SENAI CIMATEC

PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM COMPUTACIONAL E TECNOLOGIA INDUSTRIAL

Mestrado em Modelagem Computacional e Tecnologia Industrial

Dissertação de mestrado

ECOM Modelo Computacional de Seleção Automática de Termos Candidatos a partir de Mineração de Textos para Auxiliar na Construção de Ontologias

Apresentada por: Keller Santos de Araújo
Orientador: Renelson Ribeiro Sampaio

Agosto de 2013

Keller Santos de Araújo

ECOM Modelo Computacional de Seleção Automática de Termos Candidatos a partir de Mineração de Textos para Auxiliar na Construção de Ontologias

Dissertação de mestrado apresentada ao Programa de Pós-graduação em Modelagem Computacional e Tecnologia Industrial, Curso de Mestrado em Modelagem Computacional e Tecnologia Industrial do SENAI CIMATEC, como requisito parcial para a obtenção do título de **Mestre em Modelagem Computacional e Tecnologia Industrial**.

Área de conhecimento: Interdisciplinar

Orientador: Renelson Ribeiro Sampaio
SENAI CIMATEC

Salvador
SENAI CIMATEC
2013

Nota sobre o estilo do PPGMCTI

Esta dissertação de mestrado foi elaborada considerando as normas de estilo (i.e. estéticas e estruturais) propostas aprovadas pelo colegiado do Programa de Pós-graduação em Modelagem Computacional e Tecnologia Industrial e estão disponíveis em formato eletrônico (*download* na Página Web http://ead.fieb.org.br/portal_faculdades/dissertacoes-e-teses-mcti.html ou solicitação via e-mail à secretaria do programa) e em formato impresso somente para consulta.

Ressalta-se que o formato proposto considera diversos itens das normas da Associação Brasileira de Normas Técnicas (ABNT), entretanto opta-se, em alguns aspectos, seguir um estilo próprio elaborado e amadurecido pelos professores do programa de pós-graduação supracitado.

SENAI CIMATEC

Programa de Pós-graduação em Modelagem Computacional e Tecnologia Industrial

Mestrado em Modelagem Computacional e Tecnologia Industrial

A Banca Examinadora, constituída pelos professores abaixo listados, leram e recomendam a aprovação [com distinção] da Dissertação de mestrado, intitulada "ECOM Modelo Computacional de Seleção Automática de Termos Candidatos a partir de Mineração de Textos para Auxiliar na Construção de Ontologias ", apresentada no dia de agosto de 2013, como requisito parcial para a obtenção do título de **Mestre em Modelagem Computacional e Tecnologia Industrial**.

Orientador:

Prof. Dr. Renelson Ribeiro Sampaio
SENAI CIMATEC

Membro externo da Banca:

Prof. Dr. Gabriela Rezende
Universidade Estadual de Feira de Santana

Membro interno da Banca:

Prof. Dr. Eduardo Manuel de Freitas Jorge
SENAI CIMATEC

"É melhor tentar e falhar que preocupar-se e ver a vida passar.
É melhor tentar, ainda que em vão, que sentar-se fazendo nada até o final.
Eu prefiro na chuva caminhar que em casa me esconder.
Prefiro ser feliz embora louco, que em conformidade viver..."
(Martin Luther King)

Agradecimentos

Primeiramente, agradeço a Deus por conceder a mim o dom da vida e forças para concluir mais um objetivo na minha vida.

Ao meu orientador, Prof. Dr. Renelson Ribeiro Sampaio, pelos ensinamentos e apoio nos momentos mais críticos.

À Prof. Dr. Liliane de Queiroz Antônio, pelos ensinamentos, compreensão, alegria e principalmente por acreditar que era possível.

Aos meus pais, Iranildes e Francisco, pelo apoio moral e financeiro, paciência, cobranças, compreensão que tiveram neste período e por tudo o que eu consegui até hoje.

À minha irmã, Evie, por sempre acreditar e incentivar meu trabalho. Pelos momentos de descontração que tanto revigoraram os meus pensamentos.

Aos meus avós, Sebastiana e Calisio, (*in memorian*) que não puderam acompanhar essa caminhada e a sua conclusão, mas que em vida sempre me ajudaram. Sei que de alguma forma continuam me orientando e me apoiando nos percalços da vida.

Aos amigos que fiz durante o curso e que compartilharam bons e maus momentos e aos amigos antigos por entender as ausências.

Aos meus colegas do Senai Cetind pela compreensão e apoio durante os períodos que precisei me ausentar para elaboração do trabalho. Em especial, a minha colega Patricia Braga pelas ideias e aulas de C# .

Ao Programa de Pós-Graduação em Modelagem Computacional e Tecnologia Industrial e a todos os professores dos quais pude conviver durante o curso.

Por fim, a todos aqueles que compartilharam, de alguma forma, desse período de dedicação e esforço.

Salvador, Brasil — Keller Santos de Araújo

30 de Agosto de 2013

Resumo

Vivencia-se uma consolidação crescente em progressões geometricas, não só da convergência digital, mas também da criação de conteúdos totalmente já digitalizados. No entanto, grande parte dessas informações disponíveis na Web mantém como característica principal a interpretação do seu conteúdo direcionada á pessoas e não a programas de computador. Neste contexto, a construção de ontologias pode ser vista como um passo importante de evolução na especificação de conhecimento. Por isso, muitos estudos são realizados acerca da referida construção e sua estrutura. Todavia, os estudos encontrados não apresentam um modelo computacional que associe as técnicas de mineração de textos e gestão do conhecimento para construir uma lista de termos candidatos para auxiliar no processo de construção de ontologias de domínio a partir dos resultados da aplicação destas técnicas. Diante disso, e com base nos estudos realizados, o problema estudado nessa dissertação visa diminuir a intervenção manual e a dependência de especialistas do domínio na construção de ontologias de domínio unindo mineração de texto e gestão do conhecimento. Dentro desse contexto, o objetivo desse trabalho é projetar um modelo para eleição de termos candidatos para auxiliar no processo de construção de ontologias de domínio a partir dos resultados das técnicas de mineração de textos aplicadas a uma dada amostra de documentos científicos não-estruturados do domínio selecionado pelo usuário. O modelo proposto tem como finalidade subsidiar a diminuição da intervenção manual e da dependência de especialistas do dominio no processo de construção de ontologias. Apesar do modelo construir uma lista de termos candidatos e não construir ontologias, os termos candidatos podem auxiliar no melhor mapeamento do domínio e na consequente construção de novas ontologias, ou seja, o modelo proposto pode colaborar na construção de bases ontológicas, e consequentemente, no seu aumento.

Palavras-chave: Mineração de Texto, Ontologia, Descoberta do Conhecimento e Web Semântica

Abstract

Experiences is increasing consolidation in geometric progressions, not only of digital convergence, but also the creation of content already fully digitized. However, much of this information available on the Web has as main characteristic the interpretation of its contents will be directed people and not computer programs.In this context, the ontology construction can be seen as an important step in the evolution of knowledge in the specification. Therefore, many studies are conducted on the said building and its structure. However, studies have not found a computational model involving the techniques of text mining and knowledge management to build a list of candidate terms to assist in the construction of domain ontologies from the results of applying the techniques.Given this, and based on the studies conducted, the problem studied in this thesis is how to reduce manual intervention and dependence on domain experts to build domain ontologies linking text mining and knowledge management. Within this context, the aim of this work is to design a model for election of candidate terms to compose domain ontologies from the results of text mining techniques applied to a given sample scientfic papers unstructured domain selected by the user. The proposed model aims to support the reduction of manual intervention and addiction specialists in the domain ontology building process. Although the model build a list of candidate terms and do not build ontologies, the term candidates may assist in better mapping of the domain and consequently the construction of new ontologies, that is, the proposed model can collaborate in constructing ontological and, consequently, in its increase.

Keyword: Text Mining, Ontologies, Knowledge discovery e Semantic Web.

Sumário

Lista de Tabelas

Lista de Figuras

Lista de Siglas

PPGMCTI ..	Programa de Pós-graduação em Modelagem Computacional e Tecnologia Industrial
WWW	World Wide Web
URI	Universal Resource Identificator
URL	Unified Resource Locator
OWL	Web Ontology Language
RDF	Resource Description Framework
HTML	HyperText Markup Language
DARPA	Defense Advanced Research Projects Agency
DM	Data Mining
KDD	Knowledge Discovery in Database
MT	Mineração de Texto
PLN	Processamento de Linguagem Natural
W3C	World Wide Web Consortium

Capítulo Um

Introdução

Neste capítulo será apresentado o contexto que subsidiou essa pesquisa, através da descrição do problema, metodologia, objetivo e limitações.

1.1 Definição do problema

O surgimento na década de 90 da Web e a sua disseminação crescente nos dias de hoje permitem o acesso a uma imensurável quantidade de informações. No entanto, de acordo com Breitman (2005), grande parte dessas informações disponíveis nas páginas Web ainda mantêm muito de sua característica inicial, ou seja, são direcionadas para outras pessoas e não para serem processadas por programas de software. Dentro desse contexto, surge a Web semântica, com a proposta de atribuir conteúdo as páginas Web para que esses possam ser entendidos pelos computadores e, com isso seja possível fazer inferências sobre esses conteúdos de forma tal que os resultados das buscas se aproximem daqueles obtidos pelos seres humanos.

A Web semântica para atingir o seu resultado precisa fazer uso de técnicas de compreensão, classificação e recuperação da informação, como os conceitos que norteiam esse trabalho: A engenharia do conhecimento que para Maedche e Staab (2001) trata da aquisição, manutenção e acesso ao conhecimento de uma organização; Mineração de texto que Aranha e Passos (2006) define como a busca por padrões em coleções de textos em linguagem natural e, por fim, sistemas de representação do conhecimento, como as ontologias definida como "uma especificação formal e explícita de uma conceitualização compartilhada."[Gruber apud (BREITMAN, 2005)].

De acordo com Júnior (2006) a mineração de textos apresenta-se como uma ferramenta capaz de sumarizar um conjunto de documentos em agrupamentos, especificando as relações semânticas dos termos que os compõem. Dessa forma, o usuário pode ter a ideia do assunto dos textos sem precisar lê-los na integra. "Em contrapartida, a construção de ontologias pode ser vista como um passo importante de evolução na especificação de conhecimento."(FREITAS, 2003). Por isso, muitos estudos são realizados acerca da construção de ontologias de domínio e sua estrutura, a exemplo de alguns trabalhos descritos na subseção anterior. Ressalta-se que as referidas pesquisas foram realizadas no âmbito nacional e internacional com maior foco no âmbito nacional devido a problemática estudada se restringir a ontologias construídas a partir de coleções de documentos escritos na

língua portuguesa do Brasil.

Todavia, os estudos citados no paragráfo anterior não apresentam um modelo computacional que associe as técnicas de mineração de textos e gestão do conhecimento para auxiliar no processo de construção de ontologias de domínio. Diante disso, e com base nas situações abordadas nos parágrafos supracitados, o problema alvo dessa dissertação é como diminuir a intervenção direta e a dependência de especialistas do domínio no processo de construção de ontologias de domínio unindo mineração de texto e gestão do conhecimento. Não havendo a pretensão de realizar um mapeamento completo de um domínio do conhecimento.

1.2 Objetivo

O objetivo desta pesquisa é a concepção de um modelo denominado ECOM (**E**leição de **T**ermos para **C**onstrução de **O**ntologia através de **M**ineração). A proposta do ECOM é construir listas de termos candidatos para compor uma ontologia de dominio a partir de técnicas de mineração de textos e gestão do conhecimento aplicadas em coleções de documentos não estruturados escritos na língua portuguesa do Brasil.

O objetivo está estruturado nos seguintes objetivos específicos:

1.Aplicar algoritmos de mineração de textos às coleções de documentos fornecidas pelo usuário.

3. Especificar e desenvolver o método para a avaliação automática dos resultados das aplicações dos algoritmos de mineração de textos e, posterior classificação morfológica e sintática.

4. Especificar e desenvolver o método para eleição dos termos candidatos a compor a ontologia de domínio a partir dos resultados da mineração de textos e classificações citadas nos itens anteriores.

5. Submeter o ECOM a um trabalho experimental utilizando dois domínios de conhecimento, mineração de dados e impacto ambiental.

6. Realizar uma análise qualitativa e quantitativa dos resultados do trabalho experimental do item anterior.

1.3 Importância da pesquisa

O modelo proposto tem como finalidade diminuir a intervenção manual e a dependência de especialistas do domínio no processo de construção de ontologias de domínio unindo mineração de texto e gestão do conhecimento.

Apesar do modelo construir uma lista de termos candidatos e não construir ontologias, os termos candidatos podem auxiliar no melhor mapeamento do domínio e na consequente construção de novas ontologias, ou seja, o modelo proposto pode colaborar na construção de bases ontológicas e, consequentemente, no seu aumento. Ressalta-se que esta pesquisa não encontrou soluções que contemplassem a junção de mineração de textos e gestão do conhecimento para eleição de termos candidatos para a construção de ontologias, e sim foi observada a existência de ferramentas para construção de ontologias que possuem como entrada os termos definidos a partir do conhecimento de especialistas do domínio.

O ECOM também permite aos engenheiros de ontologias ter um ponto inicial para o processo de construção, pois o modelo já concede aos engenheiros um primeiro mapeamento do domínio do conhecimento. O que permite um ganho de tempo no processo tendo já um conhecimento dos termos associados ao domínio que deseja construir a ontologia.

1.4 Limites e limitações

A construção de ontologias na língua portuguesa do Brasil é algo complexo e suscetível a erros.Trabalhar com documentos não-estruturados também dificulta uma eleição de termos mais precisa e o consequente mapeamento completo de um dado domínio específico. Com isso, este trabalho se limita a desenvolver uma lista de termos candidatos a compor uma ontologia visando estudar a problemática proposta e não a construção de ontologias para o mapeamento completo de um domínio.

Uma outra limitação foi a reunião de um quantitativo de documentos de fontes confiáveis na língua portuguesa do Brasil tendo em vista que a maioria dos eventos e revistas solicitam documentos na língua inglesa. Essa limitação fica mais evidente no domínio mineração de texto, pois naturalmente as temáticas que envolvem tecnologia da informação tendem a ter uma quantidade considerável de termos em inglês e o modelo proposto nesta dissertação não considera outros idiomas diferentes da língua portuguesa do Brasil.

1.5 Aspectos metodológicos

O desenvolvimento dessa pesquisa ocorreu em três fases com três etapas cada uma, como apresenta a Figura 1.1, detalhadas a seguir.

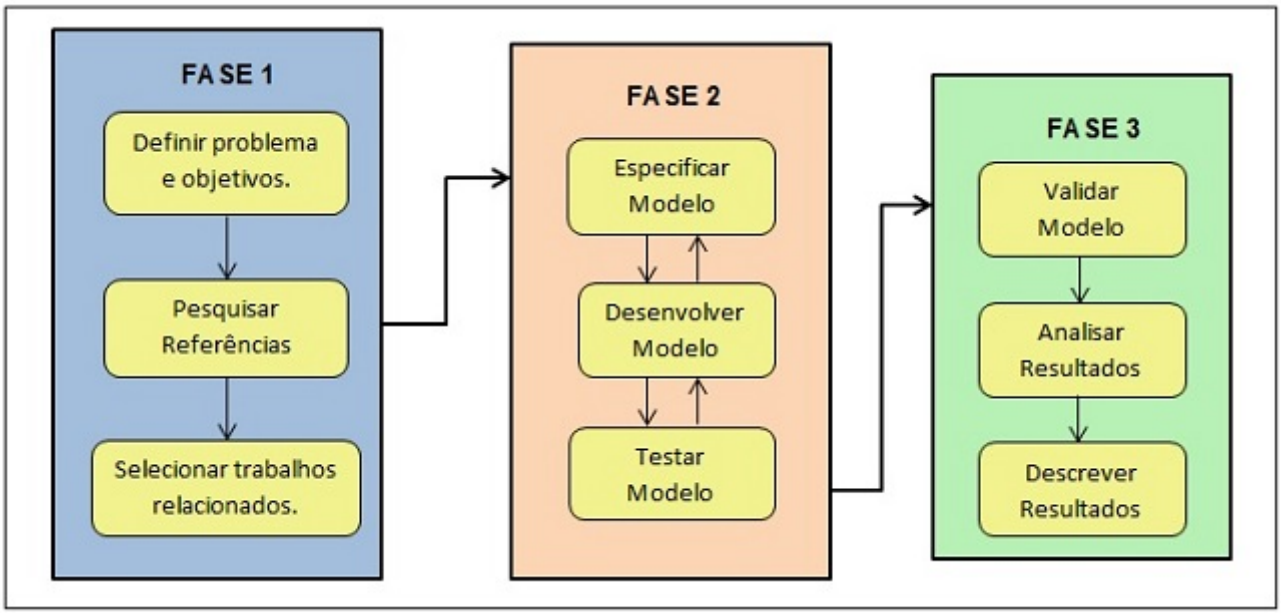

Figura 1.1: Fases da Metodologia. Fonte: Autor.

As etapas da fase um foram realizadas concomitantemente. Para definição do problema e posterior descrição dos objetivos foi necessário fazer uma pesquisa bibliográfica que de acordo com Lakatos (2007) essa pesquisa tem como objetivo trabalhar com informações levantadas e selecionadas da literatura sobre uma determinada problemática, para explicar o objeto e o(s) fenômeno(s) da pesquisa. Ao longo e ao final da referida pesquisa, foram selecionados alguns trabalhos relacionados visando mapear o que já existia de estudo que pudesse agregar conhecimento para esse trabalho, além de permitir definir a importância dessa pesquisa para o meio acadêmico. Após a construção de toda referência, iniciou-se a fase dois, composta por três etapas descritas abaixo, onde há a concepção do modelo proposto nesta dissertação denominado ECOM.

Especificar modelo - Nessa etapa, foi construído o documento de especificação dos requisitos tendo em vista que esses foram levantados em sua maioria ao longo das etapas para a definição do problema. Ao longo dessa etapa foi feito o levantamento e análise dos requisitos.

Desenvolver modelo - Nessa etapa, foi desenvolvido o modelo conforme a análise dos requisitos realizada na etapa anterior. Uma vez que todos os requisitos não eram conhecidos, ao longo do desenvolvimento outros requisitos funcionais foram identificados. Para

isso foi utilizada nesse projeto a metodologia incremental. "Nessa metodologia cada fase é desenvolvida como funcionalidade e ao final, todas são integradas, com isso caso haja algum problema este poderá ser identificado e tratado ao longo do projeto e não apenas ao seu final."(SOMMERVILLE, 2011)

Testar modelo - Nessa etapa, a autora realizou testes com o objetivo de verificar se o que foi implementado estava correspondendo aos requisitos especificados na fase 1. Fazendo uso dos preceitos da metodologia incremental, quando em algum teste era detectada uma falha ou erro, esse era novamente especificado, desenvolvido e testado mais uma vez até que a referida falha fosse sanada.

Após a finalização do desenvolvimento do ECOM, este foi submetido a validação na fase 3, composta por 3 etapas, onde foram definidos dois domínios, impacto ambiental e mineração de texto, das áreas de meio-ambiente e tecnologia da informação, respectivamente. Esses foram analisados na etapa 2 em dois momentos: ao final do processo de mineração, onde a lista de termos resultante do final do referido processo foi validada por uma especialista da área que destacou os termos sem relevância para o domínio, e ao final da construção da lista dos termos candidatos a compor a ontologia do referido domínio do conhecimento. Nesse segundo momento, houve duas análises: Verificou-se a proximidade dos termos eleitos com o domínio proposto e a comparação da lista final com a primeira lista do final do processo de mineração a fim de verificar se os termos outrora identificados pela validadora como sem relevância para o domínio foram de fato excluídos.

As formas metodológicas para condução da validaąão e análise dos resultados, tiveram como base o método de pesquisa exploratória e no que concerne a abordagem foi realizada uma análise qualitativa e quantitativa. Para Gil (1991) a abordagem qualitativa trata de uma metodologia geral para desenvolver teoria que está inserida em dados sistematicamente coletados e analisados. Essa teoria surge durante a própria pesquisa que ocorre através da interação contínua entre a coleta e a análise. Já a quantitativa supõe uma população de objetos de observação comparável entre si e enfatiza os indicadores numéricos e percentuais sobre determinado fenômeno pesquisado. A escolha da pesquisa exploratória justifica-se pela realização da coleta de dados representada pela definição dos documentos da coleção e a sua posterior submissão aos processos do modelo. Já a abordagem mista, qualitativa e quantitativa, é contemplada na forma da análise dos resultados que foi realizada por amostragem dos resultados havendo uma preocupação com as etapas do processo para o alcance desses resultados.

1.6 Organização da Dissertação de mestrado

Este dissertação de mestrado apresenta 5 capítulos e está estruturada da seguinte forma:

- **Capítulo 1 - Introdução**: Contextualiza o âmbito, que subsidiou essa pesquisa, através da descrição do problema, metodologia, objetivo e limitações, e por fim descreve como esta dissertação de mestrado está estruturada;
- **Capítulo 2 - Da Gestão a Representação do Conhecimento**: Discute os conceitos que nortearem esse trabalho de dissertação de mestrado, gestão do conhecimeto, Web semântica, mineração de texto e ontologia.
- **Capítulo 3 - Trabalhos Relacionados**: Neste capítulo serão descritos alguns trabalhos encontrados que possuem maior proximidade com a temática deste trabalho de dissertação, observados ao longo do levantamento bibliográfico.
- **Capítulo 4 - ECOM Modelo Computacional de Seleção Automática de Termos Candidatos a partir de Mineração de Textos para Auxiliar na Construção de Ontologias**: Apresenta o modelo proposto nesse trabalho e as etapas da sua construção, análise, desenvolvimento, testes e validação e, por fim discuti acerca das aplicabilidades observadas na etapa de validação;
- **Capítulo 5 - Considerações Finais**: Apresenta as conclusões, contribuições e algumas sugestões de atividades a serem desenvolvidas no futuro que foram identificadas ao longo dessa pesquisa.

Capítulo Dois

Da Gestão a Representação do Conhecimento

Neste capítulo será descrita gestão do conhecimento, a sua aplicabilidade e os conceitos relacionados com a prática de gestão, armazenamento, interpretação e disseminação do conhecimento.

"Nos dias de hoje, o conhecimento e a capacidade de criá-lo e utilizá-lo são considerados as mais importantes fontes de vantagem competitiva, sustentável de uma empresa."(TAKEUCHI; NONAKA, 2008). Por isso, antes de conceituar e descrever a gestão do conhecimento é preciso entender o conceito de conhecimento de forma isolada e aprender a diferenciá-lo de informação e dados.

De acordo com Silva (2004) vários autores buscam destacar a diferença existente entre dados, informação e conhecimento, dentre eles destacam-se (Dutta,1997; Marshall,1997; Davenport,Prusak,1998), porém não existe propriamente um consenso quanto á diferenciação ou definição entre esses três conceitos.

