimento.

Deste modo, o processo de inovação deve ser precedido pela obtenção de um conhecimento novo. Esse processo requer ônus, necessita de tempo e apresenta riscos (CHIBÁS; PANTALEÓN; ROCHA, 2013). Manter equipes de pesquisa e desenvolvimento não é simples para empresas, independentemente de seu tamanho, sendo de pequeno ou grande porte (BESSANT; TIDD, 2009). Para Chesbrough (2003) as competências internas das empresas não são mais suficientes para seu desenvolvimento, sendo necessário considerar uma abertura para que as inovações fluam entre os meios internos e externos.

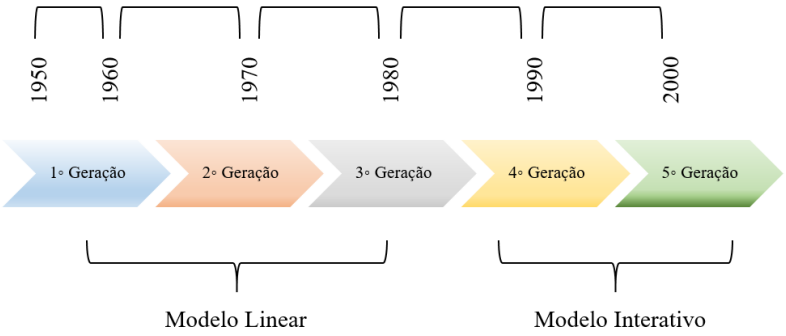
Nesse sentido, Chesbrough (2003) cunhou em seu trabalho o termo inovação aberta, que diz respeito aos limites das organizações,

movido por um sistema de relação, compreendendo a organização e seus parceiros externos, contrário ao modelo de inovação fechada. Para a diferenciação entre os modelos de inovação (aberta e fechada), Lindergaard (2011) mencionou que os paradigmas de inovação fechada e inovação aberta são diferenciados pelo modo da realização das atividades de seleção de ideias. Assim, a inovação fechada foca no ambiente interno, a inovação aberta realiza uma integração entre ideias e tecnologias externas à organização. No entanto, para chegar ao modelo de "inovação aberta", o processo de inovação passou por várias gerações com características bem específicas, que serão tratadas no tópico abaixo.

2.1.1 Processos da Inovação e seus modelos

Para Rothwell (1994), os modelos de inovação podem ser divididos em cinco gerações, em que a cada evolução a geração anterior é superada, conforme Figura 3.

Figura 3 - Gerações do Processo de Inovação, para Rothwell (1994)



Fonte: do autor, Baseado em Rothwell (1994).

Antes de começar a falar sobre a primeira geração, é importante explanar sobre o MODELO LINEAR. O modelo de inovação linear destacou-se entre os períodos que compreende os anos de 1950 a 1986 e neste período, a inovação foi reconhecida como modelo "ofertista" ou pela expressão *science push* (BARBIERI, 2004).

A **primeira geração**, segundo Rothwell (1994) se apresentou entre a década de 1950 e a metade da década de 1960, e ficou caracterizada pelo modelo *technology push* ou tecnologia empurrada. Esse modelo possui ênfase no P&D e seu processo de inovação é

sequencial, linear e simples, em que o mercado é apenas um receptor das pesquisas desenvolvidas na universidade. O desenho e a engenharia de um novo produto são encaminhados para a industrialização a partir dos resultados da pesquisa básica desenvolvida pelos cientistas, então, é sucedido pela fase de *marketing* e vendas. (ROTHWELL, 1994).

O modelo da **segunda geração** também possui características lineares, e está enquadrada entre 1960 e 1970. Este modelo considera que os novos produtos introduzidos no mercado possuem base especialmente da existência de tecnologias e o equilíbrio entre demanda e ofertas. Porém, diferente da primeira geração, esse modelo possui interesse na demanda do mercado (ROTHWELL, 1994). Esse modelo também é conhecido como *market pull*.

O período da **terceira geração** (1970 até meados de 1980) foi marcado por um crescente número de publicações de estudos empíricos a respeito do processo de inovação. Isso significa que pela primeira vez, o processo bem-sucedido de inovação pode ser modelado com base em um portfólio de serviços amplos e estudos sistemáticos, abrangendo diversos setores e países (ROTHWELL, 1994). Esse modelo de inovação, apesar de ser interativo, pois diversas atividades se realimentam, leva em consideração tanto as necessidades do mercado, quanto as tecnológicas. Ele ainda é essencialmente linear, sequencial, por possuir um loop de feedbacks (ROTHWELL, 1994).

Antes de entrar na quarta geração, cabe lembrar que a partir dela, limitações do modelo linear foram evidenciadas, consolidando então o MODELO INTERATIVO (CONDE; ARAÚJO-JORGE, 2003). Ressalta-se também que a partir deste modelo se inicia a captação de um maior número de ideias, pois envolve a diversas equipes para criação de inovação.

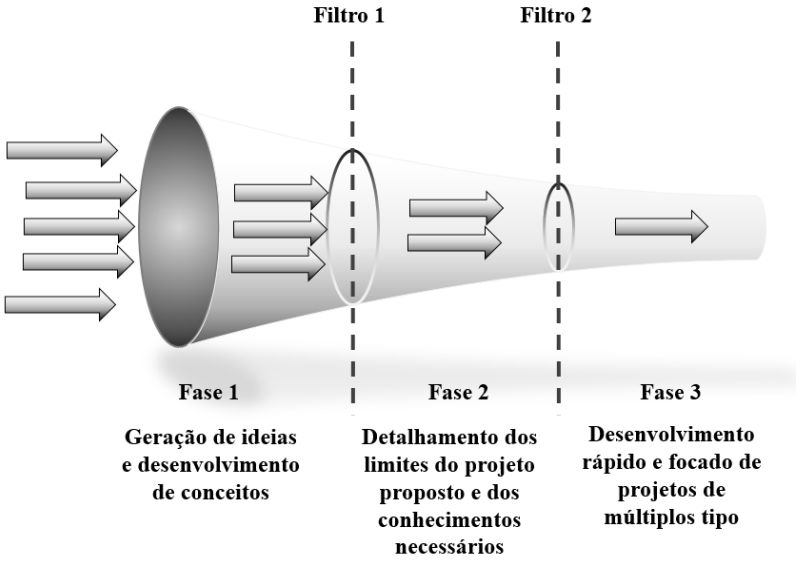
Para Rothwell (1994) a **quarta geração** (1980-1990) teve sua base em modelos de empresas japonesas, onde suas características principais são a integração e desenvolvimento em paralelo dos processos. Nesse modelo, para desenvolvimento de um novo produto, os fornecedores participam do processo, juntamente com diferentes equipes responsáveis pelo desenvolvimento. Nesta geração temos as fusões dos modelos anteriores: a inovação empurrada por pesquisa e desenvolvimento tecnológico e a inovação puxada pelas necessidades do mercado.

A **quinta geração**, e última para Rothwell (1994), estabelece de vez a integração entre as etapas, além de começar a considerar a velocidade do desenvolvimento um importante fator para a competitividade. Conhecida também como *networking model*, ela está

inserida entre 1990 e 2000, e vem de um aperfeiçoamento da quarta geração. A quinta geração tem como característica uma forte interação vertical dentro da empresa, interação horizontal externa, tais como pesquisa colaborativa, união de pesquisa, desenvolvimento e risco, alianças estratégicas para P&D de base, além de possuir desenvolvimento de processos integrados e paralelos e por fim, o uso de sofisticadas ferramentas eletrônicas (ROTHWELL, 1994). Tais processos de interação evidenciados nesta geração criam muitas ideias, e assim as barreiras para gerenciamento destas começa a ter um alto índice de complexidade.

Vários outros modelos formais são apresentados como alternativa para o processo de inovação da quinta geração, no entanto, para este trabalho evidencia-se o Funil de Desenvolvimento (CLARK; WHEELWIRGHT, 1993). A Figura 4 ilustra esse modelo.

Figura 4 - Funil de Desenvolvimento

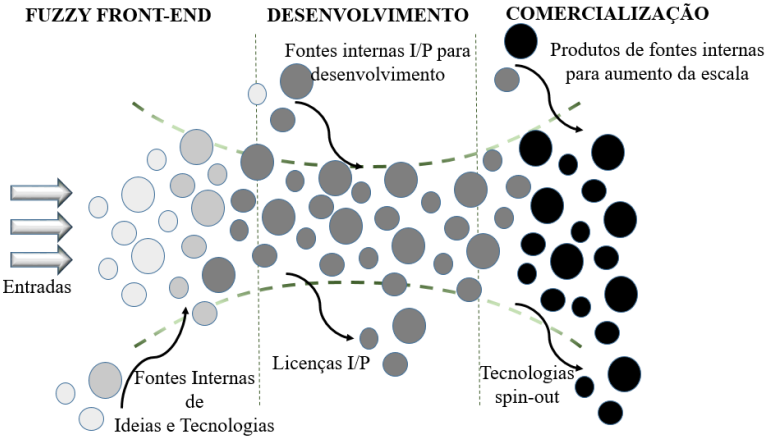


Fonte: Adaptado de Clark e Wheelwright (1993, p. 124).

O modelo acima inicia-se pelo planejamento de um conjunto de projetos, por meio de um processo com fases e avaliações, em que a organização mantém os produtos com maior probabilidade de sucesso até sua chegada ao mercado.

Importante ressaltar, que segundo Preez e Louw (2008) há ainda uma **SEXTA GERAÇÃO**, a qual é conhecida como inovação aberta ou Inovação em Redes. Essa geração tem como característica principal levar em consideração tanto ideias e caminhos internos, quanto externos à organização, construindo junto, o desenvolvimento de novas tecnologias. A proposta é buscar potencial em novas áreas de atuação, com novos conhecimentos, permitindo que a organização explore essas possibilidades. Assim, a característica dorsal dessa geração é considerar fatores externos como motores do processo de inovação (PREEZ; LOUW, 2008). A Figura 5 ilustra o modelo dessa geração, segundo Preez e Louw (2008).

Figura 5 - Modelo da Sexta Geração

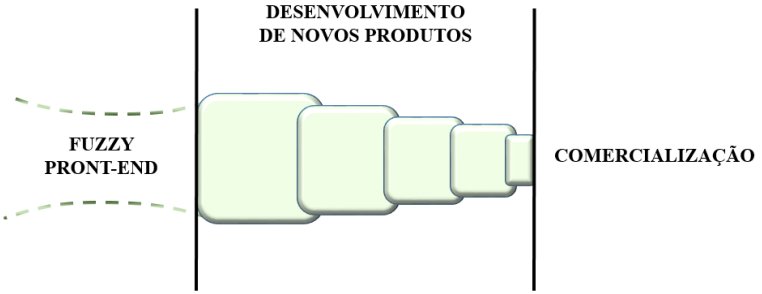


Fonte: do autor, adaptado de Preez e Louw (2008).

Para a presente dissertação, é utilizado o modelo de Koen *et al.* (2001), conforme Figura 6. Optou-se por esse modelo, pois dá ênfase ao estágio inicial do processo, o *Fuzzy Front-End* (FFE). Para este trabalho, será utilizado o termo *Front End* da Inovação, pois segundo Koen *et al.* (2001), o termo *Fuzzy Front-End* (FFE) parece difuso. Essa característica resulta em dificuldade em determinar quem é o administrador responsável por essa etapa, e indica que a parte inicial do processo não pode ser gerenciada. No tópico a seguir, será tratado de forma mais minuciosa o

Front End da Inovação, que é a primeira fase do modelo de inovação proposto.

Figura 6 - Modelo do Processo de Inovação



Fonte: do autor, Adaptado de Koen *et al.* (2001).

2.1.2 *Front End* da Inovação(FEI)

Como mencionado no tópico anterior, para a presente dissertação, adotou-se o modelo de processo de inovação de Koen *et al.* (2001). Para Teza (2012) diversos trabalhos têm verificado que as decisões tomadas na fase inicial do processo de inovação (FEI) podem influenciar as demais tomadas de decisões. A fase inicial do processo de inovação pode ser chamada de *Front End*, nesta fase são feitas as propostas de ideias ou soluções para determinado mercado, ou necessidade específica dos clientes ou mesmo da própria organização (ROCHADEL, 2016), portanto, essa fase será abordada de maneira mais detalhada para essa dissertação.

Teza (2012) destacou em seu trabalho de revisão sistemática da literatura sobre o FEI, que apesar de existirem diversos modelos do FEI, 3 elementos essenciais que se repetem. Sendo que o modelo de Koen et al. (2001) um modelo que trata dos três elementos em conjunto. Neste sentido, justifica-se a utilização do modelo de Koen para essa dissertação.

Para Koen *et al.* (2001) o *FEI* é o primeiro sub-processo do processo de inovação e envolve as atividades que ocorrem antes do desenvolvimento dos produtos, sendo as demais etapas, o processo de desenvolvimento de novos produtos e por último a comercialização.

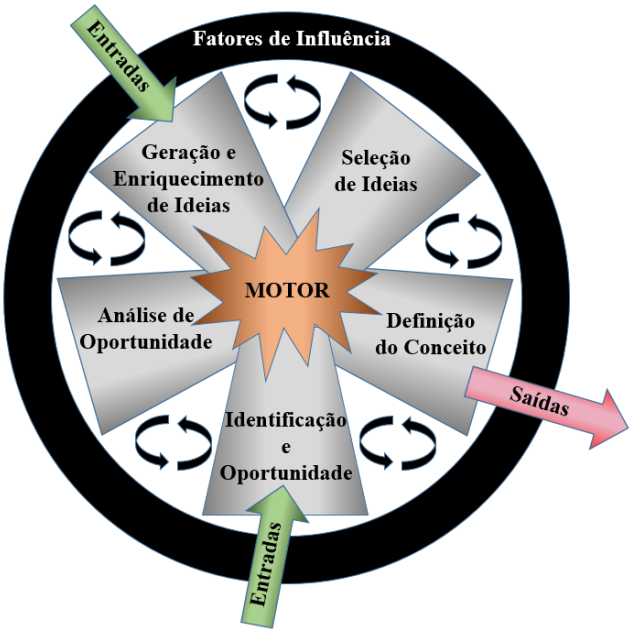
O FEI realiza o processo de descoberta de novas oportunidades, desenvolve proposição de ideias ou busca de soluções específicas para um determinado cliente, organização ou mercado (KOEN; BERTELS; KLEINSCHMIDT, 2014). Portanto, é um componente crucial para o processo de inovação, onde todas as escolhas realizadas nele podem determinar que caminhos a inovação deve tomar para o desenvolvimento e comercialização de produtos ou serviços. É composto por três elementos essenciais, sendo eles: oportunidade, ideias e conceito.

Para Koen *et al.* (2001) oportunidade é um gap de negócio ou tecnologia, percebido por uma empresa ou indivíduo, que existe entre o "hoje" e o "amanhã", a fim de capturar vantagem competitiva, estar preparado para uma ameaça, resolver um problema, ou até mesmo melhorar uma oportunidade. Ideia é a forma mais inicial de um novo produto ou serviço, muitas vezes consistindo de uma visão de solução de um problema identificado pela oportunidade. Stevanović, Marjanović, Storga (2012) corroboram com Koen *et al.* (2001) quando afirmam que ideia é o ponto inicial de qualquer processo de desenvolvimento. Além de argumentarem que ela é apenas uma apresentação de novos pensamentos, conceitos, entendimentos ou atitudes, que resultaram de atividades mentais, baseadas em conhecimentos e habilidades disponíveis.

Diferente dos termos apresentados acima, o conceito tem uma forma mais definida, com descrição escrita e visual, incluindo suas características e benefícios aos clientes em combinação com um conhecimento amplo da tecnologia necessária para desenvolvimento do conceito (KOEN *et al.*, 2001).

Os elementos acima apresentados fortalecem a escolha do modelo para o presente trabalho, já que conceitua elementos importantes para a pesquisa, além de ser um modelo interativo, que permite o dinamismo em sua implementação. Para demonstrar como se dá as etapas do FEI, Koen *et al.* (2001) propôs o modelo de Desenvolvimento de Novos Conceitos (NCD), que divide o *Front End* em três áreas: o motor, a roda e o aro conforme observa-se na Figura 7.

Figura 7 - Modelo de Desenvolvimento de Novas Ideias – fases do FEI



Fonte: do Autor, adaptado de Koen *et al.* (2001).

O motor é responsável por fornecer energia para o FEI dando suporte para as cinco atividades motoras. A roda compreende os cinco elementos de atividade do *FEI*, sendo estas a identificação de oportunidades, análise de oportunidades, geração de ideias, seleção de ideias e definição de conceitos e por fim o aro inclui os fatores que influenciam diretamente o motor e dão forma aos cinco elementos da atividade (KOEN *et al.*, 2002).

O modelo NDC, Figura 7, expõe uma forma não linear, no entanto, interativa entre os elementos. As entradas para esse modelo estão representadas pelas setas verdes, e a saída, pela seta vermelha, onde as entradas podem ser as ideias ou oportunidades, e a saída, um novo conceito para o desenvolvimento de novos produtos (NCD). O modelo circular demonstra que ideias e oportunidades estão interligadas, pois a oportunidade pode gerar ou testar uma ideia e a ideia pode levar a uma oportunidade (KOEN *et al.*, 2002).

O Modelo NCD de Koen *et al.* (2002) é dividido em cinco atividades fundamentais para o ciclo:

1) **Identificação de oportunidades:** é o momento onde organizações identificam as oportunidades que pretendem seguir. É regularmente direcionada pelos objetivos de negócios, onde uma oportunidade pode ser uma resposta a uma ameaça competitiva ou uma estratégia inovadora para se obter vantagens competitivas, permitindo acelerar, simplificar ou reduzir os custos dos processos internos da organização. A oportunidade também pode ser uma maneira de atualizar um produto existente, um novo direcionamento para negócio, uma plataforma para novos produtos, nova ofertas de serviços, novo processo de fabricação de produtos ou ainda uma nova estratégia de vendas ou marketing.

Para identificar uma oportunidade, é necessário estar alinhado com os fatores que influenciam esse processo. Nesta fase é possível utilizar ferramentas e técnicas de criatividade tais como: brainstorming, mapeamento mental e pensamento lateral, e também técnicas de resolução de problemas, que podem ser executadas por meio de diagramas de espinha de peixe, análise causal, mapeamento de processos e teoria de restrições.

2) **Análise de oportunidades:** depois de identificar a oportunidade, é necessário adquirir mais informações para julgá-las em oportunidades de negócios e tecnologia além de poder verificar se é possível realizar avaliações sobre tendências de mercado e tecnologia. Nesta fase, diante de grupos focais, são realizados muitos trabalhos, para identificar estudos de mercado e/ou experimentos científicos. O esforço gasto está relacionado diretamente com a atratividade da oportunidade, com o trabalho futuro de desenvolvimento, com o tamanho do risco para o desenvolvimento desta oportunidade e também de como ela se assemelha com a cultura organizacional e estratégia desta organização. Nesta fase a inteligência competitiva e as análises de tendências são altamente requisitadas.

3) **Geração de ideias**: o processo evolutivo que representa a construção, combinação, remodelação, modificação e atualização de ideias é denominado *Gênises*, que é o nascimento, desenvolvimento e maturação da oportunidade em uma ideia concreta. A ideia pode sofrer diversas mudanças à medida que é estudada, discutida e desenvolvida. Esse processo é reforçado pelos vínculos estabelecidos com clientes/usuários ou com equipes multifuncionais.

Para que nasça ideias novas ou modificadas para a oportunidade identificada é possível utilizar-se de um processo formal, que pode incluir sessões de brainstorming e banco de ideias, que é a Ideias *Gênises*. Uma

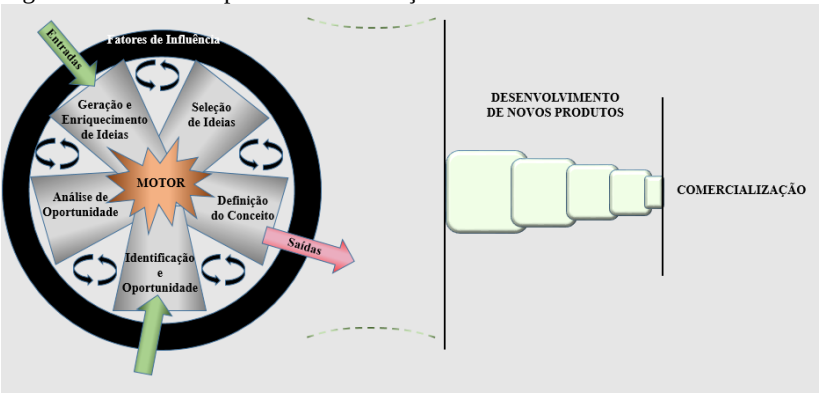
ideia nova também pode surgir de um ensaio que não funcionou, ou de um fornecedor que ofereceu um material ou de um usuário que fez um pedido incomum, ou seja, elas podem surgir fora da Ideias *Gênises*, demonstrando que os elementos NCD podem seguir de uma forma não linear, avançando e nutrindo ideias e oportunidades onde quer que elas ocorram, ou seja, Ideia *Gênises*. O retorno esperado desta fase é uma descrição mais estruturada da ideia ou do conceito do produto.

4) **Seleção de ideias**: nas organizações, existe uma infinita possibilidade de criar ideias de produtos/processos, o que torna crítica a decisão de escolha, a fim de atingir o ponto ideal, de maneira que agregue valor aos negócios. A seleção de ideias é uma importante etapa, no entanto, a pouca informação e compreensão deste momento torna difícil a seleção formalizada e a alocação de recursos no FEI. É necessário que sejam desenvolvidos modelos melhores para seleção de ideias do FEI, a fim de que os níveis de investimento, riscos de mercado e tecnologia, capacidades organizacionais, realidades competitivas, ao lado dos retornos financeiros, possam ser levadas em consideração. O processo de seleção e a Análise de Oportunidades não devem ser muito rigorosos, pois muitas ideias devem ser aceitas para que elas possam crescer e prosseguir, mesmo com incertezas de sucesso.

5) **Definição de conceitos:** o termo chave para essa etapa é o Business Case. A etapa é considerada a parte final do modelo e engloba melhorias e avanços do business case, baseando-se em estimativas de potencialidade de mercado, na necessidade do cliente, condições de investimento, estudo de concorrentes, desconhecimento de tecnologia e risco geral do projeto. Em determinadas organizações, a definição de conceitos é considerada a fase inicial do processo de desenvolvimento de novos produtos.

Então, sabendo como as atividades do FEI são formalizadas, pode-se finalizar com a Figura 8, para melhor compreensão do modelo.

Figura 8 - Modelo do processo de inovação inteiro

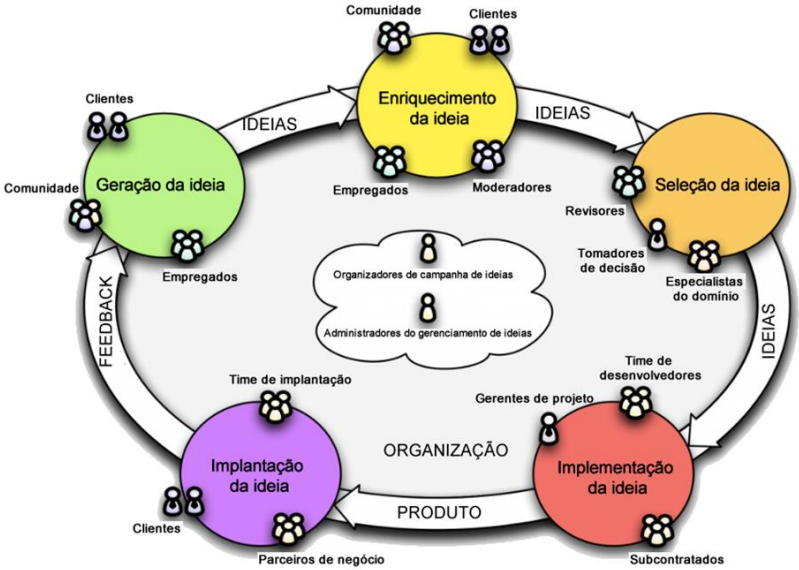


Fonte: do autor, adaptado de Koen *et al.* (2001).

2.2 GESTÃO DE IDEIAS

Para falar de gestão de ideias deve-se primeiramente saber que ideia é uma ocorrência de atividades mentais baseadas nas habilidades e conhecimentos disponíveis que resultam em novos pensamentos, conceitos, entendimentos ou atitudes (STEVANOVIĆ; MARJANOVIĆ; TORGA, 2012). Ideias podem ser representadas por meio de uma frase, um parágrafo ou até mesmo um rabisco, sem grandes detalhes (KEMPE *et al.*, 2011), concebendo um valor imprescindível para a organização, sendo necessária uma gestão, devido sua complexidade. Koen *et al.* (2001) definem ideia como uma forma embrionária de uma possível inovação. As ideias se propagam por meio de um ciclo de vida, com etapas específicas. Cabe ressaltar que esta dissertação, terá sua limitação na etapa seleção de ideia. A Figura 9 ilustra este ciclo.

Figura 9 - Ciclo de Vida das Ideias



Fonte: Rochadel (2016).

Sendo que a primeira etapa para qualquer processo de desenvolvimento é baseado em uma ideia, Stevanović, Marjanović e Ltorga (2012), buscam processos para a gestão dessas ideias, tais como; gerar, organizar, validar, classificar, descrever, armazenar e selecionar, com o objetivo de alimentar a inovação incremental e radical empregando diferentes métodos (BJORK *et al.*, 2010).

Apesar do termo gestão de ideias ainda não ser consolidado na literatura, é abordado em muitas publicações que tratam do ciclo de vida das ideias (ROCHADEL, 2016). Jansen (2012) apresenta que as pesquisas acerca de gestão de ideias começaram, nos últimos anos, investigar mais profundamente as questões sobre sistemas de gestão de ideias, tal como sua implicação organizacional e interação, pois trata-se de um grande desafio para as organizações fazer seleção de ideias potenciais quando estas estão em banco de dados de ideias.

Contudo, sistemas de gestão de ideias depara-se com desafios, tais como: sobrecarga de informações, devido aos picos de ideias triviais e redundantes, e esforço humano despendido com o processo de avaliação e seleção de ideias (WESTERSKI; DALAMAGAS; IGLESIAS, 2013).

As tecnologias da informação são empregadas pelos sistemas de gestão de ideias para auxiliar o processo de inovação (FENN; LEHONG, 2011) e possibilitar a administração, avaliação e seleção de ideias para serem empregadas pelas organizações (WESTERSKI; DALAMAGAS; IGLESIAS, 2013). E estão sendo mencionadas como um promissor ramo da indústria de software (FENN; LEHONG, 2011).

2.2.1 Sistema de Gestão de Ideias

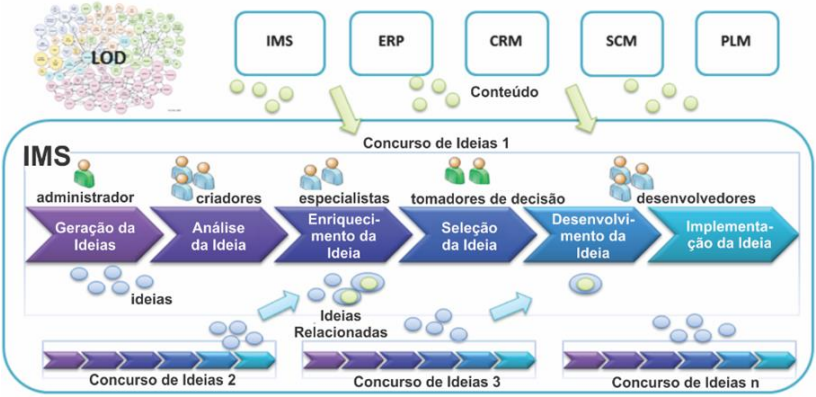
Os Sistemas de Gestão de Ideias (Idea Management System – IMS) são aplicações que proporcionam por meio de suas ferramentas criar, armazenar, procurar, editar, comentar e votar ideias (PEREZ; LARRINAGA; CURRY, 2014) de forma que devem dar suporte a diferentes tipos de concursos/campanhas e atores para desempenharem suas funções (BETTONI *et al.*, 2010).

Portanto, o objetivo destes sistemas é servir como uma ferramenta para organizar a coleta de ideias, auxiliar na avaliação e seleção proporcionando a organização uma base para o processo de tomada de decisão e aproveitar ideias com potencial para se manterem competitivas no mercado (PEREZ; LARRINAGA; CURRY, 2014).

Westerski e Iglesias (2011) definem IMS como sistemas baseados em conhecimento. Afirma também que se trata de uma categoria de sistema para a gestão do conhecimento que pode ser utilizado nas organizações para reunir ideias e fomentar a participação da equipe. Perez; Larrinaga e Curry (2014) evidenciam a necessidade de interligação com os demais sistemas da organização para que ocorra o fluxo do conhecimento impactando diretamente com o alinhamento dos objetivos e estratégias das organizações. De modo que estes sistemas podem representar um importante papel para tomadas de decisões para inovação em produto ou serviços/processos da organização.

Perez; Larrinaga e Curry (2014) propõem que um IMS deve seguir uma arquitetura com o ciclo de vida com seis etapas conforme ilustrado na Figura 10.

Figura 10 - Arquitetura de um IMS



Fonte: Rochadel (2016)

Esta arquitetura proposta por Perez; Larrinaga e Curry (2014) tem as etapas de: geração de ideias, análise da ideia, enriquecimento da ideia, seleção da ideia, desenvolvimento da ideia e implementação da ideia, de forma que cada etapa é suportada por atores com funções distintas, tais como gerentes ou administradores, especialistas, desenvolvedores, tomadores de decisão e autores. Ressalta-se que nesta arquitetura o processo de gestão de ideias segue um fluxo linear. Os autores evidenciam a importância de obter dados de várias bases externas para alimentar estas informações importantes para o ciclo da gestão de ideias, representados na parte superior da Figura 10, que representam os sistemas corporativos tais como ERP (Enterprise Resource Planning), CRM (*Customer Relationship Management*), SCM (*Supply Chain Management*), PLM (*Product Lifecycle Management*) e além de outros Sistemas para Gestão de Ideias.

Os autores Westerski e Iglesias (2011) propõe um modelo não linear e recursivo para gestão de ideias nos IMSs contendo cinco etapas. Neste modelo, as funções dos atores identificados se diferem um pouco da arquitetura já apresentada, porém envolve uma comunidade externa, se apresentando como um processo mais aberto. Cabe ressaltar que é um processo cíclico. Este fato é uma característica relevante que se difere da arquitetura proposta por Perez; Larrinaga e Curry (2014).

Quando se possibilita a participação externa, pode-se rapidamente gerar centenas de ideias provenientes de especialistas, consumidores ou funcionários (TOUBIA, 2007). Quando bem aplicado um método para gerar ideias, especialmente quando se envolve a colaboração e a

fomentação de campanhas em plataformas web podem acumular uma grande quantidade de comentários e publicações rapidamente, portanto geram um grande desafio para a gestão e tratamento de todo este conteúdo que é gerado (LUNING; PENGZHU, 2009; ELERUD-TRYDE; HOOGE, 2014; MURAH *et al.*, 2013).

Contudo, ainda possui uma característica peculiar deste processo complexo, visto que os dados gerados não são estruturados e transforma este processo muito dependente dos gestores que possuem conhecimento deste domínio específico. Como uma alternativa então inicia-se o foco na criação de sistemas computacionais para facilitar a gestão do conteúdo com intuito de analisá-lo e processá-lo (MURAH *et al.*, 2013).

Spancer (2012) relata que o uso sustentado e bem-sucedido de um sistema para a gestão de ideias, deve se preocupar como o fato de que em campanhas grandes ou executadas num curto período de tempo onde há muita iteração de usuários, há duplicação considerável e sobreposição de ideias submetidas. Isso representa uma sobrecarga de trabalho para os revisores, tendo em vista que normalmente nomeiam poucas pessoas para verificar as centenas de ideias. Assim, este processo de estruturar os bancos de dados, ainda é um desfio a ser superado pelas organizações (LUNING; PENGZHU, 2009; MURAH *et al.*, 2013).

A exemplo do que está sendo tratado, são apresentados três exemplos de empresas reconhecidas mundialmente, que utilizam sistemas de gestão de ideias:

1) a LegoIdeas[2] incentiva novas ideias a partir da sua plataforma, onde os criadores postam suas ideias. A ideia passa por uma comissão de revisão da LEGO que pode ou não implementar. Caso aprovado, os produtos fabricados mediante as ideias dos colaboradores são vendidos ao redor do mundo, e os colaboradores recebem *royaltys* e também os créditos de criador do produto. As ideias nesta aplicação possuem um contexto bem detalhado para arrecadar maior número de apoios. A ferramenta também se destaca por permitir o uso de *folksonomias* e comentários dos apoiadores que podem sugerir melhorias no projeto.
2) Outra empresa que decidiu criar a possibilidade de para enviar ideias foi a IBM para seus diversos produtos. Ela fornece feedback integrado e automatizado para conectar suas ideias com equipes responsáveis pelo desenvolvimento de produtos e engenharia da IBM.

