UNIVERSIDADE FEDERAL DE SANTA CATARINA
PROGRAMA DE PÓS-GRADUAÇÃO EM
ENGENHARIA E GESTÃO DO CONHECIMENTO

FLÁVIO CECI

UM MODELO SEMIAUTOMÁTICO PARA A CONSTRUÇÃO E MANUTENÇÃO DE ONTOLOGIAS A PARTIR DE BASES DE DOCUMENTOS NÃO ESTRUTURADOS

Florianópolis
2010

FLÁVIO CECI

UM MODELO SEMIAUTOMÁTICO PARA A CONSTRUÇÃO E MANUTENÇÃO DE ONTOLOGIAS A PARTIR DE BASES DE DOCUMENTOS NÃO ESTRUTURADOS

Dissertação submetida ao Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento da Universidade Federal de Santa Catarina para a obtenção do Título de Mestre em Engenharia e Gestão do Conhecimento. Área de concentração: Engenharia do Conhecimento. Linha de pesquisa: Teoria e prática em Engenharia do Conhecimento

Orientador: Alexandre Leopoldo Gonçalves, Dr.
Coorientador: Aran Bey Tcholakian Morales, Dr.

Florianópolis
2010

Catalogação na fonte pela Biblioteca Universitária
da
Universidade Federal de Santa Catarina

C388m Ceci, Flávio
Um modelo semi-automático para a construção e manutenção de ontologias a partir de bases de documentos não estruturados [dissertação] / Flávio Ceci ; orientador, Alexandre Leopoldo Gonçalves. - Florianópolis, SC, 2010.
131 p.: il., tabs.

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

Inclui referências

1. Engenharia e gestão do conhecimento. 2. Mineração de dados (Computação). 3. Ontologia. I. Gonçalves, Alexandre Leopoldo. II. Universidade Federal de Santa Catarina. Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento. III. Título.

CDU 659.2

FLÁVIO CECI

UM MODELO SEMIAUTOMÁTICO PARA A CONSTRUÇÃO E MANUTENÇÃO DE ONTOLOGIAS A PARTIR DE BASES DE DOCUMENTOS NÃO ESTRUTURADOS

Esta dissertação foi julgada adequada para a obtenção do Título de Mestre em Engenharia e Gestão do Conhecimento, Especialidade em Engenharia do Conhecimento, e aprovada em sua forma final pelo Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento.

Florianópolis, 8 de dezembro de 2010.

_____
Prof. Roberto Carlos dos Santos Pacheco, Dr.
Coordenador do Curso

Banca Examinadora:

_____
Prof. Alexandre Leopoldo Gonçalves, Dr.
Universidade Federal de Santa Catarina
Orientador

_____
Prof. Aran Bey Tcholakian Morales, Dr.
Universidade Federal de Santa Catarina
Coorientador

_____
Prof. Denilson Sell, Dr.
Universidade Federal de Santa Catarina

_____
Prof. Roberto Carlos dos Santos Pacheco, Dr.
Universidade Federal de Santa Catarina

_____
Prof. Ricardo Pietrobon, Ph.D.
Duke University

Dedico este trabalho aos meus pais, Altamiro e Margarida, aos meus padrinhos, Sérgio, Rita e Sidnei, e principalmente ao casal amigo Maria Odete Grah (*in memoriam*) e José Sergio Grah (*in memoriam*).

AGRADECIMENTOS

Para o desenvolvimento desta dissertação, foi necessário muito empenho e dedicação da minha parte, contudo o trabalho não seria possível sem a participação direta ou indireta de algumas pessoas, às quais eu gostaria de agradecer aqui neste espaço.

Aos meus pais, Altamiro e Margarida, que sempre estiveram ao meu lado oferecendo apoio e assistência constantes, quando não se sacrificando para garantir uma boa formação, e dando condições para o meu desenvolvimento profissional e acadêmico.

Aos meus irmãos, Rita e Sidnei, e ao meu padrinho Sérgio, pelo enorme apoio que me foi dado durante todo este período. Também agradeço aos meus sobrinhos pela cooperação e atenção.

À minha namorada Gláucia, pelas suas palavras de conforto e de apoio, e por todo o carinho e atenção que sempre teve comigo.

Ao meu grande amigo e orientador, professor Dr. Alexandre Leopoldo Gonçalves, que sempre teve paciência e disposição comigo, pelas incansáveis e esclarecedoras conversas e pela orientação deste trabalho. Também agradeço ao meu amigo e coorientador, professor Dr. Aran Bey Tcholakian Morales, que me acompanha em orientações desde a graduação, pela sua atenção e paciência, e pela coorientação neste trabalho.

Aos professores do Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento que aceitaram compor a Banca Examinadora, Dr. Denilson Sell e Dr. Roberto Carlos dos Santos Pacheco, bem como ao Ph.D. Ricardo Pietrobon, professor da Duke University. É uma grande honra tê-los como avaliadores deste trabalho.

Ao Instituto Stela, pela confiança e pela flexibilidade de horários, o que me possibilitou participar das atividades do Programa, e também pelo apoio tecnológico, pois pude contar com a disponibilidade do ambiente ISEKP®, que permitiu o desenvolvimento e a aplicação do modelo proposto.

Aos meus colegas de trabalho, Dr. Fabiano Beppler, Dr. Alessandro Bovo, M.Sc. Sandra Regina Martins, Dr. José Leomar Todesco, Lucas Nazário dos Santos e Júlio Gonçalves Reinaldo pelas conversas esclarecedoras e pelo apoio em geral.

À Coordenação e aos professores dos cursos de Ciência da Computação e Sistemas de Informação da Universidade do Sul de Santa Catarina, pela oportunidade de vivenciar a experiência de docência durante a concepção deste trabalho e pelo apoio, em especial à Dra. Maria Inés Castiñera, ao Dr. Ricardo Davalos e ao Dr. Mauro Madeira.

Ao Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento, pela oportunidade em participar desse prestigiado curso.

Agradeço também às demais pessoas que participaram direta ou indiretamente do desenvolvimento deste trabalho, em especial aos amigos de Santo Amaro da Imperatriz.

RESUMO

Considerando-se que grande parte do conhecimento de uma organização ou daquele disponível na web são documentos textuais, estes se tornam uma importante fonte para modelos de manutenção de ontologias. Nota-se ainda que o uso das ontologias como meio de representar formalmente o conhecimento vem crescendo em importância no desenvolvimento de sistemas baseados em conhecimento. Nesse sentido, o presente trabalho utiliza técnicas de extração de informação e agrupamento de documentos para explicitar entidades que podem tornar-se instâncias de uma ontologia de domínio. Para as fases de validação e classificação das instâncias encontradas, é proposta a utilização de bases de conhecimento colaborativas, contando-se com o auxílio de especialistas de domínio, o que se caracteriza como um processo semiautomático. Visando demonstrar a viabilidade do modelo proposto, foi desenvolvido um protótipo para suportar as fases de extração, validação e classificação dos resultados. O protótipo foi aplicado em um estudo de caso utilizando *résumés* de alguns pesquisadores, assim como em um estudo experimental mais amplo com *résumés* de pesquisadores da área de Biotecnologia. Por fim, foram analisados seis trabalhos similares com foco na aprendizagem e na população das ontologias com vistas a propiciar uma avaliação comparativa ante o modelo proposto. De modo geral, verificou-se que o modelo proposto auxilia tanto na construção inicial de uma ontologia de domínio, levando em consideração coleções de documentos (bases de dados não estruturadas), quanto no processo de manutenção de ontologias.

Palavras-chave: Extração de conhecimento. Reconhecimento de Entidades. Ontologia. Engenharia do Conhecimento.

ABSTRACT

The knowledge in organizations or which is available on the web as text has become an important source for ontology maintenance models. Also, the use of ontologies as a mean to formally represent knowledge is growing in importance in the development of knowledge-based systems. This work uses information extraction techniques and document clustering to recognize entities may become instances in domain ontologies. Additionally, we propose the use of collaborative knowledge bases and the help of domain experts to tackle the phases of validation and classification of instances, characterizing the process as semi-automatic. Aiming to demonstrate the viability of the proposed model it was developed a prototype to support extraction, validation and classification phases. Also, a case study and a broader experimental study taking into account résumés from researchers were prepared in order to analyze the achieved results by the prototype. Finally, we analyzed six similar studies focusing on learning and ontology population to establish a benchmark against the proposed model. Overall, it was found that the proposed model helps as much in the initial construction of a domain ontology taking into account collections of documents (unstructured databases) as in the process of ontology maintenance.