Em Ferreira (2010) encontra-se a seguinte definição para dados, informação e conhecimento: Dados elemento ou quantidade conhecido que serve de base a resolução de um problema; Informação são dados acerca de alguém ou de algo. Conhecimento amplo e bem fundamentado, resultante da análise e combinação de vários informes; Conhecimento é ideia, noção. Prática da vida, experiência. Silva (2004) define que uma informação é convertida em conhecimento quando um indivíduo consegue ligá-la a outras informações, avaliando-a e entendendo seu significado no interior de um contexto específico. Com isso, o referido autor conclui que os dados são pré-requisitos para a informação, e esta é pré-requisito para o conhecimento.

Já para Takeuchi e Nonaka (2008) a informação é um fluxo de mensagens, enquanto que o conhecimento, apesar de seguir o mesmo fluxo de informação, ele está ancorado nas crenças e no compromisso do seu portador. Com isso , o autor enfatiza que o conhecimento é essencialmente relacionado com a ação humana. Dados é uma informação desconectada, ou seja, é uma informação isolada que não emite nenhuma mensagem, pois necessita de maiores detalhes para assim formar uma informação e emitir uma mensagem passível de interpretação e aplicação com outro conjunto de dados.

De acordo com Takeuchi e Nonaka (2008), o conhecimento é formado por dois componentes dicotômicos e aparentemente opostos, descritos a seguir:

✓Conhecimento explícito: Pode ser rapidamente transmitido, formal e sistematicamente, pois é possível ser expresso em palavras, números ou sons, além de ser compartilhado na forma de dados, fórmulas científicas, recursos visuais, fitas de áudio, especificações de produtos ou manuais.

✓Conhecimento tácito: Este contêm uma importante dimensão cognitiva, e está relacionado com as intuições, palpites, percepções, ideais e valores. Por isso não é facilmente visível e explicável, sendo altamente pessoal e difícil de formalizar. Em contrapartida, é o conhecimento tácito o responsável pela forma como o indivíduo percebe o mundo ao seu redor.

"De acordo com as definições acima, o conhecimento não é explícito ou tácito. Ele é tanto explícito quanto tácito, pois é inerentemente paradoxal já que é formado do que aparenta ser dois opostos."(TAKEUCHI; NONAKA, 2008).

Após entender o conceito de conhecimento e a sua diferença entre dados e informação é possível entender a gestão do conhecimento. "Essa necessita de especialistas para identificar áreas, domínios e atividades a partir do qual se possa fazer uma gestão eficaz de todo este processo."(SILVA, 2004). Para o Group () a gestão do conhecimento é a disciplina que promove, com visão integrada, o gerenciamento e o compartilhamento de todo o ativo de informação possuído pela empresa. Esta informação pode estar em banco de dados, documentos, procedimentos, bem como em pessoas, através de suas experiências e habilidades.

"A gestão do conhecimento é a coleção de processos que governa a criação, disseminação e utilização do conhecimento para atingir plenamente os objetivos da organização."(DAVENPORT, 1998). De acordo com Takeuchi e Nonaka (2008) uma das principais abordagens da gestão do conhecimento é a conversão contínua do conhecimento tácito para o explícito. Em termos de ações em que a conversão entre o formato tácito-explícito do conhecimento normalmente ocorre, têm-se os quatro modos existentes: combinação, socialização, externalização e internalização, como mostrado na Figura 2.1.

A Figura 2.1 apresenta o espiral do conhecimento onde de acordo com Takeuchi e Nonaka (2008) a interação entre o conhecimento tácito e o conhecimento explícito tornar-se-á maior na escala à medida que sobe nos níveis ontológicos. A imagem também apresenta os quatro modos de conversão do conhecimento, explicados a seguir:

✓Socialização - De tácito para tácito. Esse conhecimento compartilhado acontece quando: Ocorre diálogo frequente e comunicação "face a face"; *Brainstorming*, *insights* e intuições são valorizados, disseminados e analisados (discutidos) sob várias perspectivas (por grupos heterogêneos); Valoriza-se o trabalho do tipo "mestre-aprendiz": observação, imitação e

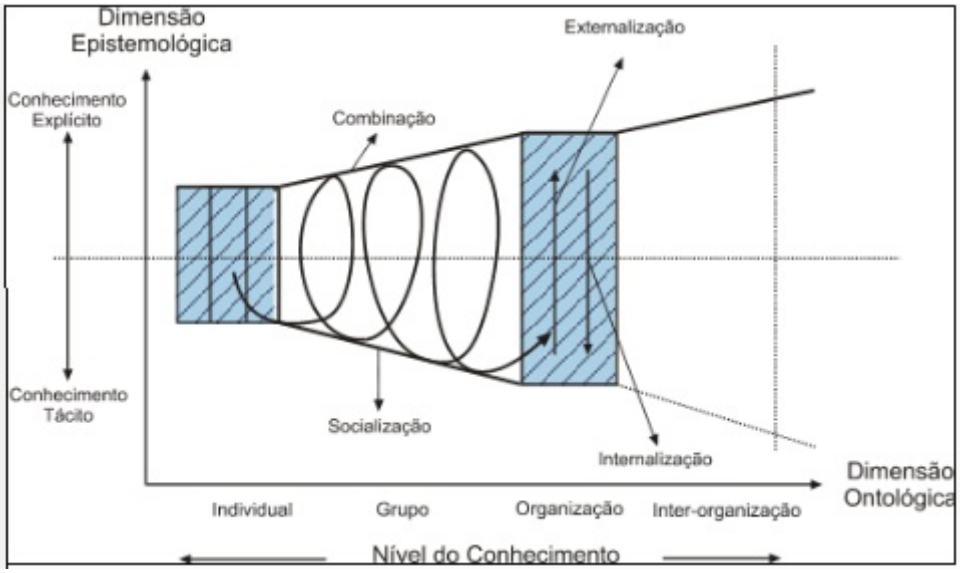

Figura 2.1: Espiral do conheciemento (TAKEUCHI; NONAKA, 2008)

prática acompanhada por um tutor; Há compartilhamento de experiências e modelos-mentais via trabalho em equipe.

✓Externalização - De tácito para explícito. Conversão de parte do conhecimento tácito do indivíduo em algum tipo de conhecimento explícito. Esse tipo de conversão é pouco abordado por outras teorias da administração.

✓Combinação - De explícito para explícito. Conversão de algum tipo de conhecimento explícito gerado por um indivíduo para agregá-lo ao conhecimento explícito da organização. Esse tipo de conversão é abordado também pelas teorias ligadas ao processamento da informação, ocorre por meio do agru-pamento (classificação, sumarização) e processamento de diferentes conhecimentos explícitos.

✓Internalização - De explícito para tácito. A abordagem ocorre pelas teorias ligadas à aprendizagem organizacional. Normalmente, esse conhecimento operacional acontece por meio de: Leitura/visualização e estudo individual de documentos de diferentes formatos/tipos (textos, imagens etc.); Prática individual (*learning by doing*); Reinterpretar/reexperimentar, individualmente, vivências e práticas.

"Os ciclos de conversão do conhecimento, passando várias vezes por esses quatro modos, formam uma espiral que serve para analisar e entender os mais diversos casos de criação e disseminação do conhecimento, sendo que cada caso terá suas particularidades ou especificidades."(SILVA, 2004)

De acordo com as definições dos autores citados acima acerca da gestão do conhecimento

e o estudo dos quatro modos de conversão do tácito para explícito é possível definir, de forma pragmática, a gestão do conhecimento como sendo o processo de adquirir, manter, mapear e criar métodos de acesso ao conhecimento organizacional.

Na gestão do conhecimento, há a indigência de difundir o conhecimento, facilitando o seu acesso através da criação de redes relacionais e informacionais que permitam estabelecer a ligação entre pessoas, para que essas possam aceder a dados e informações relevantes sempre que necessário.

De acordo com Almeida (2003), a necessidade de difundir o conhecimento, característica da gestão do conhecimento, remete a uma das preocupações da ciência da informação. A padronização da terminologia utilizada para se encontrar e se classificar a informação. Daí deriva a importância do uso de ontologias, descrita no capítulo 2.3, para caracterizar e relacionar entidades em um domínio do conhecimento.

A necessidade supracitada impulsiona as empresas á criar mecanismos e fazer uso de ferramentas para prover essa difusão. A exemplo de *wikis* corporativos e ou fóruns de discussão. O que permite que ao menos parte dos conhecimentos dos seus colaboradores possa migrar de tácito para explícito, e assim facilitar a aprendizagem dos novos colaboradores e manter o conhecimento empresarial sempre atualizado.

Como a maioria das ferramentas utilizadas para difundir o conhecimento é web, seja ela intra ou internet, faz-se necessário a aplicação de tecnologias e padrões que compreendam a semântica dessas informações disponibilizadas na web diminuindo a dependência da interpretação humana a fim de facilitar a busca dessas informações disponibilizadas em forma de textos não padronizados. Para auxiliar nesse processo, surgiu em 1969 e vêm evoluindo a websemântica, descrita na próxima secção.

2.1 WebSemântica

Em 1969, com o nome de ArphaNet, a internet foi desenvolvida com o intuito de manter a comunicação das bases militares dos Estados Unidos durante a guerra fria. Com o fim da guerra fria, foi cedida a universidades norte-americanas passando a ser utilizada por cientistas e estudantes da área de computação para compartilhar informações, o que tornou-a popular entre eles até se tornar a World Wide Web (www) ou simplesmente Web.

De acordo com Breitman (2005) inicialmente as páginas web eram desenvolvidas por programadores, e como tinham a característica de compartilhar informações de forma simples tornaram-se populares entre programadores e engenheiros de *software*. Com o

passar do tempo, ferramentas foram criadas a fim de permitir que não programadores também pudessem criar as suas próprias páginas. Nessa época, as informações disponíveis na internet só eram compreendidas pelos leitores humanos e não por máquinas e ou programas.

"A Web cresceu e continua em ascensão , mas mesmo assim grande parte das páginas disponíveis ainda mantêm muito de sua característica inicial, onde o conteúdo é direcionado para outras pessoas e não para ser processados por programa de computador."(BREITMAN, 2005). "Visto que a constante expansão de conteúdo dificulta a indexação, impossibilitando que os mecanismos de busca gerem resultados eficientes."(VILLACA, 2002) aponta um dos motivos: "o pensar e escrever humano não é binário". Para (VILLACA, 2002) o pensamento humano: "não trabalha com unidades de informação apenas, mas por figurações intuitivas e hipotéticas".

Para Pickler (2007) ao realizar uma busca por determinado termo na Web, os mecanismos de busca trarão como resultado todas as páginas da sua base de dados que possuem o termo no seu conteúdo. Entretanto, caberá ao usuário (humano) consultar cada *site* retornado para verificar qual deles possuem o termo com o significado e sentido de acordo com a sua pesquisa.

"Mecanismos de busca como *Google* e AltaVista ainda necessitam da intervenção humana para puder identificar as associações que de fato contemplem o tema buscado. Alguns sites de busca têm utilizado artifícios para melhorar essa situação, tais como os mecanismos de indexação do Google."(BREITMAN, 2005)

De acordo com Friedman (2007) o processo de indexação do Google consiste em agrupar o maior número de páginas web possível em ordem de importância que é determinada pelo algoritmo PageRank. Esse algoritmo classifica como mais relevante a página que tiver um número maior de *links* em outras páginas da web, apontando para ela. O PageRank também leva em consideração para a classificação a relevância da página que possui o *link*, pois quanto maior essa relevância maior será a pontuação da página associada ao *link*.

A figura 2.2 ilustra uma pesquisa utilizando os termos "professora Keller"no mecanismo de busca muito utilizado na internet, o google. Nesse exemplo, onde foram retornados 128.000 resultados, é possível observar três falhas no processo de busca no que concerne a semântica.

✓Falha 1. Ao longo de todo o resultado é retornado associações com o termo professora, mas não há um resultado com o termo docente que é sinônimo de professora.

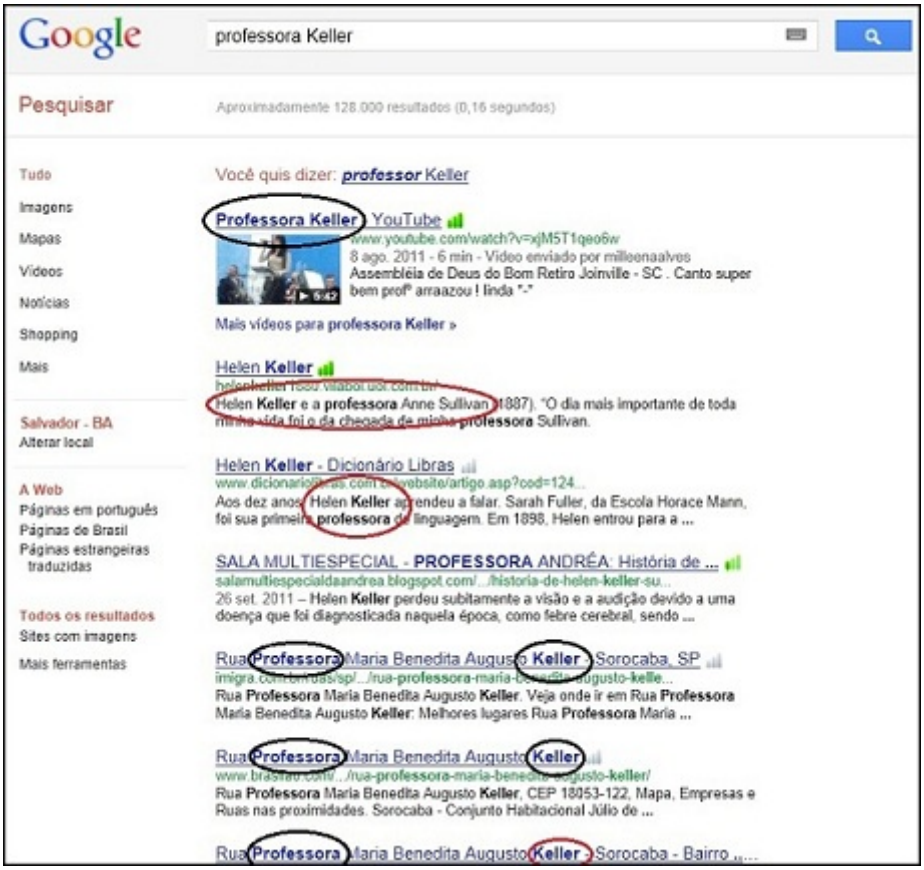

Figura 2.2: Exemplo de Pesquisa. Fonte: Autor.

✓Falha 2. Nos resultados circulados em vermelho observa-se que em alguns resultados não foi considerada a proximidade dos termos da busca, o que retornou o termo Keller dissociado de professora.

✓Falha 3: É possível observar que em alguns casos é retornado professor sem levar em consideração o gênero especificado na busca que é feminino, professora e não professor como sugere a ferramenta por possuir em sua base o termo Keller como masculino.

Devido às falhas relatadas no exemplo supracitado, onde não é levado em consideração a semântica dos termos. Retornando, dessa forma, muitos resultados de cárater irrelevante para a pesquisa, tornando o processo dispendioso, obrigando o usuário a refinar a consulta utilizando outras palavras ou a perder tempo buscando dentre os 128.000 resultados retornados aqueles que estão no domínio da sua busca. De acordo com Breitman (2005) essas dificuldades para mecanismos que trabalham com palavras-chave ocorre porque o termo consultado é comparado a um índice em uma base de dados e a busca não leva em consideração o sentido semântico, mas a sintaxe das palavras-chave

Souza e Alvarenga (2004) relata que a Web ainda é bastante insuficiente no que diz respeito a recuperação de conteúdo relevante e afirma que tal situação deve-se ao fato da falta de

uma estratégia abrangente e satisfatória para a indexação dos documentos nela contidos, pois a recuperação das informações, realizadas através de motores de busca, baseia-se, primariamente, em palavras-chave contidas no texto dos documentos originais, o que o autor afirma ser muito pouco eficaz.

Essas limitações, ainda existentes no processo de busca na internet, foram motivacionais para o surgimento de estudos que levaram ao surgimento da Web Semântica onde de acordo com Breitman (2005) a ideia central é categorizar a informação de maneira padronizada, facilitando o seu acesso.

LEE, HENDLER e LASSILA (2001) define a Web Semântica (WS) como a evolução da Web, onde o conteúdo publicado nela teria significado bem definido, permitindo que computadores e seres humanos entendam-no e trabalhem em cooperação. Além disso, eles afirmam no seu artigo de 2001, considerada a primeira publicação significativa sobre web semântica, que para a WS funcionar, os computadores devem ter acesso a coleções estruturadas de informações e conjuntos de regras de inferência para poder conduzir um raciocínio automatizado.

De acordo com Feigenbaum et al. (2007) a Web Semântica não difere da World Wide Web, pois a Web Semântica é um acessório que disponiliza a Web muito mais utilidade, pois surge quando as pessoas de diferentes campos ou vocações fazem pesquisas sobre os mais variados temas, aceitando esquemas comuns para representação das informações acerca das suas pesquisas. Como muitos grupos desenvolvem essas taxonômias, as ferramentas da Web Semântica permitem-lhes unir os seus esquemas e traduzir seus termos.

Para Koivunen (2001) a Web Semântica tem como maior motivação transformar os dados e aplicativos em elementos úteis, legíveis e compreensíveis para o software, ou melhor, para os agentes inteligentes visando facilitar a comunicação dinâmica, a cooperação e o comércio eletrônico entre empresas.

Ainda de acordo com Koivunen (2001) os príncipios da Web Semântica são implementados nas camadas e padrões de tecnologias web. A figura 2.3, mostra a arquitetura em camadas da Web que de acordo com Breitman (2005) foi proposta por Tim Berners Lee durante a conferência de XML de 2000 e é conhecida pela comunidade como "bolo de noiva". A ideia é construir uma arquitetura em cima do que já existe, evitando dessa forma a reestruturação da internet.

A arquitetura da Web Semântica apresentada na figura 2.3, é composta pela URI/Unicode, XML (*eXtensible Markup Language*) *XMLSchema/namespace*, RDF (*Resouce Description Framework*) /RDF *Schema, Ontology* (Ontologia), *logic* (lógica), *proof* (prova) e *trust*.

Pode-se agrupar essa arquitetura em três camadas. Uma relacionada ao conteúdo sintático da web semântica composto pelas camadas Unicode, URI e XML/NS (*Name Schema*) responsáveis pela identificação dos recursos; outra formada por RDF, *RDF Schemas e Ontology Vocabulary*, relacionadas à semântica e uma última camada lógica composta por *Logic, Proof, Trust e o Digital Signature* (assinatura digital).

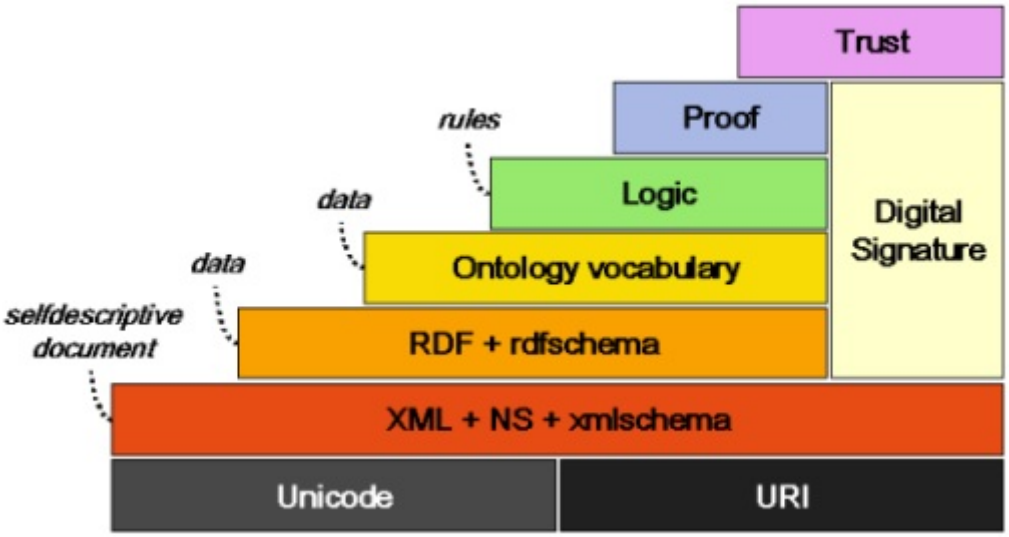

Figura 2.3: Arquitetura de Camadas da Web Semântica. Fonte: (KOIVUNEN, 2001).

Camada Base: É a base da arquitetura e é composta por URI (*Universal Resource Identifier*) e UNICODE. O URI é um padrão de identificação universal de recursos ou objetos da web. O padrão Unicode especifica a forma de representar textos em várias linguagens.

Camada 1: É composta por XML (*eXtensible Markup Language*), NameSpace (NS) e XMLschema. Namespace é um vocabulário controlado que identifica um conjunto de conceitos de forma única para que não haja ambiguidade na sua interpretação. Os namespaces XML são conjuntos de tipos de elementos e atributos possíveis para cada tipo. As triplas do RDF se baseiam em namespaces de forma que a cada recurso seja associado uma dupla de propriedade e valor. "Os namespaces podem ser referenciados por meio de uma URI que se constitui em um repositório compartilhado, e não-ambíguo, onde usuários e programas de validação de código XML podem consultar a sintaxe e propriedades semânticas dos conceitos cobertos."(SOUZA; ALVARENGA, 2004). A linguagem XML apresenta artifícios que permitem descrever a estrutura dos textos. Além disso, ela facilita a interoperabilidade entre sistemas de informação, o que permite que a XML seja utilizada como padrão de intercâmbio de documentos de dados na rede.

Camada 2: Formada por RDF (*Resource Description Framework*) e RDFschema. O RDF desenvolvido pela W3C (World Wide Web Consortium), baseado nos príncipios de redes semânticas para ser utilizada na descrição de recursos disponíveis na Web, fornece

interoperabilidade e a semântica para metadados a fim de facilitar buscas textuais simples. Para isso, o RDF possui três princípios fundamentais simples, recursos, propriedades e frases. Já o RDFschema ou RDFS que surgiu como uma extensão do RDF, oferece uma modelagem que permite a construção de hierarquias, classes, propriedades, subclasses e subpropriedades, ou seja, permite a modelagem de ontologias simples.

Camada 3: Visando atingir o nível de expressividade necessária para a Web Semântica, o que não é comtemplado na camada dois foi criada a camada ontologias, esse conceito, suas linguagens e metodologias serão discutidos no capítulo 2.3.

Camada 4,5 e 6: As camadas 4,5 e 6, lógica (logic), prova (proof) e validação (trust) , respectivamente, atuam juntas. Enquanto a camada lógica habilita a escrita das regras enquanto a prova executa juntamente com a camada validação. Por fim, a camada assinatura digital (*assignature digital*) detecta as alterações nos documentos.

Conforme visto nos parágrafos supracitados, a Web Semântica visa classificar o conteúdo da Web a partir de interpretações semânticas, para que esse possa ser interpretado por máquinas e a partir dessa interpretação seja possível fazer inferências sobre os conteúdos. Para prover essa estrutura, foram criadas varias linguagens, dentre elas: RDF, XML e OWL, descritas a seguir.

RDF - *Resource Description Framework*: De acordo com Souza e Alvarenga (2004) o RDF encerra um padrão de ontologias, para a descrição de qualquer tipo de recurso Internet, como um *site Web* e seu conteúdo. O RDF define um padrão de metadados para ser embutido na codificação XML, e sua implementação é exemplificada pelo RDF Schema, ou RDFS, que faz parte da especificação do padrão.