[2] Disponível em: < https://ideas.lego.com/> Acesso em mar. 2018.

3) A empresa Dell também usa um portal chamado IdeaStorm[3] para receber ideias sobre seus produtos e se baseada em *crowdsourcing*, contando com 28.146 ideias enviadas, 747.981 votos, 103.509 comentários e com 550 ideias implementadas. A empresa recebe ideias sobre qualquer dos seus produtos de forma aberta, porém cria tópicos específicos para incentivar a geração de ideias para seus produtos.

Outras organizações reconhecidas mundialmente também possuem sistemas de gestão de ideias que podem estar envolvidos em seus processos de inovação, a exemplo da Starbucks. Esses exemplos comprovam que com o advento da tecnologia e internet tornou-se mais fácil para comunidades externas participarem dos processos de inovação nas empresas. Entretanto, pelo grande volume de dados, fica evidente a necessidade de sistemas que tratem desses conteúdos, para possibilitar a descoberta de conhecimento.

2.3 RECONHECIMENTO DE PADRÕES

O reconhecimento de padrões é uma área de pesquisa que possui o objetivo de classificar objetos (padrões) em várias categorias ou classes, de modo que busca atribuir um padrão a um conjunto desconhecido de classes de padrões, que corresponde a clusterização, um processo não supervisionado de aprendizagem, ou ainda identificar um padrão como membro de um conjunto conhecido de classes, que corresponde a classificação, um processo supervisionado de aprendizagem (THEODORIDIS; KOUTROUMBAS, 2009). Segundo Duda (2001) reconhecimento de padrões é o ato de observar os dados brutos e tomar uma ação baseada na categoria de um padrão.

Os autores Tou e Gonzáles (1981) entendem por padrão as características que possibilitam agrupar objetos semelhantes dentro de uma determinada classe ou categoria, diante da interpretação dos dados de entrada, que permitam a extração das propriedades relevantes desses objetos. Quanto a classe, os autores definem como um padrão, um conjunto de atributos comuns entre os objetos.

Para Jain (2000) pode-se distinguir o termo classificação em supervisionada ou não supervisionada. A classificação supervisionada há a seleção de amostras representativas para cada uma das classes que se deseja classificar um novo objeto, neste cenário o padrão e as classes estão

[3] Disponível em: < http://www.ideastorm.com/> Acesso em mar. 2018.

predefinidos. Na classificação não supervisionada (clusterização) não se possui um padrão pré-estabelecido, sequer o número total de clusters a serem encontradas durante o processo de classificação.

Ainda segundo Jain (2000), o conjunto de dados é dividido em grupos, por meio de suas características específicas, tais que os objetos dentro de um grupo (cluster) sejam mais similares do que os pontos de outros grupos. Desta forma, isto nos remete a uma análise de agrupamentos. Um projeto de reconhecimento de padrões deve possuir as seguintes etapas:

1. extração de características dos objetos a classificar ou descrever;
2. seleção das características mais discriminativas;
3. construção de um classificador ou descritor.

Conforme Duda (2001), para os tipos de objetos a classificar ou descrever, se pode utilizar algumas abordagens como:

- Abordagem estatística – corresponde a uma abordagem clássica, de modo que assume que as características das classes são regidas por determinados modelos probabilísticos;
- Abordagem sintática – busca descrever a estrutura dos padrões usando inter-relações de características de descritores básicas denominadas primitivas;
- Abordagem neuronal – denominada abordagem tipo "caixa preta", de forma que procura determinar um mapeamento ótimo entre entradas e saídas inspirando-se nos modelos de neurônios do cérebro;
- Abordagem difusa - abordagem que tem em conta o grau de incerteza por vezes inerente a características e a classificações, usando a teoria dos conjuntos difusos para modelar esse grau de incerteza.

Apesar de os métodos de análise do reconhecimento de padrões terem uma tradição de longa data, apenas recentemente iniciou-se o uso destes para pesquisas voltadas à esportes, meios de transporte, reconhecimento facial, entre outras áreas de estudo (GRIMMER *et al.*, 2009). E nestes diversos contextos o reconhecimento de padrões possui características fortes que fomentam aplicações, com contribuições para diversas atividades do cotidiano. O uso da metodologia de reconhecimento de padrões atua hoje sobre os e-mails, classificando-os

como lixo eletrônico (spam) ou não (KOPRINSKA *et al.*, 2007). Outras aplicações desta metodologia de reconhecimento de padrões remetem ao reconhecimento de fala e detecção de rosto, objetos, entre outras (FURUI, 2004). De modo que o valor de tais métodos está relacionado ao fato de que estes possuem uma forma de operar imparcial e também a capacidade de trabalhar com grandes volumes de dados (GRIMMER *et al.*, 2009).

O Reconhecimento de Padrões é uma área que pode ser utilizada para a descoberta de conhecimento em bases de dados, conforme será descrito no próximo tópico desta seção.

2.4 DESCOBERTA DE CONHECIMENTO

A descoberta de conhecimento trata-se de uma atividade suportada pela engenharia do conhecimento. Ceci (2015) considera a engenharia do conhecimento como suporte às atividades intensivas em conhecimento. Ela tem por objetivo estabelecer metodologias, métodos e técnicas voltados à explicitação de conhecimento. Levando em consideração que a informação é o ativo mais importante para os negócios das organizações, torna-se algo essencial para ganho de competitividade entre as empresas de pequeno, médio e grande porte, conseguir extraí-las de forma correta, visando uma minimização na ocorrência de erros para a tomada de decisões por parte dos gestores.

Neste contexto, a engenharia do conhecimento por meio dos sistemas de descoberta de conhecimento pode oferecer à gestão de ideias métodos, técnicas e ferramentas para dar suporte as suas etapas e evidenciar o conhecimento contido nas bases de ideias.

A descoberta de conhecimento tem como objetivo principal buscar soluções para determinada situação ou problema por meio dos processos de identificação, recebimento de informações importantes, computando e agregando estas informações e assim mudando o estado de conhecimento atual, a fim de que determinada situação ou problema possa ser resolvido (WIVES, 2004).

A descoberta de conhecimento desdobra-se hoje em duas grandes áreas, a primeira é aplicada apenas em base de dados estruturadas e a segunda em bases não estruturadas de dados, mas ambas são utilizadas para mineração de dados, porém aplicadas em tipos diferentes de dados diferenciando-se por técnicas utilizadas.

Para esta pesquisa adota-se o conceito dado por Fayyad, Piatetsky-Shapiro e Smyth. (1996) em que define a mineração de dados como uma etapa do processo de Descoberta de Conhecimento e que consiste na

realização da análise de dados e na aplicação dos algoritmos ou métodos de descoberta de conhecimento, assim sendo possível inferir um conjunto de padrões sob determinados dados.

Por fim, com advento dos avanços tecnológicos a obtenção de novas informações por meio de processos de *Knowledge Discovery in Database* (KDD) têm sido facilitado, ou seja, descoberta de conhecimento em banco de dados é uma atividade suportada por processos tecnológicos. Surge também a Descoberta de Conhecimento em Texto (KDT), ambas tratadas nos tópicos abaixo.

2.4.1 Descoberta de Conhecimento em Base de Dados

O processo de extração do conhecimento é uma atividade dinâmica e evolutiva que envolve integrações com outras áreas de conhecimento como Estatística, Inteligência Artificial e Banco de Dados. A área que se dedica explorar grandes quantidades de dados com objetivo de identificar padrões úteis é conhecida como Descoberta do Conhecimento em Dados (KDD) (FELDMAN; DAGAN, 1995). Já em 1995 os autores afirmavam que era necessárias pesquisas para implementar métodos de manipulação de dados não estruturados para analisar grandes quantidades de informação, pois métodos tradicionais não eram suficientes.

Assim, os padrões extraídos devem ser além de úteis, também confiáveis e compreensíveis, para que se possa extrair e empregar o conhecimento, e também tirar proveito de alguma vantagem, seja científica ou comercial (FAYYAD; PIATETSKY-SHAPIRO; SMYTH 1996).

A KDD é reconhecida como um processo de descoberta de padrões e tendências por análise de grandes conjuntos de dados. Apresenta como principal etapa o processo de mineração. Esta consiste na execução prática de análise e de algoritmos específicos que, sob limitações de eficiência computacionais aceitáveis ou dados pré-definidos, produzem uma relação particular de padrões a partir de dados (FAYYAD; PIATETSKY-SHAPIRO; SMYTH 1996).

O campo de pesquisa que envolve KDD, para Fayyad, Piatetsky-Shapiro e Smyth (1996), compreende o desenvolvimento de métodos e técnicas que busquem fornecer significado aos dados.

A KDD é composto pela seleção dos dados, pré-processamento, mineração dos dados, validação dos resultados e análise e interpretação dos dados para aquisição do conhecimento, onde o pré-processamento é responsável pela adequação dos dados aos algoritmos e a mineração

geralmente é baseada em Inteligência Artificial (IA) ou estatística (MAIA; ROCHA, 2010).

Porém, cerca de 80% da informação produzida encontra-se em formato textual proveniente da linguagem natural (TAN, 1999; FELDMAN; DAGAN, 1995; KUECHLER, 2007), ou seja, a informação estruturada não compreende a estrutura adotada por todas as informações dispostas no contexto digital

Defronte desta afirmação, surgiu o processo de Descoberta de Conhecimento em Textos (KDT). O mesmo trabalha com uma coleção de documentos em linguagem natural, em busca de padrões e tendências para classificar e comparar documentos (SILVA; ROVER, 2011).

2.4.2 Descoberta de Conhecimento em textos

No cenário apresentado acima, surge a Descoberta de Conhecimento em Texto (KDT), que se torna mais complexa devido à falta de estruturação da informação. Compreende técnicas e ferramentas automáticas e inteligentes, responsáveis pelo auxílio na análise de grandes volumes de dados, com propósito de minerar conhecimento útil, aplicado a domínios que utilizem textos não estruturados. Feldman e Dagan (1995) descrevem a mineração de texto como técnicas de extrair informações a partir de diversas coleções de textos.

A KDT é definida por Wives (2004) como identificar, receber informações relevantes e poder computá-las e agregá-las ao seu conhecimento prévio, mudando o estado de conhecimento atual, a fim de que determinada situação ou problema possa ser resolvido. Pode-se definir que a descoberta de conhecimento em texto é um processo que envolve diversas ferramentas intensivas em conhecimento (técnicas, métodos e metodologias), a fim de procurar informações úteis a partir de uma fonte de dados.

2.4.2.1 Processamento da Linguagem Natural

O Processamento de Linguagem Natural (PLN) é uma área que se utiliza de técnicas computacionais para analisar e representar textos que ocorrem naturalmente em um ou mais níveis de análise linguística com a finalidade de alcançar processamento de linguagem semelhante a humana para uma série de tarefas e aplicações (LIDDY, 2001). Schenaider (2001) acrescenta ainda, que o objetivo geral da PLN é de processar a linguagem natural para ser compreensível por máquinas.

Essas aplicações segundo Oliveira (2002) podem ser aplicações baseadas em textos, como sistemas que procuram documentos específicos em uma base de dados. Ainda, aplicações baseadas em diálogos, as quais, referem-se às interfaces de linguagem natural para bancos de dados, sistemas tutores e os sistemas que interpretam e respondem a comandos expressados em linguagem escrita ou falada.

O PLN possui diferentes níveis, Liddy (2001) apresenta o nível: Fonológico, Morfológico, Sintático, Semântico, do Discurso e por fim o Pragmático.

O nível Fonológico trata de como interpretar aspectos do som Liddy (2001). Os demais níveis são relacionados a linguística. Como o nível Morfológico, o qual trata de analisar a composição das palavras até as menores unidades "morfemas". O nível Lexical, diz respeito a interpretação do significado das palavras individuais, principalmente com atribuição de *tags* com base no contexto (LIDDY, 2001).O nível Sintático centra-se na análise das palavras de uma frase, de modo a revelar a estrutura da sentença, para isso utiliza um analisador gramatical. É conferido ao nível Semântico a atribuição de significado as palavras, no entanto destaca Liddy (2001) que em todos os outros níveis há contribuições para determinar esse significado. Para que o processamento semântico possa determinar possíveis significados de uma frase, inclui-se tarefas como desambiguação semântica e sintática.

Enquanto a sintaxe e a semântica funcionam a nível frasal, o Discurso trabalha com unidade de textos maiores, concentrando-se na propriedade do texto como um todo. Por fim, a LPN no nível Pragmático está relacionada com o uso intencional da linguagem em situações e ela utiliza o contexto, além do conteúdo do texto, para fornecer compressão (LIDDY, 2001). Para concluir, segundo a autora é comum que sistemas implementem módulos com diferentes níveis de LPN.

Devido a esses diferentes níveis fica explícita a complexidade de se processar linguagem natural. Sendo assim, para realizar o processamento há diferentes categorias de abordagens, Liddy (2001) divide em: Simbólica, Estatística, Conexionista e Híbrida. No entanto para obter uma forma lógica adequada, o PLN pode se apoiar no conhecimento linguístico (simbólica) e em métodos estatísticos, não necessariamente de forma excludente. Inclusive, há indicação para utilização de abordagem híbrida (BOD,1995).

Gonzalez e Lima (2007) apresentam uma estratégia de processamento que envolvem o conhecimento linguístico.

a. Etiquetagem:

Para se trabalhar com o conhecimento linguístico a primeira etapa é utilizar um etiquetador. Eles podem ser tanto, gramatical, morfológica ou semântica.

Por exemplo, quando gramatical (*part-of-speech tagger*) vai identificar, com a colocação de uma etiqueta (*tag*) a categoria gramatical de cada item lexical do texto analisado (BICK, 1998). Por outro lado, um etiquetador morfológico inclui informações sobre categorias morfológicas, como substantivo e adjetivo, e um etiquetador sintático acrescenta etiquetas indicando as funções sintáticas das palavras, como sujeito e objeto direto. Já a etiquetagem semântica anexa informação relacionada ao significado, podendo indicar os papéis dos itens lexicais na sentença, como agente, processo e estado.

b. Normalização

A normalização linguística pode ser subdividida em três casos distintos morfológica, sintática e léxico-semântica.

A normalização morfológica ocorre quando há redução dos itens lexicais, sendo mais comum acontecer por:

1) ***stemming***, que reduz todas as palavras com mesmo radical a uma forma denominada *stem* (similar ao próprio radical) (ORENGO; HUYCK, 2001), sendo eliminados afixos ou sufixos oriundos de derivação ou de flexão (em alguns casos, apenas os sufixos são retirados).

2) **redução canônica** (*lemmatization*), que, geralmente, reduz os verbos ao infinitivo e os adjetivos e substantivos à forma masculina singular (ARAMPATZIS *et al*., 2000).

A normalização sintática ocorre quando há a normalização de frases semanticamente equivalentes, mas sintaticamente diferentes, em uma forma única e representativa das mesmas. Já a normalização semântica ocorre quando são utilizados relacionamentos semânticos entre os itens lexicais para criar um agrupamento de similaridades semânticas, identificado por um item lexical que representa um conceito único (GONZALEZ, LIMA, 2007).

c. Remoção das s*top words*

Stop words são palavras funcionais, como artigos, conetivos e preposições (BAEZA-YATES; RIBEIRO-NETO, 1999). A sua eliminação tem vantagens e desvantagens. Gamallo, Agustini e Lopes (2002) destacam que este tipo de termo pode exercer papel de composição de significado, mas quando separado não apresenta significado ao contrário de outras categorias gramaticais.

Para a remoção das *stop words* pode-se utilizar de gramáticas ou dados lexicais, como também de métodos estatísticos (GONZALEZ, LIMA, 2007).

2.4.2.2 Cálculo de Similaridade

As técnicas de similaridade podem ser uma das técnicas utilizadas para tratamento de textos. Os algoritmos que retornam similaridade entre documentos trabalham com métricas que permitem avaliar quantitativamente a semelhança entre dois objetos. Existem diversas medidas de similaridade como apresentado na revisão feita por Maia e Souza (2008).

No campo da estatística, o coeficiente de correlação de Pearson e do Cosseno são duas medidas de similaridade básicas que se expandem para outras áreas (MAIA, SOUZA, 2008).

A correlação de Pearson entre dois vetores retorna um valor entre -1 e 1. Se retorno for igual a 1 eles estão fortemente correlacionados, isto é, os valores de um vetor podem predizer os valores do outro. Se for igual a 0 não existe correlação. E se for -1 existe uma correlação inversamente proporcional (MAIA, SOUZA, 2008).

A métrica do cosseno é similar a correlação de Pearson, retornando valores entre 0 e 1. Ele mede o ângulo entre dois vetores num espaço vetorial. Quanto mais próximo de 1 for o valor, mais similares são os dois vetores, ou seja, quanto menor o ângulo, mais próximo de 0 será o cosseno e mais similar será o documento em relação a aquele termo, de maneira que os dois vetores são colineares (paralelos) (MAIA, SOUZA, 2008).

Quando se trabalha análise de similaridade no contexto das ideias trabalha-se com textos curtos e diferentemente dos textos longos não se pode tratar somente a frequência de termos (LI *et al.*, 2006). Ações já estão sendo empregadas neste sentido, um exemplo, no contexto das ideias, é demonstrado no artigo de Spancer (2012) com o método de similaridade de Jaccard-Tanimoto.

Nesse método a similaridade de dois vetores (V1 e V2) é denotada quando há ocorrência de igualdade entre os dois vetores, dessa forma, incrementando o contador que irá gerar o grau de similaridade. O vetor é composto pelos conjuntos A, B e C, sendo que: A contém os valores que são iguais nos dois vetores; B contém somente os valores que não coincidem; C contém os valores não coincidentes de V2. Logo, a fórmula é: Jaccard = A/(A+B+C) (CHAPMAN, 2009).

Segundo Spencer (2012) esse método é bastante utilizado na área química para medir a semelhança entre moléculas. Desta maneira o autor propõe para o contexto das ideias.

Como exemplo, na área química quando uma molécula X é interessante, outras moléculas semelhantes são buscadas nas coleções próprias, na literatura ou em patentes. O mesmo pode ser aplicado ao contexto de gestão de ideias. Se uma ideia é interessante, outras similares podem ser encontradas nos sistemas e banco de dados, assim como, encontrar pessoas que tiveram ideias semelhantes a esta, mediante da aplicação de medidas de similaridade.

2.4.2.3 Análise de Agrupamentos

A análise de agrupamento pode ser reconhecida como análise de segmentação, análise de taxonomia, data *clustering*, análise de grupos, entre outros termos, e visa identificar objetos homogêneos em um conjunto de grupos, denominados clusters, por determinados critérios (HANSEN; JAUMARD, 1997). A análise de agrupamento, ou cluster, associa um item a uma ou várias categorias (ou clusters), determinando as classes pelos dados, independentemente da classificação pré-definida.

Clusterização é uma técnica muito importante no processo de descoberta de conhecimento para seres humanos e sua história pode ser rastreada até os tempos de Aristóteles. Os clusters são definidos por meio do agrupamento de dados baseados em medidas de similaridade ou modelos probabilísticos, visando detectar a existência de diferentes grupos dentro de um determinado conjunto de dados e, em caso de sua existência, determinar quais são eles (HANSEN; JAUMARD, 1997).

A análise de agrupamentos realizada no campo da computação possui o intuito de lidar com conjuntos de dados de grande escala e complexos. Com o desenvolvimento de técnicas baseadas em computação, o agrupamento de dados em clusters tem sido amplamente utilizado em mineração de dados, processamento de imagens, aprendizado de máquina, inteligência artificial, reconhecimento de padrões, análise de redes sociais, análise de comportamento de clientes, marketing para e-business entre outros campos (HARTIGAN, 1975; JAIN; MURTY; FLYNN, 1999).

Para Carlsson (2014) a clusterização é um agrupamento de um conjunto de dados em clusters diferentes de forma que é reunido elementos que possuam alguma característica semelhante num mesmo grupo. Quando se trata de campos textuais este conjunto de dados podem

ser representados por uma coleção de artigos, notícias, trabalhos de pesquisa ou qualquer material escrito que possa ser segmentado em grupos de documentos similares, em prol de buscar algum significado semelhante que possa ser destacado entre estes documentos. Evidencia-se que a noção de similaridade é um fator essencial quando se trata de agrupamento.

2.4.2.3.1 Algoritmos Hierárquicos

Algoritmos hierárquicos juntamente com os não hierárquicos consistem em técnicas de agrupamento. A abordagem hierárquica é considerada a mais simples, pois no decorrer de divisões sucessivas entre os dados, origina uma representação baseada em árvore (JAIN; DUBES, 1988; EVERITT, 2001).

Dentro das técnicas hierárquicas há duas abordagens: as aglomerativas e as divisivas. Ambas possuem a vantagem de que determinado objetivo só pode ser atribuído a um grupo, não podendo ser realocado a outro grupo (LATTIN; DOUGLAS; PAUL, 2011). Além disso, O resultado obtido com aplicação de técnicas hierárquicas pode ser apresentado por uma árvore de classificação denominada dendograma.

A diferença básica entre as duas é que a aglomerativa busca reunir os objetos em grupos cada vez maiores, incluindo também os agrupamentos já formados. Já a abordagem divisiva os objetos partem de um único grupo que sofrerá divisões sucessivas até cada objeto estar em um agrupamento separado (WIVES, 2004).

As duas começam a partir de uma matriz de similaridade. Sendo que a similaridade pode ser feita de três formas segundo Jain e Dubes (1988):

1) Algoritmo hierárquico de ligação simples (*Single Linkage*), monta os agrupamentos de acordo com a maior similaridade entre quaisquer objetos de dois grupos.
2) Algoritmo hierárquico de ligação média (*Average Linkage*) a similaridade é obtida por meio da média de distância entre todos os objetos de dois grupos em questão.
3) Algoritmo hierárquico de ligação completa (*Complete Linkage*) obtém a similaridade por meio da menor distância entre dois objetos de grupos distintos.

2.4.2.3.2 Algoritmos de Particionamento

Os algoritmos por particionamento foram desenvolvidos para agrupar objetos em *n* grupos, definidos antecipadamente ou definidos durante a execução do processo (JOHNSON; WICHERN, 2007).

De acordo com Hair *et al.* (2010) este algoritmo designa objetos a agrupamentos, levando-se em conta a definição da quantidade de grupos. Segundo Fung (2001) os métodos por particionamento são extremamente mais rápidos que métodos hierárquicos.

Os algoritmos por particionamento possuem desvantagens, como o fato de elencar o número de agrupamentos a serem formados. Caso este número seja escolhido erroneamente, a cada iteração do algoritmo resultados diferentes podem surgir, o que poderá impor uma estrutura de dados, ao invés de identificar a estrutura inerente ao processo (FUNG, 2001; KAINULAINEN, 2002).

K-means é o mais conhecido algoritmo baseado em particionamento (JAIN; MURTY; FLYNN, 1999), foi introduzido por J. B. MacQueen em 1967 e é um dos mais simples algoritmos de aprendizagem não supervisionada.

Uma explicação é fornecida por Sérgio (2016) sobre o funcionamento do *k-means*, no qual primeiramente, o algoritmo inicia com a informação de quantos grupos serão formados durante o processo de agrupamento. Posteriormente, o algoritmo distribui um elemento para cada grupo. Inicialmente estes elementos serão a semente inicial de cada grupo e consequentemente o centróide. Durante a iteração do algoritmo, à medida que novos elementos forem atribuídos aos grupos, o centróide é recalculado, representando a média entre os elementos. O *k-means* utiliza geralmente a distância euclidiana para calcular a distância entre os elementos.

2.4.2.4 Categorização de Textos

Gerenciar o crescente número de ideia que são criadas diariamente é um desafio a ser enfrentado pelas organizações. À vista disso, as técnicas de mineração de textos podem auxiliar a extração de informações não-triviais de repositórios de documentos não estruturados, neste caso, chamados de bases de ideias. Uma destas técnicas para classificação de documentos consiste em rotular textos elaborados em linguagem natural em categorias pré-estabelecidas, conhecida como Categorização de Textos. De modo que podemos definir um documento como um objeto que contém elos e regras que o associam a outros documentos. (OLIVEIRA; MENDONÇA, 2004).

Sebastiani (2002) e Joachims (1996) definem a Categorização de Textos (CT), como a atividade de rotular documentos de textos em linguagem natural em categorias temáticas a partir de um conjunto pré-definido. Neste sentido, os Algoritmos Bayessianos têm sido utilizados com sucesso na confecção de modelos para a classificação de documentos a partir de um conjunto de amostras para treinamento.

Contudo, nota-se que a precisão destes classificadores depende diretamente do conhecimento acumulado nestes conjuntos de treinamento, sendo que isto pode demandar uma grande porção de informações rotuladas e em consequência mais tempo e dedicação de especialistas de domínio (OLIVEIRA; MENDONÇA, 2004).

Até o final da década de 1980, o processo categorização de documentos baseava-se em definir manualmente um conjunto de regras, que tinham por função representar o conhecimento de especialistas do domínio, para classificar documentos em uma categoria específica. Cabe se destacar que a CT é um dos campos da engenharia do conhecimento que tem por um de seus objetivos explicitar conhecimento de especialistas para processos de inferência.

A partir da década de 1990 esta abordagem se modificou, com o ingresso de algoritmos de aprendizado de máquina para classificação de textos (SEBASTIANI, 2002). O objetivo destas técnicas é ensinar os classificadores, a partir de exemplos, que reconheçam de forma automática as características intrínsecas de cada categoria, assim encontrando padrões para classificar os demais elementos.

2.4.2.4.1 Classificação de textos

Os algoritmos utilizados para classificação de documentos são baseados em métodos indutivos. De modo que um classificador para uma categoria *c* é construído observando as características intrínsecas de um conjunto de documentos, previamente rotulados por um especialista no domínio para uma categoria *c* (DUMAIS *et al.*, 1998). Caracterizando-se como uma abordagem de aprendizado supervisionado, no qual um novo documento é classificado de acordo com as características assimiladas por um classificador confeccionado e treinado a partir de documentos previamente rotulados (MARTINS, 2003).

Para o problema de classificação de ideias apresentado neste trabalho foi selecionado o classificador *Naive Bayes*, que necessita de um conjunto de dados de treinamento para estimar a probabilidade de um documento pertencer a uma classe.

O teorema de *Bayes*, mostrado no Equação 1, é uma ferramenta para estimar estas probabilidades (DUMAIS *et al.*, 1998).

$$\Pr(c|d) = \frac{\Pr(c)\Pr(d|c)}{\Pr(d)} \qquad \text{Equação 1}$$

De modo que temos acima:

- Pr (c | d) é a probabilidade posterior da classe (c, alvo) dada preditor (d, atributos).
- Pr (c) é a probabilidade original da classe.
- Pr (d | c) é a probabilidade que representa a probabilidade de preditor dada a classe.
- Pr (d) é a probabilidade original do preditor.

Após o cálculo das probabilidades há diversas estratégias para a execução do treinamento realização de testes. O treinamento possui o foco de demonstrar ao classificador exemplos de modo a possibilitar aprender sobre os dados textuais. A aplicação de testes possibilita a avaliação da performance, descrita a seguir. Dentre tantas técnicas serão apresentadas duas das principais estratégias descritas na literatura:

Holdout: é o processo de segmentar do conjunto de treinamento uma determinada porcentagem deste para compor o conjunto de teste. Usualmente, o teste utiliza 1/3 do conjunto total, mantendo o restante para treinamento. E segundo Junior (2007) apesar de simples e rápida de se aplicar, é criticada por não utilizar o conjunto total de amostras, o autor ainda evidencia que o conjunto de teste pode acabar sendo favorecido, assim induzindo a uma conclusão falsa sob a assertividade real do treinamento.

Cross Validation K-Fold ou Validação Cruzada: Validação Cruzada conforme introduzido por Geisser (1975) é uma metodologia de treinamento e teste que usa o conceito de folds, deste modo, o conjunto inicial de treinamento é dividido em k conjuntos. Deste total de conjuntos, um é usado para a validação do modelo (conjunto de teste) e os k-1 conjuntos irão compor o conjunto de treinamento. Este processo é repetido k vezes, onde cada um dos conjuntos K sejam utilizados no mínimo uma vez como conjunto de testes. O resultado obtido ao fim é a média de desempenho do classificador durante as k iterações.

O foco desta metodologia é aumentar a confiabilidade da avaliação. Cabe-se destacar que há a possiblidade do uso destas duas metodologias combinadas usando a holdout como mais uma forma de

validação dos resultados da validação cruzada, entretanto é necessário um com conjunto de amostras maior e se dispende de mais tempo para processamento e execução dos ciclos (JUNIOR, 2007).

Após definir o método de validação e classificar é preciso avaliar a performance do classificador que é verificar o quão capaz este é na atividade de categorizar corretamente um novo exemplo assim que proporcionado. A avaliação deve ser realizada após o treinamento, utilizando o resultado da classificação do conjunto de teste. Existem diversas métricas que sustentam esta etapa, derivadas principalmente da área de Recuperação de Informação, conforme listadas a seguir no Quadro 2 (JUNIOR, 2007):

Quadro 2 - Métricas de avaliação da classificação

Métrica	**Fórmula**
Precisão: Mede a porção de exemplos de uma classe que foi corretamente classificada.	precisão(A) = total de exemplos corretamente classificados da classe A / total de exemplos corretamente classificados
Recall (Eficiência): Proporção de amostras classificadas como sendo de uma classe em relação ao total de amostras da classe.	recall(A) = total de exemplos corretamente classificados da classe A / total de exemplos da classe A
Acurácia: Denota a proporção total de classificações corretas.	acurácia = total de amostras classificadas corretamente independente da classe / total de exemplos do conjunto de teste
F-Measure: Média harmônica entre Precisão e Recall. Bastante utilizada quando as predições de um classificador estão desbalanceadas, ou seja, eficaz para uma determinada classe e não para a outra. F-Measure também é interessante por fornecer uma medida única de comparação.	$Fmeasure = \frac{2}{\left(\frac{1}{precisão} + \frac{1}{recall}\right)}$

Fonte: do autor, baseado em Junior, 2007.

2.4.3 Ontologias

O termo “Ontologia”, com origem na filosofia, diz respeito ao estudo da existência do ser, ou aos tipos de sua existência (GRUBER, 2009), e trata-se de uma parte da metafísica que estuda a estrutura de

sistemas e, atualmente, está associada à organização e classificação do conhecimento (MCCOMB, 2004).

Atualmente o termo ontologia pode ser visto conforme duas perspectivas, sendo a primeira relacionada à filosofia, como mencionado anteriormente, e a segunda relacionada à Ciência da Computação. Inicialmente utilizado pela Inteligência Artificial, e atualmente, utilizado também pela área de EG (POLI; OBRST, 2010).

As ontologias são métodos de organização e representação do conhecimento sendo um campo de estudo da Engenharia do Conhecimento (EG) cada vez mais valorizado por possibilitar o compartilhamento e reutilização de informações (GUARINO, 1995). "As estruturas das ontologias são baseadas na descrição de conceitos e dos relacionamentos semânticos entre eles, gerando uma especificação formal e explícita de uma conceitualização compartilhada" (STUDER; BENJAMINS; FENSEL, 1998).

Para Campos (2004, p. 24), os modelos de representação do conhecimento proporcionam dentro do domínio da Ciência da Informação, "a elaboração de linguagens documentárias verbais e notacionais, visando à recuperação de informações e a organização dos conteúdos informacionais de documentos".

Segundo Café e Brascher (2008, p. 6) a organização do conhecimento, se aplica as características do pensamento e "visa à construção de modelos de mundo que se constituem em abstrações da realidade" e se apresenta como resultado desse processo, e a representação do conhecimento "é fruto de um processo de análise de domínio e procura refletir uma visão consensual sobre a realidade que se pretende representar".

A utilização de ontologias possibilita a definição de um domínio limitando uma área específica para trabalho e desta maneira melhorando o processo de extração de informação e a compartilhamento do conhecimento (GÓMEZ-PÉREZ, 1999). As ontologias, inerentes aos estudos da web semântica, objetivam o processamento automatizado da informação. Gruber (1995) cita como elementos básicos de uma ontologia: classes (organizadas em uma taxonomia); relações (representam a influência mútua entre os conceitos de um domínio); axiomas (utilizados para modelar sentenças verdadeiras); instâncias (utilizadas para representar elementos específicos, ou seja, os próprios dados). Assim ao desenvolver uma ontologia, a descrição das categorias e dos objetos e as relações entre os dados envolvidos no processo são

componentes devidamente explicitados (LULA; PALIWODA-PĘKOSZ, 2008).

2.4.3.1 Tipos de Ontologia

As ontologias podem ser classificadas, segundo diversos autores, por tipos distintos, visto que a sua utilização atualmente tornou-se frequente para possibilitar a representação do conhecimento de diversas áreas e apoiar tecnologias voltadas à gestão do conhecimento (FERNANDES *et al.*, 2011).