Keywords: Knowledge extraction. Entity Recognition. Ontology. Knowledge Engineering.

LISTA DE ILUSTRAÇÕES

LISTA DE ABREVIATURAS

CNPq – Conselho Nacional de Desenvolvimento Científico e Tecnológico
EGC – Engenharia e Gestão do Conhecimento
GC – Gestão do Conhecimento
NER – *Named Entity Recognition*
OWL – *Ontology Web Language*
POM – *Probabilistic Ontology Models*
RI – Recuperação de Informação
SBC – Sistemas Baseados em Conhecimento
SRI – Sistema de Recuperação de Informação
STC – *Suffix Tree Clustering*
XML – *eXtensible Markup Language*

SUMÁRIO

1 INTRODUÇÃO

Com o advento dos computadores e da internet, o compartilhamento de informações tem-se tornado cada vez mais rápido e eficiente, possibilitando que pessoas dispostas em vários lugares possam interagir sobre um mesmo documento ou assunto, o que aumenta a quantidade de informações disponíveis. Segundo Lastres e Albagli (1999), a informação e o conhecimento vêm desempenhando um novo papel nas economias, e isso tem provocado substantivas modificações nas relações, forma e conteúdo do trabalho, que, por sua vez, assume um caráter cada vez mais "informacional", com significativos reflexos sobre o perfil do emprego. Schreiber et al. (2002) afirmam que a vida social e econômica está tornando-se cada vez mais focada no conhecimento.

Um problema que as organizações enfrentam para trabalhar com o conhecimento é como encontrá-lo, recuperá-lo, armazená-lo e compartilhá-lo entre os seus membros.

Uma grande parte dos dados existentes nas organizações está disponível na forma textual e eletrônica. Esses arquivos textuais contêm dados, informações e ativos de conhecimentos, tais como redes de relacionamento e potenciais competências e interesses que podem auxiliar na tomada de decisão, além de possibilitar a criação de bases de conhecimento.

Para tal, é necessário primeiramente extrair os dados relevantes das bases textuais e processá-los de maneira apropriada para que façam sentido. O processo é em geral custoso e depende de um especialista no domínio.

Este trabalho apresenta um modelo que auxilia o reconhecimento de entidades a partir da coleção de documentos textuais, de modo que essas entidades reconhecidas possam servir de insumo à manutenção de estruturas formais de conhecimento, tais como taxonomias e ontologias (que é o foco da presente pesquisa), e mesmo no suporte a aplicações que se utilizem dessas estruturas.

Optou-se por armazenar as entidades reconhecidas em uma ontologia, pois elas são atualmente uma das representações de conhecimento mais utilizadas. Segundo Studer, Benjamins e Fensel (1998), uma ontologia é uma especificação explícita e formal de conceitos e relações que existem em um domínio e que são compartilhados por uma comunidade. Para Davies, Fensel e Van Harmelen (2003), ontologia é a chave para viabilizar sistemas de conhecimento e para a web semântica, pois permite que tanto as pessoas

como os sistemas entendam o contexto em questão, além de facilitar a compartilhamento e o reúso de conhecimento.

Mesmo que uma organização desenvolva uma ontologia de domínio que represente a sua visão de mundo, essa ontologia sofrerá modificações ao longo do tempo. Isso acontece uma vez que o conhecimento não é estático e as operações feitas pela organização geram insumos para novas instâncias e classes da sua ontologia de domínio. O modelo proposto neste trabalho pretende auxiliar na etapa de manutenção das ontologias de domínio de uma organização. Para tal, utilizam-se entidades reconhecidas e classificação segundo a abordagem tradicional de reconhecimento de entidades nomeadas (*named entity recognition*) e conta-se com o auxílio de bases de conhecimento colaborativas.

Segundo Zesch, Muller e Gurevych (2008), bases de conhecimento colaborativas são meios que armazenam conhecimentos construídos colaborativamente por voluntários na web. Como instâncias dessas bases podemos citar a Wikipédia e a DBpedia.

Além das entidades reconhecidas, vistas como possíveis instâncias de uma ontologia, a validação por um especialista possibilita o aprendizado do sistema e torna a base de conhecimento mais consolidada por meio de um processo colaborativo e interativo.

1.1 Definição do problema

Cada vez mais o conhecimento é mensurado como um dos bens mais valiosos das organizações. Segundo Steil (2006), pode-se analisar uma organização como um conjunto de habilidades singulares com o potencial de garantir-lhe vantagem competitiva sustentada. Nesse contexto, o conhecimento tem sido compreendido como o princípio mais importante para as organizaçãoes.

A área da Gestão do Conhecimento surge para auxiliar as organizações nos processos de criação, aquisição, representação, armazenamento, manipulação e distribuição do conhecimento organizacional, enquanto que a Engenharia do Conhecimento promove o ferramental para sistematizar e apoiar tais processos que culminam na concepção de sistemas de conhecimento (SCHREIBER et al., 2002).

Muito do conhecimento das organizações encontra-se distribuído em suas fontes de informação estruturadas (como nos sistemas de informação), semiestruturadas (como nos sites e nas páginas web), não estruturadas (como em manuais, e-mails, chats, entre outros) e também nos próprios colaboradores da organização.

Para que esse conhecimento seja manipulado e distribuído de forma a permitir sua socialização entre os membros da organização, ele deve ser adquirido (*modelo proposto*), adequadamente representado (*ontologias*) e armazenado (*memória organizacional*).

Como mencionado anteriormente, as ontologias têm sido utilizadas na representação do conhecimento e vêm auxiliando em várias áreas relacionadas à Engenharia e Gestão do Conhecimento, como, por exemplo, web semântica, redes sociais, recuperação de informação, mineração de dados, memória e aprendizagem organizacional, entre outras.

Para Souza (2003), as ontologias são usadas como forma de representação e integração do conhecimento pela sua capacidade de reúso e interoperabilidade. Outra utilização é o fato de serem empregadas como uma linguagem comum entre agentes participantes de determinado sistema ou organização, possibilitando assim a socialização do conhecimento.

A construção de uma ontologia é uma atividade que requer um especialista no domínio, o qual deverá definir os termos do domínio e suas relações semânticas. Segundo Noy e McGuinness (2001), para a construção de uma ontologia são utilizados vários processos que partem de determinar o escopo, considerar o reúso, enumerar termos, definir classes, propriedades e restrições, e por fim criar as instâncias.

Os autores Noy e McGuinness (2001) ainda afirmam que, como principais problemas na criação e manutenção de ontologias, pode-se identificar:

- a falta de agilidade no acompanhamento das classes do domínio a fim de manter a ontologia atualizada;
- o conhecimento não é estático, motivo pelo qual as ontologias precisam de constantes atualizações;
- a difícil identificação da evolução das classes e instâncias para refletirem nas ontologias;
- a necessidade de um especialista no domínio em questão para construir e manter as ontologias; e
- a grande quantidade de tempo que é despendido pelo engenheiro de ontologias para levantar as informações e poder classificá-las.

A partir do contexto acima declarado, apresenta-se a seguinte pergunta de pesquisa: como identificar e classificar elementos essenciais para a construção e manutenção de ontologias de domínio a partir de fontes de informação não estruturadas e bases de conhecimento

colaborativas de modo que o resultado desse processo possa ser integrado a sistemas de Engenharia e Gestão do Conhecimento?

1.2 Objetivos do trabalho

1.2.1 Objetivo geral

Desenvolver um modelo semiautomático que promova suporte ao processo de manutenção de ontologias a partir de informação não estruturada e de bases de conhecimento colaborativas visando auxiliar no entendimento de determinado domínio de problema.

1.2.2 Objetivos específicos

- Identificar técnicas para reconhecimento de entidades em informação não estruturada de maneira automática.
- Identificar meios automáticos e/ou semiautomáticos para auxiliar no processo de manutenção de ontologias.
- Propor um modelo de extração de entidades e manutenção de ontologias que leve em consideração os objetivos específicos anteriores.
- Demonstrar a viabilidade do modelo proposto por meio da construção de um protótipo, assim como a aplicação deste em estudos de caso.
- Realizar uma análise comparativa com outros modelos de extração de entidades e manutenção de ontologias.