XML - *Extensible Markup Language*: De acordo com Souza e Alvarenga (2004) o XML é uma recomendação formal do W3C e, em determinados aspectos, assemelha-se ao HTML *HyperText Markup Language* (Linguagem de Marcação em Hipertexto), linguagem ainda muito utilizada para construção da maioria das páginas Web, pois ambas contêm *tags* para descrever o conteúdo de um documento. Entretanto, o XML foca na descrição dos dados de um documento. Além disso, é flexível, pois permite que *tags* sejam acrescentadas à medida que forem necessárias, ou seja, qualquer comunidade de desenvolvedores pode criar suas marcações (*tags*) específicas para atender as necessidades de descrição de seus dados. O que possibilita maior significado a descrição dos dados, e, consequentemente, abri caminho para a inserção semântica em documentos da *World Wide Web* e nas intranets. Na figura 2.4, é possível observar um modelo de código XML. Nas segunda e terceira linhas de código, têm-se a referência aos namespaces utilizados pelo documento XML e o namespace do padrão RDF, respectivamente. Uma vez especificado um namespace, é possível utilizar seus descritores de forma não-ambígua ao longo do documento, fazendo sempre referência

```
<?xml version="1.0"?>
<rdf:RDF
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:dc="http://purl.oclc.org/DC#"
    xmlns:v="http://www.w3.org/2001/vcard-rdf/3.0#">
<rdf:Description about=" http://www.ucla.edu/~einstein"/>
  <dc:Creator>
      <rdf:Description about=" http://www.ucla.edu/staff/einstein"/>
        <v:Name> Isaac Einstein</v:Name>
        <v:Email="einstein@ucla.edu"/>
        <v:Orgname>UCLA</v:Orgname>
        <v:Orgunit>Department of Physics</v:Orgunit>
  </dc:Creator>
</rdf:Description>
</rdf:RDF>
```

Figura 2.4: Exemplo de código XML. Fonte: (SOUZA; ALVARENGA, 2004).

a qual está sendo utilizado.

OWL - *Web Ontology Language*: É uma linguagem para construção de ontologias que de acordo com Breitman (2005) tem a intenção de representar conceitos e seus relacionamentos na forma de uma ontologia. A OWL possui três linguagens, em ordem crescente de expressividade: OWL Lite, OWL DL e OWL Full.

Para dar subsídio a classificação de conteúdo proposta pela web semântica, é necessário a aplicação de padrões e anaálise dos conteúdos através da utilização de sistemas de representação do conhecimento e técnicas de mineração de textos, detalhados nas secções 2.2 e 2.3, respectivamente.

2.2 Mineração: Dos Dados ao Texto

Nessa secção serão apresentados o contexto histórico do surgimento da mineração de dados e da mineração de textos e suas respectivas técnicas. Tendo como foco principal a mineração de textos devido a aplicabilidade nessa dissertação de mestrado.

2.2.1 *Mineração de Dados.*

De acordo com Romao (2002) em resposta á necessidade de novas abordagens e soluções para viabilizar a análise de grandes bancos de dados surgiu a descoberta de conhecimento em banco dados - KDD (*Knowledge Discovery in Database*). O termo foi apresentado pela primeira vez em 1989 no primeiro *workshop* de KDD onde foi ressaltado o conhecimento como o produto final do processo de descoberta em banco de dados. De acordo com Fayyad, Piatetsky-Shapiro e Smyth (1996) após esse *workshop* a KDD se popularizou no campo da Inteligência artificial e aprendizagem de máquinas.

Fayyad, Piatetsky-Shapiro e Smyth (1996) define KDD como o processo geral de descoberta de conhecimento a partir de dados, incluindo como os dados são armazenados e acessados, como os algoritmos podem ser escalados para grandes conjuntos de dados, como os resultados podem ser interpretados e visualizados e como a interação homem-máquina pode ser modelada e suportada. Já a mineração de dados é uma etapa do processo, como mostra a figura 2.5.

A figura 2.5 apresenta uma visão geral do processo de KDD que tem como entrada uma base de dados e como saída o conhecimento. Esse processo envolve as seguintes etapas que se repetem em múltiplas iterações: Seleção, pré-processamento, transformação, mineração de dados, e por fim interpretação / validação. Por razões de embasamento teórico para o entendimento do modelo proposto nessa dissertação será focada a etapa de mineração de dados (*Data Mining*).

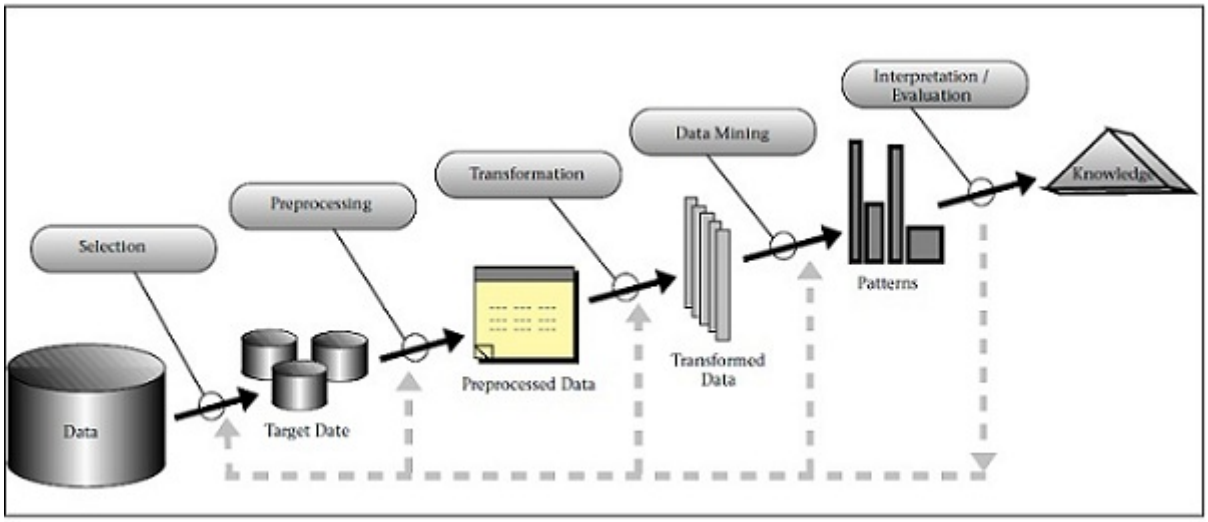

Figura 2.5: Visão geral das etapas da KDD. Fonte: (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996)

A competitividade do mundo organizacional e a nova vertente globalizada levaram as empresas a buscar uma maior qualidade dos seus produtos, precisando para isso focar

nas necessidades dos seus clientes, e consequentemente na sua satisfação. Com isso as técnicas de mineração de dados (*Data Mining*) passaram a ser utilizadas nas empresas na sumarização dos seus dados com o objetivo de separar aqueles que realmente possuíam alguma informação agregada. A análise desses dados pôde subsidiar as empresas na tomada de decisões estratégicas e, consequentemente, a tornarem-se mais competitivas.

Para Aranha e Passos (2006) a mineração de dados procura descobrir padrões em dados não estruturados, banco dados. A *Data Mining*, aplicada a análise de mercado, permite, a partir dos dados dos clientes de uma empresa X, mapear o perfil de compras desses, e assim realizar ofertas e campanhas de *marketing* personalizadas. Por exemplo, a partir da anaálise do perfil dos seus clientes, a empresa observa que 35% compram muitas fraldas descartáveis e que esses utilizam com frequência semanal o seu cartão, então a empresa pode fazer promoções com descontos em produtos para bebês visando motivar esses compradores a consumirem mais, utilizando o cartão. Em outra análise, observa-se que 45% dos clientes compram muitos ingressos de *show* e eventos em geral, de posse dessa informação, a companhia pode fazer promoções em períodos de grandes eventos, concedendo descontos nas entradas compradas com cartão ou até mesmo premiando os clientes que mais compram com os seus cartões de crédito.

De acordo com Romao (2002) a mineração de Dados pode incorporar quarenta técnicas estatísticas e/ou de Inteligência Artificial (IA) que são capazes de fornecer respostas a inúmeras questões e/ou até mesmo de descobrir novos conhecimentos em grandes base de dados. Essa é especialmente útil em casos onde não se conhece a pergunta, mas, mesmo assim, existe a necessidade de respostas.

Na figura 2.6 proposta pelo autor é possível observar os passos para obtenção de conhecimento para tomada de decisões. As diferenças entre dados, informação e conhecimento já foram abordadas no capítulo 2. Entretanto, é preciso definir decisões. Essas são especificadas a partir da análise dos indivíduos que detêm o conhecimento sobre o negócio ao qual os dados estão relacionados.

Como o gráfico ilustrado na figura 2.6 mostra, as decisões são o item mais importante. Isso deve-se ao fato de ações estratégicas serem tomadas a partir da leitura das definições dessas decisões, ou seja, a partir dessas é possível para a empresa definir qual a melhor ação para aumentar a sua competitividade no mercado, aumentar a satisfação dos clientes e, até mesmo, aumentar os seus lucros.

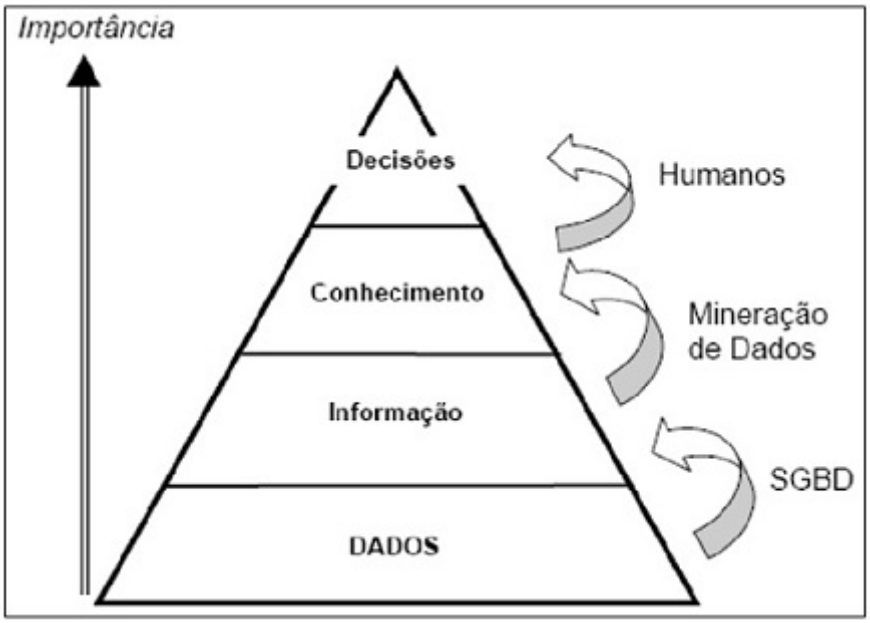

Figura 2.6: Pirâmide da obtenção do conhecimento. Fonte: (ROMAO, 2002)

2.2.2 *Mineração de Texto.*

A informação cada vez mais é registrada diretamente em meios digitais. Vivencia-se uma consolidação, não só da convergência digital, mas também da criação de conteúdos totalmente já digitalizados. Neste contexto, "a publicação e criação de conteúdos tornam-se mais fáceis e, por consequente, informações irrelevantes, de baixa qualidade e mesmo de baixa confiabilidade fazem parte de um "lixo informacional"crescente e que preocupa toda a sociedade."(MAIA, 2008)

"Estudos apontam que no ano de 2007 existiam aproximadamente 550 bilhões de documentos on-line, com aproximadamente 7,5 petabytes entre websites e base de dados on-line."(FRIZO, 2007). "Para armazenar 7,5 petabytes, em uma pilha de páginas de papel, onde cada página conteria 2500 caracteres, sendo que um byte equivale a 1 caractere, teríamos uma pilha de 300.000 km (1 cm para 100 páginas) o que daria para alcançar a lua ou dar a volta na terra 7,5 vezes. Uma pessoa lendo uma página por minuto gastaria 5.7 bilhões de anos para ler tudo."(FRIZO, 2007). "Com todo este volume informacional algum tipo de Sistema de Recuperação de Informação - SRI deverá ser necessário para que uma pessoa recupere rapidamente a informação que deseja em tempo satisfatório."(MAIA, 2008)

Dentro desse cenário, a área de mineração de textos ou text mining também conhecida como Descoberta de conhecimento em textos (*Knowlegd Discovery in Text* - KDT), surgiu com a finalidade de tratar os dados e as informações não estruturadas considerando o alto nível de complexidade envolvida neste tipo de representação de informação. Para Barçante

(2011) a mineração de textos geralmente é usada como parte de um processo que busca dar sentido a um conjunto de textos digitais dispostos nas mais diversas formas e suportes, a exemplo de frases sem verbo, bases de dados, páginas HTML e e-mails. Além disso, tem como função submeter esse conjunto a métodos para reorganizar e explorar os seus conteúdos digitais.

A mineração de texto, assim como a Mineração de Dados, também recebe influências das áreas de Processamento de Linguagem Natural, Recuperação de Informação (*Information Retrievel*), Inteligência Artificial e Ciência Cognitiva. A conjunção do conhecimento dessas áreas fez da mineração de texto uma área própria, chamada apenas de Mineração de Texto (*Text Mining*).

De acordo com Maia e Souza (2008) o processamento de linguagem natural (PLN) também é uma subárea da Inteligência artificial e da linguística que estuda os problemas da geração e tratamento automático de línguas humanas naturais.

Barion e Lago (2008) afirma que a extração da informação é usada na área de PLN com o objetivo de transformar dados semi-estruturados ou desestruturados em dados estruturados que serão armazenados em um banco de dados. Esse processo identifica palavras dentro de conceitos específicos e deve ser realizados sobre um tipo de domínio com as informações pré-definidas do que se deseja encontrar em determinados textos.

Para Maia (2008) o PLN tem como objetivo permitir ao computador comunicar-se fazendo uso da linguagem humana, nem sempre necessariamente em todos os níveis de entendimento e/ou geração de sons, palavras, sentenças e discursos. O autor ainda descreve os níveis da seguinte forma:

✓fonológico e fonético: trata do relacionamento das palavras com os sons que produzem;

✓morfológico: trata da construção das palavras a partir das unidades de significado primitivas e do como classificá-las em categorias morfológicas;

✓sintático: trata do relacionamento das palavras entre si, cada uma assumindo seu papel estrutural nas frases, e de como as frases podem ser partes de outras, constituindo sentenças;

✓semântico: trata do relacionamento das palavras com seus significados e de como eles são combinados para formar os significados das sentenças;

✓pragmático: trata do uso de frases e sentenças em diferentes contextos, afetando o significado.

"A maioria dos métodos de *Data Mining* são baseados em conceitos de aprendizagem de máquina, reconhecimento de padrões, estatística, classificação, clusterização, modelos gráficos."(BARION; LAGO, 2008)

Ainda de acordo com o mesmo autor, a mineração de textos consiste em um conjunto de métodos usados para navegar, organizar, encontrar e descobrir informações em bases de textos. A MT define técnicas de extração de padrões ou tendências que são aplicadas em grandes volumes de textos, escritos em linguagem natural, estruturados ou semi-estruturados com o objetivo de extrair conhecimentos úteis dessas coleções.

Para aplicação das muitas técnicas e métodos da mineração de texto foram desenvolvidos algoritmos computacionais, como: *Clustering, K-nearest neighbor e K-means*, descritos a seguir.

✓Clustering - De acordo com Frizo (2007), o processo de clusterização são técnicas que permitem segmentar um conjunto de objetos em grupos (cluster), tendo como objetivo que esses sejam o mais homogêneo em si e o mais heterogêneo entre si. Maia e Souza (2008) afirma que o uso de clustering para agrupar documentos envolve calcular a distância entre estes na matriz e que para isso usa-se além do co-seno de similaridade outras medidas, sendo que a distância Euclidiana é também muito utilizada. A distância Euclidiana entre dois documentos d1 e d2 é definida pela fórmula da figura 2.7

$$d(\vec{d_1}, \vec{d_2}) = \sqrt{\sum_i (w_{i,1} - w_{i,2})^2}$$

Figura 2.7: Cálculo da distância Euclidiana. Fonte: (FRIZO, 2007)

Esse algoritmo funciona basicamente através de duas formas (FRIZO, 2007):

Número de agrupamentos automático - o número de categorias é definido automaticamente, geralmente com base no número de documentos da coleção.

Número clusters pré-definido - Como o próprio nome da forma, a quantidade de cluster já é pré-definida e as categorias já se encontram definidas antes da execução do algoritmo.

✓kNN (k Nearest Neighboor) - Em uma tradução literal do inglês significa "os vizinhos mais próximos". Esse algoritmo tem como objetivo filtrar os documentos baseado na predominância dos k vizinhos mais próximos. Os vizinhos mais próximos são os documentos

que possuem maior valor de similaridade. O algoritmo é ilustrado através da equação da figura 2.8.

$$S_{c_i,d} = \sum_{d' \in N_k(d)} similaridade\,(d,d') f(c_i,d')$$

Figura 2.8: Equação KNN Fonte: (FRIZO, 2007)

Onde:

K é igual ao número de vizinhos;

Nk(d) corresponde aos documentos mais similares a k;

f(ci, d') corresponde a uma função binária que retorna se o documento **d'**pertence a uma dada categoria, representada por **ci**, ou não.

✓K-means - Em uma tradução literal do inglês significa "K-médias". Esse algoritmo tem como objetivo agrupar objetos de acordo com os seus próprios dados baseando-se em análise e comparações entre os valores numéricos dos dados fornecidos. O algoritmo analisa todas as instâncias fornecidas e as agrupa, isto é, ocorre a indicação de uma classe (*cluster*) e a determinação de que linhas pertencem a essa classe. O usuário fornece ao algoritmo a quantidade de classes que ele deseja (k).

O algoritmo *K-means* consiste nos seguintes passos:

1.Escolher k pontos como centróides;

2.Associar cada instância ao centróide mais próximo. Geralmente utiliza-se a distância euclidiana para calcular o quão 'longe' uma ocorrência esteja da outra;

3.Atualizar o centróide;

4.Repetir os passos até que nenhum ponto mude de centróide ou um número máximo de iterações seja executado.

2.2.3 *O processo de Mineração de Texto*

O processo de mineração de texto é composto por três etapas, pré-processamento, processamento e pós-processamento. Cada etapa possui diversas fases, de acordo com as técnicas e métodos de mineração e recuperação da informação utilizadas, mas algumas técnicas de PLN como *Stemming, Stop Words* e Tokenização são fundamentais no processo de mineração em coleções de documentos.

De acordo com a figura 2.9 no pré-processamento a coleção de documentos é carregada, processada e transformada numa representação numérica dos documentos que é denominada de *Bag Of Words* (BOW), traduzindo literalmente, bolsa de palavras. No processamento são aplicados métodos de mineração sobre a BOW com o objetivo de classificar e indexar os termos visando facilitar a extração do conhecimento no pós- processamento.

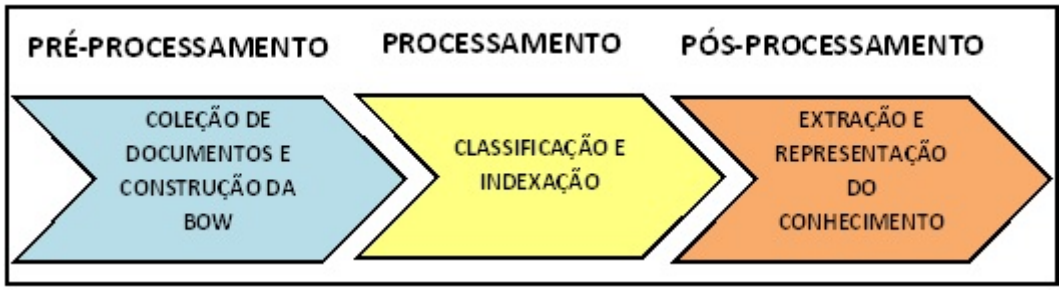

Figura 2.9: Etapas do processo de mineração. Fonte: Autor

1. Pré-processamento

Na etapa de pré-processamento há a entrada de uma coleção de documentos e essa passa por fases até gerar a BOW. Para Pires (2008) o *text mining* utiliza a BOW para representar um conjunto de documentos, com seus termos e frequência dos mesmos dentro de cada documento. Para essa representação, é utilizada uma matriz onde nas linhas da primeira coluna são dispostos identificadores de cada texto, na primeira linha são dispostos os termos e para cada par de termo e texto são associados valores que representam as frequências dos termos em cada documento. A figura 2.10 mostra um exemplo da matriz que representa a BOW.

De acordo com Pires (2008) para gerar uma BOW são necessárias quatro etapas: leitura e conversão, extração e limpeza dos termos, contagem de termos e cálculo de frequência, descritas a seguir.

Leitura Nessa etapa é definida uma coleção de documentos e cada um terá seu conteúdo carregado na memória e seguirá pelas etapas seguintes.

	Termo 1	Termo 2	Termo 3	Termo ...	Termo j
Documento 1	d_1t_1	d_1t_2	d_1t_3	$d_1t_{...}$	d_1t_j
Documento 2	d_2t_1	d_2t_2	d_2t_3	$d_2t_{...}$	d_2t_j
Documento 3	d_3t_1	d_3t_2	d_3t_3	d_3t	d_3t_j
Documento ...	$d_{...}t_1$	$d_{...}t_2$	$d_{...}t_3$	$d_{...}t_{...}$	$d_{...}t_j$
Documento i	d_it_1	d_it_2	d_it_3	$d_it_{...}$	d_it_j

Figura 2.10: Representação da matriz da BOW. Fonte: (PIRES, 2008)

Extração e Limpeza dos termos Cada documento da coleção terá o seu conteúdo dividido em termos, ou seja, cada palavra significante presente no documento. Essa etapa é composta por três sub-etapas,*tokenização*, limpeza e *Stemming*.

✓Tokenização

Na tokenização os termos considerados irrelevantes como espaços excessivos em branco entre os termos, quebras de linhas, tabulações, e alguns caracteres especiais, são removidos, conforme o exemplo abaixo.

Exemplo

O RDF possui três princípios básicos, recursos, propriedades e frases.

Após *StopWords* O RDF possui três princípios básicos recursos propriedades e frases

✓Limpeza

Após a tokenização cada termo obtido passa pela etapa de limpeza, onde são removidas as *StopWords*, verificada a existência do sinônimo do termo no dicionário e por último é realizado o *Stemming* do referido termo.

De acordo com Barion e Lago (2008) as StopWords geralmente são preposições, artigos, conjunções, alguns verbos, nomes adjetivos, advérbios. Para o processo de remoção dessas, deve ser criada uma lista, denominada *Stop-List*, no idioma em que se está trabalhando, contendo estas palavras consideradas irrelevantes. Com isso, há uma diminuição do tamanho das estruturas de indexação, facilitando a mineração.

Exemplo

O RDF possui três princípios básicos, recursos, propriedades e frases.

Após *StopWords* RDF possui três princípios básicos recursos propriedades frases

✓Stemming

Nessa fase cada termo é reduzido ao seu radical. Com sua utilização, os termos derivados de um mesmo radical serão contabilizados como um único termo. Para com Barion e Lago (2008) o processo de *Stemming* melhora o armazenamento por eliminar a quantidade de termos a serem armazenados.