Para Guarino (1998) os tipos de ontologias são divididas em:

- Ontologias gerais (*top-level ontology*): possuem significados abstratos para a compreensão do domínio de conhecimento;
- Ontologias de domínio (*domain ontology):* abordam um domínio específico de uma área genérica;
- Ontologias de tarefa (*task ontology*): abordam tarefas ou atividades genéricas;
- Ontologias de aplicação (*application ontology*): objetivam solucionar um problema específico de um determinado domínio, normalmente referenciando termos relacionados a uma ontologia de domínio.

Studer, Benjamins e Fensel (1998) considera também outros dois tipos de ontologias:

- Ontologias de representação: definem conceitos que especificam genericamente a representação do conhecimento, não se detendo a um domínio específico;
- Ontologias de método: especificam o vocabulário relativo a um método presente em um Método de Resolução de Problema (PSM).

Uma classificação ainda mais complexa é apresentada por Almeida e Bax (2003) em seus estudos. Estes autores propuseram uma especificação ainda mais completa, que leva em consideração o tipo de abordagem, sendo elas: quanto a função, quanto ao grau de formalismo; quanto à aplicabilidade, quanto ao nível, e quanto ao conteúdo. O Quadro 3 demonstra esta classificação.

Quadro 3 - Classificação das Ontologias

Abordagem	**Classificação**	**Definição**
Quanto à função Mizoguchi, Vanwelkenhuysen e Ikeda (1995)	Ontologias de domínio	Reutilizáveis no domínio fornecem vocabulário sobre conceitos, seus relacionamentos, sobre atividades e regras que os governam.
	Ontologias de tarefa	Fornecem um vocabulário sistematizado de termos, especificando tarefas que podem ou não estar no mesmo domínio.
	Ontologias gerais	Incluem um vocabulário relacionado a coisas, eventos, tempo, espaço, casualidade, comportamento, funções, etc.
Quanto ao grau de formalismo Uschold e Gruninger (1996)	Ontologias altamente informais	Expressa livremente em linguagem natural.
	Ontologias semi-informais	Expressa em linguagem natural de forma restrita e estruturada.
	Ontologias semiformais	Expressa em linguagem artificial definida formalmente. Ontologia rigorosamente formal.
	Ontologia rigorosamente formal	Os termos são definidos com semântica formal, teoremas e provas.
	Ontologias de autoria neutra	Um aplicativo é escrito em uma única língua e depois convertido para uso em diversos sistemas, reutilizando-se as informações.
Quanto à aplicação Jasper e Uschold (1999)	Ontologias como especificação	Cria-se uma ontologia para um domínio, a qual é usada para documentação e manutenção no desenvolvimento de softwares.

	Ontologia de acesso comum à informação	Quanto ao vocabulário é inacessível, a ontologia torna a informação inteligível, proporcionando vocabulário compartilhado dos termos.
Quanto à estrutura - Ontologia de alto nível Haav e Lubi (2001)	Ontologia de alto nível	Descrevem conceitos gerais relacionados a todos os elementos da ontologia (espaço, tempo, matéria, objeto, evento, ação, etc.) os quais são independentes do problema ou domínio.
	Ontologia de domínio	Descrevem o vocabulário relacionado ao domínio, como, por exemplo, medicina, ou automóveis.
	Ontologia de tarefa	Descrevem uma tarefa ou atividade, como, por exemplo, diagnósticos ou compras, mediante inserção de termos especializados na ontologia.
	Ontologias terminológicas	Especificam termos que serão usados para representar o conhecimento em um domínio (por exemplo, os léxicos).
	Ontologias de informação	Especificam a estrutura de registros de bancos de dados (por exemplo, os esquemas de bancos de dados).
	Ontologias de modelagem do conhecimento	Especificam conceituações do conhecimento, tem uma estrutura interna semanticamente rica e são refinadas para uso no domínio do conhecimento que descrevem.

Quanto ao conteúdo VanHeijist, Schreiber e Wielinga (2002)	Ontologias de aplicação	Contém as definições necessárias para modelar o conhecimento em uma aplicação
	Ontologias de domínio	Expressam conceituações que são específicas para um determinado domínio do conhecimento.
	Ontologias genéricas	Similares às ontologias de domínio, mas os conceitos que as definem são considerados genéricos e comuns a vários campos.
	Ontologias de representação	Explicam as conceituações que estão por trás dos formalismos de representação do conhecimento.

Fonte: Almeida e Bax (2003)

2.4.3.2 Aplicações Semânticas

Semântica é definida como os significados de termos e expressões. Berners-Lee *et al*. (2001) introduziram o conceito de web semântica como uma coleção de padrões e abordagens para trazer ordem e significado à informação na Internet. As tecnologias da web semântica permitem a representação explícita do conhecimento e seu processamento posterior para deduzir novos conhecimentos do conhecimento implicitamente oculto.

Além disso, o uso de técnicas semânticas na área de gestão de inovação traz a possibilidade de melhorar a eficiência do usuário final por meio de processamento automatizado e lidar com o processamento analítico avançado de metadados de inovação por meio do raciocínio. Assim, as organizações podem aumentar seus lucros com informações melhor estruturadas, integração e troca de dados entre ferramentas e plataformas, e raciocínio semântico adicional permite a estas organizações analisar ideias com base em conceitos relacionados (EL BASSITI E AJHOUN, 2014).

Outro exemplo é de Poveda, Westerski e Iglesias (2012) no qual apresentam um modelo baseado em busca semântica para sistemas de inovação aberta com foco em sistemas de Gestão de Ideias. Apresentam também uma metodologia para coleta, organização e busca de ideias, com

foco em melhorar a interação entre usuários e simplificar o processo de análise de ideias.

Haja visto, ontologias são um dos componentes fundamentais para as tecnologias semânticas, de modo que fornecem vocabulários sobre entidades dentro de um domínio e seus relacionamentos; fornecem vocabulários sobre as atividades que ocorrem no domínio; e fornecem vocabulários sobre teorias e princípios elementares que regem o domínio (GUARINO, 1998).

Além dos benefícios já explicitados a web semântica pode trazer benefícios específicos para a área de gestão da inovação, tais como mecanismos de pesquisa, filtragem de informações, anotação semântica, aprendizado contínuo e melhor tomada de decisões (EL BASSITI E AJHOUN, 2014).

2.5 TRABALHOS RELACIONADOS

A busca sistemática, detalhada no Apêndice A, permitiu verificar a existência de trabalhos correlatos que utilizam técnicas de Descoberta de Conhecimento em Texto aplicadas ao um conjunto de ideias. Nota-se que são artigos datados de 2006 até 2017, nos quais verifica-se que a técnica mais utilizada são as ontologias, como em Angeniol *et al.* (2006), Sint *et al.* (2010), Westerski e Iglesia. (2011), El Bassiti e Ajhoun (2014), Perez *et al.* (2015) e Sérgio; De Souza; Gonçalves (2017).

No Quadro 4 apresenta-se a síntese dos autores e a relação de quais ferramentas usam para análise de dados aplicados a gestão de ideias.

Quadro 4 - Passos para a construção da proposta.

Autor	**Objetivo**	**Técnica utilizadas**	**Ferramentas utilizadas**
Sérgio; De Souza; Gonçalves (2017).	Apresenta um modelo com base ontológica e a análise de cluster para apoiar a ideia de gestão, colaborando no processo de tomada de decisão.	Ontologias de Domínio e Clusterização por termos de maior peso.	Lucene e algoritmo Lingo do projeto Carrot². Não especifica qual ferramenta usou para confecção da ontologia.
Perez *et al.* (2015)	Estudo de caso onde apresenta a aplicação do	Ontologia, Wiki, Web Semântica	Neste estudo destaca apenas que foi

	sistema Gi2mo para suporte a gestão de ideias nas etapas de criação, coleta, enriquecimento, analise e suporte para seleção.	e Matrix de seleção.	desenvolvido sobre a base do Drupal para o front end do processo de inovação.
Löwer e Heller (2014)	Estudo de caso sobre um modelo proposto para a gestão ideias com enfoque na criação, coleta e armazenamento de ideias.	PDMS e PLM.	PTC Windchill 9.1 com os módulos de PartsLink e ProjectLink, banco de dados em Oracle 11g1.
El Bassiti e Ajhoun (2014)	Propõe um framework baseado em Web semântica para gestão de ideias com foco na geração, interligação, enriquecimento e validação baseado num.	Ontologias e Web Semântica.	Neste artigo apenas demonstra o framework e quais técnicas pretende usar porem não desenvolve a ferramenta.
Spencer (2012)	Propõe uma ferramenta para calcular a distância entre duas ideias, explorando a dimensionalidade.	Similaridade baseada na formula de Jaccard-Tanimoto.	Apresenta apenas os resultados e não detalha o desenvolvimento da ferramenta.
Westerski Iglesias (2011)	Estudo de caso de um experimento da construção de uma ontologia para mineração de opinião.	Ontologias e Web Semântica.	Desenvolvido sob a plataforma do Drupal e agregado ao sistemas Gi2mo onde usou as ferramentas pra construção desde modulo: Marl Ontology v0.1,

			SPARQL e OPAL.
Sint *et al.* (2010)	Propõe um framework baseado em Wiki Semântica (KiWi) para gestão da inovação.	Wiki Semântica.	Não descreve as ferramentas que usou para construção.
Paukkeri e Kotro (2009)	Desenvolve uma ferramenta para mineração textos, aplicada sobre ideias curtas.	Clusterização não supervisionada com *k-means* e métricas do cosseno.	Usa uma ferramenta para a gestão da inovação denominada NOTE, que é um bloco de notas eletrônico compartilhado, no qual os funcionários da organização podem anotar suas ideias e perguntas.
Angéniol *et al.* (2006)	Propõe uma ferramenta para reutilização de ideias Osíris (*Optimiser for Saving Idea Reuse & Information Sharing*)	Ontologias e Web Semântica; Banco de dados relacionais e	Não demonstra no artigo.

Fonte: Autor.

Angeniol *et al.* (2006) propõem uma ferramenta para reutilização de ideias denominada Osíris (*Optimiser for Saving Idea Reuse & Information Sharing)*. Para isso utilizam, além de ontologias, a web semântica e banco de dados relacionais. No entanto não demonstram a implementação da ferramenta e resultados aplicados.

Sint *et al.* (2010) propõem um framework baseado na web semântica com uso de ontologias para apoiar o processo de gestão de ideias. Com base na semântica desenvolvida no framework KiWi os autores também desenvolvem uma aplicação Enterprise 2.0, denominada

Ideator, capaz de promover suporte para a geração de ideias colaborativamente. Em seu artigo os autores não detalham quais ferramentas foram utilizadas para construir a aplicação.

El Bassiti e Ajhoun (2014) também propõem um framework baseado em Web semântica para gestão de ideias com foco na geração, interligação, enriquecimento e validação baseada numa ontologia que fornece representação semântica da inovação e uma linguagem comum para promover a interoperabilidade, declaração e serviços inteligentes entre as ferramentas de modo a apoiar o ciclo de vida da inovação. No entanto, no artigo não é demonstrado o desenvolvimento da ferramenta, ficando em proposições iniciais sem aplicações ou verificação de viabilidade, sendo estas indicações de trabalhos futuros.

Westerski, Iglesias e Rico (2010) apresentam um estudo de caso da construção de uma ontologia para mineração de opinião. A ferramenta é desenvolvida sob a plataforma de Gerenciamento de Conteúdo Drupal® agregado ao sistema GI2MO, no qual para a construção do módulo apresentado utilizou-se de ferramentas como, Marl Ontology v0.1[4] SPARQL (Structured Query Language)[5] e OPAL[6].

GI2MO é um Sistema de Gestão de Ideias de código aberto que vem se destacando em publicações. Desse modo, destaca-se outros trabalhos relacionados a estes autores que tratam do GI2MO. Westerski, Iglesias e Rico (2010) introduziram a utilização de tecnologias da Web Semântica e ontologias em Sistemas de Gestão de Ideias com a proposição de um modelo de metadados para esta integração. Em continuidade Wertersiki e Iglesias (2011) tratam da proposição de um modelo de criação de dados abertos para a *World Wide Web* para Sistemas de Gestão de ideias.

Dando continuidade, em Poveda, Westerski e Iglesias (2012) apresentam um modelo, projeto e arquitetura baseado em busca semântica com foco em sistemas de Gestão de Ideias. Com a apresentação de uma metodologia para coleta, organização e busca de ideias, melhorando a interação entre usuários e simplificando o processo de análise de ideias. Westerski, Iglesias e Garcia (2012) propõem uma série de métodos para sumarização do conjunto de dados em Sistemas de Gestão de Ideias e com

[4] Marl é uma ontologia projetada por Adam Westerski em 2011 para anotar e descrever opiniões subjetivas.
[5] SPARQL é uma linguagem de consulta semântica para bancos de dados, capaz de recuperar e manipular dados armazenados no formato Resource Description Framework (RDF)
[6] OPAL: é um plugin do Drupal que analisa os comentários postados pelos usuários e detecta se eles são positivos, negativos ou neutros.

isso demonstraram que a sua utilização pode aumentar significativamente a quantidade de relações obtidas.

Perez *et al.* (2015) também realizam um estudo de caso, no qual demonstram a aplicação no sistema GI2MO para suporte a gestão de ideias nas etapas de criação, coleta, enriquecimento, análise e suporte para seleção. Utiliza-se, além de ontologia, uma Wiki e Web Semântica e Matrix de seleção. Por utilizar como aplicação o GI2MO desenvolvido sobre a base do Drupal.

Por conta deste, e de outros trabalhos, o projeto GI2MO possibilita organizar todas as fases do processo de gestão de ideias, além disso, possibilita a configuração de tecnologias de Web Semântica no ambiente de Sistemas de Gestão de Ideias. Desse modo, o sistema permite realizar busca automática, exploração do significado semântico para melhorar ideias e possibilita incorporar ideias por meio de *Linked Data*. O grande objetivo é a interoperabilidade com as soluções existentes. Oferece ainda formato semântico para as ideias de acordo com a ontologia GI2MO (GI2MO, 2018).

Por fim, Sérgio; De Souza; Gonçalves (2017) apresentam um modelo com base ontológica e a análise de cluster para apoiar a gestão de ideias, colaborando no processo de tomada de decisão. Para isso utilizam ontologia de domínio para realizar a clusterização por termos de maior peso. No que tange a ferramentas utilizaram a Apache Lucene[7] e algoritmo Lingo do projeto Carrot[8]. Foi realizada aplicação em cenários envolvendo os bancos de dados de ideias das empresas do Dell® e Starbucks®. As aplicações permitiram verificar a formação e apresentação de grupos de ideias, o que possibilita que especialistas tenham um ferramental a fim de reduzir o tempo na análise de tendências e demandas apontadas por clientes e colaboradores.

Sérgio; De Souza; Gonçalves (2017) antes de propor sua ontologia de domínio investigaram a possível utilização da ontologia disponível no G2MO, no entanto observaram a existência de classes e subclasses descontinuadas, bem como, propriedades de dados e objetos. De modo que não promovia suporte à formação de agrupamentos.

É notório que a utilização de ontologias está sendo bem explorada neste contexto, no entanto outras formas foram verificadas, uma vez que o processamento com a utilização de ontologias pode tornar-se mais dispendioso se tiver um grande conjunto de regras para inferência devido

[7]Apache Lucene: é uma biblioteca para recuperação de informação com diversos recursos escrita em Java.
[8]Carrot: ferramenta *opensource* para Clusterização e visualização de textos.

a problemas complexos ou muito gerais, de modo que necessite do especialista de domínio a criação de um grande número de regras para expressar o conhecimento e heurística envolvidos no processo de classificação. Essas outras formas podem ser verificadas nos trabalhos de Paukkeri e Kotro (2009), Spencer (2012) e Löwer e Heller (2014).

Paukkeri e Kotro (2009) desenvolvem uma ferramenta para mineração textos, intitulada NOTE, aplicada sobre ideias curtas, ou seja, textos curtos em um banco de ideias. NOTE é um *Noteboard* eletrônico compartilhado, onde os funcionários de uma empresa podem escrever as suas observações, ideias e perguntas.

A ferramenta realiza uma clusterização não supervisionada por meio do algoritmo *k-means*. A ferramenta combina métodos estatísticos como as métricas do cosseno e mineração de texto, com o objetivo de criar e atualizar a memória coletiva de uma organização. No entanto, o artigo não descreve maiores detalhes sobre ferramenta e o *framework*, bem como os resultados não foram evidenciados no artigo.

Spencer (2012) propõe uma ferramenta para calcular a distância entre duas ideias, explorando a dimensionalidade e o tamanho de um espaço. O cálculo de similaridade é baseado na fórmula de Jaccard-Tanimoto. No artigo são apresentados os resultados, mas não detalha o desenvolvimento da ferramenta.

Löwer e Heller (2014) realizaram um estudo de caso sobre um modelo proposto para a gestão ideias com enfoque na criação, coleta e armazenamento de ideias. Abordam dos conceitos de PDMS (*Product Data Management Systems*) e PLM (*Product Lifecycle Management*). Para isto utilizaram ferramentas como PTC Windchill 9.1[9] com os módulos de PartsLink e ProjectLink[10], bem como banco de dados relacionais em Oracle 11g1[11]. Segundo os autores, inovação e gestão ideia ainda não são suficientemente compreendidos em um sistema integrado baseado em PLM para entender a lacuna entre considerações estratégicas antes dos estágios iniciais de planejamento do produto e as fases de desenvolvimento.

Sobre os artigos apresentados os que mais se assemelham aos objetivos desta dissertação são o de Paukkeri e Kotro (2009), Spancer (2012) e Sérgio; De Souza; Gonçalves (2017). Porém há diferenças

[9] PTC Windchill 9.1: é um sistema de Gerenciamento de Dados de Produto (PDM), baseado na Web.

[10] PartsLink e ProjectLink: módulos que permitem permite organizar bibliotecas do projeto interno por intermédio de mecanismos de busca de bibliotecas flexíveis e criar um espaço de trabalho virtual respectivamente.

[11] Oracle 11g1: banco de dados relacional.

notórias, quando se trata de Paukkeri e Kotro, para classificar ideias utilizam-se técnica não supervisionada e a métrica do cosseno para gerar as classes, enquanto o modelo proposto está alinhado além dos agrupamentos pela métrica do cosseno e classificação supervisionada por meio de técnicas probabilísticas. Já de Spencer (2012) diferencia-se a técnica utilizada, bem como o contexto aplicado e a finalidade. Quanto ao trabalho de Sérgio; De Souza; Gonçalves (2017). os autores adotam classificação supervisionada porem cria os rótulos por meio do algoritmo lingo identificando termos com maior peso e na presente pesquisa sugere-se que sejam utilizadas temáticas pré-estabelecidas pelas organizações e a reutilização do conhecimento gerado para criação de grupos de treinamento para o classificador.

De modo geral, esta dissertação diferencia-se dos demais trabalhos, uma vez que avança com a aplicação de outras técnicas ainda não utilizadas neste contexto para classificação das ideias guiadas por temáticas específicas das organizações orientado pelo conhecimento gerado dentro da própria organização para treinar os modelos de classificação, utilizando técnicas com bons resultados em outros contextos.

Considerando esta análise de literatura, pode-se concluir que a aplicação de técnicas de Descoberta de Conhecimento em Texto em bases de ideias ainda é um tema emergente e desafiante, que ainda necessita de pesquisas mais robustas e a realização de desenvolvimento/aplicação de ferramentas mais eficientes. Ainda, Sérgio (2016) sugere em sua dissertação, entre outros pontos, a necessidade de testar a utilização de outros algoritmos de agrupamento com o objetivo de analisar qual a melhor abordagem para lidar com informações textuais, fortalecendo a existência da necessidade de mais estudos acerca deste problema.

3 PROCEDIMENTOS METODOLÓGICOS

Este capítulo apresenta a metodologia que norteou o desenvolvimento desta dissertação. Gil (2007, p. 17) define pesquisa como:

> O procedimento racional e sistemático que tem como objetivo proporcionar respostas aos problemas que são propostos. A pesquisa desenvolve-se por um processo constituído de várias fases, desde a formulação do problema até a apresentação e discussão dos resultados.

Torna-se relevante destacar também a definição de metodologia proposta pela autora Minayo (2007, p. 44):

> a) A apresentação adequada e justificada dos métodos, técnicas e dos instrumentos operativos que devem ser utilizados para as buscas relativas às indagações da investigação; b) a "criatividade do pesquisador", ou seja, a sua marca pessoal e específica na forma de articular teoria, métodos, achados experimentais, observacionais ou de qualquer outro tipo específico de resposta às indagações específicas.

Lacerda *et al.* (2013) complementa a definição metodologia ao articular esta deve zelar pela validade do caminho escolhido para se chegar ao fim proposto pela pesquisa sendo utilizada para responder uma questão de pesquisa e que permita a avaliação da comunidade cientifica.

Segundo Rochadel (2016) a busca por uma metodologia adequada que forneça suporte ao propósito da linha tecnológica da pesquisa de engenharia do conhecimento aplicada às organizações, é de fundamental importância. A exemplo de Braga (2012) que identifica 5 propostas de metodologia para a área de pesquisa aplicada: March e Smith (1995); CommonKADS por Schreiber (2000); Metodologia CESM (acrônimo de "Composition, Environment, Structure, Mechanism") por Bunge (2003); Von Alan *et al.* (2004); e a Design Science Research Methodology (DSRM) por Peffers *et al.* (2007).

No entanto antes de realizar o enquadramento desta pesquisa buscou-se na literatura um embasamento teórico que explicite o tecnológico e demonstre aspectos importantes que tangem a metodologia *Science Research Methodology* (DSRM) ou *Design Science Research* (DSR), proposta por Peffers *et al.* (2007).

3.1 METODOLOGIA DE PESQUISA

Antes de proceder a respeito da metodologia de pesquisa *Design Science*, julga-se necessário apresentar uma fundamentação que explicite o conhecimento tecnológico para então, apresentar a abordagem de pesquisa *Design Science*, utilizada na condução dessa pesquisa.

Cupani (2006) apresenta características do conhecimento tecnológico, e argumenta, com base no livro de Carl Mitcham, que a tecnologia pode ser abordada a partir de quatro viés diferentes: 1) como artefatos; 2) saber tecnológico; 3) atividades para produção e utilização dos artefatos e 4) manifestação da vontade do ser humano em relação ao mundo.

Quanto à definição de tecnologia Cupani (2011) define que é o campo do conhecimento que se ocupa de projetar artefatos, planejar sua construção, operação, configuração, manutenção e acompanhamento com base no conhecimento científico.

Simon (1996) acredita na necessidade de criar uma ciência que proponha a concepção e construção de artefatos, que realizem objetivos, com propriedades específicas, conhecida como *Design Science*.

A *Design Science*, em português ciência do artificial ou ciência do projeto, busca projetar e produzir sistemas inexistentes e modificar situações existentes (DRESCH; LACERDA; JÚNIOR, 2015). Para Peffers *et al.* (2007) os objetivos da *Design Science* são divididos em projetar, criar e avaliar os artefatos de tecnologia da informação que são destinados a resolver problemas organizacionais, identificados em um processo rigoroso, a fim de resolver os problemas observados e então comunicar os resultados ao público interessado. Os resultados podem ser considerados em inovações sociais, novas propriedades técnicas, sociais e/ou recursos de informação (PEFFERS *et al.*, 2007).

Van Aken (2004) corrobora com essa abordagem. Para ele o principal objetivo da *Design Science* é desenvolver conhecimento para a geração e desenvolvimento de artefatos. March e Smith (1995, p. 253) ainda complementam: "*Design Research* tenta criar coisas que servem a propósitos humanos e é orientado para a tecnologia e seus produtos são avaliados de acordo com critérios de valor ou de utilidade".

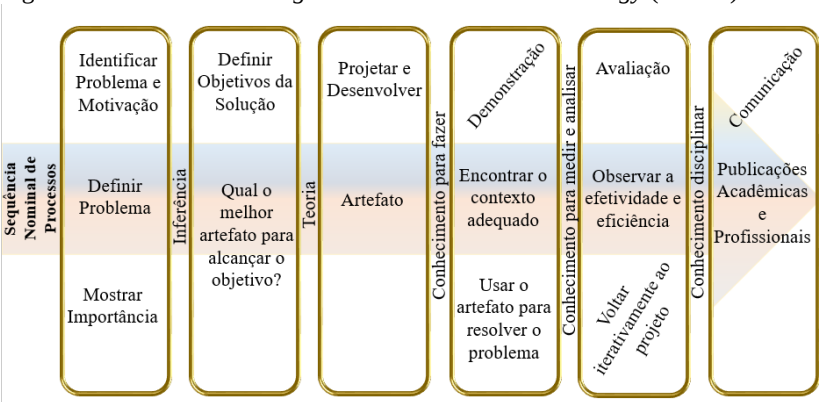
No trabalho de Simon (1996) é destacado que diferente das ciências naturais, os artefatos podem ser discutidos de maneira descritiva, no que diz respeito à comunicação e detalhamento dos componentes e

informações sobre o artefato, e em ermos imperativos, no que diz respeito a determinação das questões normativas responsáveis por envolver a construção e aplicação desse artefato.

Para Dresh *et al.* (2010) *Design Science* é a base epistemológica, já *Design Science Research* é o método que possibilita a construção do conhecimento nesse contexto. A *Design Science Research*, para Çağdaş e Stubkjær (2011), constitui-se de um rigoroso processo de projetar artefatos com objetivo de resolver problemas, avaliar o que foi projetado e comunicar os resultados que foram obtidos.

Lacerda *et al.* (2013) faz uma síntese gráfica das várias proposições na literatura a respeito da condução da *Design Science Research*. Para este trabalho, apresenta-se a proposição de Peffers *et al.* (2007), utilizado como suporte metodológico, composta de seis etapas, conforme Figura 11 abaixo.

Figura 11 - Processo de *Design Science Research Methodology* (DSRM)



Fonte: do autor, adaptado de Peffers *et al.* (2007).

1) **Identificar o problema e sua motivação:** definição do problema de pesquisa específico e justifica-se sua solução;
2) **Definir os objetivos para uma solução:** definem-se objetivos da solução que foi proposta;
3) **Projetar e Desenvolver:** aqui cria-se o artefato, que segundo March e Smith (1995) pode ser um: **constructo** ou conceitos que compõem o vocabulário de um domínio, e constituem uma conceituação utilizada para descrever os problemas dentro do domínio e para especificar as respectivas soluções; **modelo** que é um conjunto de proposições ou declarações que demonstram o

relacionamento entre os constructos, podem ser visualizados como uma representação de como as coisas são, e nas atividades de design representam situações como problema e solução; **método** que é um conjunto de passos, sendo um algoritmo ou orientação usado para executar uma tarefa e por fim **instanciação** que é concretização de um artefato em seu ambiente de modo que operacionalizam constructos, modelos e métodos. Nesta etapa de projetar e desenvolver artefatos, abordagens como algoritmos computacionais, representações gráficas, protótipos, maquetes em escala, entre outros, podem ser utilizadas;

4) **Demonstrar:** etapa responsável pela demonstração de uso do artefato, resolvendo um ou mais aspectos do problema por meio de um experimento, simulação, estudo de caso, prova formal entre outras atividades apropriadas;
5) **Avaliar:** nesta etapa, é feita observação e mensuração de como o artefato atende ã solução do problema, fazendo comparações a partir de métricas e técnicas de análises, dos objetivos propostos com os resultados observados na utilização do artefato. Ainda nesta etapa, segundo Lacerda *et al.* (2013) deve ser definido um processo de verificação do comportamento do artefato, sendo necessário: a) explicitar o ambiente interno, externo e os objetivos; b) explicitar como o artefato pode ser testado; e c) descrever os procedimentos que medem os resultados;
6) **Comunicar:** quando apropriado, faz-se a divulgação, para outros pesquisadores e outras audiências, do problema e sua relevância, do artefato que foi concebido, da sua utilidade e seu ineditismo, da efetividade e rigor do projeto.

Sendo assim, fica apresentado o embasamento teórico da metodologia que dá suporte ao desenvolvimento dessa dissertação.

3.2 DEFINIÇÃO DA PESQUISA

A luz dos conceitos de pesquisa, metodologia e conhecimento tecnológico apresentados é possível classificar esta pesquisa como tecnológica em relação a seus objetivos basilares em avançar a tecnologia na solução de um problema de um domínio de conhecimento.

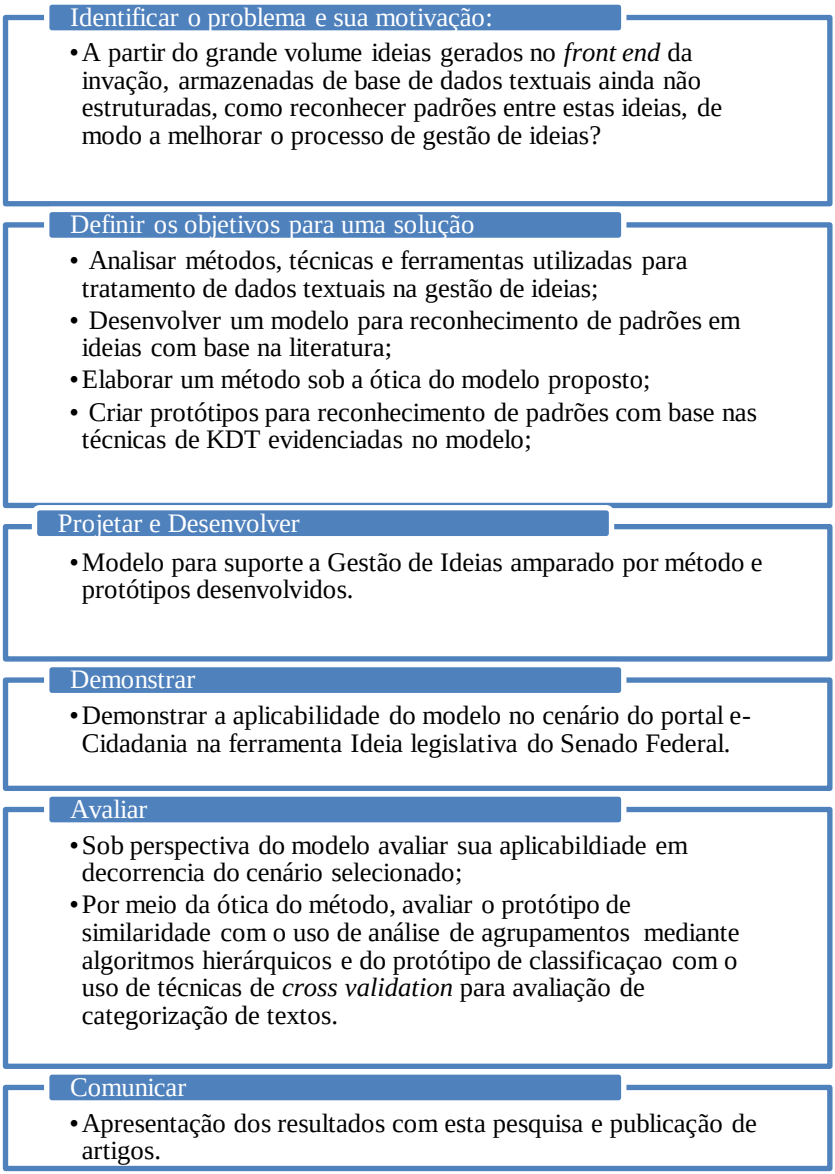
Neste sentido, a visão de mundo que se enquadra esta pesquisa é o paradigma funcionalista de Burrell e Morgan (1979) devido a direção na qual o estudo é desenvolvido, no sentido de previamente perscrutar

conhecimentos gerais e a posteriori os põe em prática, buscando-se preponderantemente uma solução funcional para um problema embasada em teorias existentes.

Esta classificação está sustentada no principal foco desta pesquisa que é disponibilizar um conjunto de artefatos que dão base a construção de protótipos combinados em um modelo para suporte a Gestão de Ideias.

Em função do objetivo de construir protótipos, esta pesquisa teve sua condução amparada na abordagem metodológica da *Design Science Research*, mais especificamente na proposição de condução de Peffers *et al.* (2007), apresentada na seção anterior, na qual segue evidenciado na Figura 12. No qual cada fase desta proposição esta correlacionada com os passos que a pesquisa talhou em busca de resolver o problema.

Figura 12 - Passos para a construção da proposta



Fonte: do autor.

Para identificação do problema e propor uma solução é necessário primeiramente uma abordagem metodológica qualitativa por meio de busca sistemática, e a procura de uma lacuna de conhecimento que correlacione a gestão de ideias e os métodos, técnicas e ferramentas que podem contribuir ou melhorar a gestão desta. Estes elementos são primordiais para construção desta pesquisa conforme apresentado no capitulo dois.

A partir do problema foram traçados os objetivos desta pesquisa, e assim o modelo que serve como base para planejamento para construção dos protótipos.