1.3 Justificativa e relevância do tema

Atualmente as organizações têm verificado que a utilização dos dados operacionais em técnicas e sistemas de apoio à decisão por si só não auxilia completamente no planejamento para o futuro organizacional. Segundo Vasconcelos, Rocha e Kimble (2003), as organizações baseadas em conhecimento possuem pessoas altamente qualificadas cujo papel é o de solucionar problemas. Esses problemas envolvem tarefas complexas de manipulação de conhecimento, tais como lidar com abstração e ambiguidade ou reconhecimento de padrões. O conhecimento nesse tipo de organização pode ser visto como um produto de inteligência, experiência e qualificação de membros e grupos

de trabalho que tornam possível a viabilidade e o sucesso da organização.

Desse modo, faz-se necessário o mapeamento do conhecimento distribuído pela organização para auxiliar na criação de uma base de conhecimento que dará insumos para a construção de sistemas baseados em conhecimento. Segundo Brachman e Levesque (2004), para as pessoas tomarem decisão sobre o que fazer numa determinada ocasião é necessário basear as soluções em algo que elas já conhecem (ou em que acreditam).

De acordo com Chaves (2007), uma grande parte do conhecimento existente atualmente está na forma de texto (a maioria não estruturado), e por esse motivo esse conhecimento precisa ser identificado, representado e manipulado, de modo a tornar-se realmente útil para as organizações. Para estruturar o conhecimento existente nas fontes de informação das organizações, são utilizadas as ontologias.

Para Navigli e Velardi (2004), o processo de criação das ontologias é algo que demanda tempo e envolve especialistas de vários campos. Verifica-se por essa perspectiva que a utilização de sistemas que auxiliem na construção e manutenção de ontologias de maneira semiautomática torna-se uma possível alternativa para a redução do tempo demandado nessa tarefa, o que possibilita ao especialista um trabalho mais voltado à avaliação do resultado encontrado nas bases não estruturadas do que ao levantamento de conceitos e relações de forma manual.

Nota-se ainda que o processo de manutenção de ontologias de domínio é contínuo, já que o conhecimento organizacional está em constante evolução. Considerando-se que o modelo proposto utiliza bases textuais como fonte de informação, a incorporação de novos documentos apontará possíveis novas instâncias e classes para determinada ontologia de domínio.

1.4 Escopo do trabalho

O foco deste trabalho é a elaboração de um modelo para manutenção de ontologias que está diretamente ligado à Engenharia do Conhecimento. Com a formalização desse conhecimento, o seu reúso e compartilhamento, objetiva-se promover um impacto positivo nas atividades organizacionais, uma vez que existe uma linguagem comum de interação entre os vários componentes de determinado sistema.

Também se pode apresentar como foco a aplicação das técnicas das áreas de reconhecimento de entidades/extração de informação,

agrupamentos e recuperação de informação para auxiliar nos processos de aquisição e criação do conhecimento (não se aplica ao conhecimento tácito). Salienta-se que não faz parte do escopo deste trabalho contribuir ou evoluir os processos acima citados, os quais serão apenas utilizados para o desenvolvimento do modelo.

Outra característica deste trabalho é o fato de estar mais centrado na parte de levantamento das entidades (instâncias de uma ontologia), as quais alimentarão a base de conhecimento, do que no reconhecimento automático das classes. Cabe salientar que a abordagem de classificação das instâncias encontradas no presente trabalho é semiautomática, tendo o usuário um importante papel na fase de validação.

Destaca-se ainda que a dissertação baseia-se no pressuposto de que o conhecimento contido em coleções de documentos não estruturados pode ser explicitado em um determinado nível por meio das técnicas apresentadas neste modelo.

1.5 Metodologia da pesquisa

A presente seção visa descrever a metodologia utilizada no trabalho a fim de classificar a pesquisa nos diversos pontos de vista. Segundo Gil (1999, p. 42, apud SILVA; MENEZES, 2001, p. 19), "o objetivo fundamental da pesquisa é descobrir respostas para problemas mediante o emprego de procedimentos científicos".

O trabalho aqui apresentado, sob o ponto de vista de sua natureza, é caracterizado por ser uma pesquisa aplicada, a qual, conforme Silva e Menezes (2001, p. 20), "objetiva gerar conhecimentos para aplicação prática dirigidos à solução de problemas específicos. Envolve verdades e interesses locais".

Para atingir os objetivos desta pesquisa, o trabalho dividiu-se nas seguintes etapas:

- levantamento de um referencial teórico para auxiliar na concepção do modelo proposto e na escrita do trabalho. Abordaram-se temas como reconhecimento de entidades nomeadas, recuperação de informação e *clustering* (agrupamento) para facilitar a etapa de extração das possíveis instâncias de uma ontologia. Adicionam-se ainda alguns temas transversais a este trabalho, tais como gestão do conhecimento, técnicas para extração e aquisição de conhecimento, inteligência coletiva, utilização de bases colaborativas, indexação, além do foco principal, que é a manutenção de ontologias;

- proposição de um modelo para atender aos objetivos deste trabalho;
- desenvolvimento de um protótipo e correspondente aplicação em um estudo de caso visando auxiliar na demonstração de viabilidade e validação do modelo proposto;
- análise comparativa do modelo proposto com outros modelos para manutenção em ontologias de maneira semiautomática, modelos estes disponíveis na literatura; e
- apresentação das conclusões e de possíveis trabalhos futuros.

1.6 Aderência ao objeto de pesquisa do programa

Este trabalho está inserido na linha de pesquisa de teoria e prática em Engenharia do Conhecimento, que tem como foco estudar as metodologias e técnicas dessa área e suas relações com a gestão do conhecimento.

No que tange ao escopo deste trabalho, o aspecto que o contextualiza na área de Engenharia do Conhecimento reside no fato de o modelo ter como objetivo a materialização, principalmente dos macroprocessos de explicitação do conhecimento, sem perder a possibilidade de promover suporte aos macroprocessos de gestão e disseminação do conhecimento.

A aderência deste trabalho ao objeto de pesquisa do Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento[1] pode ser reforçada a partir do objetivo de pesquisa e objetivo principal do Programa:

> O objeto de pesquisa do EGC refere-se aos macroprocessos de explicitação, gestão e disseminação do conhecimento. Estes incluem os processos de criação (e.g., inovação de ruptura), descoberta (e.g., redes sociais), aquisição (e.g., inovação evolutiva), formalização/codificação (e.g., ontologias), armazenamento (e.g., memória organizacional), uso (e.g., melhores práticas), compartilhamento (e.g., comunidades de prática), transferência (e.g., educação corporativa) e evolução (e.g., observatório do conhecimento). [...] Deste modo, o objetivo do EGC consiste em investigar, conceber, desenvolver e aplicar

[1] Disponível em: <http://www.egc.ufsc.br/htms/vermais_index.htm>. Acesso em: 10 out. 2010.

> modelos, métodos e técnicas relacionados tanto a processos/bens/serviços como ao seu conteúdo técnico-científico [...].

1.7 Estrutura do trabalho

Este trabalho é composto de cinco capítulos além da introdução que aqui se apresenta, sendo os demais relacionados a seguir.

- o segundo capítulo é composto de um referencial teórico no qual se apresentam as áreas de Recuperação de Informação, Reconhecimento de Entidades Nomeadas, Indexação, Agrupamento de documentos a partir de busca em documentos, Manutenção de Ontologias e Inteligência Coletiva;
- no capítulo 3, apresenta-se o modelo proposto por meio de uma descrição detalhada sobre as três etapas que o compõem, sendo: (1) reconhecimento de entidades; (2) validação; e (3) classificação;
- no capítulo 4, é apresentada a proposição de avaliação do modelo através da discussão dos resultados alcançados por meio dos estudos de casos e por uma análise comparativa com modelos correlatos ao proposto neste trabalho; e
- o quinto e último capítulo apresenta as conclusões da dissertação bem como os potenciais trabalhos vislumbrados.

2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo tem como objetivo apresentar o referencial teórico dos temas que são imprescindíveis para o desenvolvimento do presente trabalho.

2.1 Introdução

Não é novidade que as organizações utilizam seus dados operacionais para gerar informações que as auxiliem na tomada de decisão. No entanto, elas vêm percebendo que não só os dados de suas bases operacionais podem ser fonte de informação, visto que muita informação está implícita nos seus documentos textuais.