Exemplo

O RDF possui três princípios básicos, recursos, propriedades e frases.

Após Stemming: RDF pos três princípios bás recursos propri frases

Contagem dos termos Depois de todo o processo de extração e limpeza dos termos, será calculado o número de ocorrências de cada termo num documento. Depois de concluída a contagem é criada uma lista com duas colunas: termo e quantidade de ocorrência.

Cálculo da Frequência Após concluída a etapa de contagem de termos para cada documento da coleção, será calculada a frequência dos termos utilizando a fórmula da figura 2.11. Onde N é o número total de documentos do conjunto e df (ti) é o número de documentos onde o termo ti aparece, ou seja, ni diferente de 0.

$$idf(t_i) = \log \frac{N}{df(t_i)}$$

Figura 2.11: Fórmula Frequência dos termos por documento. Fonte: (PIRES, 2008)

A figura 2.12 mostra a aplicação da formula descrita acima, onde é calculada o peso do mesmo termo com frequências iguais em coleções diferentes. No exemplo 1, o termo *mining* aparece 7 vezes em um único documento, já no exemplo 2, o mesmo termo aparece 7 vezes mas em 10 documentos. Através desses exemplos é possível concluir que para o termo ser representativo para o documento, é necessário que o termo tenha um número alto de ocorrência no documento e um número baixo de ocorrência dentro da coleção.

2. Processamento

Nessa fase são aplicados métodos de mineração sobre os resultados obtidos na fase de pré-processamento. Dentre eles, podem ser aplicados os algoritmos clustering, k-means, e k-NN, descritos na seção 2.2.2. Esses foram escolhidos por terem sido utilizados no desenvolvimento do modelo proposto nessa dissertação.

Para Souza e Lindgren (2008) essa fase é muito importante, principalmente porque provê um ponto de integração com outros sistemas de mineração de dados existentes, já que nessa

Exemplo 1: Termo = "*mining*"	Exemplo 2: Termo = "*mining*"
Nº. de Documentos = 1	Nº. de Documentos = 10
Freqüência = 7 vezes	Freqüência = 7 vezes
Peso = 7 * *log2*(7/1) + 1 = 21	*Peso* = 7 * *log2*(7/10) + 1 = 5,35

Figura 2.12: Exemplo da aplicação da fórmula do cálculo da frequência. Fonte: (BARION; LAGO, 2008)

etapa existe uma visão numérica consistente de documentos. O que permite explicar ou descobrir padrões existentes em outras bases numéricas, sustentando assim um processo efetivo de Mineração de Texto.

De acordo com Goncalves (2010) as operações de mineração aplicadas nessa fase formam o núcleo de uma aplicação de mineração de textos e têm como enfoque a extração de padrões, a identificação de tendências e a aquisição de novos conhecimentos e informações previamente desconhecidos, a partir do texto pré-processado.

Ainda de acordo com o mesmo autor, nessa fase não é realizada apenas a extração de padrões ou conhecimento, mas também a comparação entre resultados, análise da distribuição e proporção de conceitos em um documento ou coleção de documentos, além disso, é realizada a avaliação do nível de relevância ou interesse dos resultados encontrados para os objetivos finais do usuário.

3. Pós-processamento

Para Souza e Lindgren (2008) o processo entra na fase de avaliação e interpretação dos resultados que envolve todos os participantes. O analista de dados tenta descobrir se o classificador atingiu as expectativas através da avaliação dos resultados fazendo uso de algumas métricas, como: taxa de erro, tempo de CPU e complexidade do modelo. O especialista no domínio irá verificar a compatibilidade dos resultados com o conhecimento disponível do domínio. E, por fim, o usuário é responsável por dar julgamento final sobre a aplicabilidade dos resultados.

De acordo com Goncalves (2010) o pós-processamento envolve as tarefas de refinar e tornar coerente os resultados da mineração para que o conhecimento extraído seja utilizado de

forma eficaz pelo usuário executor do processo.

Para verificar a compatibilidade dos resultados com o conhecimento disponível em um dado domínio, os especialistas podem fazer uso dos sistemas de representação do conhecimento, onde é possível comparar os termos minerados da coleção de documentos e as suas relações com ontologias do mesmo domínio. Além disso, os resultados da mineração de texto podem contribuir para construção de ontologias como foi desenvolvido no modelo dessa dissertação de mestrado.

Uma ontologia, segundo Gruber citado por Breitman (2005), é uma especificação explícita dos objetos, conceitos e outras entidades que se assume existirem em uma área de interesse, além das relações entre estes conceitos e restrições, expressos através de axiomas. Na próxima secção, será discutido com mais detalhes o termo ontologia.

2.3 Ontologia

Essa secção discuti o termo ontologia. Desde a sua definição na Grécia Antiga até a sua aplicação na ciência da computação, onde serão descritas algumas ferramentas e linguagens para a construção de ontologias.

2.3.1 Breve Histórico

O termo ontologia é utilizado desde a Grécia Antiga, onde era voltado ao estudo do ser e das suas relações. De acordo com Almeida e Bax (2003) o termo original é a palavra aristotélica "categoria", que pode ser usada para classificar alguma coisa. Aristóteles apresenta categorias que servem de base para classificar qualquer entidade e introduz ainda o termo "differentia"para propriedades que distinguem diferentes espécies do mesmo gênero.

Nos dias de hoje, as ontologias estão relacionadas a representação do conhecimento e, para isso a sua construção está fortemente associada a ciência da informação.

Para Mucheroni, Paiva e Netto (2009) na filosofia o termo ontologia foi empregado como referente ao conhecer o que era o ser, mas o ser em geral, tanto a sua razão como o seu logos e assim estreitamente relacionado a lógica, enquanto o sentido ôntico ligado ao ente deve ser pensado como o ser de fato e, portanto, não necessariamente ligado ao logos.

"Definir ontologias é classificar em categorias aquilo que existe em um mesmo domínio do

conhecimento."(MAEDCHE; STAAB, 2001). "A palavra ontologia vem do grego ontos (ser) + logos (palavra)."(BREITMAN, 2005). Existem várias definições para o termo especificado por diferentes autores. Algumas serão tratadas nesse capítulo.

Uma ontologia, segundo (Gruber 1995 apud (BREITMAN, 2005)), é uma especificação explícita dos objetos, conceitos e outras entidades que se assume existirem em uma área de interesse, além das relações entre estes conceitos e restrições, expressos através de axiomas.

Ainda de acordo com o mesmo autor, ontologia é uma especificação formal e explícita de uma conceitualização compartilhada. Esta definição irá nortear esse trabalho, pois de acordo com Almeida e Bax (2003) é uma das mais conhecidas.

Freitas (2003) esclarece os termos da definição de Gruber, conforme desposto abaixo:

✓Especificação explícita - refere-se as definições de conceitos, instâncias, relações, restrições e axiomas.

✓Formal - É declarativamente definida, portanto, compreensível para agentes e sistemas.

✓Conceitualização - Refere-se a um modelo abstrato de uma área de conhecimento ou de um universo limitado de discurso.

✓Compartilhada - Trata-se de um conhecimento consensual, seja uma terminologia comum da área modelada, ou acordada entre os desenvolvedores dos agentes que se comunicam.

Para o W3C uma ontologia é a definição dos termos utilizados na descrição e na representação de uma área do conhecimento.

De acordo com Breitman (2005), foi proposto por Nicola Guarino uma classificação para ontologias a partir da sua generalidade, com isso têm-se quatro tipos:

Ontologias de nível superior - são independentes de domínio, podendo ser reutilizadas na confecção de novas ontologias, pois descrevem conceitos muito genéricos, tais como espaço, tempo e eventos.

Ontologias de domínio - descrevem um vocabulário relativo a um domínio específico através da especialização de conceitos de ontologias de alto nível.

Ontologias de tarefas - descrevem um vocabulário relativo a uma tarefa genérica através

da especialização de conceitos de ontologias de alto nível.

Ontologias de aplicação - são as mais específicas, correspondem, de forma geral, a papéis desempenhados por entidades do domínio no desenrolar de alguma tarefa.

Uma Ontologia de Domínio, escrita em uma linguagem formal, possui os elementos descritos abaixo: (BREITMAN, 2005)

✓Classes: Também denominadas de conceitos. São elementos utilizados para definir as sub-áreas ou subgrupos de um dado domńio. Por exemplo: Pessoa, Imóveis ou Médico;

✓Propriedades (Relacionamentos): São formas de interligar as classes diretamente ou a classe ao seu atributo. Por exemplo, criação e criador pode ser definido como "criação é criado por criador". Já no caso de interligar conceitos a atributo, tem-se animal e o atributo reino que pode ser relacionado como "Todo animal pertence a um reino.";

✓Instâncias: São utilizados para representar a unidade materializada de uma classe, como um carro específico que possui uma placa e um chassi que identifica-o de forma única, pois cada carro tem uma placa diferente e um chassi;

✓Axiomas: São as sentenças que são sempre verdadeiras, ou seja, determinam verdade sobre um dado domínio. Por exemplo: "Todo animal mamífero tem sangue quente."

2.3.2 Linguagens para construção de ontologias

De acordo com Breitman (2005) foram propostas algumas linguagens para ontologias baseadas no RDFS - *Resource Description Framework Schema*, como a OIL, DAML, DAML+OIL e OWL, descritas a seguir.

RDF - Resource Description Framework
Proposto em fevereiro de 1999 pelo W3C, o RDF está projetado para fornecer a interoperabilidade e a semântica para metadados visando facilitar a busca por recursos na Web. Possui três princípios básicos, recursos, propriedades e frases.

Recursos são objetos ou "coisas"das quais se quer falar, como pessoas e lugares. As propriedades descrevem os relacionamentos entre os recursos, por exemplo "livro escrito por uma pessoa". Ambos podem ser representados por uma URI (*Universal Resource Identificator*) que pode ser uma URL (*Unified Resource Locator*), endereço na Web.

RDF Schema - Resource Description Framework Schema

O RDF Schema é uma linguagem que descreve propriedades e classes para os recursos do RDF. Para isso, fornece um *framework* no qual é possível descrever as classes e propriedades. Sendo as classes muito parecidas com o conceito de orientação a objetos, com isso o RDF Schema também permitir a construção de hierarquias de classes e de propriedades.

As classes permitem que os recursos sejam definidos como instâncias ou subclasses das classes presentes no RDF Schema. Há cinco tipo de classes essenciais.

✓rdfs: Resource: A classe de todos os recursos;
✓rdfs: Class: A classe de todas as classes;
✓rdfs: Literal: A classe de todos os literais (cadeias de caracteres);
✓rdfs: Property: A classe de todas as propriedades;
✓rdfs: Statement: A classe de todas as sentenças reificadas.

OIL - Ontology Inference Layer
Essa linguagem foi criada com a proposta de permitir a modelagem de ontologias na Web já que o RDF não provê a semântica necessária nem formalismo suficiente para permitir suporte a mecanismos de inferência.

A OIL é uma linguagem baseada em frames que utilizam lógica de descrição para fornecer uma semântica clara, ao mesmo tempo em que permitem implementações eficientes de mecanismos de inferência que garantam a consistência da linguagem. (Gómez-Pérez apud (BREITMAN, 2005))

DAML - DARPA Agent Markup Language
Desenvolvida pela DARPA (*Defense Advanced Research Projects Agency*), antiga ARPAnet, tem como objetivo facilitar a interação de agentes de *software* autônomos na Web. Sua primeira especificação para uma linguagem de ontologias foi lançada em outubro de 2000.

A linguagem DAML herdou muitos aspectos presentes em OIL. A exemplo do suporte a hierarquias de conceitos e propriedades, baseadas nos relacionamentos de subclasses e subpropriedades e da possibilidade da construção de conceitos a partir de outros conceitos utilizando combinações dos conectivos OR, AND e NOT.

OWL - Web Ontology Language

"A OWL é uma linguagem para representar conceitos e os seus relacionamentos na forma de ontologia."(BREITMAN, 2005). Foi definida pelo W3C como uma revisão da lingua-

gem DAML+OIL e começou como uma W3C *recomendation* em fevereiro de 2004.

De acordo com Oliveira (2009) a OWL permite a representação de conceitos através de classes e propriedades, podendo cada uma herdar características de outras classes e propriedades, respcetivamente.

De acordo com Breitman (2005) a OWL é dividida em três sub-linguagens, descritas abaixo:

OWL Lite
A OWL-Lite é a sub-linguagem sintaticamente mais simples. Destina-se a situações em que apenas são necessárias restrições simples como as de cardinalidade de valor 0 ou 1 e uma hierarquia de classe simples. A OWL-Lite tem uma menor complexidade formal que a OWL DL.

OWL-DL
A OWL-DL é mais expressiva que a OWL-Lite e baseia-se em lógica descritiva, campo de pesquisa que estudou a lógica que constitui a base formal da OWL. Essa base derivou o nome da sublinguagem. Com essa sublinguagem é possível computar automaticamente a hierarquia de classes e verificar inconsistências na ontologia.

OWL-Full
OWL-Full é a sublinguagem OWL mais expressiva. Destina-se a situações onde alta expressividade é mais importante do garantir a decidabilidade ou completeza da linguagem. Não é possível efetuar inferências em ontologias OWL-Full.

Apesar de não existir uma regra obrigatória para nomear classes OWL, utiliza-se as boas práticas de orientação a objetos para nomear as classes, ou seja, recomenda-se que todos os nomes de classes iniciem com letra maiúscula e não contenham espaços. Por exemplo: Pessoa, PessoaFisica, PessoaJuridica. Pode-se também usar o *underscore* para juntar palavras. A regra é importante para a consistência da ontologia.

Para Breitman (2005) a OWL é formada por seis elementos básicos: Namespaces, cabeçalhos, classes, indivíduos, propriedades e restrições, descritas a seguir:

✓Namespaces: São declarações que se localizam entre etiquetas rdf:RDF que possibilitam aos identificadores, presentes na ontologia, serem interpretados sem ambiguidade. Uma ontologia típica em OWL começa com um conjunto de declarações de *namespaces*.

✓Cabeçalhos: São as sentenças representadas pelas *tags* (etiquetas) owl:Ontology. Essas *tags* são responsáveis pelo registro de comentários, controle de versão e pela inclusão de

conceitos e propriedade de outras ontologias.

✓Classes: Uma classe representa um conjunto ou coleção de indivíduos (objetos, pessoas, coisas) que compartilham de um grupo de características que os distinguem dos demais.Elas são utilizadas para descrever os mais básicos de um dado domínio, que vão servir como raízes de várias taxonomias.

Todos os indivíduos em uma ontologia OWL pertence a uma classe genérica denominada owl: thing, ou seja, toda classe definida em OWL é uma subclasse de owl: thing. O que permiti que exista sempre uma única raiz para qualquer taxonomia.

✓Indivíduos: São os membros das classes. Em OWL um indivíduo é adicionado se declararmos que o mesmo é membro de uma classe.

✓Propriedades: Descrevem fatos em geral. Podem referir-se a todos os membros de uma classe, "Toda pessoa come"ou identificar um indivíduo dessa classe, "A pessoa Keller nasceu em 1982".

Existem dois tipos de propriedades:

Propriedades do tipo *object*: Relacionamento entre duas classes.
Propriedades do tipo datatype: indicam relacionamento entre instâncias de classes.

✓Restrições: As restrições são utilizadas para definir alguns limites para indivíduos que pertencem a uma classe.

2.3.3 Ferramentas para construção de ontologias

Atualmente, existem no mercado muitas ferramentas para construção de ontologias. Dentre elas, tem-se o ambiente Protégé 2000 e OilEd que serão descritas a seguir.

Protégé 2000

"Em seu projeto original, o Protégé era uma ferramenta de aquisição de conhecimento limitada a um sistema especialista para ontologia."(FREITAS, 2003)

O Protégé é uma ferramenta de software livre, desenvolvida em java que permite criar e editar ontologias e bases de conhecimento. De acordo com Freitas (2003) o protégé possui uma arquitetura aberta e graças a isso, componentes de vários matizes, elaborados por grupos de pesquisa de usuários, puderam ser adicionados ao sistema, sem necessitar do

desenvolvimento, como por exemplo, o Ontoviz, um componente gerador de gráficos com instâncias, heranças e outros tipos de relacionamento. As figuras 2.13 e 2.14, retiradas do site oficial do Protégé, mostram o Protégé o OntoViz, respectivamente.

Atualmente, o OntoViz em uma nova versão é denominado OWLViz.

A figura 2.13 apresenta a tela do Protégé onde a direita é possível ver a hierarquia de conceitos, na parte superior os usuários podem acrescentar comentários e na parte inferior vê-se as características lógicas da classe selecionada. Na figura 2.14 observa-se uma dos mais populares *plug-in*, de acordo com o *site* oficial da Protégé, que permiti visualizar graficamente ontologias OWL.

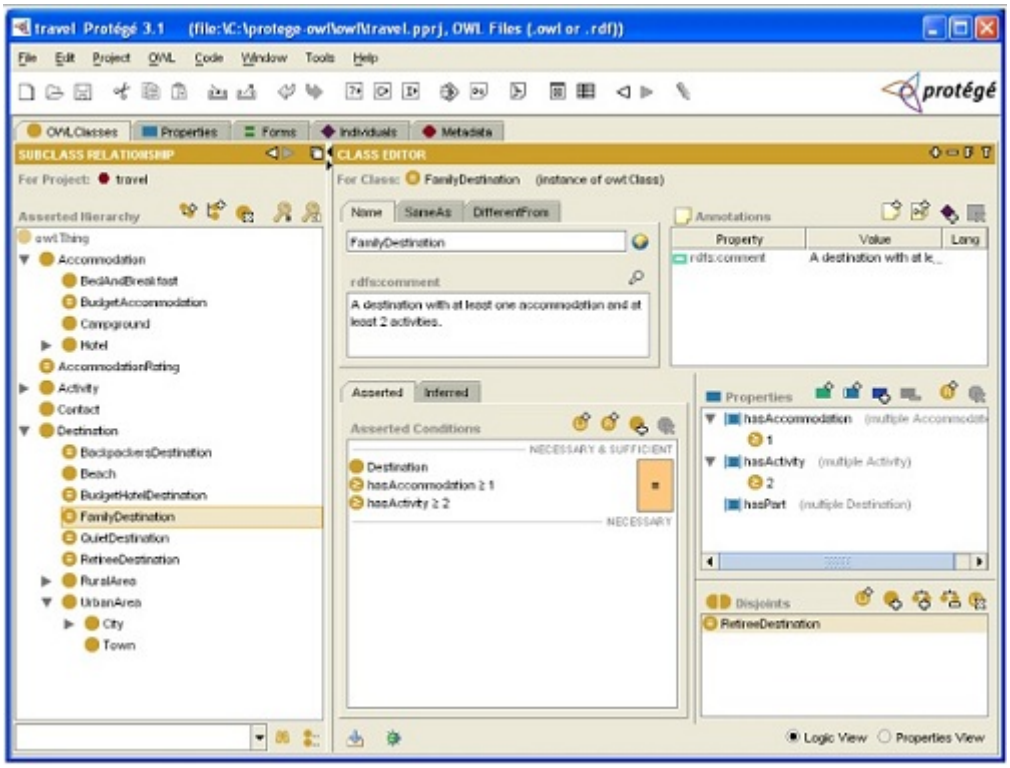

Figura 2.13: Hierarquia de heranças representada no Protégé. Fonte: (PROTEGE, 2012)

OilEd

É um editor de ontologias que permite contrui-lás utilizando a linguagem OIL, descrita na secção 2.3.2. De acordo com Breitman (2005) a OilEd é o "NotePad"dos editores de ontologias. Ela oferece suporte a ontologias desenvolvidas nas linguagens DAML+OIL e OWL.

Após as descrições dos conceitos que norteiam esse trabalho realizado neste capítulo, algumas referências foram selecionadas devido a maior relação com esta dissertação. Dessa forma, no próximo capítulo, será descrito quatro trabalhos que envolvem o estudo da

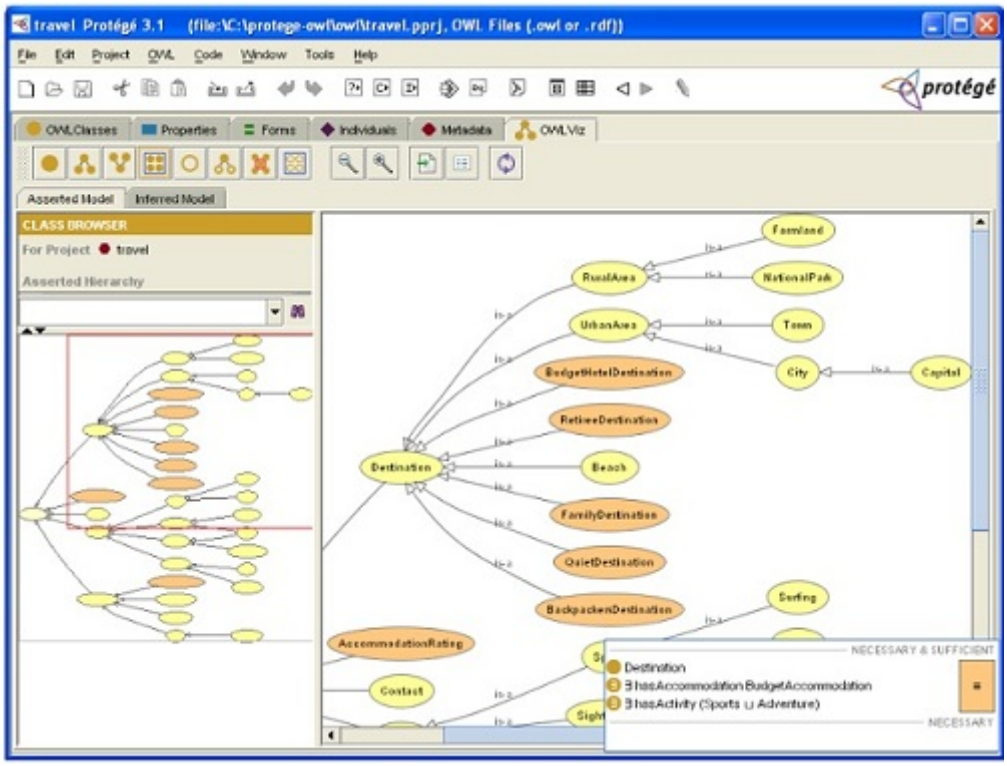

Figura 2.14: Hierarquia de heranças representada graficamente no OWLViz.Fonte:(PROTEGE, 2012)

construção de ontologias.

Capítulo Três

Trabalhos Relacionados

Neste capítulo serão descritos alguns trabalhos encontrados que possuem maior proximidade com a temática deste trabalho de dissertação, observados ao longo do levantamento bibliográfico.

3.1 Trabalhos Relacionados

Muitos trabalhos foram identificados acerca de ontologias e a sua construção a partir da aplicação de técnicas de mineração de textos e gestão do conhecimento, como é possível observar na secção referências. Entretanto, os trabalhos de (BARçANTE, 2011), (BASÉGIO, 2007), (GONCALVES, 2008) e (ALMEIDA, 2003) foram selecionados para ser descritos nesta secção devido a maior proximidade com a problemática deste trabalho.