Quanto as técnicas escolhidas para o desenvolvimento dos artefatos foram adotadas a métrica do cosseno para o cálculo de similaridade entre ideias gerando assim agrupamentos para análise, e para a classificação de ideias se definiu o uso da técnica de naive bayes pela grande utilização em outras áreas de conhecimento para classificação e também por não ser encontrado na busca sistemática seu uso para o contexto de classificação de ideias. Evidenciou-se que modelos baseados em ontologias têm sido amplamente utilizados, outro fator primordial para a escolha do método de naive bayes é pelo aproveitamento do conhecimento já gerado para se alimentar as bases de treinamento.

Para a demonstração dos resultados foi escolhida como cenário a ferramenta Ideia legislativa do portal e-Cidadania uma iniciativa do Senado Federal Brasileiro. A escolha se deu pela forma a qual as ideias são estruturadas e não ser identificado seu uso na literatura para estudos na área de gestão de ideias. Cabe destacar que a popularidade e o crescimento no uso desta ferramenta demonstram sua notoriedade perante o cenário nacional.

Consecutivamente para avaliação destes protótipos foram adotados os critérios de análise de agrupamentos por meio dos algoritmos hierárquicos para a técnica de similaridade e para naive bayes adotou-se a estratégia de *cross validation* por permitir que todo o conjunto possa fazer parte do conjunto de treinamento e conjunto de teste a cada iteração de classificação.

A comunicação dos resultados é realizada por meio desta pesquisa. Cabe-se ressaltar que o detalhamento maior das etapas de Projetar e Desenvolver, Demonstrar e Avaliar serão realizada no tópico quatro deste documento que trata da apresentação e análise do modelo.

3.3 MATERIAS E MÉTODOS

Para a construção dos protótipos desta dissertação foi amparada pelos passos ilustrados na Figura 13, sendo este um modelo bem disseminado para descoberta de conhecimento em bases de texto.

Figura 13 - Passos para a construção de protótipos de KDT

Fonte: O autor, adaptado de Schwerz e Roberto (2012).

Conforme a definição da pesquisa foi determinada o cenário para aplicação desta pesquisa, sendo escolhido a ferramenta Ideia Legislativa do portal e-Cidadania do Senado Federal Brasileiro. Após determinação do cenário de estudo inicia-se o processo de coleta de dados, pré-processamento dos documentos textuais, indexação, mineração em bases textuais e por fim análise dos resultados encontrados com os protótipos. Para a realização desta pesquisa foi coletada a base de ideias do cenário escolhido conforme descrito no próximo tópico como se deu este processo.

3.4 COLETA DOS DADOS

Para esta pesquisa os dados foram coletados automaticamente com o uso da técnica de *web scraping*, sendo utilizada a ferramenta Octoparse[12] para a captura dos dados.

O *web scraping* é o processo de solicitar automaticamente um documento da web e coletar informações deste. De modo geral *web scraping* é o processo de se movimentar nos sites em busca de dados pré-determinados. Por fim, o *web scraper* realiza as seguintes atividades a partir de site de origem definido pelo usuário, carrega as informações solicitadas e permite que estas sejam exportadas em diversos formatos para análise, e assim a coleta realizada para esta pesquisa acabou se enquadrando nesta técnica. A Figura 14 ilustra essa etapa.

[12] Disponível em: < https:// www.octoparse.com > Acesso em set. 2017.

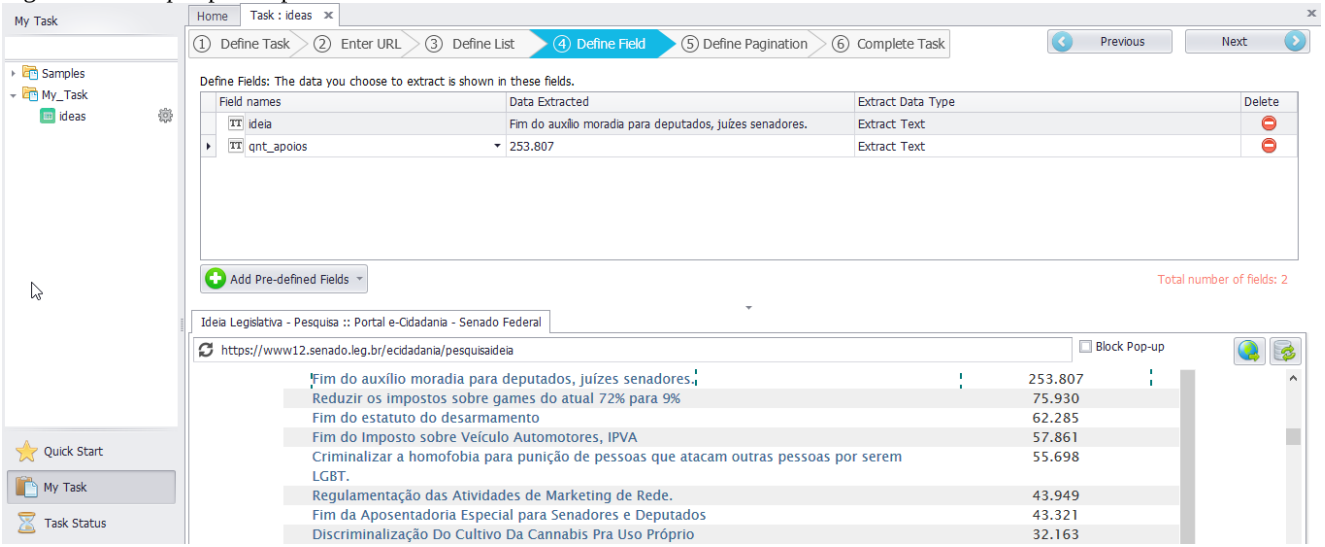
Figura 14 - Etapas do *web scraping*

Fonte: Portal ProWebScraping[13].

A primeira coleta ocorreu em setembro de 2017 e como pode ser observado na Figura 15 foram capturados os campos de Título da Ideia e Quantidade de apoios captados pelas campanhas até a data da coleta, totalizando um acervo com 25.501 ideias, que contabilizam 288.413 palavras em 23.386 tipos diferentes. Após a captura foi gerado um arquivo em formato CSV com os dados coletados.

[13] Disponível em: < http://prowebscraping.com/web-scraping-vs-web-crawling/> Acesso em abr. 2018.

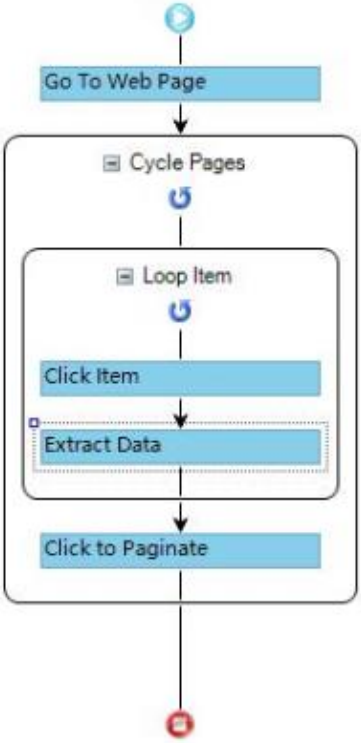
Figura 15 - *Scraper* para captura de dados



Fonte: do autor, capturado a partir do software Octoparse.

Em abril de 2018 houve uma segunda coleta de ideias para extração de mais detalhes sobre as ideias, tendo em vista que a primeira foi para testes iniciais, houve uma coleta integral de todas as ideias, onde foram capturadas 38.117 ideias, que contabilizam 415.800 palavras em 28.031 tipos diferentes, com uma média de captura que variou entre 28 a 32 registros por minuto levando cerca de 21 horas para captura de todas as ideias. A coleta ainda ocorreu utilizando o software *Octoparse* e foi necessário criar um método de looping para passar por todas as páginas e uma lista para visitar individualmente cada uma das páginas das ideias e assim extrair os dados pré-determinados (Figura 16).

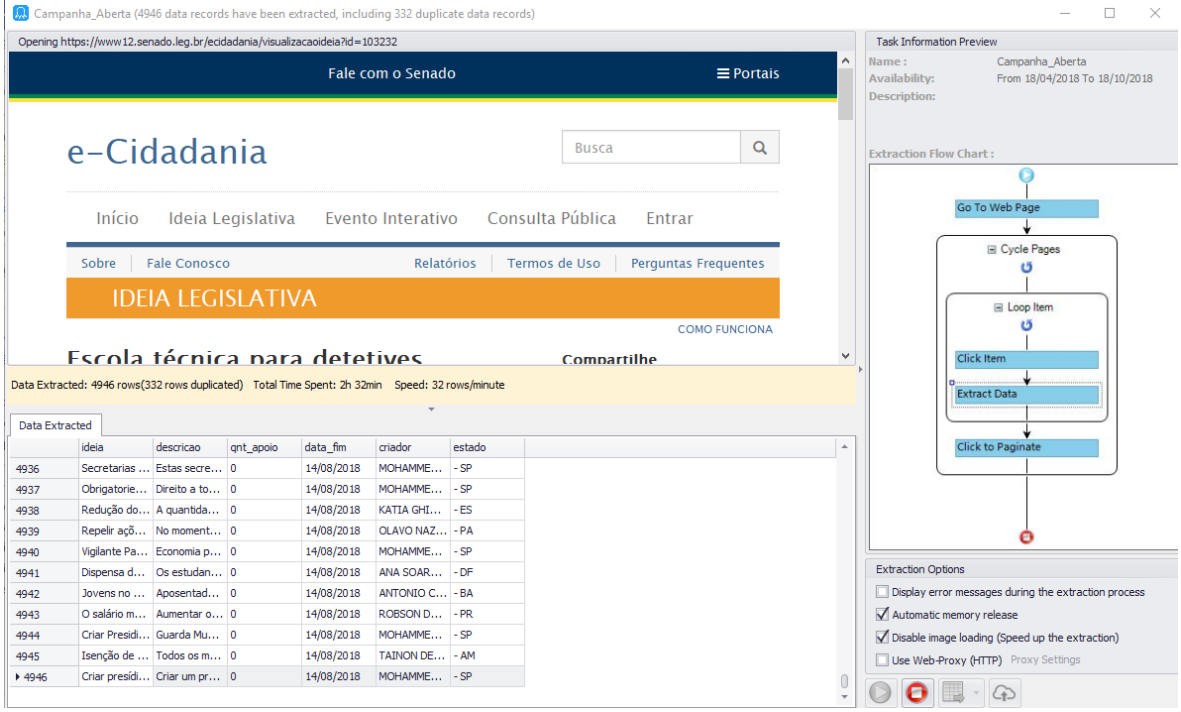
Figura 16 - Método para captura de dados



Fonte: O autor, capturado a partir do *software Octoparse.*

Nesta coleta foram extraídos mais dados sobre as ideias, tais como a descrição da ideia, autor e estado e a data final da campanha. (Figura 17).

Figura 17 - Scraper para captura de dados



Fonte: do autor, capturado a partir do *software Octoparse.*

Esta coleta foi subdividida por status da ideia capturando todas que estão em cada uma das etapas da ferramenta Ideia Legislativa como um campo adicional que é uma pré-categorização destas totalizando a recolha e 38.117 ideias.

4 APRESENTAÇÃO E ANÁLISE DO MODELO

Neste capítulo será apresentado o modelo proposto. A apresentação refere-se ao modelo lógico, sendo que o mesmo detalhará a interação decorrente entre técnicas adotadas para esta proposição. Após a apresentação detalha-se melhor o cenário adotado e aplica-se o modelo sob os dados coletados deste cenário. Por fim é realizada uma análise dos protótipos e demonstrado como o mesmo pode ser aplicado no cenário escolhido modificando o ciclo de vida destas ideias.

4.1 APRESENTAÇÃO DO MODELO PROPOSTO

A competitividade presente no cenário atual vem promovendo a busca constante por ideias, que vem sendo geradas em grandes escalas propiciadas pelas tecnologias atuais que permitem o de compartilhamento de conhecimento. Diversas organizações abrem espaço virtual para que colaboradores, clientes e demais interessados compartilhem suas ideias, com o objetivo de impulsionar o processo de inovação. Porém, isto tem gerado um grande volume de ideias submetidas, de modo que estas ideias podem ser exatamente iguais ou diferentes, mas com contextos semelhantes, ou ainda triviais para os objetivos da organização, de forma que podem representar um desafio a Gestão de Ideias. Diante desta situação exige-se mais tempo dos especialistas de domínio para analisar e tomar decisões perante estas ideias conforme evidenciado por Spencer (2012).

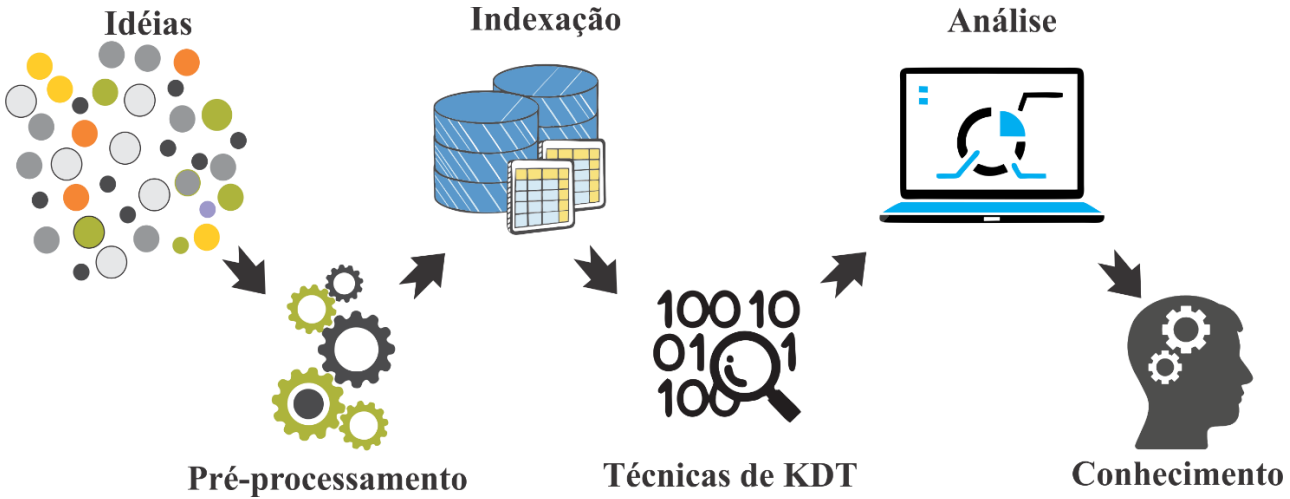
As ideias surgem nas organizações em formatos textuais, o que pode gerar um conjunto de documentos/ideias (o que pode ser organizado em um banco de ideias) com informações que podem ser essenciais para o processo de inovação ou não. Assim, é relevante filtrar este banco de ideias e também organizar de forma a facilitar o processo de gestão de ideias.

Neste contexto, o modelo proposto permite que as ideias sejam clusterizadas/classificadas conforme padrões entre elas. Ou seja, há situações em que a organização quer identificar no banco de ideias, aquelas que estão alinhadas a determinado tema de interesse configurando o aprendizado supervisionado ou ainda apenas identificar padrões comparando ideias por meio de técnicas de aprendizado não supervisionado.

O modelo proposto está divido em seis etapas fundamentais que visam fornecer suporte ao processo de gerir ideias, conforme figura 18 e previamente descritos abaixo:

- 1ª etapa: base de **ideias**, por meio de um conjunto de ideias oriundas de documentos textuais inicia-se o processo.
- 2ª etapa: no **pré-processamento** as ideias são submetidas a uma série de operações em busca de se obter uma forma de representá-las de modo estruturado;
- 3ª etapa: o processo de **indexação** é responsável pela criação de estruturas auxiliares para garantir uma maior agilidade e rapidez no processo de recuperação das ideias e seus termos;
- 4ª etapa: nesta etapa são aplicadas técnicas de **descoberta de conhecimento em textos** sobre as estruturas tradadas das ideias;
- 5ª etapa: a **análise** e avaliação e interpretação dos resultados obtidos pelo processo.
- 6ª etapa: explicitação do **conhecimento**, onde os resultados relevantes contidos nestas bases de dados textuais não estruturadas são utilizados para tomada de decisões, de modo que viabilizam suporte a gestão de ideias.

Figura 18 - Modelo para suporte a gestão de ideias



Fonte: do autor.

Desta forma, é possível adotar neste modelo tanto técnicas de clusterização como de classificação de ideia para agilizar o processo de categorização, seleção e avaliação das ideias, possibilitando-se assim que a partir de um conjunto de ideias brutas, identificar aquelas com maiores potenciais de implantação. Cabe ainda ressaltar que o modelo é genérico e pode-se adotar diversas técnicas de KDT, tendo em vista que o contexto e regras internas das organizações influenciam nesta escolha.

A partir do modelo proposto e da revisão da literatura é possível criar um método, descrevendo minunciosamente as etapas do modelo e apontando técnicas que podem ser utilizadas, métodos este voltado ao cenário escolhido, de modo que este método se divide em cinco etapas:

- 1ª etapa: **base de ideias**, por meio de um conjunto de ideias oriundas de documentos textuais inicia-se o processo.
- 2ª etapa: no **pré-processamento** as ideias são submetidas a uma série de operações baseadas nos métodos de Processamento Natural de Linguagem para se obter uma forma de representá-las de modo estruturado;
- 3ª etapa: o processo de **indexação** é responsável pela criação de estruturas auxiliares para garantir uma maior agilidade e rapidez no processo de recuperação das ideias e seus termos;
- 4ª etapa A: aplicação do **cálculo de similaridade** com a métrica do cosseno por ser uma das métricas básicas para cálculo de distancias entre vetores, neste contexto o uso aplicado para calcular o nível de similaridade entre as ideias. O processo de avaliação se dá por análise de agrupamentos por meio dos algoritmos hierárquicos;
- 4ª etapa B: aplicação da técnica de **categorização de texto** usando a técnica de Naive Bayes, no qual calcula a probabilidade de uma ideia pertencer a determinada classe amparado por um conjunto de treinamento e o processo de avaliação se dá pela metodologia de *cross validation*. Esta técnica foi adotada pois foi evidenciado mediante a revisão da literatura que ainda não havia sido aplicada ao contexto de Gestão de Ideias, destaca-se ainda a vantagem sob as demais técnicas quando aplicada a grandes volumes de dados pois exige um menor empenho de recursos para processamento;

- 5ª etapa: a análise, avaliação e explicitação do **conhecimento**, onde são realizadas a avaliação e interpretação dos resultados obtidos pelo processo.

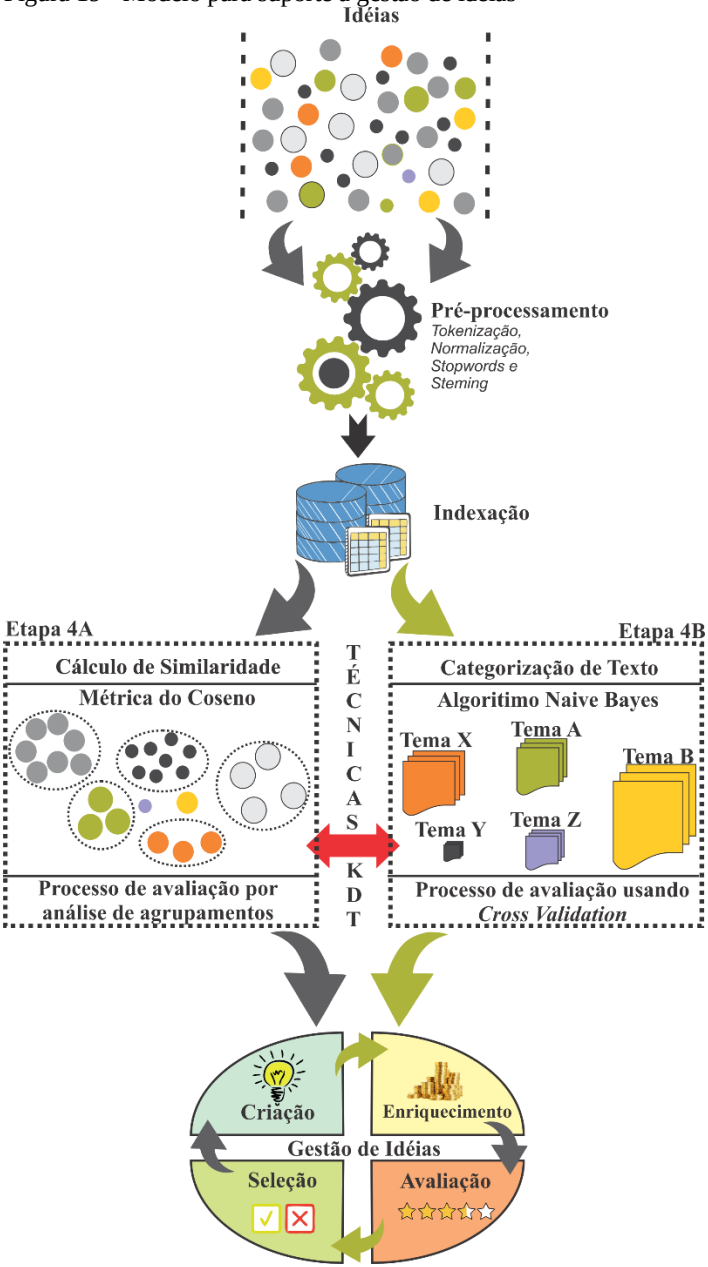
O objetivo principal deste trabalho é reconhecer padrões em ideias. Porém, ao se adotar as técnicas de descoberta de conhecimento em texto para classificação é relevante eliminar ideias iguais ou muito semelhantes, mas mantendo a informação sobre a frequência das ideias repetidas, pois é um indicador de relevância daquela ideia. Eliminar as repetidas é útil, principalmente para reduzir o tempo de processamento para a atividade de classificação.

Destacam-se nesse método dois itens na 4ª etapa, pois ambos dependem diretamente apenas da tabela de índices para serem aplicados, ofertando a possibilidade de quatro maneiras diferentes que se adaptam conforme a necessidade da organização na aplicação deste modelo, descritos abaixo:

- 1ª caminho - as etapas 4A e 4B podem ser executadas sequencialmente, de modo que primeiro realiza-se um filtro no banco de ideias eliminando ideias idênticas e similares, posteriormente pode ser efetuado a categorização destas. A vantagem desta abordagem é o menor número de ideias que chegam a categorização e, assim, exigindo um menor empenho de processamento.
- 2ª caminho - é percorrida apenas a etapa 4A para o cálculo de similaridade, em busca de se gerar clusters de ideias semelhantes.
- 3ª caminho - é percorrida apenas a etapa B a categorização das ideias, em busca de rotular as ideias da base.
- 4ª caminho - são executadas as etapas A e B separadamente, de modo que um não depende e nem influencia diretamente o outra, porém ambas são necessárias para melhorar o processo de Gestão de Ideias.

O método segue apresentado na Figura 19 e ilustra as fases supracitadas.

Figura 19 - Modelo para suporte a gestão de ideias



Fonte: do autor.

As próximas subseções apresentam primeiramente o cenário determinado para demonstração do modelo, seguido pelo detalhamento das etapas fundamentais do método, sendo apresentadas e analisadas as tecnologias utilizadas para a sua implementação.

4.2 CENÁRIO DE ESTUDO

O cenário construído para aplicação do modelo foi no portal e-Cidadania sob a ferramenta Ideia Legislativa, de modo que o objetivo é evidenciar relações e tendências abrangendo o cenário nacional e o senado.

A escolha da Base de Ideias se deu por motivos do formato estruturado da disposição das ideias, bem como a relevância da base e da ferramenta do portal e-Cidadania. Também se observou que esta base não foi utilizada em outras pesquisas envolvendo o domínio de Gestão de Ideias, como foi possível identificar na revisão sistemática e consulta na Base de Teses e Dissertações do EGC.

4.2.1 Portal e-Cidadania

A ferramenta Ideia Legislativa faz parte do portal e-Cidadania criado pelo Senado Federal em 2012 com o objetivo de possibilitar e estimular uma maior participação da sociedade nas atividades legislativas, orçamentárias, de fiscalização e de representação do Senado (BRASIL, 2018).

Segundo Brasil (2018) o portal E-Cidadania traz três ferramentas que propiciam a participação da sociedade, que são:

Evento Interativo: viabiliza a sociedade participar de audiências públicas, sabatinas e outros eventos abertos, de modo que são criadas páginas web específicas para cada evento em prol de promover a transmissão ao vivo, espaço colaborativo para publicação de comentários, apresentações, notícias e documentos atinentes ao evento;

Consulta Pública: permite ao cidadão deixar sua opinião sobre projetos de lei, propostas de emenda à Constituição, medidas provisórias e outras proposições que estão tramitando no Senado Federal até a deliberação final (sanção, promulgação, envio à Câmara dos Deputados ou arquivamento);

Ideia Legislativa: proporciona ao cidadão enviar e apoiar ideias, que podem ser sugestões de alteração na legislação em vigência ou de criação de novas leis. As ideias que recebem 20 mil apoios durante o

prazo de 4 meses quem que ficam abertas são encaminhadas para a Comissão de Direitos Humanos e Legislação Participativa (CDH) e debatidas pelos senadores, por fim podem receber parecer positivo e prosseguir para análise das comissões permanentes do senado ou se são encerradas nesta comissão.

Atualmente a ferramenta Ideia Legislativa possui 84 ideias com mais de 20.000 apoios, destas, 28 já possuem parecer da CDH e 6 ideias já foram transformadas em Projetos de leis ou Propostas de Emenda à Constituição conforme Figura 20 coletada em março de 2018.

Figura 20 - Tela inicial da ferramenta Ideia Legislativa

Fonte: Portal e-Cidadania[14].

Todo cidadão pode sugerir ideias, para cadastrar uma ideia é necessário se cadastrar no Portal, usando um e-mail ou ainda pode-se vincular seu cadastro às redes sociais do Facebook ou Google. Após o login para cadastrar a Ideia é necessário preencher alguns campos conforme a Figura 21, selecionando a área temática que podem ser: Administrativo, Econômico, Jurídico até Política Fundiária e Reforma Agrária, entre outras, além do título, o título da ideia, descrição e detalhes adicionais.

[14] Disponível em: < https://www12.senado.leg.br/ecidadania/principalideia> Acesso em mar. 2018.

Figura 21 - Tela para cadastro de Ideia da ferramenta Ideia Legislativa

Fonte: Portal e-Cidadania[15].

Após o cadastro da ideia não é possível mais editá-la e a mesma passa então por análise para verificar se está condizente com os termos de uso do portal, e é excluída caso apresente alguma destas características citadas abaixo tal como explicita Brasil (2018):

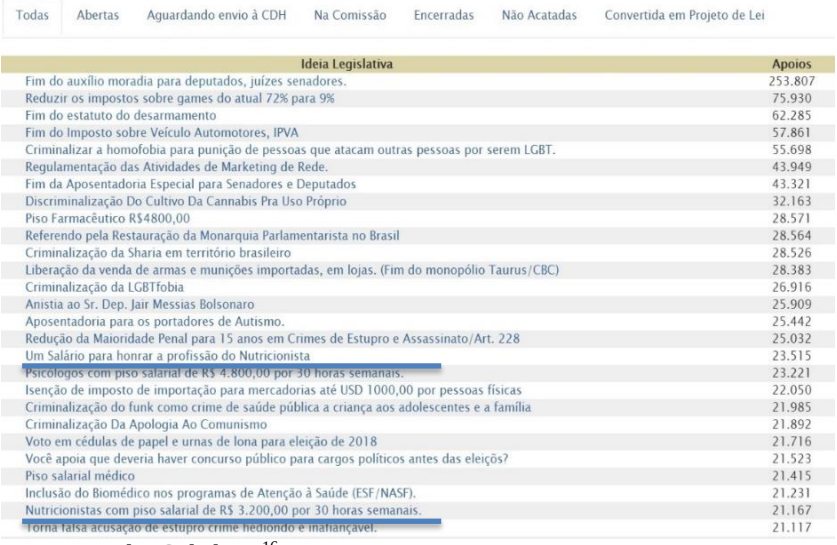
- Abordem assuntos adversos ao ambiente político, legislativo e de atuação do Senado Federal;
- Possuam qualquer tipo de declarações de cunho agressivo, pornográfico, pedófilo, racista, violento, ou ainda ofensivas à honra, à vida privada, à imagem, à intimidade pessoal e familiar, à ordem pública, à moral, aos bons costumes ou às cláusulas pétreas da Constituição;
- Sejam repetidas pelo mesmo usuário, incompreensíveis ou não estejam em português;
- Contenham dados pessoais ou referências a outras pessoas ou a páginas da internet em seu corpo.

Após esta análise, se aceitas as ideias sugeridas permanecem ativas por quatro meses em campanhas para arrecadarem apoios, sendo possível divulgá-las em mídias sociais para conseguir votos/apoios e seguir adiante na campanha.

[15] Disponível em: < https://www12.senado.leg.br/ecidadania/ideiaform> Acesso em mar. 2018.

Há a possibilidade de acompanhar o status das ideias sugeridas. A tela de acompanhamento permite visualizar todas as ideias sugeridas e a quantidade de apoios de cada uma conforme Figura 22. Além disso, é possível acompanhar as ideias abertas em campanhas, as que estão aguardando envio à CDH, as que estão nas comissões, as que foram encerradas sem apoio o suficiente, ou as que não foram acatadas e aquelas convertidas em projetos de lei. Nada garante que uma ideia, mesmo que passe por todas estas etapas, se torne efetivamente uma lei.

Figura 22 - Tela para pesquisa das Ideias


Todas | Abertas | Aguardando envio à CDH | Na Comissão | Encerradas | Não Acatadas | Convertida em Projeto de Lei

Ideia Legislativa	Apoios
Fim do auxílio moradia para deputados, juízes senadores.	253.807
Reduzir os impostos sobre games do atual 72% para 9%	75.930
Fim do estatuto do desarmamento	62.285
Fim do Imposto sobre Veículo Automotores, IPVA	57.861
Criminalizar a homofobia para punição de pessoas que atacam outras pessoas por serem LGBT.	55.698
Regulamentação das Atividades de Marketing de Rede.	43.949
Fim da Aposentadoria Especial para Senadores e Deputados	43.321
Discriminalização Do Cultivo Da Cannabis Pra Uso Próprio	32.163
Piso Farmacêutico R$4800,00	28.571
Referendo pela Restauração da Monarquia Parlamentarista no Brasil	28.564
Criminalização da Sharia em território brasileiro	28.526
Liberação da venda de armas e munições importadas, em lojas. (Fim do monopólio Taurus/CBC)	28.383
Criminalização da LGBTfobia	26.916
Anistia ao Sr. Dep. Jair Messias Bolsonaro	25.909
Aposentadoria para os portadores de Autismo.	25.442
Redução da Maioridade Penal para 15 anos em Crimes de Estupro e Assassinato/Art. 228	25.032
Um Salário para honrar a profissão do Nutricionista	23.515
Psicólogos com piso salarial de R$ 4.800,00 por 30 horas semanais.	23.221
Isenção de imposto de importação para mercadorias até USD 1000,00 por pessoas físicas	22.050
Criminalização do funk como crime de saúde pública a criança aos adolescentes e a família	21.985
Criminalização Da Apologia Ao Comunismo	21.892
Voto em cédulas de papel e urnas de lona para eleição de 2018	21.716
Você apoia que deveria haver concurso público para cargos políticos antes das eleições?	21.523
Piso salarial médico	21.415
Inclusão do Biomédico nos programas de Atenção à Saúde (ESF/NASF).	21.231
Nutricionistas com piso salarial de R$ 3.200,00 por 30 horas semanais.	21.167
Torna falsa acusação de estupro crime hediondo e inafiançável.	21.117

Fonte: Portal e-Cidadania[16].

É possível observar na Figura 22 que as ideias destacadas possuem contextos similares e estão em busca de apoios, onde ambas almejam regulamentações para a carreira de nutricionista.

Quando a ideia atinge 20 mil apoios dentro dos quatro meses que fica em campanha, é encaminhada para a Comissão de Direitos Humanos e Legislação Participativa. Então a ideia é analisada, classificada e encaminhada para outra comissão permanente, responsável por analisar, debater e aprovar os projetos apresentados pelos parlamentares do Executivo.

[16] Disponível em: < https://www12.senado.leg.br/ecidadania/pesquisaideia > Acesso em jan. 2018.