Segundo Fialho et al. (2006), a informação e o conhecimento são as principais armas competitivas da nossa era. Conforme apontam Drumond e Girardi (2010), o conhecimento é um dos grandes fatores de sucesso de uma organização. Cada vez mais ele vem ganhando destaque como um bem valioso para uma organização (STEIL, 2006, p. 4). De acordo com Schreiber et al. (2002), da mesma forma que com a Revolução Industrial ganharam foco áreas como engenharia elétrica e engenharia mecânica, na era da informação e do conhecimento áreas como a Engenharia do Conhecimento são de grande importância.

No entanto, essa área já existe há alguns anos. Surgiu na década de 70 para o desenvolvimento de sistemas especialistas e nasceu dentro da área da inteligência artificial (SCHREIBER et al., 2002, p. 6). A Engenharia do Conhecimento sofreu evoluções ao longo dos anos. Na década de 80, o desenvolvimento de Sistemas Baseados em Conhecimento (SBC) era realizado por meio do paradigma de transferência, que consiste em transferir o conhecimento humano para implementar regras heurísticas. Nessa época, tal conhecimento era apenas coletado e implementado, sendo utilizado para resolver tarefas bem específicas (STUDER; BENJAMINS; FENSEL, 1998, p. 162-163).

Com o tempo, percebeu-se que a representação de conhecimento somente por meio de regras não permitia ou limitava uma das principais características dos SBC atuais, que é a possibilidade de reúso. Essa deficiência gerou a necessidade da evolução do paradigma de transferência para o paradigma de modelagem.

Segundo Studer, Benjamins e Fensel (1998), o paradigma de modelagem, como o próprio nome sugere, preocupa-se com as atividades de modelagem de SBC e com a capacidade de resolver problemas de um domínio, ou seja, de uma área de conhecimento, não

focando apenas no conhecimento de um especialista, mas sim no da organização. Algumas características desse paradigma:

- este modelo apenas se aproxima da realidade. Pelo fato de ser cíclico, ele pode estender-se além do domínio desejado, sendo necessária a criação de limites (fronteiras);
- o processo de modelagem é cíclico, ou seja, está em constante modificação (sempre buscando o refinamento, a modificação ou a complementação do modelo existente);
- o processo de modelagem é dependente da interpretação subjetiva do engenheiro do conhecimento. Esse processo é sujeito a falhas, mas está em constante evolução e cada vez mais adequado para o domínio em questão.

Na próxima seção, são apresentados os conceitos relacionados com o conhecimento, que é o fator-chave de estudo da Engenharia e Gestão do Conhecimento.

2.1.1 Gestão do Conhecimento

Visões de negócio e estruturas organizacionais tradicionais estão sofrendo diversas mudanças. Um dos motivos para isso são as pressões de negócios e os novos mercados nesse mundo globalizado. A informação tornou-se cada vez mais importante, sendo um bem tangível para as organizações, e, de forma natural, o conhecimento passa a ser primordial para o seu sucesso (PINHEIRO, 2008, p. 289-294).

Segundo Silvia e Spitz (2006), as organizações possuem conhecimento disseminado e compartilhado, mas também existem conhecimentos pertencentes a seus membros, grupos ou áreas de trabalho. É de interesse das organizações identificar e codificar esse conhecimento para torná-lo acessível a todos.

Gestão do Conhecimento (GC) é um conjunto de processos que auxiliam na criação, aquisição, representação, no armazenamento, na manipulação, distribuição e utilização do conhecimento. A GC complementa e apresenta outras iniciativas organizacionais, como a gestão de qualidade, a reengenharia de processos e o aprendizado organizacional, trazendo consigo benefícios para a competitividade (KRUGLIANSKAS; TERRA, 2003; SUN; HAO, 2006).

De acordo com Fialho et al. (2006), como relacionado a seguir, muitos autores que estudam a GC fazem questão de distinguir dado, informação e conhecimento, pois o desentendimento sobre esses conceitos pode gerar um enorme dispêndio para uma organização.

- **Dado**: são representações simbólicas para a descrição de atributos de qualquer nível (FIALHO et al., 2006, p. 71). Segundo Pinheiro (2008), a camada de dados é responsável pela existência e pela operação dos sistemas transacionais, sistemas estes que têm como responsabilidade apoiar as operações da organização.
- **Informação**: é o conjunto de dados que são devidamente processados e tornam-se compreensíveis, ou seja, a informação é a disposição dos dados de uma forma que apresentem um significado, criando padrões e acionando significados na mente dos indivíduos (FIALHO et al., 2006, p. 71-72).
- **Conhecimento**: é a combinação completa de informação, dados e relações que levam os indivíduos à tomada de decisão, ao desenvolvimento de novas informações ou conhecimentos e à realização de tarefas (FIALHO et al., 2006, p. 72-77).

Concluindo a análise entre dado, informação e conhecimento, Fialho et al. (2006, p. 72) afirmam:

> Uma das principais causas da dificuldade de se especificar o que é conhecimento está no fato de que ele depende muito do contexto. Porém, a perspectiva com que se interpreta o conhecimento é importante na medida em que vai determinar a forma como a gestão em uma empresa é abordada. O conhecimento de uma pessoa pode ser apenas dado para outra pessoa. Os limites entre dado, informação e conhecimento não são rígidos porque dependem do contexto de uso [...].

A Gestão do Conhecimento pode ser considerada como uma gestão de recursos da organização, a qual utiliza e captura o conhecimento das pessoas, das equipes e da organização em seu conjunto (VASCONCELOS; ROCHA; KIMBLE, 2003). Segundo Schreiber et al. (2002), a gestão do conhecimento desempenha o papel de uma arquitetura para melhorar a infraestrutura do conhecimento, tendo com função adquirir o conhecimento certo para a pessoa correta em um formato e tempo adequados.

Conforme aponta Steil (2007), a gestão do conhecimento possui sete processos:

1. criação do conhecimento;
2. compartilhamento do conhecimento;
3. armazenamento da informação e do conhecimento;
4. distribuição da informação e do conhecimento;
5. aquisição da informação e do conhecimento;
6. utilização da informação e do conhecimento; e
7. reutilização da informação e do conhecimento.

Para Fialho et al. (2006), a GC trata da prática de juntar valor à informação e de distribuí-la, utilizando como tema central o aproveitamento dos recursos existentes na organização.

Para apoiar e executar esses processos da GC surge a Engenharia do Conhecimento, que utiliza todo o ferramental computacional disponível. Na próxima seção, são apresentados alguns conceitos sobre extração e aquisição do conhecimento.

2.1.2 Extração e aquisição do conhecimento

Cada vez mais as organizações estão procurando desenvolver sistemas baseados em conhecimento para auxiliar os seus processos de tomada de decisão. O grande desafio desse tipo de sistema encontra-se na criação da base de conhecimento. O engenheiro do conhecimento tem como função estudar o domínio em questão e, por meio de interações com os especialistas no domínio, criar um modelo para representar esse domínio. Como se trata de um processo muito custoso, os engenheiros do conhecimento procuram apoio computacional para desenvolver soluções automáticas ou semiautomáticas para a aquisição do conhecimento que irá compor a base de conhecimento (GARCIA; VAREJÃO; FERRAZ, 2005, p. 51-53).

Para ser adquirido, o conhecimento não precisa ter sido recentemente criado, só precisa ser novo (desconhecido) para a organização (DAVENPORT; PRUSAK, 2000, p. 53-56). Segundo Calhoun e Starbuck (2005), a aquisição do conhecimento é um processo que acessa o conhecimento existente. Essas informações e conhecimentos estão nas bases dos sistemas de informação, nas redes sociais e nos documentos da organização.

De acordo com Drumond e Girardi (2010), tradicionalmente o desenvolvimento de bases de conhecimento tem sido realizado manualmente por especialistas do domínio e por engenheiros do conhecimento. No entanto, essa é uma tarefa cara e sujeita a erros. Tal dificuldade em explicitar o conhecimento implícito nos textos e nas bases de dados é chamado de *aquisição de conhecimentos*, e superar

esse problema é crucial para o sucesso de aplicações baseadas em conhecimento.

Sobre as técnicas para a aquisição do conhecimento, Garcia, Varejão e Ferraz (2005) apresentam cinco categorias de técnicas:

1) técnicas manuais baseadas em entrevistas, em modelos ou em acompanhamento;
2) técnicas semiautomáticas baseadas em teorias cognitivas ou em modelos que já existem;
3) técnicas que utilizam aprendizado de máquina tentando induzir regras a partir de exemplos catalogados;
4) técnicas que utilizam mineração de dados, a partir da qual se busca extrair regras e comportamentos com base em análises de grandes massas de dados;
5) técnicas que aplicam mineração de texto para extrair o conhecimento de uma grande quantidade de dados não estruturados.