Propostas e metodologias de processamento automático de documentos textuais digitais: uma análise da literatura (BARçANTE, 2011)

Este trabalho de dissertação de mestrado objetivou investigar a capacidade de identificar e analisar métodos de extrair automaticamente semânticas específicas a partir de textos digitais com objetivo de reutilizá-los para outros fins diferente dos quais estes foram inicialmente produzidos. Para tanto, foram levantados e classificados artigos científicos buscando responder as seguintes questões: Em que conjunto de dados textuais o método descrito no artigo foi aplicado? E como foi especificada a semântica a ser buscada no conjunto de dados textuais? Após a referida análise, para cada texto identificado no levantamento emergiram as seguintes classes de métodos: Mineração de textos, Anotação Semântica, Análise Semântica, Análise em Linguagem Natural e Tratamento Estatístico de textos.

Com base no objetivo descrito no paragráfo supracitado, definiu-se o tema da coleção de documentos, métodos de estruturação de textos digitais. Após essa definição, foram pesquisadas propostas, experiências, projetos, entre outros, que identificassem métodos de estruturação de textos digitais. As fontes utilizadas foram artigos de periódicos nacionais e estrangeiros, trabalhos em eventos, pré-prints e documentos armazenados em repositórios das áreas de Ciência da Informação, Ciência da Computação, portal Capes, *Google Scholar*, *Citeseer*, entre outros. Tendo em vista que a questão levantada era bastante específica, temas correlatos foram também levantados: formatos textuais digitais, Web Semântica, linguagens de marcação, metadados, mineração de textos, anotação de textos, anotação

semântica, nos idiomas português e inglês. Após a coleção devidamente definida foi realizada uma análise dos dados e fichamento dos textos para só então identificar as diferentes propostas e metodologias para processamento automático do conteúdo de documentos textuais.

Com o resultado da análise efetuada, é proposto um esquema classificatório, a partir de um conjunto de critérios, com o qual são classificados/agregados os diferentes projetos, experiências, propostas e metodologias para processamento automático do conteúdo de documentos textuais encontrados na literatura. A partir dessa análise, foi possível identificar para cada texto analisado as seguintes classes de métodos: Mineração de textos, Anotação Semântica, Análise Semântica, Análise em Linguagem Natural e Tratamento Estatístico de textos. A partir dessa identificação passou-se a descrever as características gerais de cada método, cada uma das classes e a discussão e comentários dos artigos que se enquadram em cada uma das classes.

Por fim, as análises dos documentos com base nos métodos aplicados em cada texto sugerem que a combinação dos métodos mineração de texto, Tratamento Estatístico de texto e anotaccão semântica podem colaborar na obtenção de resultados mais significativos em trabalhos de pesquisa que façam uso de métodos similares aos identificados e agrupados no presente estudo.

Uma Abordagem semi-automática para Identificação de Estruturas Ontológicas a partir de Textos na Língua Portuguesa do Brasil (BASÉGIO, 2007).

Este trabalho propõe a construção de estruturas ontológicas a partir de textos na língua portuguesa do Brasil. Para isso são contruídos uma aborgagem constituída de etapas apresentadas na **Figura** 3.1 e o protótipo para construção de ontologias, descritos a seguir. Por fim, é realizada uma validação através de dois estudos de caso onde é abordado o domínio do turismo.

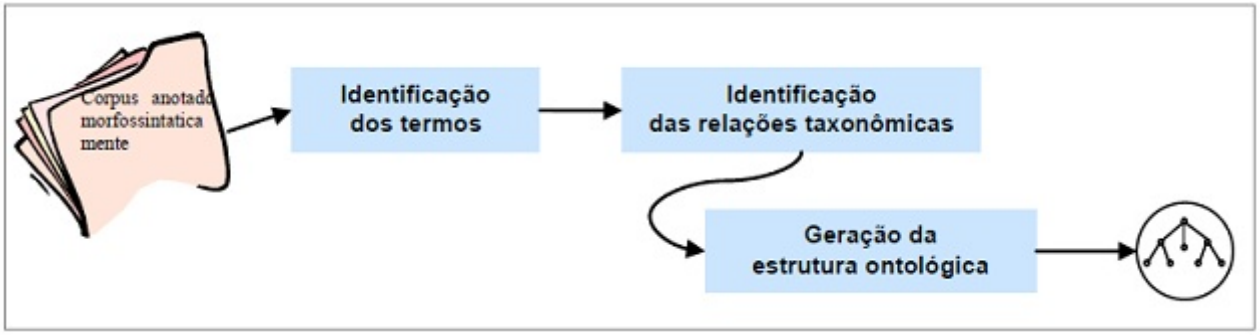

Figura 3.1: Visão simplificada da abordagem de (BASÉGIO, 2007)

De acordo com o autor, como não havia tempo hábil para desenvolver uma ferramenta

que realizasse pré-etiquetagem gramatical do corpus, o qual inclui etiquetar os textos através da tokenização, processamento léxico-morfológico e análise sintática do corpus, e não dispunha de uma ferramenta confiável que executasse o referido processo. Assumiu-se como entrada um corpus com textos já anotados linguisticamente, com as seguintes informações associadas a cada palavra do documento: A palavra no seu formato original; O lema da palavra original, ou seja, a palavra em sua forma singular e masculina e; A etiqueta gramatical da palavra (exemplo: substantivo, adjetivo, pronome, artigo definido, artigo indefinido, verbo, advérbio, entre outros). Após receber a entrada do corpus é realizada uma identificação de termos, dividida em cinco etapas, descritas a seguir:

1. Eliminar termos que não representam conceitos de domínio: Utilizando uma lista de aproximadamente 500 *stopwords* (artigos, preposições, advérbios, etc.). São excluídas do corpus palavras comuns que possuem significado semântico limitado e, portanto, não são relevantes ao domínio. Além disso, são removidos do corpus todos os termos contendo caracteres não-alfabéticos como números e símbolos. Porém o autor não desconsidera definitivamente esses termos, eles ainda podem ser utilizados em regras para identificação de termos compostos e relações taxonômicas entre esses. Em contrapartida, nessa fase há também a exclusão de nomes próprios e abreviações. E essa é definitiva.

2. Pesagem dos termos: Nessa etapa é realizado o peso das palavras candidatas a termos relevantes do domínio fazendo uso do lema das palavras já recebido como entrada, como definido anteriormente. Além disso, foram utilizadas duas medidas: TFIDF (term frequency x inverted document frequency) e Log-Likelihood. A primeira pesa os termos e apresenta-os em ordem de relevância ao engenheiro de ontologia. Já a segunda compara a frequência dos termos no corpus do domínio face a sua frequência em um corpus de referência, promovendo assim a exclusão automática de termos não relevantes ao domínio (ou seja, termos que aparecem em maior proporção no corpus de referência). O lema das palavras (disponível no corpus etiquetado) foi utilizado para evitar que um mesmo termo representado com diferentes propriedades (gênero, número e grau), receba distintos pesos como se fossem diferentes termos. Por exemplo, podem aparecer nos textos os termos "praia"e "praias". Se não fosse utilizado o lema da palavra para computar os pesos, teríamos dois termos diferentes, cada um com seu peso associado. Utilizando-se o lema, estes termos passam a ser pesados como um único termo ("praia").

3. Definição de limiar mínimo para termos: Nessa etapa é definida uma frequência (TFIDF) mínima aceitável para um termo no corpus ser considerado relevante ao domínio. Com a definição desse limiar, os termos com frequência abaixo do mesmo são excluídos. O autor orienta que essa frequência seja definida com cautela, visto que termos que aparecem poucas vezes ou apenas uma vez em um texto podem ser mais relevantes para o domínio do que termos mais frequentes.

4. Excluir e Incluir termos: Nesta etapa é permitido ao engenheiro de ontologia incluir ou excluir termos relevantes definidos previamente nas etapas anteriores. Todos os termos resultantes desta etapa serão considerados nas etapas subsequentes.

5. Identificar termos compostos: A partir da lista de termos relevantes, resultantes da execução das etapas 1 a 4, são selecionados termos compostos que contenham ao menos um termo relevante em sua composição. Nesse trabalho relacionado, a identificação dos termos compostos é realizada com base em regras expressas por sequências de etiquetas que, quando encontradas no texto, podem representar termos compostos. Para isso utiliza-se as relações expressas na **Figura** 3.2 que apresenta as sequências de etiquetas utilizadas na identificação de termos compostos. Os termos compostos resultantes desta etapa são também considerados termos relevantes do domínio e a sua validação deve ser realizada pelo engenheiro de ontologia.

Nro	Regra
1	_SU _AJ _PR _AD _SU _AJ
2	_SU _AJ _PR _AD _SU
3	_SU _PR _AD _SU _AJ
4	_SU _PR _AD _SU
5	_SU _AJ _PR _SU _AJ
6	_SU _AJ _PR _SU
7	_SU _PR _SU _AJ
8	_SU _PR _SU
9	_SU _AJ

Figura 3.2: Regras para identificação de termos compostos. Fonte: (BASÉGIO, 2007)

Após a identificação dos termos é realizada a extração de relações taxônomicas. Nessa etapa, a lista de palavras gerada na fase anterior é submetida a três abordagens diferentes para gerar as relações taxônomicas, são elas:

Identificar relações taxonômicas com base em termos compostos: Relaciona cada termo composto ao termo relevante que faz parte da sua composição. Por exemplo, se foram identificados o termo relevante "contrato"e o termo composto "contrato de venda", a idéia é relacionar taxonomicamente esses dois. Nesse exemplo, contrato exerce uma hierarquia sobre venda, pois venda é um tipo de contrato.

Identificar relações taxonômicas através dos padrões de Hearst: este segundo passo tem por objetivo a identificação de relações taxonômicas nos textos através dos padrões léxico-sintáticos propostos. A idéia é encontrar no corpus os padrões de Hearst onde exista ao menos um termo relevante envolvido. Para utilizar os padrões de Hearst o autor precisou fazer algumas adaptações visto que o método é para a linga inglesa e trabalha

com sintagma nominal (NP), informação que não estava disposta nos arquivos utilizados. Para suprir essa necessidade, a informação de NP foi substituída diretamente por um substantivo (SU). A validação das relações, definidas nessa fase, também é realizada pelo engenheiro de ontologia.

Identificar relações taxonômicas através dos padrões de Morin e Jacquemin: Esta etapa identificou relações taxonômicas através dos padrões léxico-sintáticos propostos por Morin e Jacquemin. Esse padrões foram desenvolvidos para a língua francesa, e por isso foi necessário traduzi-los/adaptá-los para a língua portuguesa do Brasil. Além disso, Morin e Jacquemin trabalham em seus padrões com informação em nível de sintagma nominal e, da mesma forma que foi feito com os padrões propostos por Hearst, oNP foi substítuido por um substantivo (SU). A validação dessas relações é realizada pelo engenheiro de ontologia.

Para subsidiar a última etapa, geração da estrutura ontológia, foi criado um protótipo, assim denominado pelo autor, para geração das estruturas ontológicas em owl a partir dos seguintes itens, gerados nas etapas anteriores: Termos simples; Termos compostos; Relações baseadas em termos compostos; Relações baseadas nos padrões de Hearst; Relações baseadas nos padrões de Morin e Jacquemin.

Para validação da abordagem, o autor aplica o protótipo em dois estudos de caso, sendo um com a validação e alteração do especialista nas listas de termos simples e compostos e o outro sem essa referida etapa de validação. Após essa aplicação o autor concluiu que solução totalmente automatizada, não levaria a um alto grau de precisão, quando se utilizando apenas técnicas estatísticas para identificação de termos e relações táxonômicas.

Construção de ontologia para suporte cognitivo a um ambiente de aprendizagem (GONCALVES, 2008)
Este trabalho objetiva construir uma ontologia para demonstrar qual a sua contribuição em um ambiente de aprendizagem para treinamentos em organizações. Para isso é elaborada uma metodologia para subsidiar a construção de uma ontologia para um dado domínio definido nesse trabalho. Após essa contrução, a ontologia é aplicada em um projeto educacional para aprendizagem a fim de validar a sua aplicabilidade e vantagens agregadas.

Para construção da ontologia proposta o autor se baseia em três metodologias, descritas, resumidamente, a seguir:

- Methontology - Proposta por pesquisadores do Laboratório de Inteligência Artificial da Universidade de Madrid auxilia na construção de uma ontologia para representação do conhecimento não estruturado. Para isso, baseia-se em representações intermediárias,

que servem de ponte entre as diferentes percepções dos indíviduos. Essa base justifica-se, pois segundo os pesquisadores diminuem a distância entre as diferentes percepções das pessoas sobre um determinado domínio.

- MDO - Modelo Processual de Desenvolvimento de Ontologias - Proposta por (BURNHAM, 2005) assim como o nome explícita , essa metodologia propõe um modelo processual para o desenho de ontologias a fim de formalizar o conhecimento. Para isso, divide-se em três etapas: Identificação dos conceitos chaves e sua relações no domínio de interesse; Produção de textos precisos e sem ambiguidades para a descrição desses conceitos e relações; Identificação de termos de referência para as descrições dos conceitos e para as relações.

- Abordagem colaborativa para construção de ontologias - Utiliza mecanismos para a construção de consenso para refletir o conhecimento e experiências de um determinado domínio. Essa abordagem divide-se em quatro etapas: Pré-definições; Ancoragem; Processo interativo; Aplicação.

Após o estudo das metodologias descritas anteriormente, são definidas as etapas que formaram a proposta de método de construção de ontologia para suporte cognitivo. De modo que ficaram definidas quatro macro-fases, descritas a seguir, subdivididas em etapas menores que utiliza itens propostos nas metodologias já estudadas, methontology, MDO e abordagem colaborativa para construção de ontologias.

Fase 1: Preparação do domínio - Nessa fase é definido o domínio de conhecimento que será detalhado e os limites da análise. Está subdividida em cinco etapas: pré-definições, análise do domínio, ancoragem, estruturação de conceitos e identificação de relações e funções e, por fim, a última etapa, progresso interativo.

Fase 2: Formalização da ontologia - Nesta etapa são criadas as tabelas de axiomas a partir das representações intermediárias definidas na etapa anterior.

Fase 3: Prototipação - Essa etapa submete a metodologia proposta a testes e tem como saída uma versão de ontologia para disponibilização de testes.

Fase 4: Fase da Análise de qualidade - O autor dessa proposta afirma que quando se trata de conhecimento, não é recomendável estabelecer um fim para a sua representação, uma vez que não existe conhecimento estático. De modo que essa fase, apesar de ser a última fase da metodologia proposta por esse trabalho, deixa claro que não tem a pretensão de afirmar que o trabalho de representação do conhecimento está concluído e sim que é necessário está em constante aperfeiçoamento. Com isso, a análise de qualidade tem por objetivo verificar os resultados dos testes da fase anterior e propor melhorias.

Após o estudo das metodologias e posterior definição do método de construção de ontologia para suporte cognitivo proposto no referido trabalho foi desenvolvida uma ontologia, seguindo as fases do método proposto, para capacitação de servidores do Ministério Público (MP) em matéria de atuação extrajudicial, mais especificamente quanto às questões relacionadas à condução de ajustamentos de conduta, por intermédio do Inquérito Civil.

O domínio do conhecimento definido para a ontologia desenvolvida foi o aprendizado de como documentar e executar os procedimentos administrativos que compõem um inquérito civil, que resulte em um Termo de Ajustamento de Conduta (TAC). Para definir os limites do domínio baseou-se na legislação que estabelece esses procedimentos. Por fim a ontologia construída a partir da metodologia proposta foi abordada em um projeto educacional que teve a sua implementação formatada para a aprendizagem dos conteúdos do domínio de conhecimento estruturado a fim de verificar de que modo a ontologia proposta poderia auxiliar na aprendizagem.

Ao final do experimento foi possível concluir que com o auxílio da ontologia um aluno pode identificar as principais etapas do processo de uma ação extrajudicial que leva a um ajustamento de conduta. Muito embora, uma ontologia represente o conhecimento de um domínio de uma forma estática, ela é um meio eficiente de representação do conhecimento e, devido aos seus critérios de codificação mínima e compartilhamento, possibilita alterações que adapte, ou melhor descreva o conhecimento representado.

Roteiro para construção de uma ontologia bibliográfica através de ferramenta automatizada (ALMEIDA, 2003).

Este artigo desenvolve um roteiro para a construção de ontologias. Para isso é utilizada uma ferramenta automatizada, que utiliza uma linguagem baseada em lógica OIL (*Ontology Interchange Language*). A aplicação do roteiro desenvolvido é demonstrada através da contrução de uma ontologia bibliográfica com a ferramenta OILEd. A construção do roteiro é baseada nos estudos de (MCGUINESS, 2001), cujas etapas são descritas a seguir:

1. Determinação do domínio e do escopo da ontologia: No exemplo em questão trata-se de uma ontologia a partir de referências bibliográficas então o autor afirma que algumas perguntas precisam ser respondidas quando se trata desse tipo de ontologia, como: Sobre autores: quantos artigos publicaram nos últimos cinco anos? Quantas são as citações referentes ao autor? Qual a titulação do autor? Sobre as publicações: quantas citações teve um artigo nos últimos X anos? Quais são os periódicos mais relevantes para a área? Quais os tipos de publicações mais produzidas? Sobre a área de conhecimento: quais são as palavras-chave da área? Que número médio de publicações é produzido na área por ano? Em que local se produz mais publicações?

2. Pesquisas sobre ontologias existente no domínio: Nessa etapa faz-se uma busca sobre as ontologias relacionadas com o escopo e domínio definido na fase anterior. Com base nessas ontologias encontradas foi criada uma lista de termos mais associados ao domínio e os seus respectivos conceitos (classes, subclasses, propriedades, entre outros.).

3. Definição das classes, hierarquias e propriedades: A partir da lista da etapa anterior é verificado os termos que são objetos e têm existência independente uns dos outros. Após isso é verificada a relação de hierarquia entre eles, definindo qual objeto é superclasse e quais são suas subclasses.

4. Definição de restrições sobre as propriedades e determinação de instâncias: Nessa etapa são definidas a cardinalidade de uma relação. Por ser ontologia bibliográfica, O autor cita o seguinte exemplo: A classe *article* tem duas relações de cardinalidade um: *citation* e *copyright*. Isso indica que um artigo tem sempre, no mínimo, uma citação e um direito autoral.

Após as etapas, a ferramenta OILEd é utilizada para construir a ontologia que ilustra a aplicação do roteiro proposto nesse trabalho. Por fim, o autor conclui que para tornar a ontologia útil, pesquisas adicionais são necessárias e que o roteiro apresentado pode auxiliar na construção de pequenas ontologias.

Ao final dos estudos, não apenas dos trabalhos relacionados, mas de toda a referência pesquisada no âmbito nacional e internacional, foi possível definir o escopo do problema e a posterior solução, descritos nas subsecções 1.1 e 1.2, respectivamente.

Capítulo Quatro

ECOM Modelo Computacional de Seleção Automática de Termos Candidatos a partir de Mineração de Textos para Auxiliar na Construção de Ontologias

Conforme descrito no capítulo 1, o objetivo do modelo proposto neste trabalho é aplicar técnicas de mineração de textos e gestão do conhecimento a coleções de documentos de um dado domínio, definido pelo usuário, para associá-los semanticamente (contextualizar) e a partir dessa associação construir uma lista de termos candidatos a compor uma ontologia do domínio selecionado.

Neste sentido, para verificar a eficiência do sistema proposto, esse foi testado no domínio de "mineração de textos"e "impactos ambientais". Detalhes sobre os processos de desenvolvimento, a utilização da aplicação, testes realizados e resultados serão descritos nesse capítulo nas seções 4.1, 4.2, 4.3 e 4.4, respectivamente.

4.1 Análise e desenvolvimento do modelo

O modelo proposto nessa dissertação possui dois módulos denominados mineração de textos e eleição de termos. O primeiro trata da aplicação de técnicas de processamento de linguagem natural e mineração , descritas no capítulo 3, em coleções de textos na língua portuguesa do Brasil. O segundo é constituído das etapas para eleição de termos para auxiliar no processo de construção de ontologias de domínio, tendo como entrada do processo os resultados obtidos com a execução do primeiro módulo.

Para o desenvolvimento do modelo, foi necessário primeiramente entender as técnicas de mineração e construção de ontologias. Após o referido entendimento, foram realizadas as fases de análise e desenvolvimento, descritas nas subseções 4.1.1 e 4.1.2, respectivamente.

4.1.1 Análise

De acordo com Sommerville (2011), os requisitos de um sistema são descrições dos serviços fornecidos pelo sistema e as suas restrições operacionais. Esses requisitos refletem as ne-

cessidades dos clientes de um sistema que ajuda a resolver algum problema, por exemplo, controlar um dispositivo, enviar um pedido ou encontrar informações. Os requisitos precisam ser elicitados e especificados e, é na fase de análise de requisitos que ocorre a elicitação e a especificação para só depois iniciar o fase de desenvolvimento. Seguindo os preceitos da engenharia de *software*, na fase de análise do sistema ECOM foram definidos os requisitos devidamente contemplados pelo *software* e elaborados os diagramas UML (*Unified Modeling Language*) de caso de uso e de sequênca.

REQUISITOS FUNCIONAIS (RF)

"Os requisitos funcionais definem, detalhadamente, as funções, os serviços e as restrições operacionais do sistema."(SOMMERVILLE, 2011). Diante dessa definição, os requisitos do modelo ECOM foram levantados a partir da identificação do problema estudado, dos estudos bibliográficos realizados e da solução proposta ao referido problema.

✓RF01 - Permitir que o domínio da ontologia seja informado pelo usuário.

✓RF02 - Realizar *upload* de arquivos com extensões .pdf.

✓RF03 - Criar árvore de diretórios "c://ontologia//nomeDodomínio"para salvar todos os arquivos e a lista de termos candidatos.

✓RF04 - Minerar a coleção de textos na língua protuguesa do Brasil, disponibilizada pelo usuário, rertornando uma lista dos termos mais frequentes na coleção.

✓RF05 - Construir a lista de termos candidatos a compor a ontologia, tomando como base a lista de termos do RF04.

✓RF06 - Visualizar a lista de termos candidatos.

REQUISITOS NÃO FUNCIONAIS (RNF)

"Os requisitos não funcionais não estão diretamente relacionados ás funções específicas fornecidas pelo sistema. Podem definir restrições, como a capacidade dos dispositivos de entrada e saída (E/S)."(SOMMERVILLE, 2011). Seguindo a definição do RNF, esses foram levantados e descritos, conforme listados abaixo.

✓RNF01 - Para instalação do sistema e posterior execução é necessária uma máquina com 4gb de RAM, processador DualCore.

✓RNF02 - Para instalação do sistema e posterior execução é necessário o sistema opera-

cional Windows a partir da versão Vista.

Com base nos requisitos funcionais supracitados foram desenvolvidos os diagramas das figuras 4.1 e 4.2.