Conforme o Regimento Interno do Senado Federal compilado em dezembro de 2017 (BRASIL,2018) as comissões permanentes do Senado são as seguintes:

I - Comissão de Assuntos Econômicos (CAE), com 27 membros;

II - Comissão de Assuntos Sociais (CAS), com 21 membros;

III - Comissão de Constituição, Justiça e Cidadania (CCJ), com 27 membros;

IV - Comissão de Educação, Cultura e Esporte (CE), com 27 membros;

V - Comissão de Transparência, Governança, Fiscalização e Controle e Defesa do Consumidor (CTFC), com 17 membros;

VI - Comissão de Direitos Humanos e Legislação Participativa (CDH), com 19 membros;

VII - Comissão de Relações Exteriores e Defesa Nacional (CRE), com 19 membros;

VIII - Comissão de Serviços de Infraestrutura (CI), com 23 membros;

IX - Comissão de Desenvolvimento Regional e Turismo (CDR), com 27 membros;

X - Comissão de Agricultura e Reforma Agrária (CRA), com 17 membros;

XI - Comissão de Ciência, Tecnologia, Inovação, Comunicação e Informática (CCT), com 17 membros;

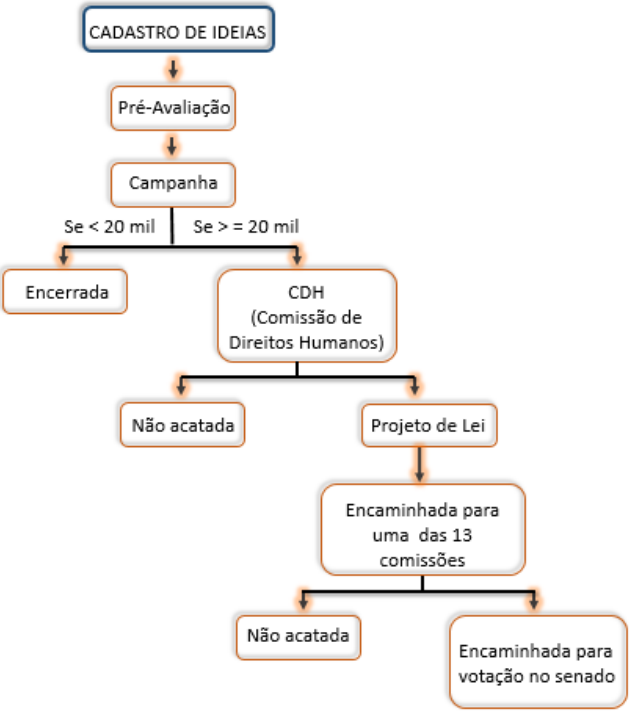
XII - Comissão Senado do Futuro (CSF), com 11 membros;

XIII - Comissão de Meio Ambiente (CMA), com 17 membros.

Ainda segundo o Regimento Interno do Senado Federal (BRASIL, 2018) cabe ressaltar que cada senador poderá fazer parte de no máximo três comissões como membro titular e outras três como suplente, e a indicação destes para as comissões são pelo Presidente, por indicação dos respectivos líderes, assegurada, tanto quanto possível, a participação proporcional das representações partidárias ou dos blocos parlamentares com atuação no Senado Federal.

Então conforme a estrutura do portal apresentada pode-se inferir que o ciclo de vida das ideias segue conforme a Figura 23:

Figura 23 - Ciclo de vida das ideias na ferramenta Ideia Legislativa



Fonte: do autor.

Identificando-se assim as possíveis etapas em que as ideias podem estar alocadas, de modo que é possível segmentar a base de ideias por status da ideia como uma forma de pré-categorização destas, conforme o Tabela 1 pode-se visualizar o total de ideia que estão em cada uma das etapas da ferramenta Ideia Legislativa.

Tabela 1 - Coleta de ideias

Status da ideia	**Quantidade de ideias**
Campanhas abertas	5058
Campanhas encerradas	32972
Aguardando envio à CDH	5
Na comissão CDH	51
Não Acatadas	24
Convertida em Projeto de Lei	7
Total --->	**38.117**

Fonte: do autor.

4.3 PRÉ-PROCESSAMENTO DAS IDEIAS

A partir da base de ideias, a etapa de pré-processamento representa a primeira ação do modelo proposto ilustrado na Figura 24 na subseção 4.1 deste capítulo, etapa essencial para descoberta de conhecimento para dados não estruturados, por definir uma estrutura que possibilite a aplicação de técnicas sob estes dados. Na etapa de pré-processamento dos dados textuais, neste caso das ideias, usou-se a abordagem do Processamento de Linguagem Natural, e envolveu atividades que foram definidas como: *tokenização*, remoção das *stopwords*, normalização *e steming*.

Para esta etapa utilizou-se a plataforma de desenvolvimento em *Python*® com a biblioteca NLTK (Natural Language Toolkit) que trabalha com processamento de linguagem natural e possui um conjunto de bibliotecas com funções que permite a *tokenização, stemming, tagging*, análise e raciocínio semântico, processamento de texto para classificação, dentre outras.

1. Tokenização[17]: A tokenização, é a primeira etapa que deve ser realizada e se constitui na função responsável pela transformação do texto em termos, portanto a partir de cada ideia divide-se as frases em *tokens* dando suporte para que os passos seguintes como a normalização, remoção das *stopwords*, dentre as demais atividades possam ser realizados. No Quadro 5 é apresentado o código fonte utilizado para identificar e separar todos os termos de determinada ideia dando suporte para as etapas posteriores.

Quadro 5 - Tokenização

Entrada de dados:

```
[('Fim do auxílio moradia para deputados, juízes senadores.' , '253.804') ,
('Revogação da Lei 8313/1991 (Lei Rouanet) com redução de impostos na mesma proporção' , '134.114') ,
('Fim do imposto  de renda sobre o salário de professores.' , '65.815') ,
```

[17] O processo de tokenização tem como objetivo separar palavras ou sentenças em unidades, de modo que neste contexto foi separada dentro de uma ideia cada palavra como um *token*, identificando-a mesmo se tiver encostada em alguma pontuação. Processo essencial para PLN e aplicação de métodos de descoberta de conhecimento utilizados no modelo, por tratarem da classificação, relação ou probabilidade entre palavras.

('Proibam fogos de artifício COM RUÍDOS (rojões, morteiros, bombas, etc)' , '53.361') , ('Fim da Aposentadoria Especial para Senadores e Deputados' , '43.319')]

Função desenvolvida em Phyton:
 	
def tokenizarideias(texto):
frases = []
for (ideia, votos) in texto:
tokenideias = [p for p in nltk.word_tokenize(ideia,'portuguese')]
frases.append((tokenideias, votos))
return frases

Retorno: [(['Fim', 'do', 'auxílio', 'moradia', 'para', 'deputados', ',', 'juízes', 'senadores', '.'], '253.804'), (['Revogação', 'da', 'Lei', '8313/1991', '(', 'Lei', 'Rouanet', ')', 'com', 'redução', 'de', 'impostos', 'na', 'mesma', 'proporção'], '134.114'), (['Fim', 'do', 'imposto', 'de', 'renda', 'sobre', 'o', 'salário', 'de', 'professores', '.'], '65.815'), (['Proibam', 'fogos', 'de', 'artifício', 'COM', 'RUÍDOS', '(', 'rojões', ',', 'morteiros', ',', 'bombas', ',', 'etc', ')'], '53.361'), (['Fim', 'da', 'Aposentadoria', 'Especial', 'para', 'Senadores', 'e', 'Deputados'], '43.319')]

Fonte: do autor.

2. Remoção das s*topwords*: nesta etapa são retiradas as *stopwords* que são palavras como conjunções, preposições, pronomes, ou seja, palavras que, neste contexto, não promovem significado relevante ao texto. Foi usado o conteúdo tokenizado para remoção das *stopwords* com uma lista adotado pelo autor, o resultado é apresentado no Quadro 6. O resultado encontrado utilizando a lista personalizada apresentou uma limpeza mais eficiente do que usando a lista padrão da biblioteca do NLTK. Esta etapa poderia ser executa posterior a normalização, porém foi adiantada pelo motivo de reduzir o processamento, pois reduz de forma significativa a quantidade de *tokens* e também porque a lista já apresentarem acentos e caracteres especiais.

Quadro 6 - Remoção das *stopwords* utilizando lista dos autores

[(['Fim', 'auxílio', 'moradia', 'deputados', ',', 'juízes', 'senadores', '.'], '253.804'),

```
(['Revogação', 'Lei', '8313/1991', '(', 'Lei', 'Rouanet', ')', 'redução', 'impostos',
'proporção'], '134.114'),
(['Fim', 'imposto', 'renda', 'sobre', 'salário', 'professores', '.'], '65.815'),
(['Proibam', 'fogos', 'artifício', 'RUÍDOS', '(', 'rojões', ',', 'morteiros', ',',
'bombas', ', ')'], '53.361'),
(['Fim', 'Aposentadoria', 'Especial', 'Senadores', 'Deputados'], '43.319')]
```

Fonte: do autor.

3. Normalização: A normalização trata de questões como: conversão de letras maiúsculas e minúsculas, remoção de acentos, pontos, números dentre outros. Para a normalização foi utilizada a biblioteca *unicodedata* com o intuito de remover caracteres especiais e acentuação utilizada no texto. No Quadro 7 observa-se um exemplo de normalização.

Quadro 7 - Normalização

Entrada de dados:

```
[(['Fim', 'auxílio', 'moradia', 'deputados', ',', 'juízes', 'senadores', '.'], '253.804'),
(['Revogação', 'Lei', '8313/1991', '(', 'Lei', 'Rouanet', ')', 'redução', 'impostos',
'proporção'], '134.114'),
(['Fim', 'imposto', 'renda', 'sobre', 'salário', 'professores', '.'], '65.815'),
(['Proibam', 'fogos', 'artifício', 'RUÍDOS', '(', 'rojões', ',', 'morteiros', ',',
'bombas', ',')'], '53.361'),
(['Fim', 'Aposentadoria', 'Especial', 'Senadores', 'Deputados'], '43.319')]
```

Função desenvolvida em Phyton:

```
        def limparpalavra(palavra):
            # Unicode normalize transforma um caracter em seu equivalente
em latim.
            nfkd = unicodedata.normalize('NFKD', palavra)
            palavraSemAcento = u"".join([c for c in nfkd if not
unicodedata.combining(c)])
            # Usa expressão regular para retornar a palavra apenas com
números, letras e espaço
    return re.sub('[^a-zA-Z0-9 \\\]', '', palavraSemAcento)

        def normalizar(texto):
            frases = []
            for (ideia, votos) in texto:
                ideianormalizada = [limparpalavra(p) for p in ideia ]
                frases.append((ideianormalizada, votos))
    return frases
```

Retorno:

[(['fim', 'auxilio', 'moradia', 'deputados', '', 'juizes', 'senadores', ''], '253.804'), (['revogacao', 'Lei', '83131991', '', 'lei', 'rouanet', '', 'reducao', 'impostos', 'proporcao'], '134.114'), (['fim', 'imposto', 'renda', 'salario', 'professores', ''], '65.815'), (['proibam', 'fogos', 'artificio', 'ruidos', '', 'rojoes', '', 'morteiros', '', 'bombas'], '53.361'), (['fim', 'aposentadoria', 'especial', 'senadores', 'deputados'], '43.319')]

Fonte: do autor.

4. *Steming*: por fim nesta etapa aplica-se o *steming* que consiste em converter os termos para sua raiz gramatical, eliminando os plurais, sufixos e prefixos dependendo do método utilizado. Para este trabalho foi adotado o método *SnowBall* que é amplamente reconhecido e apresentou resultados semelhantes para este fim comparado ao método RSLPStemmer que apresenta suporte ao português. Existem vários outros métodos, por exemplo o WordNet Lemmatizer e Porter Stemmer. O resultado da aplicação desta fase de steming é observado no Quadro 8.

Quadro 8 - *Steming* utilizando o método SnowBall

[(['fim', 'auxili', 'morad', 'deput', 'juiz', 'senador'], '253.804'),
(['revog', 'lei', '83131991', 'lei', 'rouanet', 'reduca', 'impost', 'proporca'], '134.114'),
(['fim', 'impost', 'rend', 'salari', 'professor'], '65.815'),
(['proib', 'fog', 'artifici', 'ruid', 'roj', 'morteir', 'bomb'], '53.361'), (['fim', 'aposentador', 'especial', 'senador', 'deput'], '43.319')]

Fonte: do autor.

4.4 INDEXAÇÃO

A indexação consiste na segunda etapa do modelo proposto, e necessita do resultado apresentado no decorrer da etapa anterior, de modo que é criado um *corpus* tratado com essa série de documentos, e nesta etapa a partir das palavras tratadas gerou-se um índice que possibilita a leitura mais ágil destas. O armazenamento deste índice pode ser realizado em um banco de dados relacional ou em outros meios. A biblioteca NLTK não possui método pronto para indexação, mas com a criação de algumas funções é possível extrair os termos e gerar a tabela de índices.

Para se criar a tabela de índices primeiramente foi definido os elementos que representam as colunas, palavras ou termos, porém há uma

preocupação para que não exista colunas repetidas, da maneira que um termo não pode ser relacionado como coluna duas vezes. De modo que o tamanho da matriz binária gerada, com todas as ideias possui o tamanho de 17.875 colunas por 38.117 linhas.

Desta forma se aplicado sobre o conjunto de ideias usadas no exemplo da etapa de pré-processamento para formar as colunas da tabela de índices é possível identificar os elementos dispostos na Tabela 2. A segunda parte desta atividade consiste em identificar dentro de cada ideia se as palavras estão presentes em cada uma das colunas, identificando 1 como verdadeiro se o termo corresponde a coluna e 0 como falso se o termo não pertence aquela coluna, conforme demonstrada na Tabela 2, usando as ideias a seguir:

1. ('Fim do auxílio moradia para deputados, juízes senadores.');
2. ('Revogação da Lei 8313/1991 (Lei Rouanet) com redução de impostos na mesma proporção');
3. ('Fim do imposto renda sobre o salário de professores.');
4. ('Proibam fogos de artifício COM RUÍDOS (rojões, morteiros, bombas, etc)') ;
5. ('Fim da Aposentadoria Especial para Senadores e Deputados').

Tabela 2 - Tabela de índices de termos x ideias

	fim	auxili	morad	deput	juiz	senador	revog	lei	83131991	rouanet	reduca	impost	proporca	rend	salari	professor	proib	fog	artifici	ruid	roj	morteir	bomb	aposentador	especial
1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
3	1	0	0	0	0	0	0	0	0	0	0	1	0	1	1	1	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	0	0
5	1	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1

Fonte: O autor

4.5 TÉCNICAS DE KDT

Neste tópico serão analisados os dados coletados usando alguns algoritmos diferentes em busca de extrair conhecimento para serem utilizados no auxílio de tomadas de decisão entre outras atividades.

Antes de aplicar as técnicas de descoberta de conhecimento, contabilizaram-se os termos com maior frequência, gerando assim uma tabela com estes termos e uma nuvem de palavras com os termos que mais apareceram nas ideias. Observa-se os dez primeiros itens na Tabela 3, esta atividade foi oportunizada por meio da etapa de pré-processamento dos dados.

Tabela 3 - Termo frequência base de ideias

Termo	**Frequência**
Fim	3.556
Todo	2.340
Lei	2.321
Público	2.127
Político	2.099
Imposto	2.019
Salário	1.799
Cargo	1.377
Redução	1.368
Pública	1.293

Fonte: do autor.

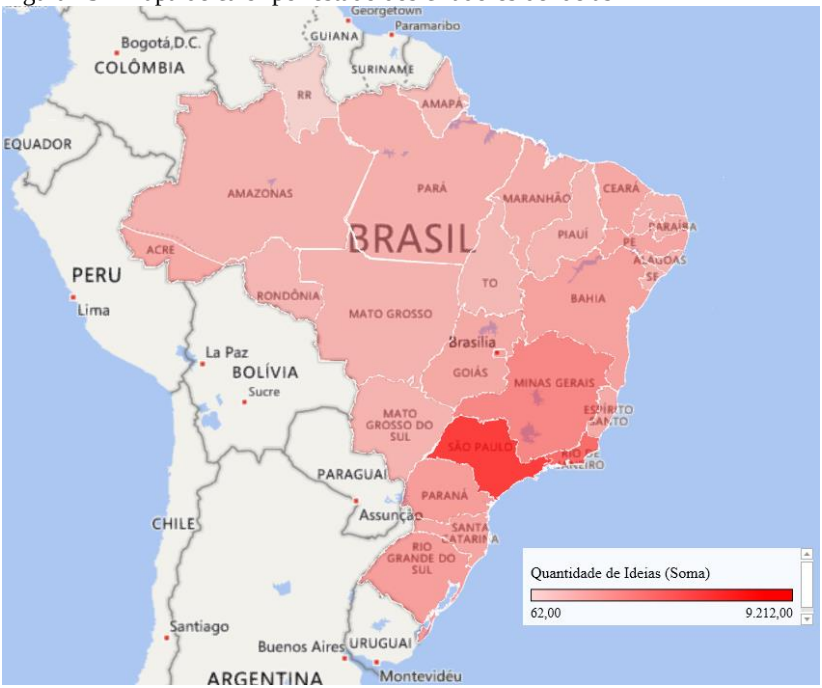
Após gerado a tabela criou-se a nuvem de palavras apresentada na Figura 24, demonstrando palavras-chave que aparecem em uma grande porcentagem das ideias cadastradas no Portal e-Cidadania. Destacando-se aqui as palavras como “fim” presente em 3.556 ideias que representa 9,33% de todas a ideias no portal, seguida de palavras como “todo” com 2.340, “lei” com 2.321, “público” com 2.127 e “político” com 2.099 aparições, evidenciando-se assim palavras que fazem parte de muitas ideias que podem expressar algum desejo comum dos colaboradores.

Figura 24 - Nuvem de palavras base de ideias

Fonte: do autor.

Ainda sobre os dados coletados foram identificados quais são os estados que mais criam ideias para a ferramenta Ideia Legislativa, destacando-se o estado de São Paulo com 9.212 ideias no portal seguido do Rio de Janeiro com 5.601 e Minas Gerais com 3.568 ideias. Os estados que menos participaram foram Roraima e Amapá com 62 e 91 ideias, respectivamente. A Figura 25 ilustra essa afirmativa.

Figura 25 - Mapa de calor por estado dos criadores de ideias



Fonte: do autor.

Nesta segunda etapa foi adotado apenas as 5.087 ideias que estão em campanha aberta para receber apoio, na busca de se ter uma perspectiva de quais são os termos estão vigorando no portal em 2018, contabilizado os termos com maior frequência *tidf*, gerando assim uma planilha com estes termos e uma nuvem de palavras. A Figura 26 ilustra essa nuvem de palavras.

Figura 26- Nuvem de palavras das ideias em campanha aberta

Fonte: do autor.

Nesta segunda nuvem de palavras é possível identificar que as palavras que aparecem são semelhantes às da primeira nuvem de palavra, destaca-se que o portal está no ar desde 2016, entretanto o número de ideias postadas no portal vem crescendo exponencialmente com o passar dos anos, mas mesmo assim os anseios de quem cria as ideias parecem permanecer os mesmos, sendo possível evidenciar pela aparição dos termos: imposto, público, político, deputado, redução, entre outros.

4.5.1 Cálculo de similaridade para agrupamento de ideias

Após o texto ser indexado iniciou-se o processo de mineração, neste passo da pesquisa buscou-se encontrar documentos que são similares, usando o cálculo de similaridade de cosseno por possibilitar de forma simples a comparação entre documentos textuais, neste caso ideias. Esta função se torna útil para calcular a relevância de palavras em documentos, por meio da métrica do cosseno entre dois vetores. O algoritmo utilizado para realizar o cálculo é apresentado no Quadro 9.

Quadro 9 - Cálculo de similaridade baseado no cosseno

```
def cosseno_similaridade(vet1, vet2):
```

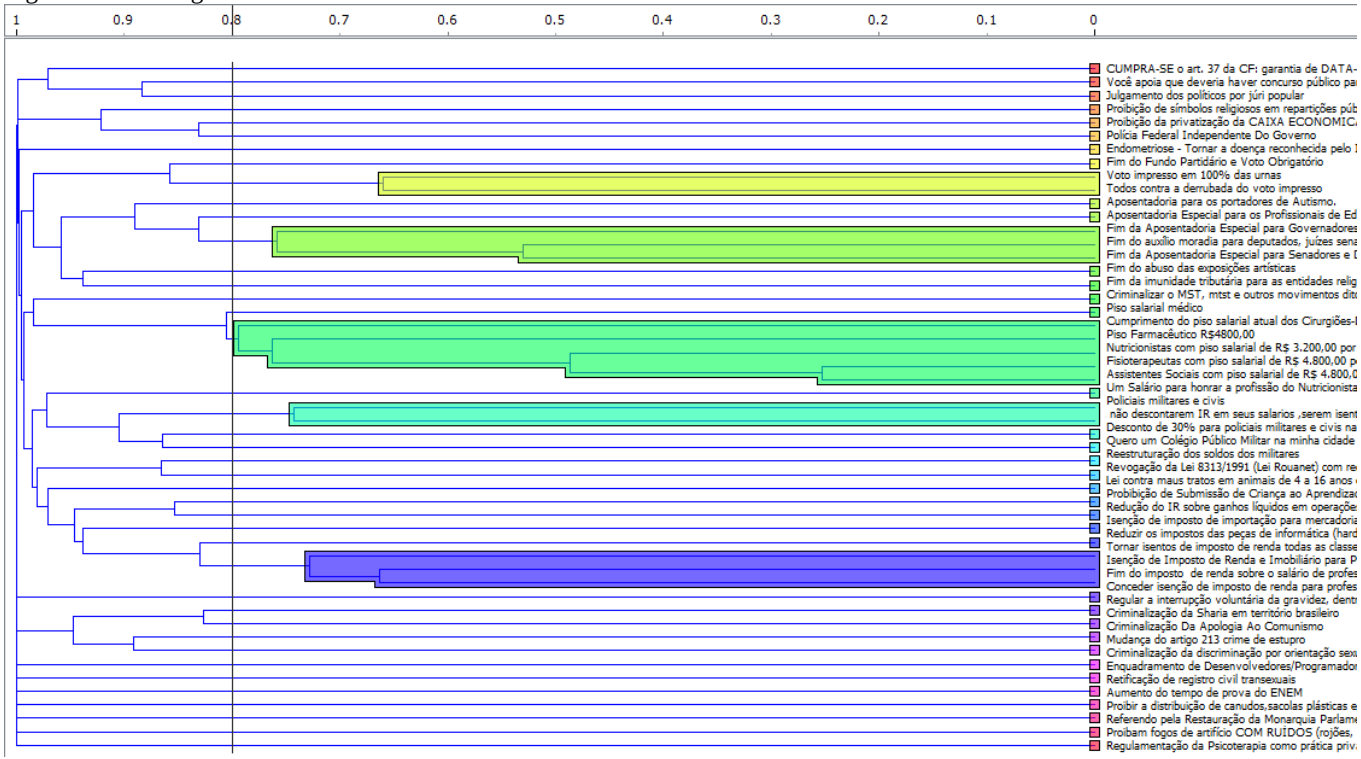
```
    interseccao = set(vet1.keys()) & set(vet2.keys())
    numerador = sum([vet1[x] * vet2[x] for x in interseccao])
    sum1 =    sum([vet1[x]**2 for x in vet1.keys()])
    sum2 = sum([vet2[x]**2 for x in vet2.keys()])
      denominador =      math.sqrt(sum1) * math.sqrt(sum2)

if not denominador:
   return 0.0
else:
   coef = float(numerador) / denominador
return coef
```

Fonte: do autor.

Para o primeiro exemplo foi adotado o grupo de ideias que estão na CDH (Comissão de Direitos Humanos) aguardando parecer do relator para prosseguir ou se extinguir. Utilizando a função apresentada foi calculado à similaridade entre todas as ideias que compõe esta categoria, e por fim gerado uma matriz de distância entre todos os elementos. A posteriori foi gerado um cluster hierárquico que calcula o agrupamento hierárquico de tipos arbitrários de objetos a partir da matriz de distâncias e demonstrado num dendrograma os resultados.

No dendograma foram aplicadas as seguintes propriedades para gera-lo: para medir distâncias entre clusters foi adotado a *Average linkage* que calcula a distância entre os elementos mais próximos entre dois clusters e o aspecto de seleção com o limiar de 0.8 para análise. A ferramenta visual se demonstra muito útil para análise pois dimensiona próximo as ideias similares e os clusters que possuem algum aspecto similar àquele grupo de ideias. Na Figura 27 pode ser observado o dendograma das ideias que estão na CDH.

Figura 27 - Dendograma das ideias estão na CDH



Fonte: do autor.

Outra maneira de se visualizar ideias similares é utilizando Multidimensional Scaling (MDS) que é uma técnica que encontra uma projeção de pontos de baixa dimensão (neste caso, bidimensional), onde ela tenta ajustar as distâncias entre os pontos. Usando configuração similar definida para gerar o dendograma anterior, criou-se o MDS para a categoria de ideias na CDH e as cores foram definidas de acordo com a quantidade de apoios recebidos, iniciado entre 0 a 50 mil apoios até 200 a 250 mil apoios (Figura 28).

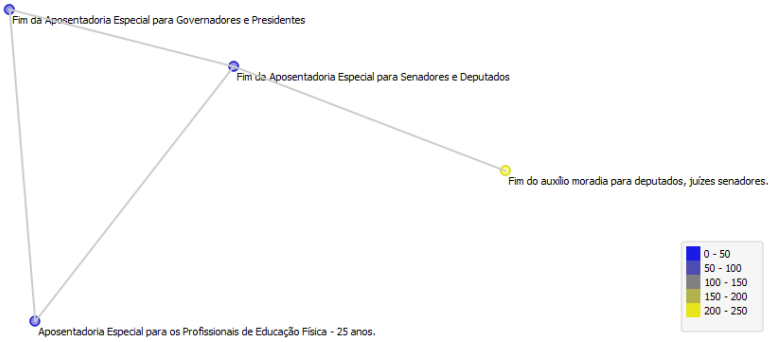
Figura 28 - MDS das ideias estão na CDH

Fonte: O autor.

Por meio da Figura 27 e Figura 28 destaca-se a existência de cinco clusters com a limiar definida. No primeiro cluster as ideias são: “Voto impresso em 100% das urnas” e “Todos contra a derrubada do voto impresso” com 24.487 e 20.843 respectivamente, são duas ideias que esquadrinham pelo mesmo contexto e ambas deverão ser analisadas pela CDH.

No cluster 2 temos as ideias: “Fim do auxílio moradia para deputados, juízes senadores.”, “Fim da Aposentadoria Especial para Senadores e Deputados” e “Fim da Aposentadoria Especial para Governadores e Presidentes”. Estas representam um contexto similar entre as ideias que buscam extirpar alguns benefícios concedidos, porém para categorias e tipos de privilégios diferentes, conforme Figura 29.

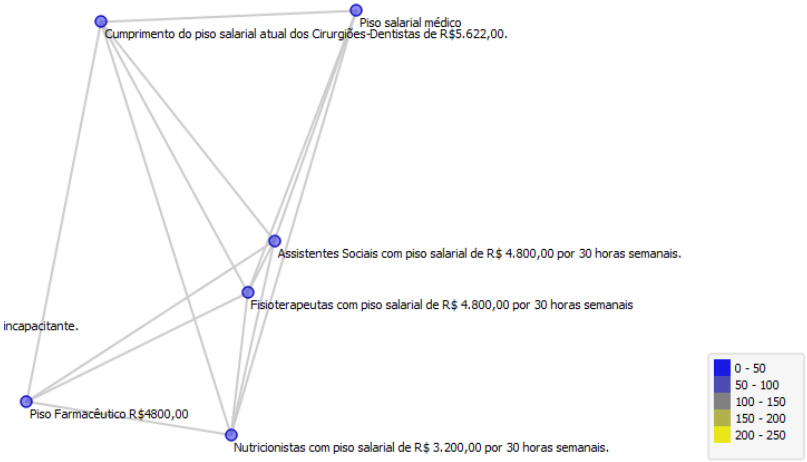
Figura 29 - Cluster 2, MDS das ideias estão na CDH



Fonte: do autor.

Os demais clusters se assemelham ao cluster 2, pois tem apenas um objeto em comum (Cluster 3: piso salarial; Cluster 4: policiais militares e civis e Cluster 5: imposto de renda). Todavia, se diferem pelas categorias beneficiadas com a ideia ou pelo direito que estão almejando, conforme apresenta na Figura 30 do cluster 3.

Figura 30 - Cluster 3, MDS das ideias estão na CDH



Fonte: do autor.

Com objetivo de criar um segundo exemplo, foi selecionado o grupo de ideias que estão em campanha aberta para coleta de apoios. Na sequência, calculado à similaridade entre todas as ideias que compunham este corpus e assim gerado uma matriz de distância entre todos os elementos.

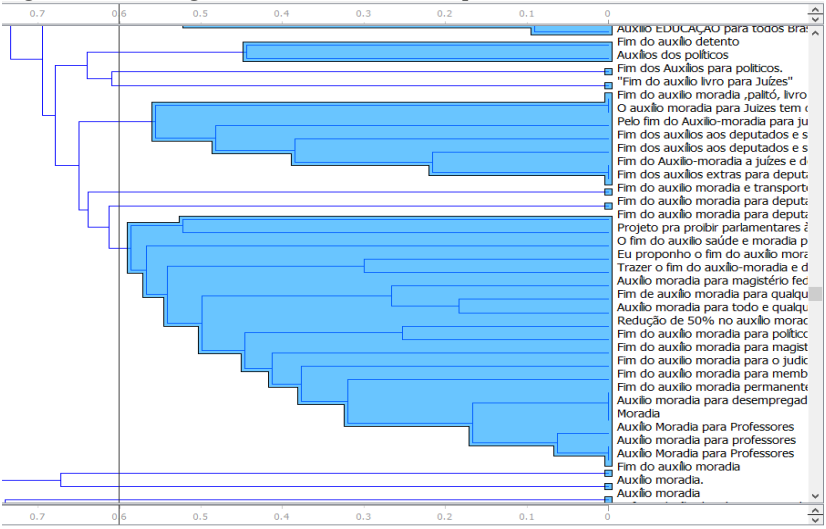
A *posteriori* foi gerado um dendograma utilizando a abordagem *Average linkage*, para medir distâncias entre clusters, com limiar de 0.6 para análise. Cabe ressaltar que este grupo possui 5.058 ideias ficando inviável o tamanho da imagem para inserção desta nesta pesquisa, contudo foram selecionados alguns clusters que se destacaram para demonstrar a análise.

Assim, pode-se verificar que há muitas ideias com contextos similares em busca do mesmo objetivo, a primeira selecionada trata sobre o fim do auxílio moradia tanto para políticos tanto quanto para membros do judiciário conforme figura 31. Nos *clusters* menores, que circundam os maiores, é possível evidenciar duas situações, uma em que o contexto é o mesmo e a outra diferente, já que propõe a criação de uma lei para se implantar auxílio moradia aos professores.

Portanto, ideias criadas com palavras diferentes, não são reconhecidas diretamente como similares com a utilização do cálculo do cosseno. Mesmo que tenham as mesmas palavras não significa que o contexto é exatamente o mesmo, mas cabe destacar que com uma

ferramenta de visualização de dados, tal como o dendograma, torna-se mais ágil esta verificação.

Figura 31 - Dendograma 1 das ideias em campanha aberta



Fonte: do autor.

Dentre os 3.231 *cluters* de ideias similares gerados diante das configurações do dendograma, o cluster demonstrado na ilustração acima tange o contexto do fim do auxílio moradia para políticos e juízes, a ideia com maior número de apoios possui cinco e varia até nenhum apoio.

Ressalta-se que está mesma ideia já havia sido cadastrada: “Fim do auxílio moradia para deputados, juízes senadores” e possui 253.807 apoios, e já está na Comissão de Direitos Humanos aguardando parecer do relator. Agora tomando essa ideia como *query* para consulta, foi realizada uma busca dentro de toda a base de ideias, no qual o resultado é apresentado no Tabela 4 informando inicialmente a ideia utilizada como argumento de busca e as demais ideias similares com limiar maior ou igual a 0,8.

Tabela 4 - Resultado para ideias similares para “Fim do auxílio moradia para deputados, juízes senadores” com o limiar de 0,8

Ideia	Qnt. Apoio	Criador	Estado	Status
Fim do auxílio moradia para deputados, juízes senadores.	253.804	Idealizador 1	RJ	Na Comissão
Fim Do Auxílio Moradia Para Deputados, Senadores E Juízes	51	Idealizador 1	RJ	Encerrada
Fim do auxílio moradia para Senadores e Deputados	16	Idealizador 2	DF	Encerrada
Fim de auxílio de moradia p/juízes e deputados e senadores	10	Idealizador 3	SP	Encerrada
Reajuste do Auxílio Moradia para Deputados, senadores e juízes.	9	Idealizador 4	RJ	Encerrada
O fim do auxílio moradia para deputados, senadores e juízes.	5	Idealizador 5	BA	Encerrada
Fim do auxílio moradia para deputados, juízes senadores.	5	Idealizador 6	RS	Em Campanha
Fim do auxílio moradia para Deputados, senadores, e juízes	3	Idealizador 7	MG	Encerrada
Fim do auxílio moradia para juízes e deputados	2	Idealizador 8	MT	Encerrada
Fim do auxílio moradia para deputados, senadores e juízes	2	Idealizador 9	RS	Encerrada
Fim Do Auxílio-Moradia Aos Deputados Federais E Aos Senadores	1	Idealizador 10	PI	Encerrada
Fim do auxílio moradia para deputados e juízes	1	Idealizador 11	PB	Encerrada
Fim do auxílio moradia para deputados e juízes	1	Idealizador 12	PB	Encerrada
Fim do auxílio moradia e transporte dos senadores e deputados.	1	Idealizador 13	ES	Encerrada
Fim do auxílio moradia para deputados, juízes senadores.	1	Idealizador 14	SP	Em Campanha
Auxílio moradia para deputados e senadores	1	Idealizador 15	ES	Encerrada

Fonte: do autor.