Este trabalho possui como foco as técnicas apresentadas na categoria 5 para fazer a aquisição e a extração do conhecimento. O detalhamento das técnicas utilizadas é apresentado na seção 2.3.

Depois de adquirido o conhecimento implícito nas bases textuais da organização, o grande problema é como representá-lo de tal modo que possa ser utilizado, compartilhado e reutilizado. A próxima seção aborda a representação do conhecimento.

2.1.3 Representação e armazenamento do conhecimento

Uma vez que o conhecimento tenha sido adquirido de fonte humana ou extraído automaticamente, é necessário representá-lo para que possa ser armazenado e compartilhado (DAVIES; FENSEL; VAN HARMELEN, 2003, p. 1-9).

Vasconcelos, Rocha e Kimble (2003) afirmam que, para manter e estruturar grandes quantidades de informação heterogênea, é necessário aplicar uma metodologia que permita classificar, reconhecer e reutilizar os recursos de conhecimento das organizações.

Ao longo das décadas, como afirmam Studer, Benjamins e Fensel (1998), a Engenharia do Conhecimento trocou de paradigma, visto que antes era focada em transferência de conhecimento e, posteriormente, em modelagem. Em ambos os paradigmas, o processo de representação e armazenamento do conhecimento está presente. No paradigma de transferência era identificado o conhecimento de um especialista, e depois esse conhecimento era transformado em regras que fariam parte

da base de conhecimento do sistema. Com o tempo, percebeu-se que esse processo era bastante custoso e possibilitava pouco reúso do conhecimento identificado, motivo pelo qual surge o paradigma voltado à modelagem do conhecimento.

As ontologias são usadas como forma para representação do conhecimento, logo possibilitam a reutilização e transmissão deste, além de ser uma forma estruturada para o seu armazenamento com a utilização do conceito de classes, relações, atributos, etc. (GÓMEZ-PÉREZ; FERNÁNDEZ-LÓPEZ; CORCHO, 2004, p. 47-56).

Como já foi dito anteriormente, além de representar o conhecimento, as ontologias podem ajudar o armazenamento desse conhecimento de forma estrutural e formal, auxiliando na construção de uma memória organizacional (VASCONCELOS; ROCHA; KIMBLE, 2003).

Percebe-se a importância do uso das ontologias para auxiliar os processos da Gestão do Conhecimento. Nesse sentido, na próxima seção são apresentados o conceito de ontologias, as etapas para a sua construção, seu uso e seus benefícios.

2.2 Ontologias

Vasconcelos, Rocha e Kimble (2003) definem ontologias como uma especificação formal de alto nível de um domínio de conhecimento. Para Almeida e Bax (2003), uma ontologia define as regras que regulam a combinação entre os termos e as relações em um domínio do conhecimento, sendo geralmente desenvolvida por especialistas.

Uma ontologia é constituída por partes semelhantes às que compõem o paradigma orientado a objeto de desenvolvimento de software. Segundo Noy e McGuinness (2001), as ontologias possuem conceitos de domínios por meio de classes e subclasses. As propriedades dos conceitos são representadas através dos atributos (slots), as restrições sobre as propriedades são demonstradas por meio dos tipos (cardinalidade) e as relações entre os conceitos, através das igualdades e disjunções.

As ontologias são divididas em vários tipos de acordo com o seu grau de generalidade. Guarino (1998) ilustra a seguir alguns desses tipos.

- Ontologias gerais (*top-level ontology*): possuem definições abstratas para a compreensão de aspectos do mundo, como, por exemplo, processos, espaços, tempo, coisas, seres, etc.

- Ontologias de domínio (*domain ontology*): tratam de um domínio específico de uma área genérica, como, por exemplo, uma ontologia sobre família.
- Ontologias de tarefa (*task ontology*): tratam de tarefas genéricas ou atividades (como diagnosticar ou vender).
- Ontologias de aplicação (*application ontology*): têm como objetivo solucionar um problema específico de um domínio, normalmente referenciando termos de uma ontologia de domínio.

Freitas (2003) apresenta outros tipos de ontologias a partir do trabalho de Guarino apresentado anteriormente:

- Ontologias de representação: definem as primitivas de representação, tais como *frames*, atributos, axiomas, etc., na forma declarativa. Com isso, são abstraídos os formalismos de representação, o que também traz desvantagens.
- Ontologias centrais (genéricas de domínio): definem os ramos de estudo de uma área ou conceitos mais abstratos dessa área.

A seção a seguir traz algumas aplicações e benefícios no uso das ontologias para a Gestão do Conhecimento.

2.2.1 Uso e benefícios das ontologias

Sabe-se que as ontologias têm como principal função a representação do conhecimento e que essa representação pode auxiliar uma série de aplicações e sistemas de conhecimento. Segundo Almeida (2003), as ontologias podem ser aplicadas sobre uma fonte de dados, o que proporciona uma maior organização e, consequentemente, uma recuperação mais eficiente. Essa aplicação possibilita uma compreensão de um domínio compartilhando entre pessoas e sistemas, adicionando estruturas semânticas a uma fonte de dados e desenvolvendo um intercâmbio de informações.

Sobre os benefícios do uso das ontologias, Freitas (2003) faz as seguintes afirmações:

- há a possibilidade de se reusarem as ontologias e bases de conhecimento pelos desenvolvedores mesmo com adaptações e extensões, já que, segundo o autor, a fase de construção de bases de conhecimento é a etapa mais cara e demorada;

- as ontologias permitem aos usuários efetuarem consultas, integrações, comparações e checagem de consistência do conteúdo destas;
- há a possibilidade de tradução entre diferentes linguagens e formalismos de representação de conhecimento;
- existe uma vasta quantidade de ontologias disponíveis em bases de conhecimento na web para reúso e que possibilitam um vocabulário uniforme;
- há mapeamento entre formalismo de representação de conhecimento que é inspirado no componente de conectividade para sistemas gerenciadores de banco de dados.

Além disso, as ontologias auxiliam em várias áreas do conhecimento. Morais (2006) cita algumas dessas áreas abaixo.

- Recuperação de informação: permite a reutilização de ontologias na web semântica, provendo estrutura de buscas em banco de dados de ontologias e em outros documentos semânticos na internet.
- Processamento de linguagem natural: pode auxiliar em processos de tradução de textos de uma área específica, como, por exemplo, nos significados dos termos médicos.
- Gestão do conhecimento: possibilita o armazenamento da memória corporativa da empresa por meio do uso das ontologias.
- Web semântica: através das ontologias, é possível dar o sentido semântico à web.

Na seção a seguir são descritos o processo de construção de ontologias e suas partes principais.

2.2.2. Construção de ontologias

Antes de descrever o processo de criação de ontologias, deve-se entender alguns conceitos relacionados a elas. Os conceitos a seguir foram sumarizados a partir de Fensel, (2001), Gómez-Pérez, Fernández-López e Corcho (2004), Gruber (1993) e Morais (2006).

- Classes: geralmente são organizadas em forma de taxonomia e representam algum tipo de interação da ontologia com o domínio.
- Relações: representam o tipo de interação entre as classes (elementos) do domínio.

- Axiomas: são utilizados para modelar sentenças verdadeiras.
- Instâncias: representam elementos específicos, os próprios dados das ontologias (geralmente estão ligadas a uma classe, como instância de uma classe).
- Funções: eventos que podem ocorrer no contexto da ontologia.

Sobre a construção das ontologias, Noy e McGuinness (2001) listam as principais etapas:

- determinar o domínio e o escopo da ontologia;
- reutilizar ontologias existentes;
- enumerar termos importantes da ontologia;
- definir as classes e as suas hierarquias;
- definir as propriedades das classes (*slots*);
- definir as restrições; e
- criar as instâncias.

2.3 Recuperação de Informação

A área da recuperação de informação (RI) nasceu entre as décadas de 40 e 50, quando surgiu a necessidade de se encontrarem documentos de maneira eficiente. Um dos autores mais clássicos dessa área é Salton (1968), que afirma que a recuperação de informação preocupa-se com os principais processos para lidar com a informação, como, por exemplo:

- estrutura;
- análise;
- organização;
- representação; e
- recuperação e busca das informações.