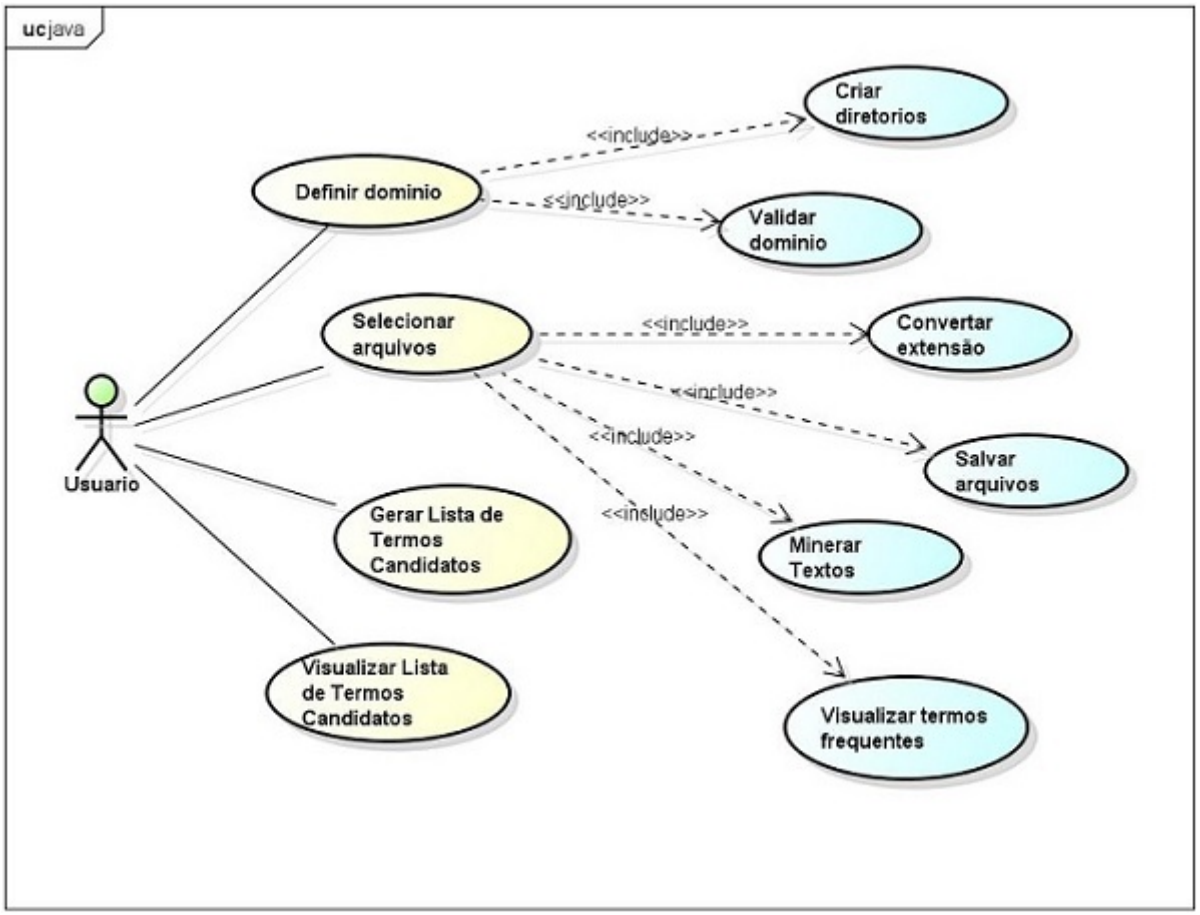

Figura 4.1: Diagrama de caso de uso. Fonte: Autor

O diagrama de caso de uso apresenta as funcionalidades que serão executadas pelo ator (usuário) e algumas atividades *include*(executadas obrigatoriamente após a atividade a qual ela está associada), ambas descritas a seguir:

✓Definir domínio: Associado ao RF01, onde o usuário irá informar o domínio desejado para que o modelo construa a lista de termos candidatos. Após informar o nome serão executados os casos de uso incluídos (*include*) validar domínio e criar diretórios, descritos a seguir:

Validar domínio: Esse nome deve possuir mais do que 5 caracteres, caso possua menos será exibida uma mensagem solicitando que seja informado um domínio válido com mais do que 5 letras. Foi determinada essa quantidade de caracteres de acordo com os exemplos das ontologias observadas nas referências, onde a maioria possuia muito mais do que cinco

caracteres.
Criar diretórios: Associado ao RF03, caso o domínio seja válido, o sistema irá criar a pasta ontologia e dentro dela todas as pastas referentes a cada novo domínio solicitado pelo usuário, onde serão salvos os respectivos documentos de cada domínio.

✓Selecionar arquivos: Associado ao RF02. Após validar o domínio, o sistema mostrará a tela de *upload* de arquivos, onde o usuário pode acrescentar os documentos da coleção que só podem possuir extensão .pdf. Após realizar o *upload* dos arquivos, serão executados os casos de uso *include* a seguir:

Converter extensão: todos os documentos da coleção, independente da extensão são convertidos para .txt. Essa conversão foi definida devido a facilidade de processamento e a padronização das extensões, pois o ECOM em uma versão futura poderá permitir o *upload* de textos em outras extensões e, nesse caso, já estará contemplando uma padronização.
Salvar arquivos: Associado ao RF04. Após converter os arquvivos, esses são salvos na pasta criada com o nome do domínio.
Minerar Textos: Associado ao RF04. Nesse caso de uso, as técnicas de mineração de textos são aplicadas.
Visualizar termos frequentes: Associado ao RF04. Após finalizar a mineração dos textos da coleção de documentos na língua protuguesa do Brasil disponibilizada pelo usuário, uma lista com os termos mais frequentes é retornada ao usuário.

✓Gerar lista de termos candidatos: Associado ao RF05. Após visualizar a lista de termos mais frequentes da coleção, o usuário irá clicar no botão "Construir Candidatos"e a lista de termos candidatos será construída tendo como base os termos retornados do caso de uso minerar textos e a coleção de documentos disponibilizada pelo usuário.

✓Visualizar lista de termos candidatos: Associado ao RF06. Depois de gerada a lista, o usuário poderá visualizá-la na pasta criada no caso de uso Criar diretórios, ou clicar no botão visualizar lista para que o sistema retorne os termos candidatos.

O diagrama de sequência, Figura 4.2 demonstra graficamente a interação entre o sistema e o ator.

O ator informa o nome do domínio. O sistema retorna a tela para que o usuário possa realizar o *upload* dos arquivos. O ator realizar o *upload*. O sistema minera os arquivos e retorna a lista dos termos mais frequentes para o usuário. O ator solicita que a lista de termos candidatos seja gerada. O sistema gera a lista e informa o final do processo. O ator solocita a visualização da lista. O sistema retorna tela com a lista dos termos candidatos a compor a ontologia do respectivo domínio do conhecimento.

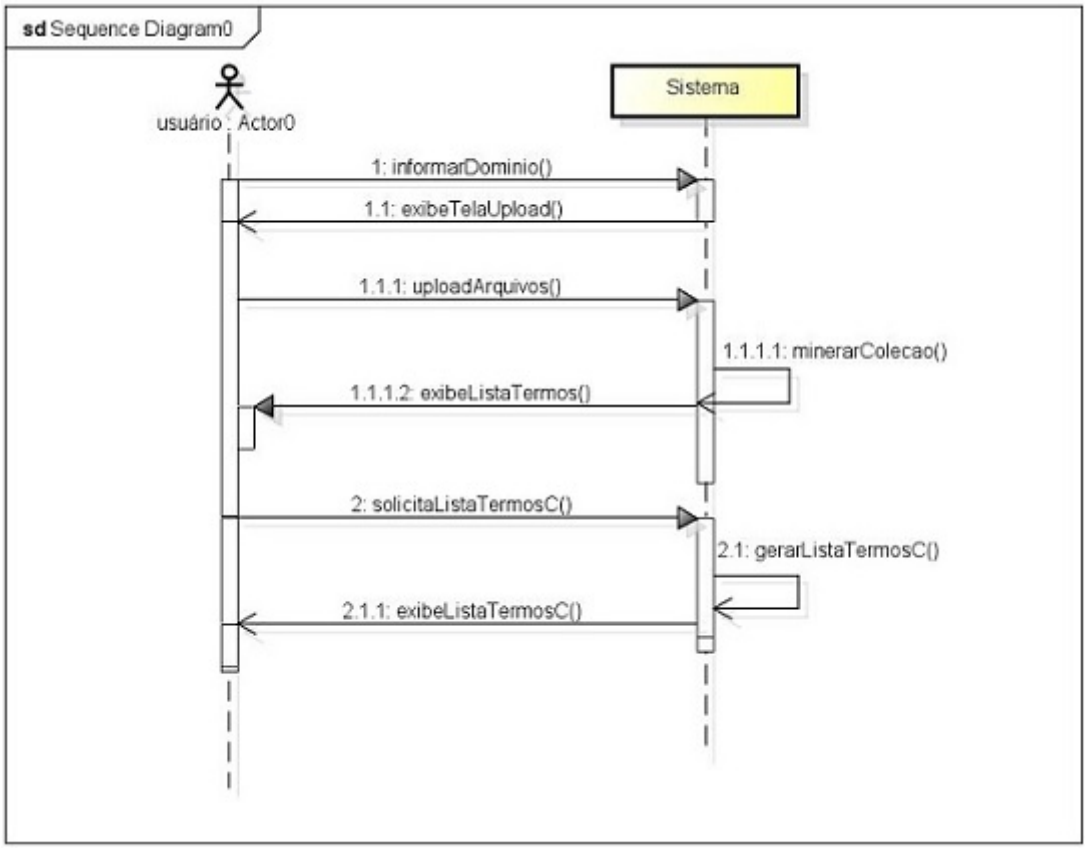

Figura 4.2: Diagrama de sequência. Fonte: Autor

4.1.2 Desenvolvimento

Após completar a análise dos requisitos descritos na subseção anterior, o sistema foi implementado. Vale ressaltar que todos os requisitos não eram conhecidos. Por isso, ao longo do desenvolvimento outros requisitos funcionais, a exemplo da restrição da extensão para PDF e visualização da lista dos termos candidatos no próprio sistema, foram identificados. Para isso foi utilizada nesse projeto a metodologia incremental, onde de acordo com Sommerville (2011) cada fase é desenvolvida como funcionalidade e, ao final, todas são integradas, pois caso haja algum problema este poderá ser identificado e tratado ao longo do projeto e não apenas ao seu final.

O modelo foi desenvolvido na plataforma .NET utilizando a sua linguagem C# e o ambiente de desenvolvimento Visual Studio 2010, todos da *Microsoft Corporation*. "O Visual Studio é um pacote de ferramentas de desenvolvimento baseadas em componentes e outras tecnologias para a criação de aplicativos avançados de alto desempenho. Além disso, o Visual Studio é otimizado para o *design*, o desenvolvimento e a implantação de soluções empresariais com base em equipes."(MSDN,). A escolha do referido ambiente de desenvolvimento justifica-se por dois motivos: A maior familiaridade e experiência da pesquisadora com a linguagem C# e a utilização da forma de aplicação do algoritmo k-

means do trabalho de dissertação da pesquisadora (BRAGA, 2010) que foi concebido na linguagem C#.

O modelo implementado possui três fases, conforme a figura 4.3, descrita a seguir.

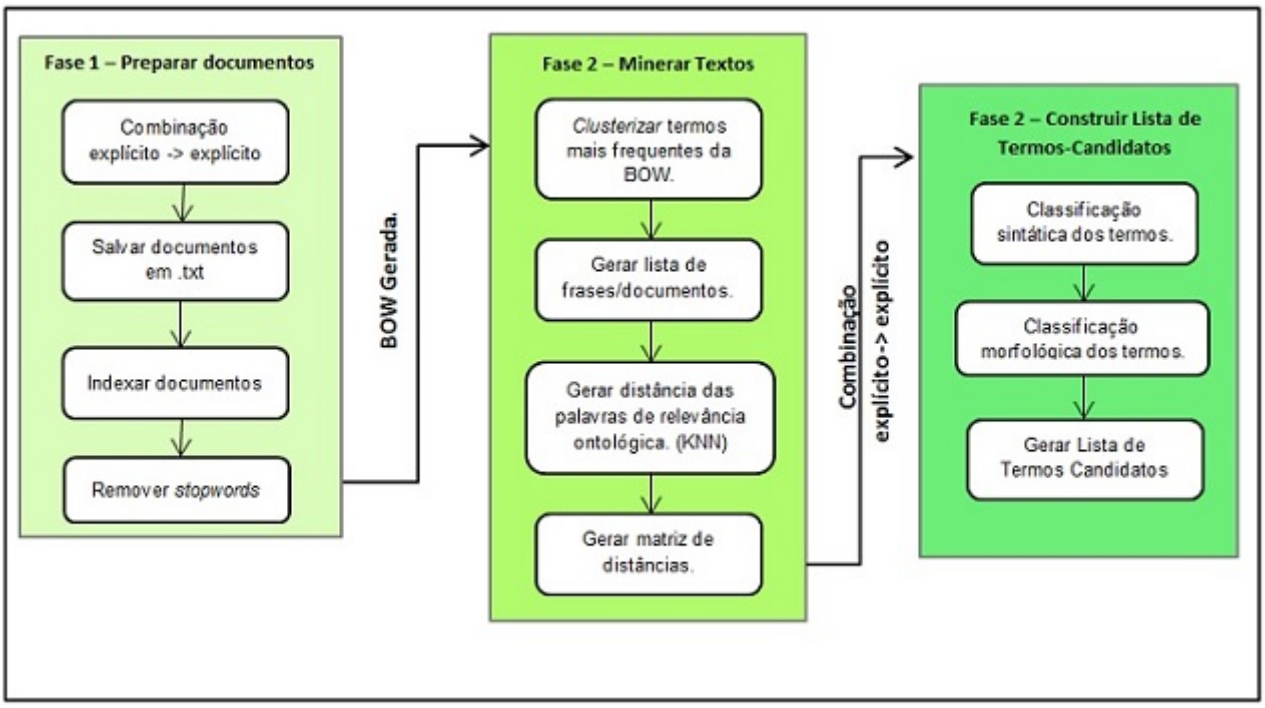

Figura 4.3: Fases do modelo. Fonte: Autor

Fase 1: Preparar documentos.

Os objetivos da primeira fase são: Preparar os documentos da coleção que o usuário realizou o *upload* para aplicação das técnicas de mineração e gerar a BOW(*Bag of Words*), descrita na subseção 2.2.3. Essa fase está subdividida em quatro etapas,combinação, salvar documentos em .txt, indexar documentos e remover *stopwords*.

✓Combinação - Essa fase coloca em prática um dos modos de conversão do conhecimento, combinação (vide capítulo 2) Esse tipo de conversão é abordado também pelas teorias ligadas ao processamento da informação, ocorre por meio do agrupamento (classificação, sumarização) e processamento de diferentes conhecimentos explícitos. Nesse momento, os documentos são selecionados pelo conhecedor do domínio que utiliza do seu conhecimento adquirido para analisar os documentos candidatos a fazer parte da coleção. Para só então, após a leitura dos mesmos, definir os documentos que formarão a coleção que será analisada pelo modelo. Dessa forma, há uma conversão do conhecimento explícito para o explícito.

✓Salvar documentos em .txt - Após o *upload* dos textos, esses são convertidos para o formato .txt e salvos na pasta Arquivos_Originais, conforme figura 4.4, criada dentro da árvore de diretório descrita no RF03. Esse processo é feito para estabelecer um padrão para as outras fases e ganhar em performance, pois com a manipulação de arquivos apenas em .txt há uma econômia no processamento.

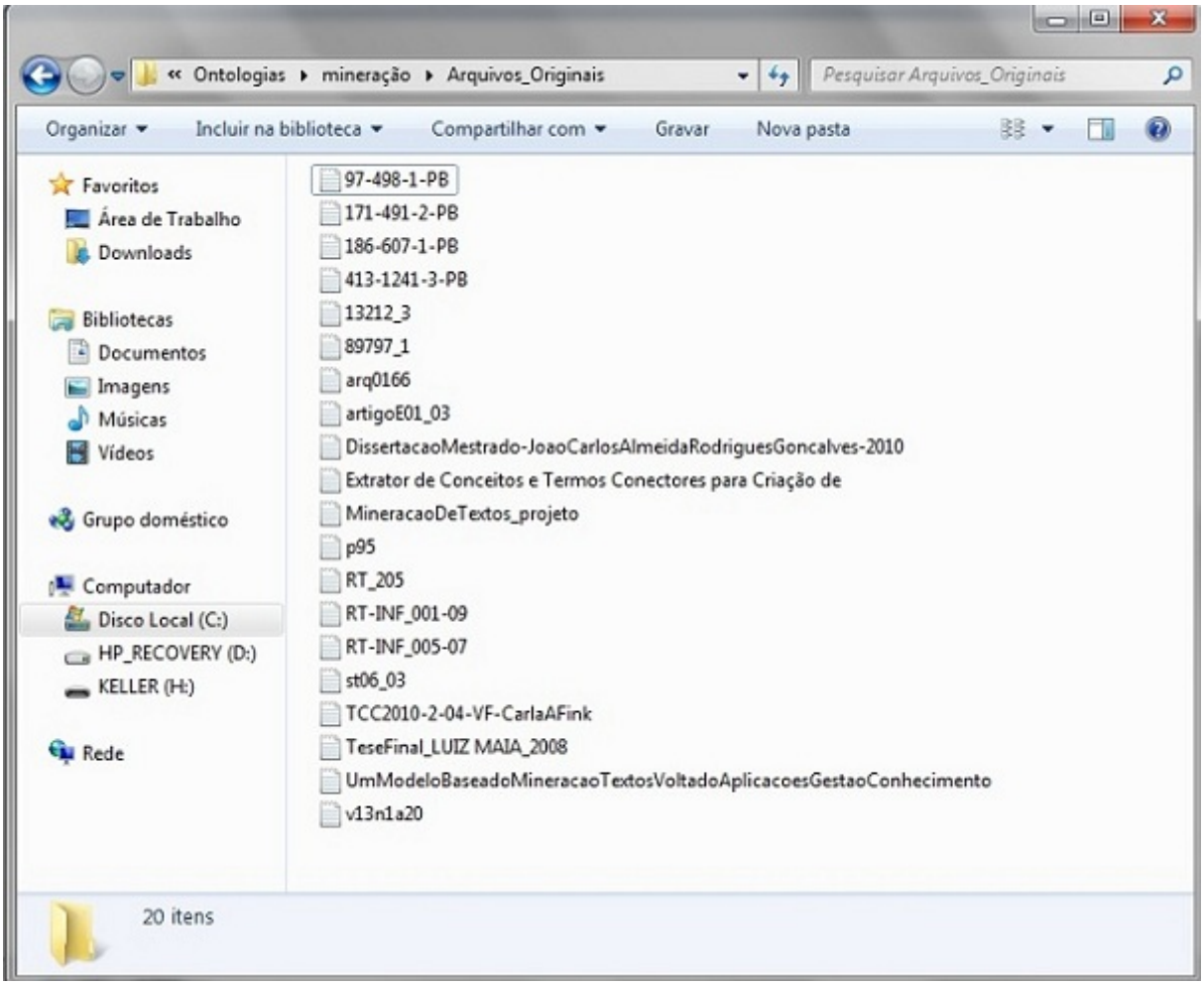

Figura 4.4: Arquivos da coleção convertidos em .txt. Fonte: Autor

✓Indexar documentos - O processo de indexação de documentos é necessário para subsidiar o processo, realizado na fase 2, de construção da BOW (*Bag of Words*). É criado um arquivo indexDoc.txt que possui a lista de todos os arquivos da coleção associado a um índice estabelecido em ordem crescente e iniciando em 0, conforme figura 4.5. Através desse indexador cada documento passa a ser identificado por um número, o que facilita a manipulação dos arquivos na coleção.

✓Remover *stopwords* - Observa-se na figura 4.5, arquivos com o nome [n_ListaPalavras], onde n varia de 0 a (quantidade de arquivos da coleção -1) e corresponde ao índice do documento da coleção formado pela lista de palavras salvas no referido arquivo.

Para obter uma lista de palavras com significados capaz de estabelecer relações semânticas e sintáticas entre elas, é preciso excluir termos como: preposições, números, caracteres especiais,como: (,[,),],{,&,%,), entre outros. Para subsidiar esse processo, nessa etapa são aplicadas técnicas de processamento de linguagem natural (PLN), descritos no capítulo 3. É aplicada inicialmente a *tokenização* para substituir os caracteres especiais por espaços em branco, essa substituição foi necessária porque em muitos casos existem termos separados por caracteres especiais, como no exemplo abaixo, e ao retirar o caractere iria ocorrer a junção das palavras, o que impediria que essa fosse identificada como um termo válido da lingua portuguesa do Brasil e acabaria sendo desconsiderada.

Exemplo:
mineração(Descoberta do conhecimento)
mineraçãoDescobertaconhecimento. Sem a inclusão dos espaços. mineração Descoberta conhecimento. Com a inclusão dos espaços.

Após o processo de tokenização, é aplicada a *StopWords* retirando todas as preposições e artigos e também subsituindo cada *StopWords* por um espaço em branco.

Fase 2: Minerar Textos.

O objetivo dessa fase é aplicar as técnicas de mineração para construir a lista dos termos mais frequentes de cada documento e da coleção. Posteriormente, é elaborada uma matriz de distâncias entre as palavras mais frequentes e os termos com distância 1 ou 2 associadas a cada termo. Conforme exemplo abaixo, o termo mais frequente em análise é mineração então o termo com distância 1 é **dados** e com distância 2 é **recebe**. Vale ressaltar que a preposição "de"não foi considerada, pois nessa fase já foram retiradas as *stopWords*. Ao final dessa fase, ocorre também uma nova combinação, pois o conhecimento explícito pré-processado na fase 1 volta a ser refinado, mas, dessa vez, de forma automática.

Exemplo:
A mineração de Dados recebe influências das áreas de Processamento de Linguagem Natural.