Destaca-se que as duas primeiras ideias na Tabela 4 foram criadas pela mesma pessoa e concorriam entre si, durante o tempo de quatro meses que estas permanecem em campanha. Analisando o número de

apoios, se todas as ideias similares apresentadas fossem reunidas, o total seria de 253.870 apoios, apesar da diferença não ser discrepante, a análise evidencia a quantidade de ideias com o mesmo contexto nesta base.

Neste caso também foi possível identificar que mesmo aquelas com o grau de similaridade mais baixo que o estabelecido no exemplo anterior, tratavam do mesmo contexto da ideia de busca. Na Tabela 5 apresenta-se alguns exemplos dessas ideias.

Tabela 5 - Resultado para ideias similares para “Fim do auxílio moradia para deputados, juízes senadores” com o limiares menores entre 0,6 a 0,8

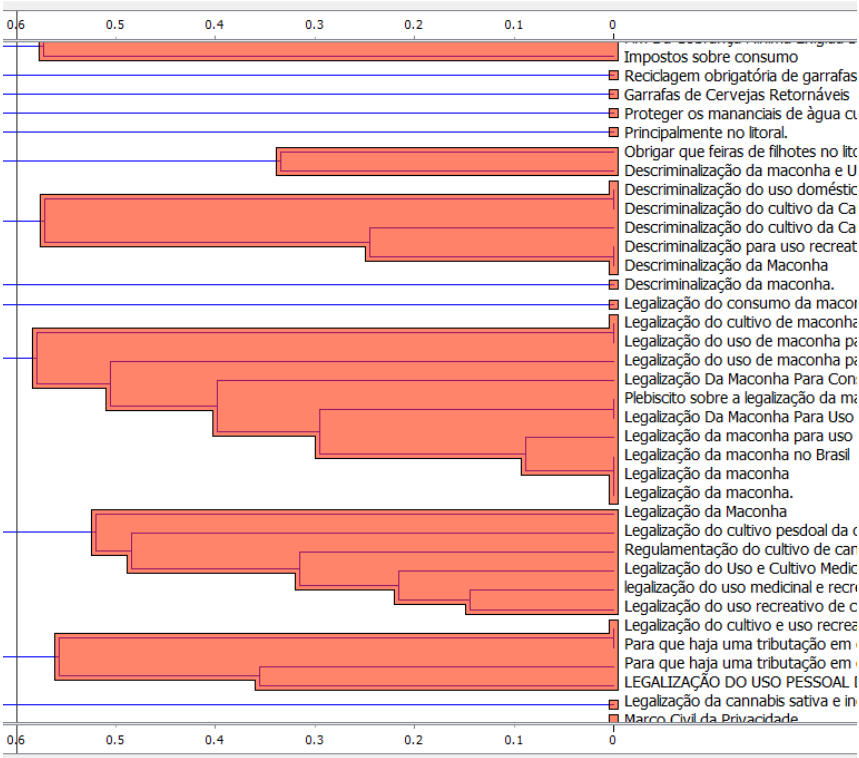
Ideia	Qnt. Apoio	Estado	Status
Fim do auxílio moradia para políticos.	2.246	DF	Encerradas
Fim do auxílio moradia para políticos e juízes.	89	PR	Encerradas
Fim do auxílio moradia	45	RS	Encerradas
Fim do auxílio moradia para políticos.	29	PE	Encerradas
Fim do auxílio moradia para políticos	22	SC	Encerradas
Auxilio Moradia para Militares	17	RJ	Encerradas
Proibição do auxílio moradia para cargos públicos	13	MG	Encerradas
Fim dos auxilio para Políticos e Juízes.	12	MG	Encerradas
Fim do auxílio moradia e outros auxílios para, juízes, e todos os políticos.	11	SP	Encerradas
Auxílio moradia para professores	11	RJ	Em campanha
Fim do auxilio moradia e outros auxílios para, juízes, e todos os políticos.	9	SP	Encerradas
Fim do auxílio-moradia.	7	RJ	Encerradas
Fim do auxílio moradia	5	TO	Encerradas
Fim de auxílio,moradia e escola para políticos e juízes	5	RS	Encerradas
Fim do auxílio moradia para deputados, juízes senadores.	5	RS	Em campanha
Auxílio Moradia para Professores	5	MA	Em campanha

Fonte: O autor.

Dentre os demais *clusters* encontrados que estão em campanha foi verificado um grande número de ideias com contexto similares como o exemplo da Figura 32, no qual ilustra a existência de 5 *clusters* sobre a legalização da maconha. Diversos outros contextos foram identificados, tal como, porte de arma de fogo, privatização dos correios, redução da

maioridade penal, criminalização do *fake news*, fim do voto obrigatório, redução de impostos, entre dezenas de outros contextos.

Figura 32 - Dendograma 2 das ideias em campanha aberta



Fonte: do autor.

Por fim ficou evidente a quantidade de ideias com contextos semelhantes que concorrem entre si por apoio e que poderiam unir-se. Essa união, resultaria na redução significativa do tamanho total da base e facilitaria o processo de tomada de decisão, além de promover maior apoio a ideias que apresentam maior frequência.

4.5.2 Categorização de texto

Para o segundo protótipo definido no modelo apresentado neste estudo, será classificado ideias de acordo com as comissões permanentes do senado. A realização desta abordagem só foi possível após realizar a

fase de pré-processamento em busca de reduzir a dimensionalidade dos dados e se obter resultados mais expressivos, com o uso do algoritmo de Naive Bayes para classificação.

Para validar os resultados deste algoritmo é necessário dois conjuntos de dados antes de aplicar a classificação na base de ideias conforme descrito por Sebastiani (2002). O primeiro conjunto é o de treinamento, o qual contém as ideias e as classes as quais elas se enquadram. O segundo conjunto possui a mesma configuração, entretanto é usado para teste do conjunto de treinamento avaliando assim de forma estatística o índice de assertividade do algoritmo (DUMAIS *et al.*, 1998).

Para montar estes conjuntos foi visitado o portal do senado em busca de temáticas apreciadas pelas comissões atualmente. Entretanto foi identificado que as matérias encaminhadas e apreciadas pelas comissões permanentes muitas vezes também podem fazer parte de outra comissão, de modo que podem ser criadas comissões mistas para analise destas, ou ainda serem encaminhadas precipitadamente, de modo que não estão em conformidade com regimento que define os objetivos de cada comissão. A Figura 33 ilustra a relação dos assuntos da comissão e o tipo de matéria.

Figura 33 - Quadro de Assuntos x Tipo de Matéria - CDH

Quadros Assuntos X Tipo de Matéria

Assunto Geral	Assunto Específico	PLS	ECD	PLC	SCD	
Administrativo	Servidores públicos	2				2
		2				2
Jurídico	Defesa do consumidor	1				1
	Direito civil e processual civil	2		2		4
	Direito penal e processual penal	4		1		5
	Trânsito	1		1		2
		8		4		12
Social	Assistência social	2				2
	Direitos humanos e minorias	24	1	2		27
	Educação	6				6
	Família, proteção a crianças, adolescentes, mulheres e idosos	23		4	1	28
	Meio ambiente	2				2
	Saúde	3				3
	Trabalho e emprego	2		1		3
		62	1	7	1	71
Total de matérias na Comissão:		72	1	11	1	85

Fonte: Portal do Senado[18].

Na figura pode-se identificar que no quadro Jurídico o primeiro grupo que contém uma ideia trata sobre defesa do consumidor que corresponde diretamente aos objetos da Comissão de Transparência, Governança, Fiscalização e Controle e Defesa do Consumidor e dois grupos subsequentes pertencem a Comissão de Constituição, Justiça e Cidadania, de modo que foi buscado junto ao regimento da casa quais são os objetivos e abordagens de cada comissão conforme Anexo A. No regimento consta quais são os assuntos que cada uma das comissões trata

[18] Disponível em: < http://www8d.senado.leg.br/dwweb/sgmDoc.html?docId=92615> Acesso em abr. 2018.

e define que se há uma matéria que pertença há duas comissões estas deverão ser discutidas em comissões mistas.

A partir disso, manualmente foram identificadas palavras chaves dos objetivos das comissões para configurar a classe, e na sequência separar o conjunto de treinamento e teste.

Após os passos do processo de pré-processamento dos dados para gerar a matriz de frequência dos termos são aplicados sob os conjuntos, cria-se a tabela de índices, destacando que esta tabela se difere um pouco da utilizada no cálculo de similaridade, sendo que agora é representada por termo x classe.

Com a tabela de índices gerada o algoritmo de Naive Bayes cria uma tabela de estimativa com as probabilidades de que cada termo ocorrer para uma determinada classe, conforme função da biblioteca NLTK demonstrada no Quadro 10.

Assim têm-se duas formas de se estimar onde uma nova ideia será classificada. A primeira forma é dada pela associação da quantidade de termos que ocorrem com uma maior frequência para uma mesma classe, de modo que quanto maior a frequência deste, maior a probabilidade deste termo pertencer a determinada classe. A segunda maneira é dada pela associação de não ocorrência do termo para determinada classe, pois se este não ocorre nenhuma vez as chances de ideias que contenham este termo pertencer a esta categoria são nulas. Entretanto para a categorização da ideia é necessário a somatória entre a probabilidade de todos os termos perante a classe.

Quadro 10 - Constrói a tabela de probabilidades e impressão dos rótulos e *tokens* mais significativos

```
classificador = nltk.NaiveBayesClassifier.train(basecompletatreinamento)
print(classificador.labels())
print(classificador.show_most_informative_features(20))
```

Resultados:

Classes encontradas no conjunto de treinamento:
['CCJ', 'CAE', 'CDH', 'CE', 'CAS', 'CMA', 'CRA', 'CTFC', 'CSF']

Termos mais significativos:

soc = True	CRA : CCJ	=	20.0 : 1.0
regim = True	CMA : CCJ	=	13.3 : 1.0
crim = True	CDH : CCJ	=	10.5 : 1.0
diminu = True	CTFC : CAE	=	10.0 : 1.0
crimin = True	CDH : CCJ	=	8.6 : 1.0
lei = True	CMA : CCJ	=	8.0 : 1.0

contr = True	CMA : CCJ = 8.0 : 1.0
proib = True	CMA : CCJ = 8.0 : 1.0
uso = True	CMA : CCJ = 8.0 : 1.0
criminal = True	CRA : CDH = 7.0 : 1.0

Fonte: do autor.

O algoritmo do Quadro 10 foi treinado usando as 100 primeiras ideias com mais apoios, no qual todas foram classificadas em alguma determinada comissão. Na primeira linha de resultado é identificado todas as classes encontradas no conjunto de treinamento que a variável classificador recebe do método *NaiveBayesClassifier.train*. Nota-se que, entre o conjunto de ideias selecionado para o exemplo, têm-se apenas nove das treze comissões permanentes representadas.

Ainda no Quadro 10, na segunda parte do resultado, é apresentado os termos mais significativos dentro deste conjunto de ideias, por exemplo, na primeira linha é apresentado o radical "soc" que corresponde a palavra "sociais" e é calculado pelo algoritmo que se o termo "soc" aparece há uma probabilidade 20 vezes maior da ideia pertencer a Comissão de Agricultura e Reforma Agrária (CRA) do que da Comissão de Constituição, Justiça e Cidadania (CCJ).

Por meio desta demonstração, observa-se que um conjunto pequeno de treinamento para tantas classes ainda se apresenta ineficaz para classificação, pois apresenta probabilidades altas para termos que são comuns tais como "sociais" apresentado no exemplo anterior. Destaca-se também os termos "crimin" e "lei" possuírem probabilidades relativamente baixas quando relacionados a CCJ e índices altos para Comissão de Diretos Humanos e Legislação Participativa e Comissão de Meio Ambiente. Evidencia-se também que neste conjunto de termos significativos o resultado apresenta apenas situações verdadeiras em são dadas pela frequência em que os termos estão alocados na tabela de índice não ocorrendo situações de não frequência de termos.

A aplicação do algoritmo de Naive Bayes em um conjunto de documentos textuais, neste caso ideias, consiste em estimar a probabilidade das ideias pertencerem a todas as classes possíveis, representadas neste cenário como comissões permanentes do senado. Nesta linha após aplicação do algoritmo é apresentado como resultado a classe que apresente a maior probabilidade que ela pertença.

Para o próximo exemplo, foram utilizadas 200 ideias para o conjunto de treinamento, e buscou-se identificar a probabilidade de uma ideia pertencer a todas as classes. Neste exemplo o objetivo foi de

compreender melhor o funcionamento deste algoritmo e evidenciar como ele pode auxiliar no processo de categorização das ideias do senado quando uma ideia possa pertencer há mais de uma comissão, e se verifique a possibilidade de criação comissões mistas com junção de duas ou mais. Desta maneira, no Quadro 11 tem-se primeiramente, o algoritmo e após os resultados deste.

Quadro 11 - Calcula a probabilidade de uma ideia pertencer a todas as classes

```
distribuicao = classificador.prob_classify(novo)
for classe in distribuicao.samples():
    print("%s: %f" % (classe, distribuicao.prob(classe)))
```

Resultados:

```
['extinc', 'curs', 'human', 'univers', 'public']
CE: 0.703301
CAE: 0.018547
CDH: 0.002667
CCJ: 0.575479
CAS: 0.000006
CMA: 0.000000
CRA: 0.000000
CTFC: 0.000000
```

Fonte: do autor.

É possível observar, no Quadro 11 que por meio da função *prob_classify,* foi calculado para a ideia "Extinção dos cursos de humanas nas universidades públicas" a probabilidade de pertencer a todas as classes. Pelos resultados, observa-se que a ideia foi classificada corretamente ao rótulo Comissão de Educação, Cultura e Esporte (CE) (CE: 0.703301), a qual é responsável por tratar de assuntos da educação. No entanto, ficou próximo também do rótulo CCJ (CCJ: 0.575479), isto se dá devido ao primeiro e último termo da ideia serem similares aos termos utilizados por outras ideias que são matérias de análise da CCJ, sendo que a CCJ é a detentora do maior número de ideias no conjunto de treinamento.

Apresentado detalhes do funcionamento da técnica de classificação, e agora utilizando um conjunto de treinamento com 1.586 ideias previamente categorizadas, aplicou-se o algoritmo na base de ideias utilizando o método de validação do *Cross Validation* com 10 *folds.* De modo que o conjunto de amostras se divide em 10 partes e cada uma das amostras foi utilizada ao menos uma vez para o conjunto de teste nas 10

iterações realizadas. Para maior detalhamento é apresentado a Figura 34 com a Matriz de Confusão da classificação com *Naive Bayes*.

Figura 34 - Matriz de Confusão, instâncias da Classificação - *Naive Bayes*

	CAE	CAS	CCJ	CCT	CDH	CDR	CE	CI	CMA	CRA	CRE	CTFC	∑
CAE	206	1	1	24	0	106	1	6	0	3	2	1	351
CAS	2	55	4	17	0	24	0	1	0	0	0	0	103
CCJ	1	8	155	13	3	23	0	0	0	0	1	0	204
CCT	0	0	0	25	0	2	0	0	0	0	0	0	27
CDH	0	1	1	41	102	88	1	1	2	0	1	0	238
CDR	0	0	0	0	0	26	0	0	0	0	0	0	26
CE	3	2	1	73	3	112	100	9	5	1	1	0	310
CI	0	0	0	1	0	2	0	36	0	1	0	0	40
CMA	0	0	0	0	0	2	0	0	26	0	0	0	28
CRA	3	0	0	6	0	33	0	1	3	22	0	0	68
CRE	3	0	0	4	0	15	0	0	0	0	61	0	83
CTFC	0	0	0	14	0	53	0	6	1	0	1	33	108
∑	218	67	162	218	108	486	102	60	37	27	67	34	1586

Fonte: do autor.

Na matriz de confusão, apresentada na Figura 35 é possível destacar duas classes com o maior número de instâncias classificadas corretamente, sendo elas a CAE com 206 classificações corretas e a CJJ com 155. Importante observar que a CAE possui um alto nível de acerto, porém esta mesma classe possui 106 instancias classificadas como CDR de forma imprecisa, mas somando o total de instâncias classificadas em outras classes temos 145, ainda que esta classe tenha o maior número total de instâncias dentre as classes, contando com 351 ideias.

A classe CDR foi também a classe que mais recebeu instâncias classificadas incorretamente somando 460 instâncias. Fator que pode estar associado ao fato de ser uma classe formada por apenas 26 ideias e por possuir muitos termos que podem ser similares a outras classes.

O resultado final do processo de avaliação é apresentado na Tabela 6, sendo esta formada pela média de desempenho de todas as classes do classificador após as 10 iterações.

Tabela 6 - Média da avaliação dos resultados da classificação de todas as classes

Métrica	**Resultado**
Precisão:	0,898
Recall:	0,534
Acurácia:	0,534
F-Measure:	0,626

Fonte: do autor.

Ao observar a Tabela 6 destaca-se que a precisão da classificação neste conjunto de treinamento foi de 89,8% que representa a porcentagem de ideias que foram corretamente classificadas, dentre todas as ideias. Quando comparado a acurácia da classificação fica numa média de 53,40% de acerto, que denota a proporção total de classificações corretas. De forma que com uma precisão alta e uma acurácia média significa que nosso conjunto de treinamento ainda necessita de mais ideias classificadas para atingir padrões de excelência, todavia o resultado está bem próximo do de níveis aceitáveis de assertividade.

Esta baixa taxa de assertividade pode ser explicada por causa das comissões CCT e CDR possuírem poucos exemplos na base de treinamento, de forma que nestes exemplos alguns termos podem possuir uma frequência mais alta do que comparada a outras classes, por exemplo a CAE, de modo que pode induzir o classificador ao erro.

Por meio da Figura 35 demonstra-se a porcentagem de acertos e erros para cada classe por meio da mesma Matriz de Confusão da classificação com *Naive Bayes*.

Figura 35 - Matriz de Confusão, índices da Classificação com Naive Bayes

	CAE	CAS	CCJ	CCT	CDH	CDR	CE	CI	CMA	CRA	CRE	CTFC	∑
CAE	58.7 %	0.3 %	0.3 %	6.8 %	0.0 %	30.2 %	0.3 %	1.7 %	0.0 %	0.9 %	0.6 %	0.3 %	351
CAS	1.9 %	53.4 %	3.9 %	16.5 %	0.0 %	23.3 %	0.0 %	1.0 %	0.0 %	0.0 %	0.0 %	0.0 %	103
CCJ	0.5 %	3.9 %	76.0 %	6.4 %	1.5 %	11.3 %	0.0 %	0.0 %	0.0 %	0.0 %	0.5 %	0.0 %	204
CCT	0.0 %	0.0 %	0.0 %	92.6 %	0.0 %	7.4 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	27
CDH	0.0 %	0.4 %	0.4 %	17.2 %	42.9 %	37.0 %	0.4 %	0.4 %	0.8 %	0.0 %	0.4 %	0.0 %	238
CDR	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	100.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	26
CE	1.0 %	0.6 %	0.3 %	23.5 %	1.0 %	36.1 %	32.3 %	2.9 %	1.6 %	0.3 %	0.3 %	0.0 %	310
CI	0.0 %	0.0 %	0.0 %	2.5 %	0.0 %	5.0 %	0.0 %	90.0 %	0.0 %	2.5 %	0.0 %	0.0 %	40
CMA	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	7.1 %	0.0 %	0.0 %	92.9 %	0.0 %	0.0 %	0.0 %	28
CRA	4.4 %	0.0 %	0.0 %	8.8 %	0.0 %	48.5 %	0.0 %	1.5 %	4.4 %	32.4 %	0.0 %	0.0 %	68
CRE	3.6 %	0.0 %	0.0 %	4.8 %	0.0 %	18.1 %	0.0 %	0.0 %	0.0 %	0.0 %	73.5 %	0.0 %	83
CTFC	0.0 %	0.0 %	0.0 %	13.0 %	0.0 %	49.1 %	0.0 %	5.6 %	0.9 %	0.0 %	0.9 %	30.6 %	108
∑	218	67	162	218	108	486	102	60	37	27	67	34	1586

Fonte: do autor.

Na figura 35, na qual é apresentada os percentuais de acerto e erro para cada classe de ideias, destaca-se novamente a CDR que possui 100% de acerto na classificação real de suas instâncias. Porém, do mesmo modo, é a que mais recebe instâncias classificadas erroneamente. Nota-se também que a CCT com 27 instâncias, CI com 40 instâncias e CMA com 28 instâncias possuem índices de acerto acima dos 90% na classificação de seus itens. Dentre estes rótulos, destaca-se que a classes CI e CMA têm índices muito baixos de classificações com imprecisão e CCT muito se assemelha a CDR com um alto índice de classificações erradas.

Cabe ainda destacar que a classe CAE mesmo possuindo a maior porcentagem de instancias que compõe o conjunto de treinamento possui um índice de acerto de 58,7% indica que um balanceamento entre as classes do conjunto pode afetar o processo de classificação.

Assim, encerra-se este tópico que tratou de aplicar técnicas de mineração de dados sob a base de ideias do portal e-Cidadania, em busca de soluções para o problema proposto.

4.6 ANÁLISES E DISCUSSÕES

Neste tópico será discutido os resultados encontrados na aplicação das técnicas apresentadas. A primeira constatação feita é em relação ao aumento circunstancial na quantidade de ideias coletas em 2017 para 2018 são cerca de 12 mil ideias criadas em menos de 8 meses. Este fato corrobora com a literatura uma vez que segundo Kampa e Cziulik (2016) o processo de ideação amparado no *crowndsourcing* pode gerar um

grande número de ideias num curto espaço de tempo o que dificulta os processos de gestão, ressaltando a importância das técnicas utilizadas na presente dissertação.

A mineração de dados textuais inicia-se com o uso do cálculo da similaridade para dar suporte aos processos da Gestão de Ideias, no qual Poveda, Westerski e Iglesias (2012) salientam sobre a importância do uso de técnicas e ferramentas em prol de facilitar este trabalho. Deste modo, ao aplicar esta técnica foi possível destacar a quantidade de ideias similares contidas no portal e-Cidadania, e que por muitas vezes possuem o mesmo objetivo.

Neste caso, algumas ainda estão na situação em campanha aberta, ou seja, para receberem apoio da população por meio de votação, e assim competem entre si pelo mesmo apoio de uma determinada comunidade que almeja esta melhoria, e podem dividir apoios e não conseguindo os 20 mil necessários para próxima etapa mesmo que o objetivo destas ideias sejam o mesmo.

Tal afirmação é evidenciada na Figura 36, mas principalmente, pelas análises demonstradas no tópico 4.5.1 que trata do uso do cálculo de similaridade sob os dados do portal. Na Figura 36 toma-se como exemplo as ideias que abordam sobre uma proposta para que seja criado para os nutricionistas um teto salarial base e destaca-se que a segunda ideia determina um valor para este teto e acrescenta uma carga horária para a jornada de trabalho para este valor base.

Figura 36 - Tela para pesquisa das Ideias

Ideia Legislativa	Apoios
Fim do auxílio moradia para deputados, juízes senadores.	253.807
Reduzir os impostos sobre games do atual 72% para 9%	75.930
Fim do estatuto do desarmamento	62.285
Fim do Imposto sobre Veículo Automotores, IPVA	57.861
Criminalizar a homofobia para punição de pessoas que atacam outras pessoas por serem LGBT.	55.698
Regulamentação das Atividades de Marketing de Rede.	43.949
Fim da Aposentadoria Especial para Senadores e Deputados	43.321
Discriminalização Do Cultivo Da Cannabis Pra Uso Próprio	32.163
Piso Farmacêutico R$4800,00	28.571
Referendo pela Restauração da Monarquia Parlamentarista no Brasil	28.564
Criminalização da Sharia em território brasileiro	28.526
Liberação da venda de armas e munições importadas, em lojas. (Fim do monopólio Taurus/CBC)	28.383
Criminalização da LGBTfobia	26.916
Anistia ao Sr. Dep. Jair Messias Bolsonaro	25.909
Aposentadoria para os portadores de Autismo.	25.442
Redução da Maioridade Penal para 15 anos em Crimes de Estupro e Assassinato/Art. 228	25.032
Um Salário para honrar a profissão do Nutricionista	23.515
Psicólogos com piso salarial de R$ 4.800,00 por 30 horas semanais.	23.221
Isenção de imposto de importação para mercadorias até USD 1000,00 por pessoas físicas	22.050
Criminalização do funk como crime de saúde pública a criança aos adolescentes e a família	21.985
Criminalização Da Apologia Ao Comunismo	21.892
Voto em cédulas de papel e urnas de lona para eleição de 2018	21.716
Você apoia que deveria haver concurso público para cargos políticos antes das eleições?	21.523
Piso salarial médico	21.415
Inclusão do Biomédico nos programas de Atenção à Saúde (ESF/NASF).	21.231
Nutricionistas com piso salarial de R$ 3.200,00 por 30 horas semanais.	21.167
Torna falsa acusação de estupro crime hediondo e inafiançável.	21.117

Fonte: Portal e-Cidadania[19].

Desta forma, evidencia-se que são ideias similares e com anseio por regulamentações similares, porém não idênticas, mas que poderiam estar unidas em uma única ideia com um maior número de apoios dando um maior peso no *ranking* para estas e também facilitando o trabalho da CDH que terá de analisar ambas.

A ferramenta Ideia Legislativa passou por mudanças no mês de abril/2018 por conta dos problemas com ideias idênticas competindo por apoio, assim foram criadas algumas regras na tentativa de sanar este problema. Nesta atualização uma mesma ideia não pode ser cadastrada por mais de um usuário e nem pelo mesmo usuário enquanto a ideia estiver em uma campanha para arrecadar apoios.

Este fato torna-se evidente a preocupação devido a quantidade de dados que repetidamente está sendo cadastrado no portal. Cabe destacar que no início desta dissertação, não havia nenhuma providência quanto ao tratamento de ideias similares. A iniciativa do portal é uma afirmação da importância e necessidade deste tipo de tratamento, indo de encontro aos diversos autores (SPENCER, 2012; POVEDA; WESTERSKI; IGLESIAS 2012) que defendem a utilização de técnicas de similaridade de texto.

[19] Disponível em: < https://www12.senado.leg.br/ecidadania/pesquisaideia > Acesso em jan. 2018.

Neste sentido, fora implantado uma ferramenta para detectar ideias idênticas. No entanto, o método encontra-se frágil, uma vez que, ainda encontra apenas ideias exatamente iguais de modo que se houver um ponto ou uma letra maiúscula diferente, não serão mais idênticas.

Este é um primeiro passo, mas ainda não explora as técnicas de processamento da linguagem natural e de cálculo de similaridade de texto, como apresentado nesta dissertação. De forma, pode-se destacar neste aspecto a relevância do pré-processamento dos dados textuais para realizar quaisquer procedimentos com dados não estruturados.

Esta nova função está sendo aplicada na etapa posterior de cadastro conforme pode ser visualizado na Figura 37 com o ciclo de vida das ideias atualizado.

Figura 37 - Ciclo de vida das ideias na ferramenta Ideia Legislativa

Fonte: do autor.

Os ganhos com implantação de ferramenta de similaridade, como apresentados nesta dissertação, trarão vantagens tanto para os usuários quanto para os especialistas que avaliam estas ideias. Do lado dos usuários a possiblidade de unir esforços para criar ideias mais robustas que atendam o maior número de usuários e a possibilidade de fortalecer a

rede para atraírem mais apoios. Esta situação corrobora com o autor Spancer (2012) que diz que quando um usuário encontra uma ideia interessante nos sistemas de dados ele também pode encontrar as pessoas que criaram estas ideias semelhantes.

Pelo lado dos especialistas de domínio que avaliam estas ideias o uso da similaridade abre a possibilidade de ideias com contextos mais elaborados por meio da criação em rede conforme evidenciado na literatura por Spancer (2012), assim reduzindo o tempo empenhado pelo relator da CDH para elaborar contextos para defesa de porquê tal ideia deva se tornar uma lei.

Deve-se levar em consideração os apontamentos dos autores que afirmam que os especialistas de revisão muitas vezes não têm tempo para examinar centenas de ideias e evidenciam que o uso de técnicas para agrupar ideias pode favorecer o trabalho dos especialistas de domínio no processo de examinar e avaliar as ideias coletadas a partir de comunidades *online*, e que agrupamentos por similaridade possibilitam ser analisadas um grande número de ideias em conjunto (POVEDA; WESTERSK; IGLESIAS, 2012; SPANCER, 2012).

Este é outro fator que merece destaque, pois pode-se ter ideias com o mesmo teor e objetivo que podem chegar a CDH num mesmo período ou com pequena diferença de tempo e serem designadas a relatores diferentes assim tomando tempo de análise para ideias com objetivos similares. O uso de tais técnicas pode evitar neste cenário que dois relatores diferentes analisem a mesma ideia, podendo assim realizarem outras tarefas.

Sobre o enriquecimento de ideias Perez *et al.* (2015) evidenciam que quanto maior for o conhecimento do contexto melhor a qualidade das ideias geradas. Neste sentido o cálculo de similaridade pode ser usado também como uma ferramenta para criar esta rede aproximando pessoas que tentam criar ideias semelhantes, resultando assim em ideias mais robustas, pois pode aproximar pessoas conforme explicitado por Spancer (2012), e estas podem conhecer bem os contextos propiciando discussões sobre estas temáticas.

Ao se analisar a maneira como é realizada a distribuição hoje dos projetos de leis, ocorre uma sobrecarrega a CDH, devido a quantidade de ideias que vem sendo cadastradas e vem atingindo a quantidade de apoios necessários para serem analisadas. Murah *et al*. (2013) destaca que quando há um aumento significativo no volume de dados nas bases de ideias, estas se tornam um desafio à gestão, e cria uma certa dependência de gestores com conhecimento específico para a tomada de decisão.

Os autores apontam ainda como alternativa o foco na criação de sistemas computacionais, com objetivo de facilitar a gestão do conteúdo, sendo mais rápida sua análise, classificação e agrupamento, para que estejam disponíveis no momento certo (MURAH *et al.*, 2013).

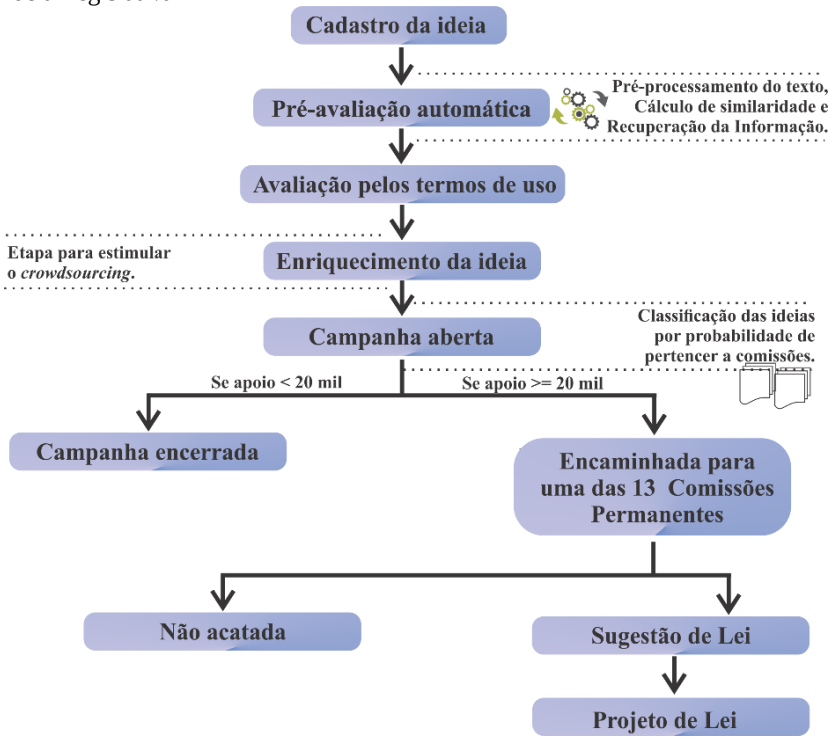
Neste sentido um sistema de classificação semi-automatizado aliado com ideias que contenham contextos mais robustos pode reduzir o trabalho realizado na justificativa e contextualização destas, e assim criar respaldo para se tornarem projetos de lei, hoje realizados pela CDH, de forma que se encaminhadas diretamente para a comissão fim poderiam resultar numa maior agilidade ao processo de análise da ideia, assim pode-se reduzir o tempo de ciclo de vida desta ideia e respectivamente o trabalho dos especialistas de domínio.