O objetivo principal da RI é tornar o acesso mais fácil aos documentos de maior relevância conforme a necessidade de informação do usuário. Essa necessidade normalmente é simbolizada por meio de um busca por palavra-chave. A recuperação de informação, nesse contexto, consiste basicamente na determinação de quais documentos de uma coleção contêm as palavras-chave da consulta realizada pelo usuário. A dificuldade está não somente em extrair a informação, mas também em decidir a sua relevância.

Segundo Russell e Norvig (2004), um sistema de recuperação de informação (SRI) tem como características:

- possuir uma coleção de documentos;

- ter como entrada uma consulta apresentada em uma linguagem de consulta;
- encontrar um conjunto de resultados; e
- apresentar um conjunto de resultados que atenda à consulta.

Para que os sistemas de recuperação de informação armazenem e recuperem os documentos baseados em palavras-chave, primeiramente os SRIs devem representar e organizar esses documentos. Nesse sentido, existe um processo de transformação de um documento em um vetor de palavras-chave que representa o documento. O sistema de recuperação de informação deve usar outras técnicas para auxiliar a sua busca, como as apresentadas a seguir.

- Stopwords: para Korfhage (1997), as palavras classificadas como *stopwords* pertencentes a um documento trazem consigo duas influências para os SRIs. A primeira delas é que devido à sua grande ocorrência nos documentos, há uma grande influência no grau de frequência das palavras; já a segunda é o processamento desnecessário dessas palavras que não auxiliam na busca do usuário. O autor sugere ainda que, antes de se submeter o documento ao processo de indexação, deve-se limpar todas as ocorrências dessas *stopwords* nos documentos. Esse mesmo processo é realizado para os termos que forem pesquisados.
- Stemming: como Ebecken, Lopes e Costa (2005) explicam, os algoritmos de *stemming* processam separadamente todas as palavras do texto, tentando trabalhar com a sua possível palavra-raiz. Eles não se apegam ao contexto da palavra, pois os ganhos obtidos em precisão não justificam a grande quantidade de erros decorridos de uma análise de sentido equivocado.

Para avaliar a eficiência de um sistema de recuperação de informação, é necessário levar em consideração duas medidas: (1) precisão e (2) revocação.

A primeira mede a proporção de documentos no conjunto de resultados que são relevantes. Para auxiliar o entendimento da medida de precisão, é apresentada a fórmula abaixo.

$$precisão = \frac{|\{documentos\_recuperados\_relevantes\} \cap \{documentos\_recuperados\}|}{|\{documentos\_recuperados\}|}$$

Por exemplo, considere uma consulta em que são recuperados 60 documentos relevantes num total de 300 documentos recuperados. Nesse caso, a precisão é de 0,2.

$$precisão = \frac{60 \cap 300}{300} = 0{,}2$$

Já a revocação mede a proporção de todos os documentos relevantes para a coleção de documentos que estão no conjunto de resultados. Para auxiliar o entendimento da medida de revocação, é apresentada a fórmula a seguir.

$$revocação = \frac{|\{documentos\_recuperados\_relevantes\} \cap \{documentos\_recuperados\}|}{|\{documentos\_relevantes\}|}$$

Para a revocação, considere o seguinte exemplo. Em uma consulta são recuperados 60 documentos relevantes num total de 300 documentos recuperados, sendo que o total de documentos relevantes para a consulta em questão é de 75. Nesse caso, a revocação é de 0,8.

$$revocação = \frac{60 \cap 300}{(60+15)} = 0.8$$

Quando se lida com uma base muito grande de informações (como no caso da internet), é necessário avaliar a recuperação utilizando-se amostragem (RUSSELL; NORVIG, 2004, p. 816-817).

Para os sistemas de recuperação efetuarem as buscas, são utilizados alguns modelos de recuperação de informação, os quais são apresentados na próxima seção.

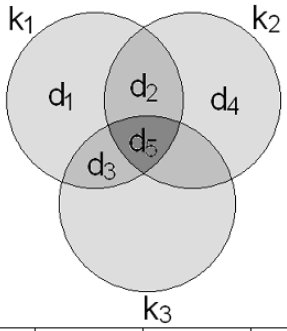
2.3.1 Modelos de recuperação de informação

Nesta seção, são abordados os principais modelos de recuperação de informação que dão suporte ao motor de busca utilizado neste trabalho.

2.3.1.1 Modelo booleano

Bastante simples para se trabalhar com recuperação de informação, o modelo booleano é baseado nos conceitos da lógica ou álgebra booleana, que utilizam os operadores lógicos E, OU e NÃO para

refinar as consultas (KORFHAGE, 1997, p. 51-62). A Figura 1 a seguir procura exemplificar o funcionamento do modelo booleano.



	k_1	k_2	k_3
d_1	1	0	0
d_2	1	1	0
d_3	1	0	1
d_4	0	1	0
d_5	1	1	1

$q = k_1$ and (k_2 or (not k_3))

Figura 1 - Representação do modelo booleano

Considerando-se que *q* representa a busca feita pelo usuário, verifica-se que são solicitados todos os documentos que possuam a palavra *k1* e a palavra *k2* ou que não possuam a palavra *k3*. O resultado dessa busca seriam os documentos *d1* e *d2*.

Para exemplificar de forma prática como seria uma busca utilizando-se o modelo booleano, imagina-se que o usuário deseja comprar um DVD ou CD. Ele poderia solicitar na busca desta forma: *(dvd or cd)*. Ele ainda poderia definir o artista do CD e do DVD da seguinte maneira: *(dvd or cd) and Titãs*. Utilizando essa expressão, o usuário teria todos os CDs e DVDs dos Titãs. Se o usuário desejar limitar mais ainda essa busca, ele pode excluir os álbuns que já tem por meio da seguinte expressão: *(dvd or cd) and Titãs and not (titanomaquia or domingo)*.

As consultas baseadas em lógica são bastante simples de serem montadas e possibilitam uma boa representação do resultado esperado, mas este modelo possui alguns problemas já conhecidos que Ebecken, Lopes e Costa (2005) citam a seguir.

- O tamanho do resultado não pode ser controlado, podendo conter milhares de itens.
- Como não é possível utilizar pesos nos termos na consulta, os resultados não são ordenados de acordo com a relevância.
- A seleção dos termos que irá compor a consulta, quando não feita por um especialista no domínio, pode ser bastante complicada.

Um modelo que é mais elaborado e que se preocupa com os problemas listados acima é o modelo espaço vetorial, o qual é apresentado na seção a seguir.

2.3.1.2 Modelo espaço vetorial

O modelo espaço vetorial (ou apenas modelo vetorial) representa cada documento como um vetor ou uma lista ordenada de termos e um peso. Esse peso pode ser considerado como o grau de importância aplicado num espaço euclidiano de *n* dimensões, em que *n* é o número de termos (KORFHAGE, 1997, p. 63-69).

Segundo Cardoso (2000), para identificar o resultado da consulta, o modelo vetorial faz um cálculo de similaridade que permite gerar o vetor resultado. Cada vetor é a representação de um termo da consulta ou de um documento, e entre esses vetores forma-se um ângulo denominado Θ. É por meio do cosseno desse ângulo que podemos identificar a semelhança existente.

Segundo Manning, Raghavan e Schutze (2008), os termos mais relevantes de um documento são calculados normalmente pelo *TF-IDF*, em que *DF* é a frequência dos documentos (ou *Document Frequency*) que contêm o termo a ser pesquisado. Outro dado que precisa ser extraído é o *IDF*, que é calculado a partir do próprio *DF*. Trata-se da quantidade de vezes em que o termo é encontrado no conjunto de documentos. A fórmula será ilustrada a seguir.

$$IDF = \log\left(\frac{|D|}{DF}\right)$$

O "D" representado na fórmula é a cardinalidade dos documentos armazenados. Agora que se têm todos os termos para calcular o TF-IDF(d,v), onde "d" é dimensão e "v" o vetor, devemos aplicar os valores na seguinte fórmula:

$$TF - IDF(d,v) = TF(d,v) \times IDF(d)$$

O modelo espaço vetorial assume que um documento textual *di* é representado por um conjunto de palavras (*t1, t2, t3, ..., tn*), em que cada *ti* é uma palavra que aparece no documento textual *di* e *n* representa o número total de várias palavras usadas para identificar o significado do documento textual. A palavra *ti* possui um correspondente peso *wi* calculado como resultado da combinação estatística *TF(wi,di)* e *IDF(wi)*, portanto *di* pode ser representado como um vetor específico *n*-dimensional *di = (w1, w2, ..., w3)*. Peso é a medida que indica a importância estatística das palavras correspondentes. O peso *wi* da palavra *ti* pode ser determinado pelo valor de *TF(wi,di) x IDF(wi)*. O valor TF é proporcional à frequência das palavras no documento enquanto que o valor do IDF é inversamente proporcional à frequência do documento no *corpus* (ZHANG; GONG; WANG, 2005, p. 49-55).