✓Clusterizar termos mais frequentes dos documentos (K-*means*) - Nessa etapa aplica-se o algoritmo k-*means* sobre a lista de termos da coleção, mas antes dessa aplicação é gerada uma lista única de todos os termos já eliminando palavras repetidas e palavras no plural, mantendo apenas a sua forma sem a flexão do plural que corresponde as seguintes terminações: s, ões, ães, ãos, es, is, eis. Esse tratamento fez-se necessário, pois observou-se que o algoritmo de clusterização retorna grupos de palavras sumarizados pela frequeência no texto sem fazer qualquer tipo de análise dos termos. Foi estabelecido que o algoritmo *k-means* trabalhará sempre com cinco *clusters* realizando 500 iterações. Esses valores foram

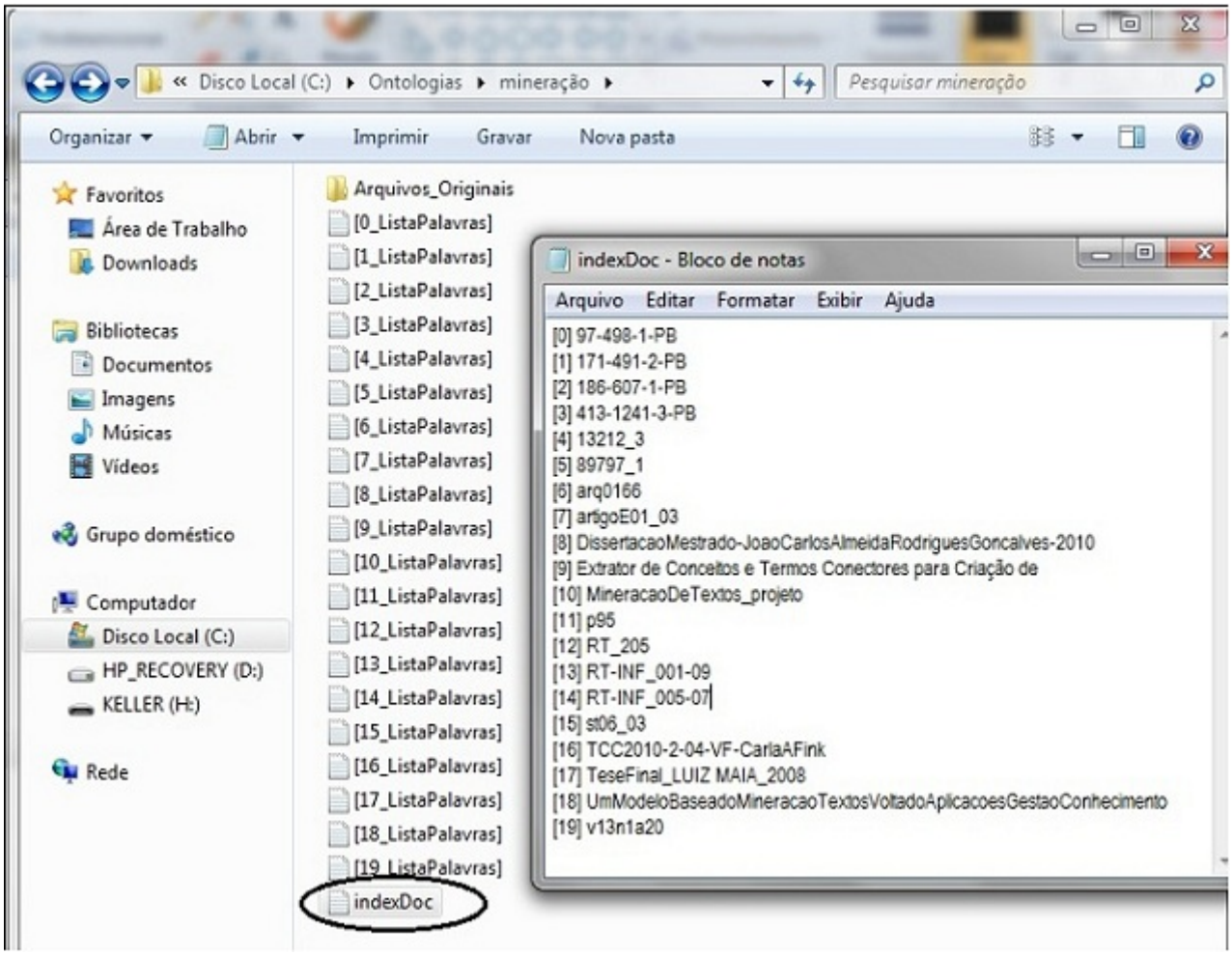

Figura 4.5: Indexador de documentos e lista de palavras por documentos. Fonte: Autor

definidos a partir de testes com coleções de documentos de até 60 arquivos e verificou-se que trabalhar com mais do que 5 clusters ou 1000 iterações haveria o risco de termos relevantes ficarem de fora da análise por serem agrupados com termos menos relevantes. Após a clusterização, o cluster mais relevante é definido através do cálculo da média entre o somatório das frequências e a quantidade de palavras do *cluster*. Mesmo com o tratamento da lista dos termos antes da clusterização e a execução do algoritmo propriamente dito, ainda observou-se alguns termos repetidos e algumas falhas como: palavras iniciando com ¬¡, ¡ , ",entre outros. Para isso, foi criado um método para verificar cada caractere dos termos retornados, evitando assim que termos irrelevantes do corpus fossem mantidos e, consequentemente, seguissem para etapa de análise. Após essa etapa, observou-se nos testes realizados ao longo do desenvolvimento do ECOM que do total de 13938 termos da coleção, após o *cluster* restaram 102 e 25 após a aplicação do método de verificação.

✓Gerar Lista de frases/documentos - Nessa etapa, retorna-se aos documentos em sua forma completa para criar uma lista das frases de todos os documentos da coleção. As

frases são identificadas por coleções de palavras ponto(.) a ponto, ou seja, a cada ponto encontrado entende-se o fim de uma frase e início da outra, dessa forma cada conjunto de palavras entre pontos é armazenada em uma posição da lista.

✓Gerar distância das palavras de relevância ontológica - Em resumo uma ontologia de um dado domínio é formada pelos termos que possuem relações, seja ela semântica ou hierárquica, ou seja, para uma ontologia sobre mineração entende-se que filtra, descobrir_Conhecimento possuem relação semântica com o domínio, já dados e texto são tipos de mineração portanto possuem relação hierárquica. A partir dessa busca de relações para formar a lista de termos candidatos para a construção de ontologias de domínio, essa etapa busca a relação dos termos com base na sua proximidade. Para isso, a lista de termos mais frequentes da etapa um é confrontada com a lista de frases da etapa 2 e é gerada uma matriz de palavras e frases, onde cada termo é associado as frases que o mesmo aparece. Sendo mantidas duas palavras que antecedem e sucedem o termo.

✓Gerar matriz de distâncias - A partir da matriz gerada na fase anterior, é gerada uma segunda matriz onde a coluna 1 possui os termos já trazidos da etapa anterior e a coluna dois a respectiva matriz com as palavras mais frequentes da coleção de frases de cada termo e a sua distância, 1 ou 2.

Fase 3: Construir lista de termos candidatos.

✓Classificação sintática dos termos - Nessa etapa, cada termo da matriz de matrizes gerada na fase 2 é classificado em sujeito, verbo e objeto direto ou indireto. Para isso, incialmente é utilizada a técnica *stemming* não em sua totalidade, mas apenas quando verbos são identificados. Para o reconhecimento dos verbos foi utilizada um lista com 10000 verbos definida no trabalho de (BRAGA, 2010), e se ao reduzir a palavra ao seu radical, esta for encontrada na referida lista, então é classficada como verbo. Após a identificação dos verbos, verfica-se a palavra que o antecede e a que sucede. Além da lista de verbos também foi utilizada uma lista com 45000 substantivos da mesma fonte dos verbos. Dessa forma, para identificar um sujeito foi utilizada a regra gramatical de posicionamento dos termos que possui algumas exceções que não foram tratadas em sua totalidade nesse trabalho. Dessa forma, o que antecede o verbo e está na lista de substantivos é considerado sujeito e o termo sucessor é classificado como objeto direto se estiver com distância um do verbo, objeto indireto se estiver separado do verbo por uma preposição e predicativo do sujeito se o verbo estiver na lista dos verbos de ligação.

✓Classificação morfológica dos termos - Essa etapa é realizada concomitantemente com a anterior, pois na anterior já identifica-se substantivo e verbo. Nessa etapa, são verificadas a presença de adjetivo nas frases. Adjetivo é todo termo que acrescenta característica ao substantivo. Como: casa bonita (adjetivo), processo rápido (adjetivo), entre outros.

Para reconhecer um adjetivo na frase existem duas formas: É adjetivo quando está ligado diretamente a um substantivo ou ligado a um substantivo por um verbo de ligação. Entretanto, existem muitas exceções e regras que até o momento só podem ser aplicadas pelos seres humanos, como o exemplo abaixo onde substitui-se o verbo por um de ligação e verifica se o contexto indica um termo que acrescenta característica, se sim o termo que acrescenta a característica é classificado como adjetivo. Dessa forma, para que fosse possível identificar os adjetivos em um algoritmo computacional foi criada uma lista de 100 adjetivos e uma palavra foi dita adjetivo quando estiver ligada a um substantivo diretamente ou por um verbo de ligação e estiver na referida lista.

Exemplo: Eram exercícios difíceis. Pergunta-se: O exercício é o que? difícil. Portanto, difícil é adjetivo.

✓Gerar lista de termos candidatos - Após as classificações das etapas anteriores, o algoritmo analisa os termos mais frequentes relacionados com os termos principais trazidos da clusterização da etapa 1 da fase 2 e a sua classificação. Dessa forma, uma ordem de execução é estabelecida. Primeiro são selecionados todos os termos que mais se relacionam com distância 1, depois a classificação de cada um é analisada e se tiver etiquetado como verbo é eleito como termo candidato juntamente com o seu sucessor que será na classificação morfológica um substantivo ou adjetivo, caso não seja um verbo só será eleito se no texto original estiver separado do termo principal por uma preposição, pois indicará um complemento do termo principal ou até mesmo um tipo desse, a exemplo de mineração de dados e mineração de texto onde dados e texto dão mais sentido ao termo principal mineração, pois especializam o termo. Após a análise dos termos com distância 1 são analisados os termos com distância 2 onde são considerados apenas termos etiquetados como substantivos ou adjetivos.

Após as etapas descritas acima terem sido desenvolvidas e testadas pela pesquisadora, a construção do ECOM foi finalizada. A partir desse momento, ele pode ser utilizado pelos usuários conforme demonstrado na secção 4.2.

4.2 A utilização do modelo

O modelo desenvolvido com *windowsForm*, ou seja, para ser utilizado instalando-se na máquina do usuário que possua sistema operacional Windows da *microsoft corporation* possui três passos para ser utilizado e atingir o seu objetivo que é mapear os conceitos mais frequentes na coleção e construir uma lista de termos candidatos a compor uma ontologia.

Ao executar o *software*, será exibida a primeira tela do modelo apresentada na figura 4.6 , onde será informado o domínio da coleção de documentos selecionada pelo usuário na próxima tela. Na figura 4.6 é informado o domínio mineração.

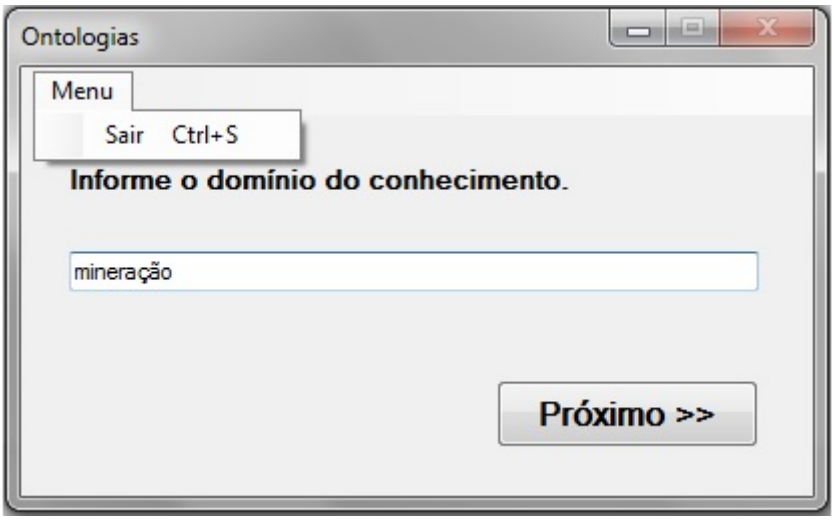

Figura 4.6: Tela inicial do modelo. Fonte: Autor

Após informar o domínio, clica-se no botão próximo e exibirá uma tela para selecionar os arquivos, comforme figura 4.7. Na tela de *upload* há uma área em branco, nessa será exibida, ao final da aplicação das técnicas de mineração, a lista de palavras mais frequentes do *corpus* (coleção dos documentos utilizados como entrada informado pelo usuário.), e um botão que inicialmente está desabilitado porque apenas ao final do processo de mineração é dada a opção para o usuário de utilizar os resultados ou continuar no processo de representação do domínio do corpus através da construção da lista de termos candidatos. Escolhendo a segunda opção, será executada a fase três do modelo, detalhada na subseção 4.1.2, e ao final é aberta uma tela, conforme Figura 4.8 onde serão listados todos os termos candidatos eleitos pela ferramenta.

Ressalta-se que o o usuário também poderá ter acesso a lista de termos candidatos em formato .txt, pois ao final da construção da lista é gerado um arquivo na pasta do respectivo domínio com a lista dos termos.

4.3 Validação do ECOM

Para validar o modelo ECOM proposto neste trabalho, trabalhou-se com duas coleções de textos com temas principais distintos, impactos ambientais e mineração, das áreas de meio ambiente e tecnologia da informação, respectivamente.

As coleções de textos dos domínios citados no paragráfo anterior foram previamente se-

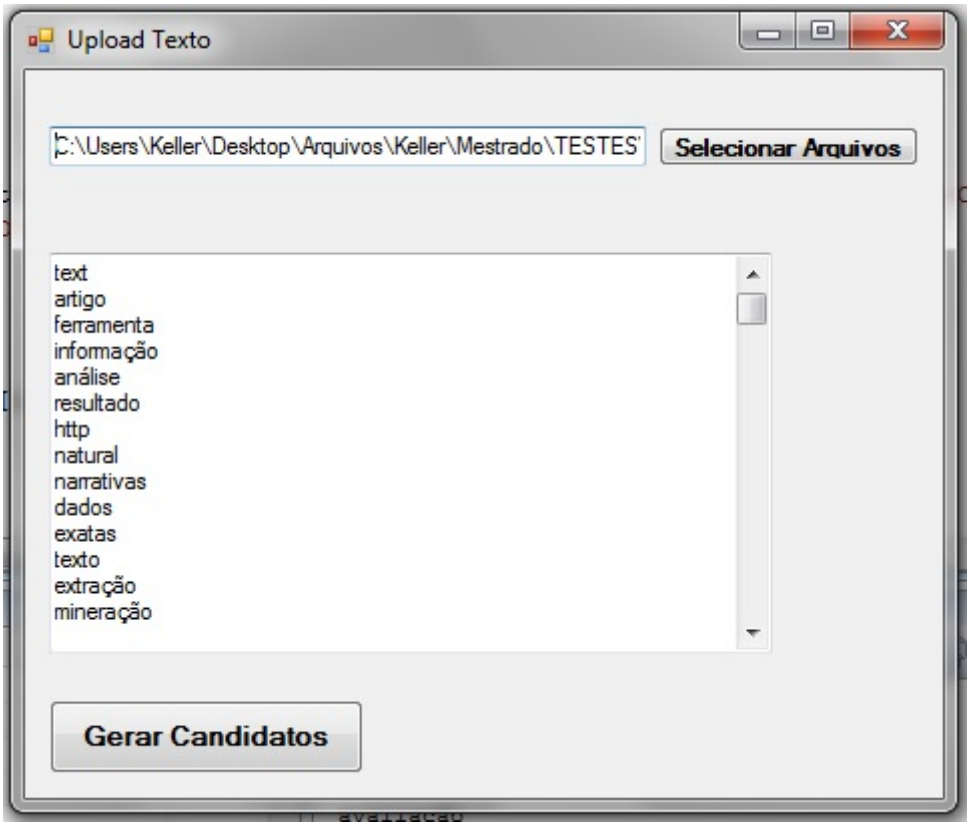

Figura 4.7: Tela de seleção dos arquivos. Fonte: Autor

lecionadas por conhecedores do domínio. O termo conhecedores foi utilizado em lugar de especialistas, pois esses possuem conhecimento nas suas respectivas áreas, meio-ambiente e tecnologia da informação, mas não são especialistas nos domínios definidos para realização dos testes.

A descrição dos testes realizados nos domínios de mineração e impacto ambiental será realizada nas subsecções 4.3.1 e 4.3.2, respectivamente.

4.3.1 Aplicação do modelo no domínio de mineração

O primeiro domínio testado foi o de mineração, e assim foi definido devido ao amadurecimento do conhecimento nessa linha de pesquisa durante a construção desta dissertação de mestrado. A coleção foi validada por Patricia Freitas Braga, profissional com nível de mestrado e doutoranda em modelagem computacional, que será citada ao longo do texto como validadora_A . Para realizar os testes, a autora e a validadora_A selecionaram uma amostra de 30 (trinta) documentos composta por dissertações e artigos. A seleção desses textos foi realizada através de pesquisas feitas na internet no portal Capes, scielo e *Google Scholar*.

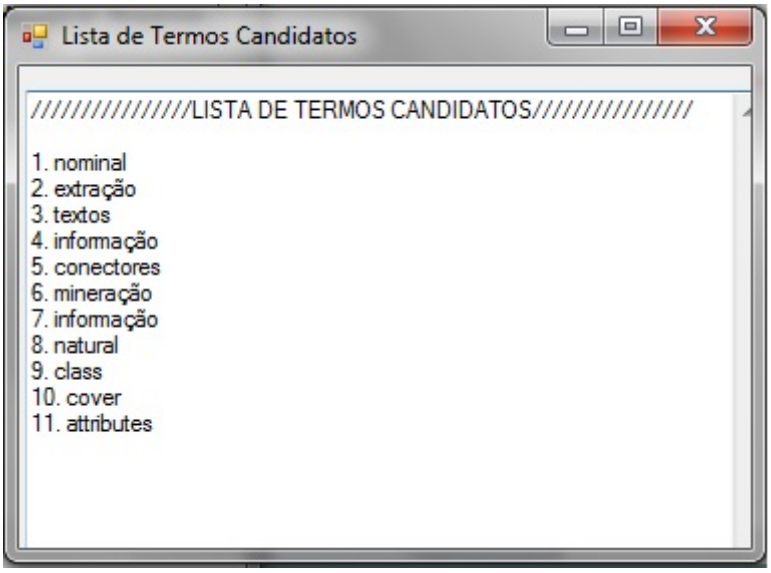

Figura 4.8: Lista dos termos eleitos. Fonte: Autor

Após a validação da coleção de documentos, os mesmos foram minerados utilizando a ferramenta, o que resultou na lista de palavras apresentadas na tabela 4.9. Essa lista foi submetidsa a validação da validadora_A que destacou em amarelo 73,33% dos termos considerados irrelevantes, resultando em uma confiabilidade do resultado da mineração, neste domínio, realizada pela ferramenta de 26,67%. Entretanto, foram mantidas no processo todas as palavras já que o objetivo final é demonstrar a capacidade do modelo de eleger termos para a construção de uma ontologia de domínio.

Vale ressaltar que apesar do usuário visualizar apenas a lista já refinada pelos processos do algoritmo, representada pela tabela 4.9. A diferença entre a lista inicial de termos da coleção de documentos (total de 13927 termos) e a lista dos termos de retorno após as fases de mineração, processamento e pré-processamento (vide capítulo 3), (total de 45 termos) é de 13882.

Após a mineração, todos os 45 termos resultantes foram submetidos ao processo de classificação sintática, morfológica e a seleção dos termos candidatos, conforme descrito na subsecção 4.1.2. O que resultou em uma lista de 19 termos candidatos para construção de uma ontologia do domínio de conhecimento mineração, conforme apresentado na Figura 4.10. Após a eleição automática dos termos candidatos para construção da ontologia, esses foram analisados pela validadora A e comparados com a lista inical dos 45 termos, onde foi possível observar que os termos selecionados pelo algoritmo estão relacionados com o domínio e, além disso, os termos considerados irrelevantes pela validadora na primeira análise na etapa de mineração foram excluídos da lista de termos candidatos. Entretanto, alguns termos em inglês permaneceram, mas foram desconsiderados, pois o escopo deste trabalho contempla apenas a língua portuguesa do Brasil. Contudo, ressalta-se que as

1. sintagma	24. natural
2. incidente	25. cadastrados
3. histórias	26. dados
4. análise	27. extração
5. usos	28. nominal
6. matrícula	29. mineração
7. atualizar	30. from
8. texto	31. souza
9. language	32. informação
10. pontuado	33. administrador
11. prospectivo	34. turismo
12. gramáticas	35. etiqueta
13. experimento	36. teses
14. processos	37. gene
15. aninhados	38. simplekmeans
16. rules	39. acessa
17. quadro	40. inscrição
18. comentário	41. unsupervised
19. desk	42. filters
20. notícia	43. attributes
21. workflow	44. instances
22. cadastro	45. class
23. conectores	

Figura 4.9: Lista de palavras. Domínio Mineração. Fonte: Autor

palavras em inglês também estão fortemente associadas ao domínio.

4.3.2 Aplicação do modelo no domínio de impacto ambiental

O domínio impacto ambiental foi escolhido por ser fora do contexto de tecnologia da informação (TI), pois julgou-se importante aplicar o modelo em um tema fora do contexto de TI. Salienta-se que o domínio impacto ambiental é um dos que menos possuem termos em outros idiomas, como inglês. Além disso, no referido domínio foi possível reunir o maior número de documentos na língua portuguesa do Brasil. A coleção de documentos foi definida por Clelia Nobre de Oliveira, profissional mestre em engenharia ambiental urbana, que será citada ao longo do texto como validadora_B. Para realizar os testes, a validadora_B selecionou uma amostra de 20 (vinte) documentos composta por artigos.

Após a validação da coleção de documentos, os mesmos foram minerados utilizando a ferramenta, o que resultou na lista de palavras apresentadas na figura 4.11. Essa lista

Figura 4.10: Lista de palavras eleitas para compor a ontologia. Domínio Mineração. Fonte: Autor

foi novamente validada pela validadora_B onde 44,7% dos termos foram considerados irrelevantes, resultando em uma confiabilidade do resultado da mineração nesse domínio realizada pela ferramenta de 55,3%. As palavras apontadas como irrelevantes estão destacadas em amarelo na figura da tabela 4.11. Vale ressaltar que os termos definidos como irrelevantes foram mantidos, pois o processo para eleição dos termos candidados a construção da ontologia ainda submetia a lista a outros refinamentos. Dessa forma, todos os termos foram mantidos com o objetivo de verificar se os algoritmos implementados no modelo seriam capazes de alcançar ou se aproximar da filtragem feita por um ser humano conhecedor do domínio. Haja vista que o objetivo desse trabalho foi diminuir a intervenção manual e a dependência de especialistas do domínio no processo de construção de ontologias de domínio.

Ressalta-se que a diferença entre a lista inicial de termos da coleção de documentos (total de 6083 termos) e a lista dos termos de retorno após as fases de mineração, processamento e pré-processamento (vide capítulo 3), (total de 123 termos) é de 5960.