Quanto ao artefato de classificação, este pode representar vantagem perante aos demais pela maneira em que o senado trabalha quando se trata da criação de comissões mistas. Estas comissões mistas são criadas quando uma matéria pertence a duas comissões distintas. O uso do teorema de *Naive Bayes* facilitaria o trabalho da criação de comissões mistas. Uma vez que, como apresentado neste trabalho, o algoritmo estima a probabilidade de uma ideia pertencer primeiramente a todas comissões e a *posteriori* as classifica para uma determinada comissão permanente, retornando assim, as probabilidades de pertencimento para todas as classes.

Nota-se que perante o conjunto de treinamento, apresentado nos resultados expostos, ainda é possível observar um grande percentual de imprecisão, porém cabe evidenciar que segundo Junior (2007) quanto mais robusto for o conjunto de treinamento com uma maior quantidade de exemplos para treinamento maior índice de acertos apresentados pela técnica. Mesmo diante do fato da necessidade de tornar mais robusto o conjunto de treinamento, o artefato se mostrou de grande valia para o processo de categorização de texto devido as especificidades do cenário escolhido.

Perante a análise demonstrada neste tópico, foi possível construir uma sugestão de modelo para o ciclo de vida das ideias para o cenário acrescendo as técnicas adotadas pelo modelo apresentado nesta dissertação, assim justiçando a aplicabilidade deste em diversos contextos dentro das organizações. A proposição teve embasamento na literatura de gestão de ideias e no modelo proposto, suportado pelas técnicas de descoberta de conhecimento aplicado aos artefatos. A proposta de modelo para o ciclo de vida das ideias na ferramenta Ideia Legislativa é ilustrada na Figura 38.

Figura 38 - Proposta de novo modelo para ciclo de vida das ideias na ferramenta Ideia Legislativa



Fonte: do autor.

Conforme o modelo proposto embasado na literatura e nos resultados dos protótipos desenvolvidos tem-se 8 etapas previamente descritas na sequência:

1. **Cadastro das ideias:** nesta etapa o usuário irá seguir os passos pré-definidos na ferramenta Ideia Legislativa, porém deve ser evidenciado o poder de contextualização mais ricas das ideias e valor desta para as demais etapas e sucesso da ideia;
2. **Pré-avaliação automática:** ao confirmar o cadastro da ideia primeiramente seria realizado o pré-processamento dos dados textuais para assim realizar o cálculo de similaridade com outras ideias que estão no portal. Caso a ideia apresente radicais com a limiar de similaridade igual a 1, não seria permitido o cadastro desta no portal, caso

contrário seriam apresentadas as ideias com limiar acima 0,65 que estão em campanha e também agora as que estão em fase de enriquecimento;

3. **Avaliação pelos termos de uso:** esta etapa permaneceria a mesma seguindo os mesmos padrões adotados pelo portal;
4. **Enriquecimento da Ideia:** etapa na qual sugere-se adotar o meio de funcionamento de outras plataformas de gestão de ideias tal como Legoideias, habilitando a possibilidade de melhoria na ideia antes de iniciar a coleta de apoios criando uma rede de autores de ideias com o mesmo propósito. Isto seria viável por meio do cálculo de similaridade para encontra-los, além da criação do campo de comentários nestas ideias permitindo nesta fase edição da ideia;
5. **Campanha aberta:** nesta etapa continuaria com a campanha por quatro meses, após este período, seria bloqueado a edição da ideia tal como funciona hoje. Além disso, neste novo ciclo de vida, nesta fase deve ser realizada a classificação da ideia apontando as duas classes com maior probabilidade de se enquadrarem, desse modo, se adaptando as condições de comissões mista do senado.
6. **Encaminhada para uma das 13 comissões permanentes:** a ideia que conseguiu captar os 20 mil apoios deve ser encaminhada a comissão, na qual a probabilidade seja maior. Assim o relator designado pode criar uma comissão mista se achar necessário ou criar diretamente o processo de sugestão de lei e por meio do parecer levar a comissão para votação e assim transforma-la num projeto de lei.
7. **Projeto de lei:** final do ciclo da ideia dentro do portal, de modo que agora a ideia já possui um contexto robusto e previamente validada por várias etapas, dessa forma é encaminhada a mesa diretora do senado para ser votada podendo ou não se tornar uma lei.

Finalmente, sobre as alterações sugeridas cabe ainda destacar que a ferramenta Ideia Legislativa do portal e-Cidadania já possui dispositivos que favorecem o *crowdsourcing* porém ainda são passiveis de melhoria,

outrossim que os resultados poderiam ser mais promissores com a implantação de técnicas como foi evidenciado na literatura e demonstrados neste trabalho.

5 CONSIDERAÇÕES FINAIS

Apresentam-se neste capítulo as contribuições geradas pelo presente estudo, as limitações da pesquisa e as recomendações para trabalhos futuros.

5.1 CONSIDERAÇÕES FINAIS

A presente dissertação apresentou um modelo para suporte a Gestão de Ideias utilizando protótipos de mineração de dados em texto para reconhecer padrões em ideias. Por meio da busca sistemática constatou-se que a grande maioria dos trabalhos estão voltados a utilização de técnicas de web semântica para classificação de ideias.

Na revisão de literatura foi encontrado apenas dois trabalhos que aplicam técnicas de similaridade para a gestão de ideias, sendo do autor Spancer (2012) que utiliza apenas a métrica de Jaccard e os autores Paukkeri e Kotro (2009) que utilizam *k-means* e a métrica do cosseno. Contudo, ambos não apresentam em seu artigo se há um pré-processamento das ideias. Assim, identificou-se o *gap* de pesquisa para utilização de técnicas probabilísticas com aprendizado supervisionado aplicadas sob ideias.

Quanto aos objetivos específicos, consideram-se que foram devidamente alcançados:

1 - *Analisar métodos, técnicas e ferramentas utilizadas para tratamento de dados textuais na gestão de ideias*. Foi alcançado uma vez que foram identificados na literatura por meio de uma busca sistemática, os métodos, técnicas e ferramentas aplicadas para tratamento de dados textuais utilizados na Gestão de Ideias, no qual foi apresentado o resultado no capítulo 2 na seção 2.4.

2 - *Criar protótipos para reconhecimento de padrões com base nas técnicas de KDT evidenciadas no modelo*. A fim de atingir os objetivos, foram desenvolvidos dois protótipos para classificação de ideias um baseado na métrica da similaridade do cosseno e outro no método de categorização de textos usando o algoritmo de Naive Bayes. Entretanto conforme estudos realizados sobre métodos de mineração de dados em bases textuais foi identificado que os resultados são mais promissores quando se tem um melhor pré-processamento destes dados de modo que para ambos protótipos foram incluídos o processamento de linguagem natural para tratamento dos dados antes da aplicação das técnicas e assim atingindo este objetivo.

3 - *Verificar viabilidade do modelo proposto a partir de uma aplicação em um cenário*. Para atingir este objetivo foi escolhido o cenário da ferramenta Ideia Legislativa do portal e-Cidadania do Senado Federal brasileiro, onde foram testados os protótipos desenvolvidos e apresentados no capítulo 4.

Considera-se igualmente alcançado o objetivo geral *propor um modelo de reconhecimento de padrões em ideias amparado por técnicas de descoberta de conhecimento em texto*, todavia tornou-se necessário percorrer os objetivos específicos em prol de se criar sustentação para propor um modelo com base na literatura de forma que possa se adaptar à diversos cenários e que respeite suas particularidades de cada um destes cenários. O modelo é apresentado no final do item 4.5 que trata da análise e discussões dos artefatos.

Os principais resultados da pesquisa mostram que mediante ao uso de técnicas de descoberta de conhecimento é possível dar suporte a gestão de ideias, tendo em vista a quantidade de ideias geradas nestas plataformas web propiciados pelo advento da tecnologia. Destaca-se que técnicas de similaridade podem favorecer em diversos aspectos a gestão de ideias tanto no suporte a decisão para a avaliação e seleção de ideias quanto no fortalecimento da rede entre os colaboradores para criação de ideias mais robustas.

Evidencia-se também que o pré-processamento reduz a quantidade de termos e a dimensionalidade dos dados textuais de forma considerável. Assim, minimiza o esforço de processamento exigido para aplicação da métrica da similaridade do cosseno e do algoritmo de *Naive Bayes*, o que conduz a um aumento significativo nos índices de desempenho e respectivamente impactando nos acertos.

A contribuição científica desta dissertação foi cooperar no avanço de pesquisas de ferramentas da Engenharia do Conhecimento do campo de descoberta do conhecimento e reconhecimento de padrões que podem ser aplicadas ao contexto de Gestão de Ideias fornecendo assim suporte as suas atividades.

Do ponto de vista prático e aplicado, destaca-se como o detalhamento da aplicação do algoritmo de *Naive Bayes* se adapta a este modelo levando em consideração que no cenário analisado muitas ideias podem pertencer a suas comissões permanentes e o protótipo consegue pré-determinar a comissões podem ser rotuladas estas ideias.

A pesquisa apresentou algumas limitações, a primeira que cabe destacar é que o método criado é adaptado as regras da organização e que para uso em outros tipos de organização pode ser necessário adequações,

pois as particularidades de cada organização impactam diretamente na forma desenvolvem o processo de inovação.

O segundo aspecto é quanto ao esforço de processamento necessário ao se tratar de textos, conforme apresentado no tópico de coleta de dados a quantidade de *tokens* para campo ideias de todo o conjunto coletado é muito alto e mesmo tratados possuem uma alta dimensionalidade e de modo que ao aplicar técnicas de descoberta de conhecimento exige-se um alto poder de processamento para extração de resultados, dificultando a análise de dados utilizando todo o conjunto.

5.2 PERSPECTIVAS DE TRABALHOS FUTUROS

Para trabalhos futuros, vislumbram-se a evolução do algoritmo, usando a abordagem da web semântica e *folksonomias* juntamente com técnicas estatísticas para uma maior efetividade e eficiência do modelo. Para que tal resultado seja alcançado é indicado ainda o estudo de mais técnicas de descoberta de conhecimento que possam ser aplicadas em dados não estruturados e também formas de estruturá-los diminuindo a dimensionalidade e o esforço de processamento exigido e assim facilitar a tomada de decisões.

Sugere-se a aplicação das técnicas em outros contextos em prol de se construir um framework que possa ser submetido a diversos cenários, destacando ferramentas que podem auxiliar a construção de modelos, indo ao encontro das limitações evidentes de um modelo.

Como principal perspectiva percebida para continuidade da pesquisa, destaca-se a necessidade de estudos sobre aprendizado de máquina correlacionado o sucesso das ideias deste cenário e conteúdos em destaque em redes sociais, em busca de compreender como se dá o sucesso de ideias populares e assim poder criar sistemas para a geração de ideias.

REFERÊNCIAS

ALMEIDA, M. B.; BAX, M. P. Uma visão geral sobre ontologias: pesquisa sobre definições, tipos, aplicações, métodos de avaliação e de construção. **Ci. Inf., Brasília**, v. 32, n. 3, Dec. 2003.

ANGÉNIOL, S. *et al*. Supporting cost saving ideas reuse with an ontology based tool. **ASME International Design Engineering Technical Conferences and Computers and Information In Engineering Conference**, DETC2006, 2006, Philadelphia, PA.

ARAMPATZIS, A. T.; VAN DER WEIDE, T. P.; KOSTER, C. H. A.; VAN BOMMEL, P. **Linguistically-motivated Information Retrieval**. Encyclopedia of Library and Information Science, V.69, 2000. p.201-222.

BAEZA-YATES, R.; RIBEIRO-NETO, B. **Modern Information Retrieval**. New York: ACM Press, 1999. 513 p.

BAILEY, B. P.; HORVITZ, E. What's your idea? A case study of a grassroots innovation pipeline within a large software company. 28th **Annual CHI Conference on Human Factors in Computing Systems**, CHI.2010, 2010, Atlanta, GA. p.2065-2074.

BANERJEE, C. The Human Factor: The Fundamental Driver of Innovation. In: DUTTA, S.; LANVIN, B.; WUNSCH-VINCENT, S. (eds.). **The Global Innovation Index2014**: The Human Factor in Innovation. Cornell University, INSEAD, and WIPO: Fontainebleau, Ithaca, and Geneva, 2014.

BARBIERI, J. C. **Organizações inovadoras:** Estudos e casos brasileiros. 2. ed. Rio de Janeiro: FGV, 2004.

BAREGHEH, A.; ROWLEY, J.; SAMBROOK, S. Towards a multidisciplinary definition of innovation. **Management Decision**, v.47, n. 8, p. 1323-1339, 2009.

BESSANT, J. *et al.* Managing innovation beyond the steady state. **Technovation,** Amsterdam, vol. 25, nº. 12, p. 1366-1376, 2005.

BESSANT, J.; TIDD, J. **Inovação e empreendedorismo: administração.** Bookman Editora, 2009.

BETTONI, M. *et al*. Idea management by role based networked learning. **11th European Conference on Knowledge Management, ECKM 2010**, Famalicao. 2010. ISSN 20488963. p. 107-116.

BETTONI, M.; BERNHARD, W.; BITTEL, N. Collaborative solutions quick&clean: The SFM method. **14th European Conference on Knowledge Management, ECKM 2013**, 2013, Kaunas. p.44-51.

BICK, E. Structural Lexical Heuristics in the Automatic Analysis of Portuguese. **11th Nordic Conference on Computational Linguistics**, Copenhagen, 1998. p.44-56.

BJÖRK, J.; BOCCARDELLI, P.; MAGNUSSON, M. G. Ideation capabilities for continuous innovation. **Creativity & Innovation Management, Malden**, v. 19, n. 4, p. 385-396, 2010.

BOD, R. **Enriching Linguistics with Statistics: Performance Models of Natural Language**. Tese de doutorado. Institute for Logic, Language and Computation (ILLC), Universidade de Amsterdã, 1995.

BOTHOS, E.; APOSTOLOU, D.; MENTZAS, G. Collective intelligence with web based information aggregation markets**:** The role of market facilitation in idea management. **Expert Systems with Applications**, Amsterdam, vol. 39, nº. 1, p. 1333-1345, 2012.

BRAGA, M. C. G. **Diretrizes para o Design de Mídias em Realidade Aumentada: Situar a Aprendizagem Colaborativa Online.** 2012. 243 f. Tese (Doutorado) - Curso de Engenharia e Gestão do Conhecimento., Programa de Pós-graduação em Engenharia e Gestão do Conhecimento, Universidade Federal de Santa Catarina, Florianópolis, 2012.

BRASIL. **Senado Federal - Regimento Interno.** Disponível em: < https://www25.senado.leg.br/web/atividade/regimento-interno#/>. Acesso em: 08 jan. 2018.

BRASIL. **Senado Federal. Sobre O Portal E-Cidadania.** Disponível em: <https://www12.senado.leg.br/ecidadania/sobre>. Acesso em: 08 jan. 2018.

BUNGE, M. **Emergence and convergence: Qualitative novelty and theunity of knowledge.** University of Toronto Press, 2003.

BURRELL, G.; MORGAN, G. **Sociological paradigms and organisational analysis.** London: Heinemann, 1979.

ÇAĞDAŞ, V.; STUBKJÆR, E. Design research for cadastral systems**.** **Computers, Environment and Urban Systems**, v. 35, p. 77-87, 2011.

CARLSSON, G. Topological pattern recognition for point cloud data**.** **Acta Numerica**, v. 23, p. 289-368, 2014.

CECI, F. **Um Modelo Semi-automático para a Construção e Manutenção de Ontologias a partir de bases de documentos não estruturados**. Dissertação, 2015. 177 f. 211 Dissertação (Mestrado) - Curso de Engenharia e Gestão do Conhecimento, Programa de Pós-graduação em Engenharia e Gestão do Conhecimento, Universidade Federal de Santa Catarina, Florianópolis, 2015.

CHAPMAN, S. **SimMetrics**. 2009 Disponível em: < http://www.dcs.shef.ac.uk/ ~sam/stringmetrics.html>. Acesso em: 08 jan. 2018.

CHESBROUGH, H. W. Open Innovation: the new imperative for creating and profiting from technology. **Harvard Business School Press**, 2003.

CHIBÁS, F. O.; PANTALEÓN, E. M.; ROCHA, T. A. **Gestão da Inovação e da Criatividade Na Atualidade**. 2013, v. 3, p. 12, 2013-08-02 2013. Disponível em:<http://www2.ifrn.edu.br/ojs/index.php/HOLOS/article/view/1082/678 >. Acesso em 05 de janeiro de 2018.

CLARK, K. B.; WHEELWRIGHT, S. C. **Managing new product andprocess development: text and cases**. New York: The Free Press, 1993.

CONDE; M. V. F.; ARAÚJO-JORGE, T. C. Modelos e concepções de inovação: a transição de paradigmas, a reforma da C&T brasileira e as concepções de gestores de uma instituição pública de pesquisa em saúde. **Ciência & Saúde Coletiva.** 8(3):727-741, 2003.

COOPER, R. G.; EDGETT, S. J. Ideation for product innovation: what are the best methods? ***PDMA* Visions**, v. 32, n. 1, p. 12-17, 2008.

CROSSAN, M. M; APAYDIN, M. A Multi-Dimensional Framework of Organizational Innovation: A Systematic Review of the Literature. **Journal Of Management Studies**. [s.i], p. 1154-1191. set. 2010.

CUPANI, A. **Filosofia da Tecnologia: um convite**. Florianópolis: Editora da UFSC, 2011.

CUPANI, A. La peculiaridad del conocimiento tecnológico. **Scientia Studia**, São Paulo, v.4, n.3, p.353-71, 2006.

DOROW, P. F. **Processo de Seleção de Ideias em Empresas Inovadoras.** 2013. 158 f. Dissertação (Mestrado) - Curso de Engenharia

e Gestão do Conhecimento, Centro Tecnológico, Universidade Federal de Santa Catarina, Florianópolis, 2013.

DRESCH, A.; LACERDA, D. P.; JÚNIOR, J. A. V. A. **Design Science Research: Método de Pesquisa para Avanço da Ciência e Tecnologia**. Bookman Editora, 2015.

DUDA, R.; HART, P.; STORK, D. Pattern classification. **Pattern Classification and Scene Analysis:** Pattern Classification. Wiley, 2001.

DUMAIS, S. T.; PLATT, J.; HECKERMAN, D.; SAHAMI, M. Inductive learning algorithms and representations for text categorization**.** In Proceedings of CIKM-98, **7th ACM International Conference on Information and Knowledge Managemen**t. Bethesda, MD, 1998.

EL BASSITI, L.; AJHOUN, R. Semantic-Based Framework for Innovation Management. In: VIVAS, C.; SEQUEIRA, P. (Ed.). Proceedings of the 15th European Conference on Knowledge **Management. Nr Reading: Acad Conferences Ltd**, 2014. p.1173-1182.

ELERUD-TRYDE, A.; HOOGE, S. Beyond the generation of ideas: Virtual idea campaigns to spur creativity and innovation**. Creativity and Innovation Management**, v. 23, n. 3, p. 290-302, 2014.

EVERITT, B. S. **A handbook of statistical analyses using S-Plus**. CRC Press, p. 376. United States 2001.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From Data Mining to Knowledge Discovery in Databases. **AL Magazine**, Califórnia, v. 17, n. 3, p.37-54, 1996.

FELDMAN, R.; DAGAN, I. Knowledge discovery in textual databases.**Knowledge Discovery and Data Mining**, p. 112–117, 1995.

FENN, J.; LEHONG, H. **Hype cycle for emerging technologies.** Gartner, July, 2011.

FERNANDES, R. P.; GROSSE, I. R.; KRISHNAMURTY, S.; WITHERELL, P.; WILEDEN, J. C. Semantic methods supporting engineering design innovation. **Advanced Engineering Informatics**. v. 25, n. 2, p. 185-192, 2011.

FREEMAN, C. **La Teoría Económica de La Innovación Industrial.** Madrid: Alianza, 1975.

FUNG, G. A comprehensive overview of basic *cluster*ing algorithms.**19th International Conference, CN 2012,** Szczyrk, Poland p. 01: 37 2001.

FURUI, S. Fifty years of progress in speech and speaker recognition. **The Journal of the Acoustical Society of America, 16(4)**, 2497-2498. 2004.

GAMALLO, P.; AGUSTINI, A.; LOPES, G. P. Using Co-Composition for Acquiring Syntactic and Semantic Subcategorisation.**Acl Special Interest Group On The Lexicon (siglex)**, Philadelphia, p.34-41, 2002

GEISSER, S. The predictive sample reuse method with applications. **Jornal of the American Statistical Association**, 70(350), 1975.

GIBSON, R.; SKARZYNSKI, P. **Inovação: prioridade nº 1: o caminho para a transformação nas organizaçõe**s. Elsevier: Rio de Janeiro, 2008.

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

GÓMEZ-PÉREZ, A. Ontological engineering**:** A state of the art.Expert Update: **Knowledge Based Systems and Applied Artificial Intelligence**, v. 2, n. 3, p. 33-43, 1999.

GONZALEZ, M.; LIMA, V. de. The PUCRS NLP-group participation in CLEF2006**:** Information retrieval based on linguistic resources. In: PETERS, C. *et al.* (Eds.). Evaluation of Multilingual and Multi-modal Information Retrieval . **Springer Berli**n / Heidelberg, 2007, (Lecture Notes in Computer Science, v. 4730). p. 66-73, 2007.

GRAVES, A. Comparative Trends in Automotive Research and Development. DRC Discussion Paper .No. 54. **Science Policy Research Unit**, Sussex University, Brighton, Sussex, 1987.

GRIMMER, R., ESKOFIER, B., SCHLARB, H. & HORNEGGER, J. Comparison and classification of 3d objects surface point clouds on the example of feet. **Machine Vision and Applications, Article in press**. 2009.

GRUBER, T. Ontology. In: Liu, L., Zsu, M. T. (eds.) **Encyclopedia of Database Systems**, pp. **1963–1965**. Springer US, New York. 2009. http://dx.doi.org/10.1007/ 978-0-387-39940-9_1318

GRUBER, T. R. Toward principles for the design of ontologies used for knowledge sharing? **International journal of human-computer studies.** v. 43, n. 5, p. 907 -928, 1995.

GUARINO, N. Formal Ontology in Information Systems: Proceedings. **1st International Conference** June 6-8, 1998, Trento, Italy. 1998.

GUARINO, N. Formal ontology, conceptual analysis and knowledge representation. **International journal of human-computer studies**, v. 43, n. 5-6, p. 625-640, Italy.1995.

GUPTA, A. S. K.; WILEMAN, D. L. Accelerating the Development of Technology-based New Product**. California Management Review**, Vol. 32 No. 2, Winter, pp. 24-44, 1990.

HAIR, J. F. *et al.* **Multivariate Data Analysis**. 7. ed. Pearson Prentice Hal, 2010. 593 p.

HANSEN, P.; JAUMARD, B. Cluster analysis and mathematical programming. **Mathematical programming**, v. 79, n. 1-3, p. 191-215, 1997. ISSN 0025-5610.

HARTIGAN, J. A. **Clustering algorithms**. New York: Wiley. 1975.

HERSTATT, C. *et al.* "Fuzzy front end" practices in innovating Japanese companies. **International Journal of Innovation and Technology Management,** v. 3, n. 01, p. 43-60, 2006. ISSN 0219-8770.

HORTON, G.; GOERS, J. Mining Hidden Profiles in the Collaborative Evaluation of Raw Ideas**. System Sciences** (HICSS)**.** 2014.

HRASTINSKI, S. *et al.* **A review of technologies for open innovation**: Characteristics and future trends. **43rd Annual Hawaii International Conference on System Sciences,** HICSS-43, 2010, Koloa, Kauai, HI.

JAIN, A. K.; DUBES, R. C. **Algorithms for clustering data.** Prentice Hal PTR.1988.

JAIN, A. K.; MURTY, M. N.; FLYNN, P. J. Data clustering**:** a review. **Acm Computing Surveys**, v. 31 nº 3, p. 264-323, 1999.

JAIN, A. K.; DUIN, R. P. W.; MAO, J. Statistical pattern recognition: A review. **IEEE Trans. Pattern Anal. Mach. Intell**., 22(1):4–37. 2000.

JANSEN, M. Noise reduction by wavelet thresholding. **Springer Science & Business Media**, New York 2012.

JOACHIMS, T. **A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization**. Carnegie-mellon univ pittsburgh pa dept of computer science, 1996.

JOHNSON, R. A.; WICHERN, D. W. **Applied multivariate correspondence analysis.** Pearson, New York 2007.

JUNIOR, J. R. C. **Desenvolvimento de uma Metodologia para Mineração de Textos.** Departamento de Engenharia Elétrica, Pontífica Universidade Católica do Rio de Janeiro, 2007.

KAINULAINEN, J. **Clustering algorithms: basics and visualization**. Helsinki University of Technology, Laboratory of Computer and Information Science, 2002.

KAMPA, J. R.; CZIULIK, C. **Discussion on generic innovation models and new product opportunity identification.** In: VERHAGEN, W. J. C.;PERUZZINI, M., et al, 23rd ISPE Inc. International Conference on Transdisciplinary Engineering, TE 2016, 2016, IOS Press BV. p.67-76.

KANTER, R. M. Innovation-The Only Hope for Times Ahead? **Sloan management review**, v. 25, n. 4, p. 51, 1984.

KEMPE, N. *et al.* An Optimal Algorithm for Raw Idea Selection under Uncertainty. System Science (HICSS), 2011. **45th, Hawaii International Conference on,** 2012, 4-7 Jan. 2012. p.237-246.

KOEN, P. A. *et al.* **Fuzzy front end: effective methods, tools, and techniques**. Wiley, New York, NY, 2002.

KOEN, P. A.; BERTELS, H. M.J.; KLEINSCHMIDT, E. J. Managing the Front End of Innovation-Part II: Results from a Three-Year Study: EffectiveFront-End activities were found to be significantly different for incremental and radical projects. **Research-Technology Management**. V. 57, n.3, p. 25-35, 2014.

KOEN, P. *et al.* Providing clarity and a common language to the "fuzzy front end". **Research-Technology Management,** v. 44, n. 2, p. 46-55, 2001.

KOPRINSKA, I., POON, J., CLARK, J. & CHAN, J. **Learning to classify email. Information Sciences**, 177(10), 2167-2187. 2007.

KUECHLER, W. L. Business applications of unstructured text**. Communications of ACM**, vol. 50, n. 10, p. 86-93, 2007.

LACERDA, D. P. *et al.* **Design Science Research:** método de pesquisa para a engenharia de produção. Gest. Prod., São Carlos, v. 20, n.4, p. 741-761, 2013.

LATTIN, J. M.; DOUGLAS C.; PAUL E. G**. Análise de dados multivariados**. São Paulo: Cengage Learning, 2011. 455 p.

LI, X; LI, L; CHEN, Z. Toward extenics-based innovation model on intelligent knowledge management. **Annals of Data Science**, v. 1, n. 1, p. 127-148, 2014.

LI, Y. *et al*. (2006). Sentence similarity based on semantic nets and corpus statistics**. IEEE Transactions on Knowledge and Data Engineering**. v. 18, n. 8, p. 1138-1150. ISSN 10414347 (ISSN).

LIDDY, E. D. **Natural language processing.** In Encyclopedia of Library and Information Science, 2nd Ed. NY. Marcel Decker, Inc. 2001.

LINDERGAARD, S. **A revolução da inovação aberta:** a chave da nova competitividade nos negócios. São Paulo: Évora, 2011.

LÖWER, M.; HELLER, J. E. PLM reference model for integrated idea and innovation management. IFIP Advances in Information and Communication Technology: **Springer**.New York. LLC. 442**:** 257-266 p. 2014.

LULA, P.; PALIWODA-PĘKOSZ, G. An Ontology-Based Cluster Analysis Framework. **Proceedings Of The First International Workshop On Ontology-supported Business Intelligence**, New York, p.1-6, 2008.

LUNING, X; PENGZHU, Z. A three phase idea selection approach for team creation**. International Seminar On Business And Information Managemen**t, Isbim 2008. Wuhan, p. 326-329. 2009.

MAGNUSSON, P. R.; NETZ, J; WÄSTLUND, E. Exploring holistic intuitive idea screening in the light of formal criteria**, Technovation**, vol. 34, n. 5–6, May–June 2014, Pages 315-326, 2014.

MAIA, L. C. G.; ROCHA, R. S. Uso de sintagmas nominais na classificação automática de documentos eletrônicos**. Perspectivas em Ciência da Informação,** v. 15, n. 1, p.154-172, 2010.

MAIA, L. C. G; SOUZA, R. R. Medidas de similaridade em documentos eletrônicos. **IX ENANCIB- Diversidade cultural e políticas de informação**. Universidade de São Paulo. São Paulo.(2008).

MARCH, S. T.; SMITH, G. F. Design and natural science research on information technology**. Decision Support Systems**, v. 15, n. 4, p. 251–266, 1995.

MARTINS, J. **Classificação de páginas na internet.** Trabalho de Conclusão (Mestrado). Instituto de Ciências Matemáticas e de Computação. USP. São Carlos, 2003.

MCCOMB, D. **Semantics in business systems:** the savvy manager's guide: the discipline underlying web services, business rules, and the semantic web. Morgan Kaufmann, 2004.

MIGUEZ, V. B. **Uma Abordagem de Geração de Ideias para o Processo de Inovação.** 2012. 187 f. Dissertação (Mestrado) - Curso de Engenharia e Gestão do Conhecimento, Centro Tecnológico, Universidade Federal de Santa Catarina, Florianópolis, 2012.

MIKELSONE, E.; LIELĀ, E. **Discussion On the Terms Of Idea Managament And Idea Managament Systems.** DISKUSIJA DĖL IDĖJŲ VALDYMO IR IDĖJŲ VALDYMO SISTEMŲ TERMINŲ., n. 17, p. 97-111, 2015. ISSN 20299370. Disponível em: < http://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=1103592 50&lang=pt-br&site=eds-live&authtype=cookie,ip,uid >.

MINAYO, M. C. S. **O desafio do conhecimento. Pesquisa qualitativa em saúde**. São Paulo: Hucitec, 2007.

MURAH, M. Z. *et al.* Kacang cerdik**:** A conceptual design of an idea management system. **International Education Studies**, v. 6, n. 6, p. 178-184, 2013. ISSN 19139020

OECD – Organização De Cooperação E Desenvolvimento Econômico. Oslo Manual**:** Guide-line for collecting and interpreting innovation data, 2005. 3. **Ed. European Comission:** OECD. Disponível em: www.oecd.org. Acesso em: Dez. 2017. Acesso em 05 dezembro de 2017.

OLIVEIRA, F. A. D. de. Processamento de linguagem natural: princípios básicos e a implementação de um analisador sintático de sentenças da

língua portuguesa**.** In: **Revista de Ciência da Informação**. Rio de Janeiro. n. 5. Maio 2002.

OLIVEIRA, G.; MENDONÇA, M. ExperText**:** Uma Ferramenta de Combinação de Múltiplos Classificadores Naive Bayes. Anales de la 4ª Jornadas **Iberoamericanas de Ingeniería de Software e Ingeniería de Conocimiento**. Madrid, v. 1, p. 317-32, 2004.

ORENGO, V. M.; HUYCK, C. R. A Stemming Algorithim for The Portuguese Language**.** In: **Proceedings of the SPIRE Conference**. Laguna de San Raphael: [s.n.], 2001, p. 13-15.

PACHECO, R. C. D. S. **Dados e Governo Abertos na Sociedade do Conhecimento.** Linked Open Data - Brasil. Florianópolis - SC 2014.

PAUKKERI, M. S.; KOTRO, T. Framework for analyzing and clustering short message database of ideas**. 9th International Conference on Knowledge Management and Knowledge Technologies**, I-KNOW 2009 and 5th International Conference on Semantic Systems, I-SEMANTICS 2009, 2009, Graz. p. 239-247.

PEFFERS, K. *et al.* A design science research methodology for information systems research**. Journal of management information systems**, v. 24, n. 3, p. 45-77, 2007. ISSN 0742-1222.

PEREZ, A. *et al*. Innoweb**:** Gathering the context information of innovation processes with a collaborative social network platform. **19th International Conference on Engineering, Technology and Innovation, ICE 2013 and IEEE International Technology Management Conference**, ITMC 2013, 2015.

PEREZ, A.; LARRINAGA, F.; CURRY, E. The role of linked data and semantic-technologies for sustainability idea management. **1th International Conference on Software Engineering and Formal Methods**, SEFM 2013 - Collocated Workshops: BEAT2, WS-FMDS, FM-RAIL-Bok, MoKMaSD, and OpenCert. Madrid: Springer Verlag. 8368 LNCS**:** 306-312 p. 2014.