Para demonstrar a utilização desse modelo, vamos chamar a consulta de "cons" e o documento de "doc". Imagina-se que a consulta solicitada pelo usuário seja "vento onda vida" e que o documento que irá ser comparado tem o seguinte conteúdo: "Posso ouvir o vento passar, assistir a onda bater, mas o estrago que faz, a vida é curta para ver [...]". Então temos os seguintes vetores: cons = {(vento, 0.4), (onda, 0.3), (vida, 0.3)} e doc = {(vento, 0.5), (onda, 0.4), (vida, 0.3)}. Os vetores devem ser representados da seguinte maneira: vetor = {(termo 1, peso 1),..., (termo n, peso n)}, com o seu peso entre 0 e 1.

Como foi citado anteriormente, por meio do cálculo do cosseno do ângulo Θ é possível medir a proximidade do termo com o documento. Caso houvesse mais um documento a ser comparado, o plano teria mais um vetor representado. Esse cálculo é chamado de cálculo de similaridade.

Segundo Ferneda (2003), o cálculo de similaridade permite estabelecer o grau de semelhança entre dois documentos ou ainda entre os documentos com os termos a serem pesquisados. Abaixo temos a fórmula utilizada para efetuar o cálculo, levando em consideração o cosseno do ângulo formado entre os dois vetores a serem calculados.

$$sim(x,y) = \frac{\sum_{i=1}^{t}(w_{i,x} \times w_{i,y})}{\sqrt{\sum_{i=1}^{t}(w_{i,x})^2 \times \sum_{i=1}^{t}(w_{i,y})^2}}$$

em que "*w*" é o peso do elemento "i" nos vetores "x" e "y".

Vamos considerar os vetores usados anteriormente: cons = {(vento, 0.4), (onda, 0.3), (vida, 0.3)} e doc = {(vento, 0.5), (onda, 0.4), (vida, 0.3)}. Chamaremos o doc de doc2 e acrescentaremos mais um documento que possui o seguinte conteúdo: "[...] Vento, ventania, agora que estou solto na vida. Me leve pra qualquer lugar [...]". Obtivemos o seguinte vetor, que será chamado de doc1: doc1 = {(vento, 0.3), (onda, 0.0), (vida, 0.5)}.

Aplica-se o vetor cons (da consulta feita pelo usuário) e o vetor doc1 na fórmula de similaridade:

$$sim(cons, doc1) = \frac{(0.4 \times 0.3) + (0.3 \times 0.0) + (0.3 \times 0.5)}{\sqrt{0.3^2 + 0.0^2 + 0.5^2} \times \sqrt{0.4^2 + 0.3^2 + 0.3^2}}$$

$$sim(cons, doc1) = \frac{0.27}{0.34} = 0.79$$

$$sim(cons, doc1) = 79\% \text{ de similaridade}$$

O mesmo é feito com o vetor cons (da consulta feita pelo usuário) e o vetor doc2 na fórmula de similaridade:

$$sim(cons, doc2) = \frac{(0.4 \times 0.5) + (0.3 \times 0.4) + (0.3 \times 0.3)}{\sqrt{0.5^2 + 0.4^2 + 0.3^2} \times \sqrt{0.4^2 + 0.3^2 + 0.3^2}}$$

$$sim(cons, doc2) = \frac{0.41}{0.4123} = 0.99$$

$$sim(cons, doc2) = 99\% \text{ de similaridade}$$

Na Figura 2 a seguir, é apresentada a representação no espaço "n" vetorial:

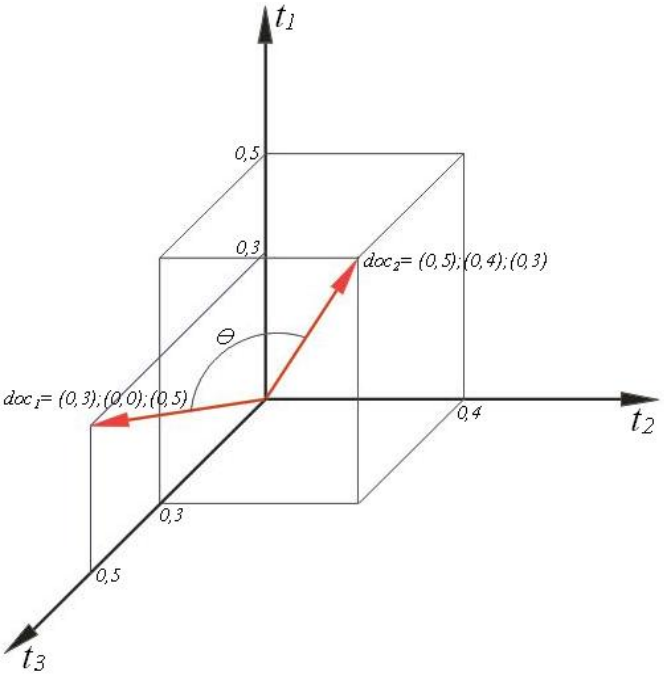


Figura 2 - Representação do modelo de espaço vetorial

Com a aplicação dos vetores na fórmula de similaridade, chega-se ao valor de 79% do grau de similaridade do "documento 1" com a consulta. Aplicando-se os valores do documento 2 com os da consulta na mesma fórmula, chega-se ao valor de 99%. Com isso, constata-se que o documento 2 é mais relevante que o 1.

2.3.1.3 Outros modelos de recuperação de informação

O modelo *fuzzy* (ou difuso) é baseado no conceito *fuzzy* estudado pela matemática. Na teoria dos conjuntos *fuzzy*, cada elemento possui um grau de membros associados com relação a um determinado conjunto, o que representa, em certo sentido, a força ou o grau de crença em sua associação ao conjunto. Esses graus de associação são valores geralmente dados entre 0.0 e 1.0 (KORFHAGE, 1997, p. 69-70).

Baeza-Yates e Ribeiro-Neto (1999) explicam que o modelo *fuzzy* nada mais é do que uma extensão do modelo booleano, mas que possui um método de ordenação em que o retorno da consulta em relação aos

documentos é aproximado, formando a ideia de uma nuvem de documentos aproximados.

Segundo Freitas e Pereira (2005), para recuperar os documentos por meio deste modelo é utilizada uma matriz de correlação em que é apresentado o quanto dois termos coocorrem dentro de um conjunto de documentos. A partir dessa matriz, são gerados os índices *fuzzy* para cada termo indexado. A seguir, é apresentada a fórmula para obtenção do índice:

$$\mu_{doc,j} = \otimes W_{j,k}$$

ou

$$\mu_{doc,j} = 1 - \prod_{K_k \in A_{doc}} (1 - W_{j,k})$$

Nesta fórmula, *doc* é um documento, *j* uma palavra-chave, *k* é uma palavra-chave que está presente em *doc*, *W* é o grau de correlação entre os termos em questão e A_{doc} é o conjunto de palavras-chave do documento *doc* (FREITAS; PEREIRA, 2005, p. 2).

Cita-se ainda o modelo probabilístico, que utiliza métodos matemáticos aplicados à distribuição de termos da coleção, ou seja, para um dado documento disposto em uma coleção de termos, o método bayesiano é o mais utilizado nesse caso (teorema de Bayes). Dado um termo de busca *q* e um documento d_j num conjunto de documentos, o modelo probabilístico estima a probabilidade de o documento d_j possuir o termo de busca (BAEZA-YATES; RIBEIRO-NETO, 1999, p. 24-39).

Utilizando o Teorema de Bayes citado anteriormente e as deduções das estimativas de relevância nas buscas efetuadas, Cardoso (2000, p. 4) propõe a seguinte equação:

$$sim(d,q) = W_{d|q} = \sum_{i=1}^{t} x_i \times W_{qi}$$

Onde:

- $x_i \in \{0,1\}$;
- $W_{qi} = \log r_{qi}(1 - s_{qi}) / s_{qi}(1 - r_{qi})$;
- r_{qi} é a probabilidade de que um termo de indexação *i* ocorra no documento, dado que o documento é relevante para a consulta *q*; e
- s_{qi} é a probabilidade de que um termo de indexação *i* ocorra no documento, dado que o documento não é relevante para a consulta *q*.