1.drenagem	25.questões	49.energia	73.serrote	97.brasileira
2.consumo	26.econômicos	50.planeta	74.todos	98.pluviais
3.helmintíases	27.emprego	51.braço	75.social	99.industrial
4.atendida	28.cerrado	52.acesse	76.tempo	100.redução
5.bairros	29.causal	53.etapas	77.proliferação	101.emissão
6.doença	30.condições	54.expert	78.ministério	102.regiões
7.ambiental	31.direto	55.agricultura	79.participação	103.terra
8.floresta	32.incidências	56.agrícola	80.dezembro	104.crescimento
9.degradação	33.incubadora	57.hidrelétricas	81.resíduos	105.cerca
10.rios	34.pública	58.perda	82.município	106.ambientes
11.agronegócio	35.sistema	59.instituto	83.mata	107.metano
12.efeito	36.aquecimento	60.econômica	84.Paulo	108.febre
13.emissões	37.malária	61.turbinas	85.cabral	109.expansão
14.amazônia	38.classes	62.barra	86.classificação	110.modelo
15.água	39.tonelada	63.lançamento	87.média	111.produção
16.pressão	40.hídrica	64.alagadiço	88.barreiro	112.poluição
17.municipal	41.causais	65.planejamento	89.esgotos	113.milhões
18.gases	42.contaminação	66.reservatório	90.relação	114.matos
19.infecciosas	43.serviço	67.sólidos	91.empoçamento	115.saúde
20.plano	44.saneamento	68.disponibilidade	92.conflito	116.casos
21.desmatamento	45.superfície	69.amarela	93.pelas	117.concentração
22.renda	46.ingepro	70.dengue	94.fonte	118.atmosfera
23.ocorrência	47.quais	71.sociedade	95.ferreira	119.carbono
24.alagoinhas	48.vetores	72.tucuruí	96.urbana	120.estufa
				121.civil
				122.valor
				123.contato

Figura 4.11: Lista de palavras. Domínio Impacto Ambiental. Fonte: Autor

Após a mineração, os 123 termos resultantes foram submetidos ao processo de classificação sintática, morfológica e a eleição dos termos, conforme descrito na subsecção 4.1.2. O que resultou em uma lista de 31 termos candidatos para construção de uma ontologia do domínio de conhecimento impacto ambiental, conforme mostrado na Figura 4.12. Os termos-candidatos foram validados pela validadora B e todos foram considerados relevantes para o domínio impacto ambiental. Além disso, a tabela da Figura 4.11 foi comparada com a tabela da Figua 4.12, onde foi possível observar que os termos considerados irrelevantes na primeira análise foram excluídos da lista final dos termos candidatos. A partir da referida análise observou-se que os termos eleitos pelo algoritmo para a construção da ontologia estão, em sua totalidade, relacionados com o domínio. Entretanto, muitos termos considerados relevantes na lista anterior foram descartados.

4.4 Resultados

Os testes realizados mostram que o modelo atinge o objetivo de minerar uma coleção de documentos e a partir desses resultados construir uma lista de termos candidatos a compor uma ontologia. Sendo esses processos realizados de forma totalmente automatizada o que contempla a solução do problema estudado nesta dissertação, como diminuir a intervenção direta manual e a dependência de especialistas do domínio no processo de construção de ontologias de domínio? Entretanto, alguns pontos devem ser considerados:

O resultado da construção da lista de termos candidatos está fortemente associado ao resultado do processo da mineração e, esse, por sua vez, depende da criteriosidade na seleção dos documentos que irão compor a coleção. Devido a essa necessidade, observou-se que os testes na área de meio-ambiente foram mais satisfatórios do que na área de tecnologia da informação. Isso deve-se ao fato de no segundo haver muitos termos em inglês e textos mais abrangentes como dissertações que envolvem vários assuntos não mantendo o foco no domínio como ocorre com artigos.

Diante dos pontos observados, verifica-se que o modelo auxilia no processo de construção de ontologias, tornando menor a intervenção manual dos especialistas e dos engenheiros de ontologias, a partir do momento que seleciona corretamente os termos para compor tal ontologia. Entretanto, ainda há muita dependência do especialista, principalmente no início do processo onde faz-se necessária uma boa seleção dos documentos, e nesta dissertação esse processo é feito exclusivamente pelos especialistas, pois o modelo não contempla a validação dos documentos de acordo com o domínio definido para a construção da ontologia.

Figura 4.12: Lista de palavras eleitas para compor a ontologia. Domínio Impacto Ambiental. Fonte: Autor

Capítulo Cinco

Considerações finais

Nas pesquisas realizadas nesta dissertação de mestrado não foram encontrados trabalhos correlatos que propusessem modelos computacionais associando técnicas de mineração de textos e gestão do conhecimento para eleição de termos para uma ontologia de domínio. Diante disso, e com base nos estudos realizados, descritos nos capítulos do aporte teórico, o problema abordado nessa dissertação foi como diminuir a intervenção manual e a dependência de especialistas do domínio na construção de ontologias de domínio unindo mineração de texto e gestão do conhecimento.

5.1 Conclusões

O presente trabalho propôs um modelo para eleição de termos candidatos a compor uma ontologia de domínio a partir dos resultados das técnicas de mineração de textos aplicadas a uma dada amostra de documentos científicos do domínio selecionado pelo usuário. A concepção do trabalho dividiu-se em três fases. Na fase 1, foi realizado um levantamento bibliográfico dos conceitos que norteiaram esse trabalho, técnicas de mineração, gestão do conhecimento, construção de ontologias de domínio e processamento da linguagem natural (PLN), descritos nos capítulos do aporte teórico.

A fase dois que começou a ser realizada logo após a definição do problema e objetivos dividiu-se em três etapas: **Especificar modelo**, **Desenvolver modelo** e **Testar modelo**.

Especificar modelo - Nessa etapa foi construído o documento de especificação dos requisitos tendo em vista que esses foram levantados em sua maioria ao longo das etapas para a definição do problema. Ao longo dessa etapa foi feita análise de requisitos e projeto.

Desenvolver modelo - Nessa etapa foi desenvolvido o modelo conforme a análise dos requisitos realizada na etapa anterior. Uma vez que todos os requisitos não eram conhecidos, ao longo do desenvolvimento outros requisitos funcionais e não funcionais foram identificados. Para isso foi utilizada nesse projeto a metodologia incremental.

Testar modelo - Nessa etapa foram realizados testes com o objetivo de verificar se o que foi implementado estava correspondendo aos requisitos especificados na fase 1. Fazendo uso dos preceitos da metodologia incremental, quando em algum teste era detectada uma

falha ou erro, esse era novamente especificado, desenvolvido e testado mais uma vez até que a referida falha fosse sanada.

Na fase três, o modelo proposto foi aplicado em domínios das áreas de meio-ambiente e tecnologia, conforme descrito no capítulo 4 e foi analisado de duas formas:

1 - Resultados do processo de mineração de texto: A lista de termos retornada do processo de mineração foi validada pelos conhecedores do domínio, onde a média de termos relevantes entre os dois cenários foi de 48% . O que permite concluir que o modelo possui relevância dentro do esperado tendo em vista que o processo é por si só falho, pois pesquisas, como as citadas nesse trabalho, comprovam que o processo manual ainda é mais exato. Vale ressaltar que esse valor é uma média dos resultados dos dois cenários, pois no domínio impacto ambiental o modelo demonstrou-se muito mais preciso ao retornar 55,3% de termos relevantes. Em contrapartida, o conhecimento mineração possuiu um fator limitante ao utilizar-se de muitos termos considerados de maior relevância na língua inglesa que foi retornado corretamente, mas descartado por não ser o idioma foco desta pesquisa.

Outro fator limitante para o resultado da mineração foi a definição de trabalhar com 5 clusters e 1000 iterações, pois observou-se que caso aumentasse esses valores haveria o risco de termos relevantes ficarem de fora da análise por serem agrupados com termos menos relevantes. Devido a definição desses parâmetros, muitos termos considerados irrelevantes para o domínio foram mantidos já que a quantidade definida de cluster fez com que fossem constuídos grupos maiores e que nesses houvesse uma maior concentração de termos variando entre menos relevantes e mais relevantes.

2- Eleição dos termos candidatos: A lista de termos candidatos gerada ao final de todo o processo foi validada manualmente pelas respectivas validadoras conhecedoras dos seus respectivos domínios e foi possível observar que os termos eleitos estão em sua totalidade associados ao domínio mapeado. Além disso, observou-se que os termos considerados irrelevantes na primeira etapa do processo, mineração de texto, foram eliminados nessa segunda etapa.

Considerando a análise realizada dos resultados do modelo e o nível de complexidade existente no processo para extração de dados textuais e relações entre esses, o resultado do processo de mineração e a posterior eleição dos termos candidatos a compor a ontologia foram conferidos e considerados corretos. O que permitiu concluir que o modelo proposto nesse trabalho atingiu os seus dois objetivos principais, diminuir a intervenção direta manual no processo de construção de ontologias e eleger termos candidatos para compor ontologia a partir da aplicação de técnicas de mineração e gestão do conhecimento em coleções de documentos. Tendo em vista que apesar das validações da coleção

de documentos e da lista de termos terem sido feitas de forma manual pelos validadores, o processo de mineração e posterior eleição dos termos transcorreu sem a intervenção desses.

5.2 Contribuições

A essência da solução proposta está no desenvolvimento de um modelo para apoiar no processo de construção de ontologias de domínio diminuindo a intervenção direta manual e a dependência de especialistas do domínio.

O aspecto relevante dessa pesquisa é o fato de não ter encontrado soluções que contemplassem a junção de mineração de textos, gestão do conhecimento e construção de ontologias, e sim foi observada a existência de ferramentas para construção de ontologias que possuem como entrada os termos definidos a partir do conhecimento de especialistas do domínio e, muitas vezes, as técnicas de construção dos engenheiros de ontologias. Além disso, nos moldes atuais, para unir as técnicas de mineração de texto faz-se necessário o uso de uma segunda ferramenta para execução das referidas técnicas.

Ressalta-se que o modelo permite a análise automatizada dos resultados da aplicação das técnicas de mineração de texto na coleção de documentos, pois muitos métodos foram desenvolvidos para analisar os termos obtidos com o resultado da mineração e as suas relações. Entretanto, essa análise também pode ser feita pelo próprio usuário, pois ao final do processo de mineração, é retornada a lista dos termos mais frequentes para que essa possa ser analisada pelo usuário e a partir deste ponto decida se irá solicitar ao modelo a eleição dos termos candidatos ou não. Em resumo, o modelo proposto também pode ser aplicado unicamente para minerar textos na língua portuguesa do Brasil.

Apesar do modelo eleger termos para compor uma ontologia e não construir ontologias, a lista de termos auxilia no mapeamento do domínio e consequente construção de novas ontologias, ou seja, o modelo proposto pode colaborar na construção de bases ontológicas, e consequentemente, no seu aumento.

5.3 Atividades Futuras de Pesquisa

Para trabalhos futuros recomenda-se melhorias e inclusões de funcionalidades no modelo proposto, como:

- Permitir a retroalimentação dos termos, onde o usuário a partir do retorno dos termos

mais frequentes já apresentado no modelo dessa dissertação poderá optar por excluir alguns termos e estabelecer relações entre esses, pois é sabido que o processo de construção de ontologia depende bastante das relações estabelecidas entre os termos, o que ainda é muito complexo de ser definido de forma automatizada.

- Realizar a identificação das relações hierárquicas entre os termos para construir redes semânticas para que o usuário possa visualizar as relações geradas a partir dos resultados da mineração na coleção de documentos. Essas redes podem permitir visualizar a relação entre os documentos da coleção e os seus termos.

Apêndice A

ANEXOS

A.1 ANEXO A

ANEXO A

TERMO DE CONSENTIMENTO LIVRE E ESCLARECIDO

Eu, Patricia Freitas Braga, autorizo minha participação no projeto de pesquisa intitulado **ECOM - Modelo Computacional de Seleção Automática de Termos-Candidatos a partir de Mineração de Textos para Auxiliar na Construção de Ontologias,** sob a responsabilidade da consultora e pesquisadora Keller Santos de Araújo, RG 30.789.284-0 SSP SP, vinculada como aluna do mestrado MCTI da Faculdade SENAI CIMATEC, sob orientação do Prof. Dr. Renelson Ribeiro Sampaio.

Declaro que fui informada tratar-se de um estudo que pretende analisar a eleição de termos para a construção de uma ontologia. Sendo a lista de termos gerada pela ferramenta desenvolvida pela pesquisadora.

Fui informada de que estarei participando de experimento, definindo a coleção de documentos, dentro da minha área de conhecimento, que servirá de entrada para o ECOM e validando a saída do processo, a lista minerada dos termos. O experimento supracitado será realizado no SENAI CIMATEC na cidade de Salvador - Bahia.

Fui informada que os dados coletados nesta pesquisa serão divulgados única e exclusivamente para fins acadêmico-científico, ressaltando inclusive que não há riscos profissionais e nem socioemocionais para os participantes deste projeto de pesquisa.

Fui informada que os resultados serão encaminhados para publicação em revistas especializadas e apresentações em eventos científicos com o propósito de contribuir para o desenvolvimento da ciência e da sociedade. Contudo, fica firmada a garantia de sigilo das informações que possam identificar os participantes, assegurando o anonimato a eles.

A pesquisadora garantiu que acompanhará todo o desenvolvimento da pesquisa e estará à disposição para qualquer esclarecimento adicional, que se fizer necessário, antes, durante ou depois da realização da pesquisa, deixando para contato, telefone, e e-mail (Telefone: (71) 9112-5363 , e-mail: kellersa@ig.com.br.

Fui informada que este termo de consentimento é emitido em duas vias, para que eu possa ficar com uma via e a pesquisadora com a outra.

A pesquisadora esclareceu que eu posso cancelar, se assim desejar, a presente autorização, sem qualquer tipo de prejuízo sobre mim.

Estou ciente de que a participação neste projeto é livre e voluntária, assino abaixo confirmando a autorização solicitada.

Salvador, 31 de Julho de 2013.

Assinatura do Participante

Keller Santos de Araújo
Pesquisadora Responsável

Figura A.1: Termo de Consentimento Livre e Esclarecido

ANEXO A

TERMO DE CONSENTIMENTO LIVRE E ESCLARECIDO

Eu, Clelia Nobre de Oliveira, autorizo minha participação no projeto de pesquisa intitulado **ECOM - Modelo Computacional de Seleção Automática de Termos-Candidatos a partir de Mineração de Textos para Auxiliar na Construção de Ontologias,** sob a responsabilidade da consultora e pesquisadora Keller Santos de Araújo, RG 30.789.284-0 SSP SP, vinculada como aluna do mestrado MCTI da Faculdade SENAI CIMATEC, sob orientação do Prof. Dr. Renelson Ribeiro Sampaio.

Declaro que fui informada tratar-se de um estudo que pretende analisar a eleição de termos para a construção de uma ontologia. Sendo a lista de termos gerada pela ferramenta desenvolvida pela pesquisadora.

Fui informada de que estarei participando de experimento, definindo a coleção de documentos, dentro da minha área de conhecimento, que servirá de entrada para o ECOM e validando a saída do processo, a lista minerada dos termos. O experimento supracitado será realizado no SENAI CIMATEC na cidade de Salvador - Bahia.

Fui informada que os dados coletados nesta pesquisa serão divulgados única e exclusivamente para fins acadêmico-científico, ressaltando inclusive que não há riscos profissionais e nem socioemocionais para os participantes deste projeto de pesquisa.

Fui informada que os resultados serão encaminhados para publicação em revistas especializadas e apresentações em eventos científicos com o propósito de contribuir para o desenvolvimento da ciência e da sociedade. Contudo, fica firmada a garantia de sigilo das informações que possam identificar os participantes, assegurando o anonimato a eles.

A pesquisadora garantiu que acompanhará todo o desenvolvimento da pesquisa e estará à disposição para qualquer esclarecimento adicional, que se fizer necessário, antes, durante ou depois da realização da pesquisa, deixando para contato, telefone, e e-mail (Telefone: (71) 9112-5363 , e-mail: kellersa@ig.com.br.

Fui informada que este termo de consentimento é emitido em duas vias, para que eu possa ficar com uma via e a pesquisadora com a outra.

A pesquisadora esclareceu que eu posso cancelar, se assim desejar, a presente autorização, sem qualquer tipo de prejuízo sobre mim.

Estou ciente de que a participação neste projeto é livre e voluntária, assino abaixo confirmando a autorização solicitada.

Salvador, 06 de agosto de 2013.

Clelia Nobre de Oliveira
Assinatura do Participante

Keller Leal de Araújo
Keller Santos de Araújo
Pesquisadora Responsável

Figura A.2: Termo de Consentimento Livre e Esclarecido

Referências Bibliográficas

ALMEIDA, M. B. Roteiro para construção de uma ontologia bibliográfica através de ferramenta automatizada. *Perspecitivas em Ciência da Informática*, v. 8, p. 164–179, 2003. 2, 3.1, 3.1

ALMEIDA, M. B.; BAX, M. P. Uma visão geral sobre ontologias pesquisa sobre definições, tipo, aplicações, métodos de avaliaccão e construção. v. 32, 2003. 2.3.1

ARANHA, C.; PASSOS, E. A tecnologia de mineração de textos. *RESI-Revista Eletrôténica de Sistemas de Informação*, v. 2, 2006. 1.1, 2.2.1

ARISTOTELES. *Metafisica: O livro Alfa*. [S.l.: s.n.], 1978.

BARçANTE, E. *Propostas e metodologias de processamento automático de documentos textuais digitais: uma análise da literatura*. Dissertação (Mestrado) — Universidade Federal Fluminense, 2011. 2.2.2, 3.1

BARION, E. C. N.; LAGO, D. Mineração de textos. *Revista de Ciências Exatas e Tecnologia*, v. 3, 2008. (document), 2.2.2, 2.2.3, 2.12

BASÉGIO, T. L. *Uma Abordagem Semiautomática para Identificação de Estruturas Ontológicas a partir de Textos na Língua Portuguesa do Brasil*. Dissertação (Mestrado) — Universidade Católica do Rio Grande do Sul, 2007. (document), 3.1, 3.1, 3.2

BRAGA, P. F. *Um Modelo Computacional para Extração Textual e Construção de Redes Sociais e Complexas*. Dissertação (Mestrado) — SENAI CIMATEC, 2010. 4.1.2, 4.1.2

BREITMAN, K. K. *Web semântica a internet do futuro*. [S.l.: s.n.], 2005. 1.1, 2.1, 2.1, 2.1, 2.2.3, 2.3.1, 2.3.2, 2.3.3

BURNHAM, G. P. P. Definição de uma ontologia para os canais preferenciais do conhecimento técnico-científico: Fase de preparação. *VI CINFROM*, 2005. 3.1

DAVENPORT, T. H. Managing customer support knowledge. 1998. 2

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. *AI Magazine*, v. 17, 1996. (document), 2.2.1, 2.5

FEIGENBAUM, L. et al. The web semantic in action. *Scientific American*, v. 297, p. 90–97, 2007. 2.1

FERREIRA, A. B. de H. *Novo dicionário da língua portuguesa*. [S.l.: s.n.], 2010. 2

FREITAS, F. L. G. de. Ontologias e a web semân. *Anais do XXIII Congresso da Sociedade Brasileira de Computação*, v. 8, 2003. 1.1, 2.3.1, 2.3.3

FRIEDMAN, V. *Google PageRank: What Do We Know About It?* 2007. Disponível em: <http://www.smashingmagazine.com/2007/06/05/google-pagerank-what-do-we-really-know-about-it/>. 2.1

FRIZO, J. *CLUSTERING OF SCIENTIFIC FIELDS BY INTEGRATING TEXT MINING AND BIBLIOMETRICS*. Tese (Doutorado) — KATHOLIEKE UNIVERSITEIT LEUVEN, 2007. (document), 2.2.2, 2.7, 2.2.2, 2.8

GIL, A. C. *Como elaborar projetos de pesquisa*. [S.l.: s.n.], 1991. 1.5

GONCALVES, G. C. *Construção de ontologia para suporte cognitivo a um ambietne de aprendizagem*. Dissertação (Mestrado) — Universidade Federal do Rio Grando do Sul, 2008. 3.1, 3.1

GONCALVES, J. ao Carlos de A. R. *Story Mining Elicitação de Processos de Negócio a partir de Group Storytelling e Técnicas de Mineração de Texto*. Dissertação (Mestrado) — Universidade Federal do Estado do Rio de Janeiro, 2010. 2.2.3

GROUP, G. Disponível em: <http:\\www.gartner.com>. 2

JÚNIOR, R. H. de A. Precisão no processo de busca e recuperação da informação uso da mineração de textos. 2006. 1.1

KOIVUNEN, E. M. M.-R. *W3C Semantic Web Activity*. 2001. (document), 2.1, 2.3

LAKATOS, M. M. E. *Metodologia Científica*. [S.l.: s.n.], 2007. 1.5

LEE, T. B.; HENDLER, J.; LASSILA, O. The semantic web. *Scienti*, 2001. 2.1

MAEDCHE, A.; STAAB, S. Ontology learning for the semantic web. 2001. 1.1, 2.3.1

MAIA, L. C. G. *Uso de sintagmas nominais na classificação automática de documentos eletrônicos*. Tese (Doutorado) — Universidade Federal de Minas Gerais, 2008. 2.2.2

MAIA, L. C. G.; SOUZA, R. R. *MEDIDAS DE SIMILARIDADE EM DOCUMENTOS ELETRONICOS*. 2008. 2.2.2

MCGUINESS, N. F. N. D. L. Ontology development 101: A guide to creating your first ontology. 2001. Disponível em: <http://liris.cnrs.fr/alain.mille/enseignements-/Ecole_Centrale/What is an ontology and why we need it.htm>. 3.1

MSDN. *Microsoft Corporation*. Disponível em: <http://msdn.microsoft.com/pt-br-/vstudio/ff431702>. 4.1.2

MUCHERONI, M. L.; PAIVA, D. C. de; NETTO, M. L. Três ontologias e a web semântica. *Ponto de Acesso*, v. 3, p. 281 a 298, 2009. 2.3.1

OLIVEIRA, L. H. G. de. *Estração de Metadados utilizando uma ontologia de domínio*. Dissertação (Mestrado) — Universidade Federal do Rio Grande do Sul, 2009. 2.3.2

PICKLER, M. E. V. Web semântica ontologias como ferramentas de representação do conhecimento. *Perspectivas em Ciência da Informação*, v. 12, 2007. 2.1

PIRES, M. M. *Agrupamento Incremental e Hierárquico de documentos*. Dissertação (Mestrado) — Universidade Federal do Rio de Janeiro, 2008. (document), 2.2.3, 2.2.3, 2.10, 2.11

PROTEGE. 2012. Disponível em: <http://stanford.edu.br>. (document), 2.13, 2.14

ROMAO, W. *DESCOBERTA DE CONHECIMENTO RELEVANTE EM BANCO DE DADOS SOBRE CIÊNCIA E TECNOLOGIA*. Tese (Doutorado) — Universidade Federal de Santa Catarina, 2002. (document), 2.2.1, 2.2.1, 2.6

SILVA, S. L. da. Gestaão do conhecimento: uma revisão crítica orientada pela abordagem da criação do conhecimento. *Ciência da Informação Brasília*, v. 33, p. 143–151, 2004. 2, 2

SOMMERVILLE, I. *Engenharia de Software*. 9. ed. [S.l.: s.n.], 2011. 1.5, 4.1.1, 4.1.2

SOUZA, R. R.; ALVARENGA, L. A web semântica e as suas contribuições para a ciência da informação. *Ciência da Informação*, v. 33, 2004. (document), 2.1, 2.1, 2.1, 2.4

SOUZA, T.; LINDGREN, A. Mineração de texto. *Revista Eletrônica de Informatica*, v. 1, 2008. 2.2.3

TAKEUCHI, H.; NONAKA, I. *Gestão do conhecimento*. [S.l.: s.n.], 2008. (document), 2, 2, 2.1

VILLACA, N. *Impresso ou eletrônico: Um trajeto de leitura*. [S.l.: s.n.], 2002. 2.1

ECOM Modelo Computacional de Seleção Automática de Termos Candidatos a partir de Mineração de Textos para Auxiliar na Construção de Ontologias

Keller Santos de Araújo

Salvador, Agosto de 2013.