POLI, R.; OBRST, L. The interplay between ontology as categorial analysis and ontology as technology. In: **Theory and applications of ontology:** Computer applications. Springer Netherlands, 2010. p. 1-26
.

POVEDA, G.; WESTERSKI, A.; IGLESIAS, C. A. Application of semantic search. Idea Management Systems. **International Conference for Internet Technology And Secured Transactions**, 2012, vol., no., p.230 - 236, 10-12 Dec. 2012

PRADA, C. A. **Proposta de Modelo para o Gerenciamento de Portfólio de Inovação: Modelagem do Conhecimento na Geração de Ideias.** 2009. 161 f. Dissertação (Mestrado) - Curso de Engenharia e Gestão do Conhecimento, Centro Tecnológico, Universidade Federal de Santa Catarina, Florianópolis, 2009.

PREEZ, N. D.; LOUW, L. A framework for managing the innovation process**. Management of Engineering & Technology***, 2008. PICMET 2008*. Portland International Conference on. IEEE, 2008.p. 546-558.

QUINTANE, E.; CASSELMAN, R. M.; REICHE, B. S.; NYLUND, P. A. Innovation as a knowledge-based outcome**. Journal of Knowledge Management**, v. 15, n. 6, p. 928-47, 2011.

REPKO, A. F. **Interdisciplinary Research: Process and Theory**. SAGE Publications, 2011. ISBN 9781412988773. Disponível em: < https://books.google.com.br/books?id=I0PiSIgmp38C >. Acesso em 08 de dezembro de 2017.

ROCHADEL, W. **Identificação de Critérios para Avaliação de Ideias: Um Método Utilizando Folksonomias.** 2016. 177 f. Dissertação (Mestrado) - Curso de Engenharia e Gestão do Conhecimento, Programa de Pós-graduação em Engenharia e Gestão do Conhecimento, Universidade Federal de Santa Catarina, Florianópolis, 2016.

ROTHWELL, R. Towards the fifth generation innovation process**. International Marketing Review**, v. 11, n. 1, 1994.

SCHNEIDER, M. **Processamento de linguagem natural (PLN).** Master´s thesis, PUC- Campinas, 3.2001.

SCHUMPETER, J. A. The creative response in economic history. The journal of economic history, 7(2), 149-159. (Reprinted in Joseph

Schumpeter (Edited by Richard Swedberg, 1991, **The Economics and Socialism of Capitalism, Princeton University Press**, Princeton, New Jersey)

SEBASTIANI, F. Machine learning in automated text categorization**. ACM computing surveys** (CSUR), v. 34, n. 1, p. 1-47, 2002.

SÉRGIO, M. C.; DE SOUZA, J. A.; GONCALVES, A. L. Idea identification model to support decision making**. IEEE Latin America Transactions**, v. 15, n. 5, 2017. ISSN 15480992 (ISSN).

SÉRGIO, M. C. **Um Modelo Baseado em Ontologia e Análise de Agrupamento para Suporte à Gestão de Ideias.** 128 f. Dissertação (Mestrado) - Curso de Engenharia e Gestão do Conhecimento, Centro Tecnológico, Universidade Federal de Santa Catarina, Florianópolis, 2016.

SHIMA, W.; ESTEVÃO, J. S. B. Uma análise bibliométrica da produção acadêmica sobre o tema inovação (Innovation Studies) em língua portuguesa**. Blucher Engineering Proceedings**, v. 3, n. 4, p. 1445-1465, 2016.

SILVA, E. R. G.; ROVER, A. J. **O** Processo de descoberta do conhecimento como suporte à análise criminal: minerando dados da Segurança Pública de Santa Catarina. **Anais da International Conference on Information Systems and Technology Managemen**t. São Paulo: FEA, 2011. v. 8.

SILVA, D. C. **Uma Arquitetura de Business Intelligence para Processamento Analítico baseado em Tecnologias Semânticas e em Linguagem Natural.** 2011. 163 f. Dissertação (Mestrado) - Curso de Engenharia e Gestão do Conhecimento, Centro Tecnológico, Universidade Federal de Santa Catarina, Florianópolis, 2011.

SIMON, H. A. The Sciences of the Artificial. 3rd ed.Cambridge/Massachussets: **MIT Press**, 1996 [1961].

SINT, R. *et al*. Ideator - A collaborative enterprise idea management tool powered by KiWi? , 5th Workshop on Semantic Wikis - Linking Data and People, - **7th Extended Semantic Web Conference,** ESWC 2010 , Hersonissos, Heraklion, Crete. p.41-48.

SMITH, P. G.; REINERTSEN, D. G. **Developing products in half the time**. New York: Van Nostrand Reinhold, 1991.

SPENCER, R. W. The size and shape of "idea space". **International Journal of Innovation Science**, 2012. Vol. 4 Issue: 2, pp.71-76,_.

STEVANOVIć, M; MARJANOVIć, D; STORGA, M. Decision Support System For Idea Selection**. International Design Conference - Desig**n 2012. Dubrovnik, p. 1951-1960. 21 maio 2012.

STUDER, R.; BENJAMINS, V. R.; FENSEL, D. Knowledge engineering: principles and methods. **Data & knowledge engineering**, v. 25, n. 1, p. 161-197, 1998.

TAN, A.-H. ***Text mining:*** The state of the art and the challenges**.** In: Proceedings of the Pacific Asia Conference on Knowledge Discovery and Data Mining – **PAKDD'99 Workshop on Knowledge Discovery from Advanced Databases**, Beijing, p. 65–70, 1999.

THEODORIDIS, S. & KOUTROUMBAS, K. **Pattern recognition**. Elsevier Academic Press. Amsterdam, 2009.

TEZA, P. **Front end da Inovação: proposta de um modelo conceitual.**. 147 f. Dissertação (Mestrado) - Curso de Engenharia de Produção, Centro Tecnológico, Universidade Federal de Santa Catarina, Florianópolis, 2012.

TEZA, P. **Fatores Determinantes da Adoção de Métodos, Técnicas e Ferramentas para Inovação.** Tese.. Curso de Engenharia e Gestão do Conhecimento, Centro Tecnológico, Universidade Federal de Santa Catarina, Florianópolis, 2018.

TIDD, J.; BESSANT, J. **Gestão da inovação-5**. Bookman Editora, 2015.

Tou, J. T., Gonzalez, R. C. **Pattern Recognition Principles.** Addison-Wesley Publishing Company. Massachusetts, 1981.

VALDATI, A. B. **Processo de Seleção de Ideias em Empresas Inovadoras.** 2017. 2016 f. Dissertação (Mestrado) - Curso de Engenharia e Gestão do Conhecimento, Centro Tecnológico, Universidade Federal de Santa Catarina, Florianópolis, 2017.

VAN AKEN, J. E. Management Research Based on the Paradigm of the Design Sciences: The Quest for Field- Tested and Grounded

Technological Rules. **Journal of Management Studies**, v. 41, n. 2, p. 219-246, 2004.

VANDENBOSCH, B.; SAATCIOGLU, A.; FAY, S. Idea Management**:** A Systemic View. **Journal of Management Studies**, v. 43, n. 2, p. 259-288, 2006. ISSN 1467-6486.

VON ALAN, R. H. *et al*. Design science in information systems research. **MIS quarterly**, v. 28, n. 1, p. 75-105, 2004.

WESTERSKI, A.; DALAMAGAS, T.; IGLESIAS, C. A.Classifying and comparing community innovation. **Idea Management Systems**, **Decision Support Systems,** 2013.

WESTERSKI, A.; IGLESIAS, C. A. Exploiting Structured Linked Data in Enterprise Knowledge Management Systems: An Idea Management Case Study. EDOCW. p. 395-403, **IEEE Computer Society**, 2011.

WESTERSKI, A.; IGLESIAS, C. A.; GARCIA, J. E. Idea relationship analysis in open innovation crowdsourcing systems. In: **Collaborative Computing:** Networking, Applications and Worksharing. **8th International Conference on IEEE**, 2012. p. 289-296.

WESTERSKI, A.; IGLESIAS, C. A.; RICO, F. T. A Model for Integration and Interlinking of Idea Management Systems**. 4th Metadata and Semantics Research Conference** (MTSR 2010), Alcalá de Henares, Spain, 2010.

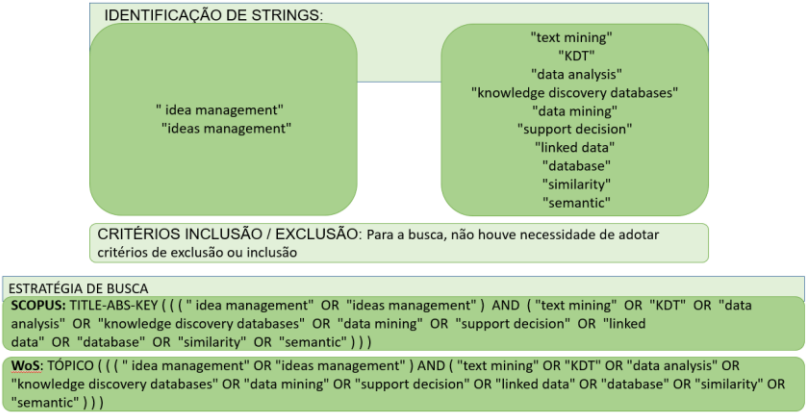
WESTERSKI, A.; IGLESIAS, C. A.; RICO, F. T. Linked opinions: Describing sentiments on the structured web of data. 4th International Workshop on Social Data on the Web. **In Conjunction with the International Semantic Web Conference**, ISWC 2011, 2011, Bonn.

WIVES, L. K. **Utilizando conceitos como descritores de textos para o processo de identificação de conglomerados (*clustering*) de documentos** – Tese (doutorado) – Universidade Federal do Rio Grande do Sul. Programa de Pós-graduação em Computação, Porto Alegre, BR – RS, Brasil, 2004.

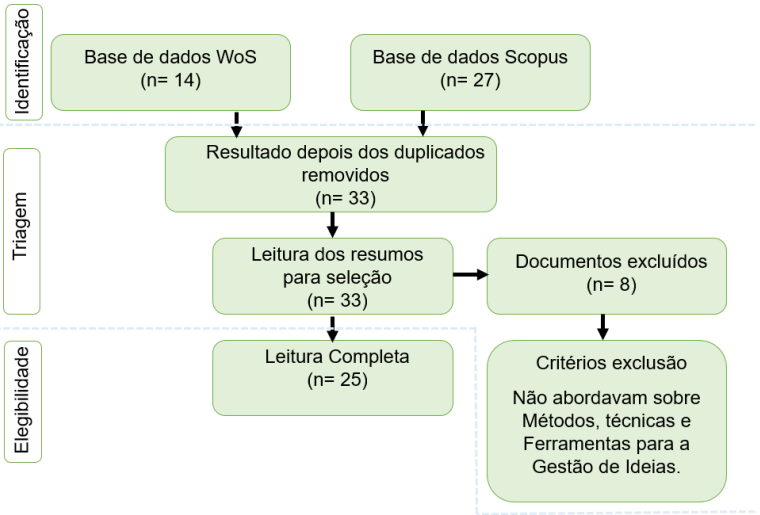
XIE, L., ZHANG, P. Idea Management System for Team Creation. **Journal of Software**, North America, 5, nov. 2010.

APÊNDICE A – Protocolo da busca sistemática

A partir da questão e dos objetivos da pesquisa, foram definidas as strings e então, formulou-se as estratégias de busca para as bases de dados, conforme figura XX abaixo.



Após a realização da busca nas bases de dados SCOPUS e WoS, que ocorreram no dia 28 de janeiro de 2018, realizou-se a triagem e elegibilidades dos artigos, conforme figura XX abaixo.



ANEXO A – Objetivos das Comissões permanentes

Tabela 1 – Objetivos das Comissões permanentes.

Comissão	Objetivo
I - Comissão de Assuntos Econômicos (CAE), com 27 membros;	I – aspecto econômico e financeiro de qualquer matéria que lhe seja submetida por despacho do Presidente, por deliberação do Plenário, ou por consulta de comissão, e, ainda, quando, em virtude desses aspectos, houver recurso de decisão terminativa de comissão para o Plenário; II – (Revogado.) III – problemas econômicos do País, política de crédito, câmbio, seguro e transferência de valores, comércio exterior e interestadual, sistema monetário, bancário e de medidas, títulos e garantia dos metais, sistema de poupança, consórcio e sorteio e propaganda comercial; IV – tributos, tarifas, empréstimos compulsórios, finanças públicas, normas gerais sobre direito tributário, financeiro e econômico; orçamento, juntas comerciais, conflitos de competência em matéria tributária entre a União, os Estados, o Distrito Federal e os Municípios, dívida pública e fiscalização das instituições financeiras;
II - Comissão de Assuntos Sociais (CAS), com 21 membros;	I – relações de trabalho, organização do sistema nacional de emprego e condição para o exercício de profissões, seguridade social, previdência social, população indígena e assistência social; II – proteção e defesa da saúde, condições e requisitos para remoção de órgãos, tecidos e substâncias humanas para fins de transplante, pesquisa, tratamento e coleta de sangue humano e seus derivados, produção, controle e fiscalização de medicamentos, saneamento, inspeção e fiscalização de alimentos e competência do Sistema Único de Saúde; III – (Revogado.) IV – outros assuntos correlatos. (NR)
III - Comissão de Constituição, Justiça e Cidadania (CCJ), com 27 membros;	I – opinar sobre a constitucionalidade, juridicidade e regimentalidade das matérias que lhe forem submetidas por deliberação do Plenário, por despacho da Presidência, por consulta de qualquer comissão, ou quando em virtude desses aspectos houver recurso de decisão terminativa de comissão para o Plenário; II – ressalvadas as atribuições das demais comissões, emitir parecer, quanto ao mérito, sobre as matérias de competência da União, especialmente as seguintes: a) criação de Estado e

	Territórios, incorporação ou desmembramento de áreas a eles pertencentes; b) estado de defesa, estado de sítio e intervenção federal (Const., art.49, IV), requisições civis e anistia; c) segurança pública, corpos de bombeiros militares, polícia, inclusive marítima, aérea de fronteiras, rodoviária e ferroviária; d) direito civil, comercial, penal, processual, eleitoral, aeronáutico, espacial, marítimo e penitenciário; e) uso dos símbolos nacionais, nacionalidade, cidadania e naturalização, extradição e expulsão de estrangeiros, emigração e imigração; f) órgãos do serviço público civil da União e servidores da administração direta e indireta do Poder Judiciário, do Ministério Público e dos Territórios; g) normas gerais de licitação e contratação, em todas as modalidades, para as administrações públicas diretas, autárquicas e fundacionais da União, Estados, Distrito Federal e Municípios, obedecido o disposto no art. 37, XXI, da Constituição, e para as empresas públicas e sociedades de economia mista, nos termos do art. 173, § 1º, III, também da Constituição (Const., art. 22, XXVII); h) perda de mandato de Senador (Const., art. 55), pedido de licença de incorporação de Senador às Forças Armadas (Const., art. 53, § 7º); i) escolha de Ministro do Supremo Tribunal Federal, dos Tribunais Superiores e de Governador de Território, escolha e destituição do Procurador- Geral da República (Const., art. 52, III, a, c e e); j) transferência temporária da sede do Governo Federal; l) registros públicos, organização administrativa e judiciária do Ministério Público e Defensoria Pública da União e dos Territórios, organização judiciária do Ministério Público e da Defensoria Pública do Distrito Federal; m) limites dos Estados e bens do domínio da União; n) desapropriação e inquilinato; o) criação, funcionamento e processo do juizado de pequenas causas, assistência jurídica e defensoria pública, custas dos serviços forenses; p) matéria a que se refere o art. 96, II, da Constituição Federal; III – propor, por projeto de resolução, a suspensão, no todo ou em parte, de leis declaradas inconstitucionais pelo Supremo Tribunal Federal (Const., art. 52, X); IV – opinar, em cumprimento a despacho da Presidência, sobre as emendas apresentadas como de

	redação, nas condições previstas no parágrafo único do art. 234; V – opinar sobre assunto de natureza jurídica ou constitucional que lhe seja submetido, em consulta, pelo Presidente, de ofício, ou por deliberação do Plenário, ou por outra comissão; VI – opinar sobre recursos interpostos às decisões da Presidência; VII – opinar sobre os requerimentos de voto de censura, aplauso ou semelhante, salvo quando o assunto possa interessar às relações exteriores do País. § 1º Quando a Comissão emitir parecer pela inconstitucionalidade e injuridicidade de qualquer proposição, será esta considerada rejeitada e arquivada definitivamente, por despacho do Presidente do Senado, salvo, não sendo unânime o parecer, recurso interposto nos termos do art. 254. § 2º Tratando-se de inconstitucionalidade parcial, a Comissão poderá oferecer emenda corrigindo o vício. (NR)
IV - Comissão de Educação, Cultura e Esporte (CE), com 27 membros;	I – normas gerais sobre educação, cultura, ensino e desportos, instituições educativas e culturais, diretrizes e bases da educação nacional e salário-educação; II – diversão e espetáculos públicos, criações artísticas, datas comemorativas e homenagens cívicas; III – formação e aperfeiçoamento de recursos humanos; IV – (Revogado). V – (Revogado). VI – outros assuntos correlatos. (NR)

V - Comissão de Transparência, Governança, Fiscalização e Controle e Defesa do Consumidor (CTFC), com 17 membros;	I - exercer a fiscalização e o controle dos atos do Poder Executivo, incluídos os da administração indireta, podendo, para esse fim: a) avaliar a eficácia, eficiência e economicidade dos projetos e programas de governo no plano nacional, no regional e no setorial de desenvolvimento, emitindo parecer conclusivo; b) apreciar a compatibilidade da execução orçamentária com os planos e programas governamentais e destes com os objetivos aprovados em lei; c) solicitar, por escrito, informações à administração direta e indireta, bem como requisitar documentos públicos necessários à elucidação do ato objeto de fiscalização; d) avaliar as contas dos administradores e demais responsáveis por dinheiros, bens e valores públicos da administração direta e indireta, incluídas as fundações e sociedades instituídas e mantidas pelo poder público federal, notadamente quando houver indícios de perda, extravio ou irregularidade de qualquer natureza de que resulte prejuízo ao Erário; e) providenciar a efetivação de perícias, bem como solicitar ao Tribunal de Contas da União que realize inspeções ou auditorias de natureza contábil, financeira, orçamentária, operacional e patrimonial nas unidades administrativas da União e demais entidades referidas na alínea d; f) apreciar as contas nacionais das empresas supranacionais de cujo capital social a União participe de forma direta ou indireta, bem assim a aplicação de quaisquer recursos repassados mediante convênio, acordo, ajuste ou outros instrumentos congêneres, a Estado, ao Distrito Federal ou a Município; g) promover a interação do Senado Federal com os órgãos do Poder Executivo que, pela natureza de suas atividades, possam dispor ou gerar dados de que necessite para o exercício de fiscalização e controle; h) promover a interação do Senado Federal com os órgãos do Poder Judiciário e do Ministério Público que, pela natureza de suas atividades, possam propiciar ou gerar dados de que necessite para o exercício de fiscalização e controle; i) propor ao Plenário do Senado as providências cabíveis em relação aos resultados da avaliação, inclusive quanto ao resultado das diligências realizadas pelo Tribunal de Contas da União; II - opinar sobre matérias pertinentes aos seguintes

	temas: (Redação dada pela Resolução nº 3, de 2017) a) prevenção à corrupção; (Redação dada pela Resolução nº 3, de 2017) b) acompanhamento e modernização das práticas gerenciais na administração pública federal direta e indireta; (Redação dada pela Resolução nº 3, de 2017) c) prestação eficaz, efetiva e eficiente de serviços públicos; (Redação dada pela Resolução nº 3, de 2017) d) transparência e prestação de contas e de informações à população, com foco na responsabilidade da gestão fiscal e dos gastos públicos, bem como nas necessidades dos cidadãos; (Redação dada pela Resolução nº 3, de 2017) e) difusão e incentivo, na administração pública, de novos meios de prestação de informações à sociedade, tais como redes, sítios e portais eletrônicos, e apoio a Estados e Municípios para a implantação desses meios; (Redação dada pela Resolução nº 3, de 2017) III - opinar sobre assuntos pertinentes à defesa do consumidor, especialmente: (Redação dada pela Resolução nº 3, de 2017) a) estudar, elaborar e propor normas e medidas voltadas à melhoria contínua das relações de mercado, em especial as que envolvem fornecedores e consumidores; (Redação dada pela Resolução nº 3, de 2017) b) aperfeiçoar os instrumentos legislativos reguladores, contratuais e penais, referentes aos direitos dos consumidores e dos fornecedores, com ênfase em condições, limites e uso de informações, responsabilidade civil, respeito à privacidade, aos direitos autorais, às patentes e similares; (Redação dada pela Resolução nº 3, de 2017) c) acompanhar as políticas e as ações desenvolvidas pelo Poder Público relativas à defesa dos direitos do consumidor, à defesa da concorrência e à repressão da formação e da atuação ilícita de monopólios; (Redação dada pela Resolução nº 3, de 2017) d) receber denúncias e denunciar práticas referentes a abuso do poder econômico, qualidade e apresentação de produtos, técnicas de propaganda e publicidade nocivas ou enganosas; (Redação dada pela Resolução nº 3, de 2017) e) avaliar as relações entre custo e preço de produtos, bens e serviços, com vistas a estabelecer normas de repressão à usura, aos lucros excessivos, ao aumento indiscriminado de preços e à cartelização

	de segmentos do mercado; (Redação dada pela Resolução nº 3, de 2017) f) analisar as condições de concorrência com ênfase na defesa dos produtores e dos fornecedores nacionais, considerados os interesses dos consumidores e a soberania nacional; (Redação dada pela Resolução nº 3, de 2017) g) gerar e disponibilizar estudos, dados estatísticos e informações, no âmbito de suas competências. (Redação dada pela Resolução nº 3, de 2017) Parágrafo único. No exercício da competência de fiscalização e controle prevista no inciso I do caput, a Comissão de Transparência, Governança, Fiscalização e Controle e Defesa do Consumidor: (Redação dada pela Resolução nº 3, de 2017) I - remeterá cópia da documentação pertinente ao Ministério Público, a fim de que este promova a ação cabível, de natureza cível ou penal, se for constatada a existência de irregularidade; II - poderá atuar, mediante solicitação, em colaboração com as comissões permanentes e temporárias, incluídas as comissões parlamentares de inquérito, com vistas ao adequado exercício de suas atividades. Art. 102-B. A fiscalização e o controle dos atos do Poder Executivo, inclusive os da administração indireta, pela Comissão de Transparência, Governança, Fiscalização e Controle e Defesa do Consumidor obedecerão às seguintes regras: (Redação dada pela Resolução nº 3, de 2017) I - a proposta de fiscalização e controle poderá ser apresentada por qualquer membro ou Senador à Comissão, com específica indicação do ato e fundamentação da providência objetivada; II - a proposta será relatada previamente, quanto à oportunidade e conveniência da medida e ao alcance jurídico, administrativo, político, econômico, social ou orçamentário do ato impugnado, definindo-se o plano de execução e a metodologia de avaliação; III - aprovado o relatório prévio pela Comissão, o relator poderá solicitar os recursos e o assessoramento necessários ao bom desempenho da Comissão, incumbindo à Mesa e à Administração da Casa o atendimento preferencial das providências requeridas. Rejeitado o relatório, a matéria será encaminhada ao Arquivo; IV - o relatório final da fiscalização e controle, em termos de comprovação da legalidade do ato, avaliação

	política, administrativa, social e econômica de sua edição, e quanto à eficácia dos resultados sobre a gestão orçamentária, financeira e patrimonial, obedecerá, no que concerne à tramitação, as normas do artigo 102-C. Parágrafo único. A Comissão, para a execução das atividades de que trata este artigo, poderá solicitar ao Tribunal de Contas da União as providências ou informações previstas no art. 71, IV e VII, da Constituição Federal. Art. 102-C. Ao termo dos trabalhos, a Comissão apresentará relatório circunstanciado, com suas conclusões, que será publicado no Diário do Senado Federal e encaminhado. I - à Mesa, para as providências de alçada desta, ou ao Plenário, oferecendo, conforme o caso, projeto de lei, de decreto legislativo, de resolução ou indicação; II - ao Ministério Público ou à Advocacia-Geral da União, com cópia da documentação, para que promova a responsabilidade civil ou criminal por infrações apuradas e adote outras medidas decorrentes de suas funções institucionais; III - ao Poder Executivo, para adotar as providências saneadoras de caráter disciplinar e administrativo decorrentes do disposto no art. 37, §§ 2º a 6º, da Constituição Federal, e demais disposições constitucionais e legais aplicáveis; IV - à comissão permanente que tenha maior pertinência com a matéria, a qual incumbirá o atendimento do prescrito no inciso III; V - à Comissão Mista de Planos, Orçamentos Públicos e Fiscalização e ao Tribunal de Contas da União, para as providências previstas no art. 71 da Constituição Federal. Parágrafo único. Nos casos dos incisos II, III e V a remessa será feita pelo Presidente do Senado. Art. 102-D. Aplicam-se à Comissão de Transparência, Governança, Fiscalização e Controle e Defesa do Consumidor as normas regimentais pertinentes às demais comissões permanentes, no que não conflitarem com os termos das disposições constantes dos arts. 102-A a 102-C. (Redação dada pela Resolução nº 3, de 2017) § 1º Ocorrendo a hipótese de exercício concorrente de competência fiscalizadora por duas ou mais comissões sobre os mesmos fatos, os trabalhos se desdobrarão em reuniões conjuntas, por iniciativa do Presidente de um dos órgãos ou de um ou mais de seus membros. § 2º A Comissão de Transparência,

	Governança, Fiscalização e Controle e Defesa do Consumidor poderá, se houver motivo suficiente, comunicar fatos investigados à comissão correspondente da Câmara dos Deputados, para que esta adote a providência que considerar cabível. (Redação dada pela Resolução nº 3, de 2017).

VI - Comissão de Direitos Humanos e Legislação Participativa (CDH), com 19 membros;	I – sugestões legislativas apresentadas por associações e órgãos de classe, sindicatos e entidades organizadas da sociedade civil, exceto partidos políticos com representação política no Congresso Nacional; II – pareceres técnicos, exposições e propostas oriundas de entidades científicas e culturais e de qualquer das entidades mencionadas no inciso I. III – garantia e promoção dos direitos humanos; IV – direitos da mulher; V – proteção à família; VI – proteção e integração social das pessoas portadoras de deficiências e de proteção à infância, à juventude e aos idosos; VII – fiscalização, acompanhamento, avaliação e controle das políticas governamentais relativas aos direitos humanos, aos direitos da mulher, aos direitos das minorias sociais ou étnicas, aos direitos dos estrangeiros, à proteção e integração das pessoas portadoras de deficiência e à proteção à infância, à juventude e aos idosos. Parágrafo único. No exercício da competência prevista nos incisos I e II do caput deste artigo, a Comissão de Direitos Humanos e Legislação Participativa observará: I – as sugestões legislativas que receberem parecer favorável da Comissão serão transformadas em proposição legislativa de sua autoria e encaminhadas à Mesa, para tramitação, ouvidas as comissões competentes para o exame do mérito; II – as sugestões que receberem parecer contrário serão encaminhadas ao Arquivo; III – aplicam-se às proposições decorrentes de sugestões legislativas, no que couber, as disposições regimentais relativas ao trâmite dos projetos de lei nas comissões, ressalvado o disposto no inciso I, in fine, deste parágrafo único. (NR)

VII - Comissão de Relações Exteriores e Defesa Nacional (CRE), com 19 membros;	I – proposições referentes aos atos e relações internacionais (Const., art. 49, I) e ao Ministério das Relações Exteriores; II – comércio exterior; III – indicação de nome para chefe de missão diplomática de caráter permanente junto a governos estrangeiros e das organizações internacionais de que o Brasil faça parte (Const., art. 52, IV); IV – requerimentos de votos de censura, de aplauso ou semelhante, quando se refiram a acontecimentos ou atos públicos internacionais; V – Forças Armadas de terra, mar e ar, requisições militares, passagem de forças estrangeiras e sua permanência no território nacional, questões de fronteiras e limites do território nacional, espaço aéreo e marítimo, declaração de guerra e celebração de paz (Const., art. 49, II); VI – assuntos referentes à Organização das Nações Unidas e entidades internacionais de qualquer natureza; VII – autorização para o Presidente ou o Vice-Presidente da República se ausentarem do território nacional (Const., art. 49, III); VIII – outros assuntos correlatos. Parágrafo único. A Comissão integrará, por um de seus membros, as comissões enviadas pelo Senado ao exterior, em assuntos pertinentes à política externa do País.
VIII - Comissão de Serviços de Infraestrutura (CI), com 23 membros;	I – transportes de terra, mar e ar, obras públicas em geral, minas, recursos geológicos, serviços de telecomunicações, parcerias público-privadas e agências reguladoras pertinentes; II – outros assuntos correlatos. (NR)
IX - Comissão de Desenvolvimento Regional e Turismo (CDR), com 27 membros;	I – proposições que tratem de assuntos referentes às desigualdades regionais e às políticas de desenvolvimento regional, dos Estados e dos Municípios; II – planos regionais de desenvolvimento econômico e social; III – programas, projetos, investimentos e incentivos voltados para o desenvolvimento regional; IV – integração regional; V – agências e organismos de desenvolvimento regional; VI – proposições que tratem de assuntos referentes ao turismo; VII – políticas relativas ao turismo; VIII – outros assuntos correlatos. (NR)

X - Comissão de Agricultura e Reforma Agrária (CRA), com 17 membros;	I – direito agrário; II – planejamento, acompanhamento e execução da política agrícola e fundiária; III – agricultura, pecuária e abastecimento; IV – agricultura familiar e segurança alimentar; V – silvicultura, aquicultura e pesca; VI – comercialização e fiscalização de produtos e insumos, inspeção e fiscalização de alimentos, vigilância e defesa sanitária animal e vegetal; VII – irrigação e drenagem; VIII – uso e conservação do solo na agricultura; IX – utilização e conservação, na agricultura, dos recursos hídricos e genéticos; X – política de investimentos e financiamentos agropecuários, seguro rural e endividamento rural; XI – tributação da atividade rural; XII – alienação ou concessão de terras públicas com área superior a dois mil e quinhentos hectares, aquisição ou arrendamento de propriedade rural por pessoa física ou jurídica estrangeira, definição da pequena e da média propriedade rural; XIII – uso ou posse temporária da terra e regularização dominial de terras rurais e de sua ocupação; XIV – colonização e reforma agrária; XV – cooperativismo e associativismo rurais; XVI – emprego, previdência e renda rurais; XVII – políticas de apoio às pequenas e médias propriedades rurais; XVIII – política de desenvolvimento tecnológico da agropecuária, mediante estímulos fiscais, financeiros e creditícios à pesquisa e experimentação agrícola, pesquisa, plantio e comercialização de organismos geneticamente modificados; XIX – extensão rural; XX – organização do ensino rural; XXI – outros assuntos correlatos.
XI - Comissão de Ciência, Tecnologia, Inovação, Comunicação e Informática (CCT), com 17 membros;	I – desenvolvimento científico, tecnológico e inovação tecnológica; II – política nacional de ciência, tecnologia, inovação, comunicação e informática; III – organização institucional do setor; IV – acordos de cooperação e inovação com outros países e organismos internacionais na área; V – propriedade intelectual; VI – criações científicas e tecnológicas, informática, atividades nucleares de qualquer natureza, transporte e utilização de materiais radioativos, apoio e estímulo à pesquisa e criação de tecnologia; VII – comunicação, imprensa, radiodifusão, televisão, outorga e renovação de concessão, permissão e autorização para serviços de

	radiodifusão sonora e de sons e imagens; VIII – regulamentação, controle e questões éticas referentes a pesquisa e desenvolvimento científico e tecnológico, inovação tecnológica, comunicação e informática; IX – outros assuntos correlatos.
XII - Comissão Senado do Futuro, com 11 membros.	À Comissão Senado do Futuro compete promover discussões sobre grandes temas e o futuro do País, bem como aprimorar a atuação do Senado nessas questões
XIII - Comissão de Meio Ambiente (CMA), com 17 membros.	I - proteção do meio ambiente, controle da poluição, conservação da natureza e defesa do solo, dos recursos naturais e genéticos, das florestas, da caça, da pesca, da fauna, da flora e dos recursos hídricos; II - política e sistema nacional de meio ambiente; III - preservação, conservação, exploração e manejo de florestas e da biodiversidade; IV - conservação e gerenciamento do uso do solo e dos recursos hídricos, no tocante ao meio ambiente e ao desenvolvimento sustentável; V - fiscalização dos alimentos e dos produtos e insumos agrícolas e pecuários, no tocante ao meio ambiente e ao desenvolvimento sustentável; VI - direito ambiental; VII - agências reguladoras na área de meio ambiente, inclusive a Agência Nacional de Águas (ANA); VIII - outros assuntos correlatos.

Fonte: organizado pelo autor, Regimentos das Comissões do Senado Federal BRASIL(2018).