Segundo Ferneda (2003), no ano de 1976 foi proposto um modelo probabilístico por Robertson e Jones, o qual posteriormente ficou conhecido como *Binary Independence Retrieval*. Esse modelo tenta representar o processo de recuperação de informação por uma perspectiva probabilística, tendo um *corpus* com vários documentos que podem ser divididos em quatro subconjuntos distintos, a saber:

1) conjunto dos documentos relevantes (Rel);
2) conjunto dos documentos recuperados (Rec);
3) conjunto dos documentos relevantes que foram recuperados (RR); e
4) conjunto dos documentos não relevantes e não recuperados.

A Figura 3 a seguir ilustra o relacionamento entre esses conjuntos.

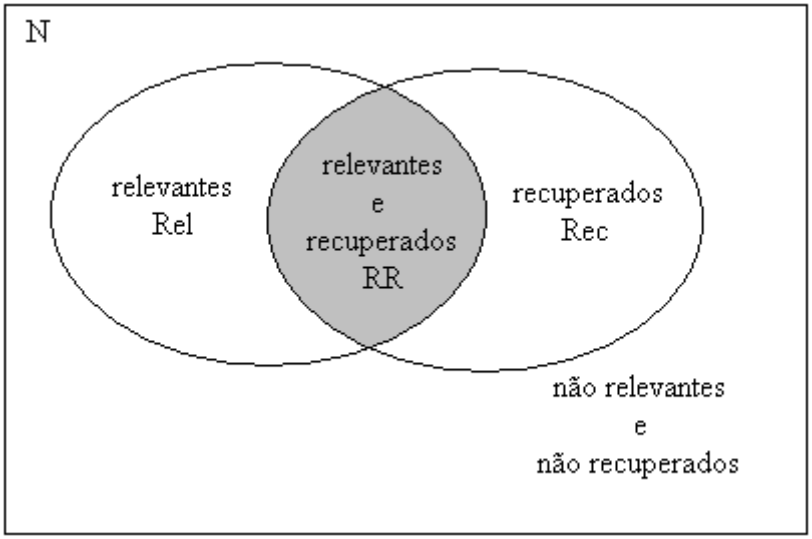


Figura 3 - Representação dos subconjuntos após uma busca
Fonte: adaptado de Ferneda (2003, p. 39).

Korfhage (1997) desenvolveu o seguinte exemplo para demonstrar a utilização deste modelo: supondo-se que tenhamos uma pergunta (termos para a busca) e uma base de dados. Para se chegar ao valor de a probabilidade de algum documento dessa base ser relevante a uma pergunta de valor 0.1, é necessário fazer a seguinte simulação: P(Rel)/P(-Rel) = 0.1/0.9 = 0.111... onde -Rel são os documentos não relevantes. Agora, considerando-se um documento específico D e

supondo-se que este documento é representado por quatro termos, os valores de cada termo estão representados abaixo.

Termo	P(t|Rel)	P(t|-Rel)
t1	0.8	0.4
t2	0.6	0.1
t3	0.2	0.9
t4	0.9	0.6

$$dis(D) = \frac{0.8x0.6x0.2x0.9x0.1}{0.4x0.1x0.9x0.6x0.9} = \frac{0.00784}{0.01944} = 0.403$$

Uma vez que este valor é inferior a um, o documento não deve ser recuperado. Para que todos os modelos encontrem as informações no texto de maneira mais rápida, são utilizados índices gerados a partir do processo de indexação, o qual será visto na próxima seção.

2.3.2 Indexação

O processo de indexação é uma das tarefas mais importantes para a recuperação de informação. Segundo Ebecken, Lopes e Costa (2003), a indexação tem como função permitir que se efetue uma busca em texto sem a necessidade de varrer o documento inteiro, similarmente ao que acontece com o processo "homônimo" utilizado convencionalmente em bancos de dados. Os autores classificam a indexação em quatro tipos distintos, a saber:

1) indexação de texto completo;
2) indexação temática;
3) indexação semântica latente; e
4) indexação por tags.

Para Baeza-Yates e Ribeiro-Neto (1999), as principais técnicas para a construção de arquivos de indexação são:

- arquivos invertidos;
- arquivos de assinatura; e
- árvores e vetores de sufixos.

De acordo com Baeza-Yates e Ribeiro-Neto (1999), a técnica de arquivo (ou índice) invertido trabalha com uma lista de palavras-chave ordenadas, em que cada palavra está ligada ao documento que a possui. Esse documento é associado a uma lista invertida de palavras-chave, que passa a ser ordenada de forma alfabética. Essas palavras-chave possuem

um peso. Após o processamento, a lista fica dividida em dois arquivos, um de vocabulário e outro de endereçamento.

Um índice invertido, segundo Manning e Schutze (1999), é uma estrutura de dados em que se relaciona cada palavra com todos os documentos que a contêm, e também armazena a frequência com que a palavra ocorre no documento. A utilização do índice invertido torna mais fácil a busca de informação em documentos. O autor afirma que as versões mais sofisticadas de índice invertido também armazenam as posições do documento. Para Manning, Raghavan e Schutze (2008), a técnica para inversão de índice divide-se em duas fases: (1) primeiramente faz-se um levantamento dos pares de termo por documento, armazenando as posições onde os termos aparecem no documento em questão; e (20 na última etapa, é feita a combinação dos documentos pelas palavras. Por exemplo: a palavra *casa* ocorre nos documentos d1, d5 e d23.

2.3.3 Extração de informação

A extração de informação faz parte da área de Processamento de Linguagem Natural (PLN), e tem como foco identificar informações importantes em bases textuais. Essa é uma área já bastante estudada e que ainda atrai muitos pesquisadores. Wilks e Catizone (1999) comentam que extração e gerenciamento de informação sempre tiveram uma grande importância para as agências de inteligência, mas está claro que atualmente e nas próximas décadas essa é uma área crucial para a educação, a medicina e o comércio. É estimado que 80% das informações estão no formato textual, e por esse motivo essa é uma área tão importante.

Para Weiss et al. (2005), textos digitais em linguagem natural são fontes muito importantes de informação, em que cada informação é apresentada de forma não estruturada. Segundo Konchady (2006), sistemas de extração de informação geralmente convertem textos não estruturados em algo que possa ser carregado para uma base de dados, sendo que esses dados são normalmente nomes de pessoas, lugares ou organizações mencionadas no texto. Finn e Kushmerick (2004) afirmam que a extração de informação é um processo de identificação de itens relevantes em documentos textuais que verifica as “fronteiras” entre os termos que formam as entidades.

Segundo Nédellec e Nazarenko (2005), a utilização de sistemas de extração de informação pode ser usada para popular ontologias. Textos são uma fonte de conhecimento para desenhar e enriquecer essas

ontologias, em que os termos extraídos da base de documentos podem ser possíveis instâncias e classes das ontologias.

Uma das técnicas usadas para extrair termos de documentos é a de reconhecimento de entidades nomeadas (em inglês *Named Entity Recognition* – NER), que será abordada na próxima seção. O motivo principal para a apresentação desta seção é que essa técnica é utilizada pelo modelo proposto para a etapa de reconhecimento de entidades.

2.3.4 Reconhecimento de entidades nomeadas

As classes que irão compor as ontologias serão extraídas diretamente dos documentos que compõem a base de dados. Para identificar quais termos são classes, pode-se utilizar uma técnica chamada de reconhecimento de entidades nomeadas.

Segundo Zhu, Gonçalves e Uren (2005), reconhecimento de entidades nomeadas – em inglês *Named Entity Recognition* (NER) – é uma técnica da área de extração de informação (EI) que tem como função reconhecer entidades em textos de diferentes tipos e de diferentes domínios.

Ceci et al. (2010) explicam que existem muitas técnicas automáticas que podem auxiliar no processo de reconhecimento de entidades, tais como a aplicação de expressões regulares (técnica muito usada para identificar datas, e-mails, etc.), o uso de dicionários (thesauros), as heurísticas, que são regras conforme o padrão léxico e sintático do idioma, os modelos estatísticos e também o uso de ontologias.

Para Negri e Magnini (2004), NER tem como tarefa identificar e categorizar entidades mencionadas (pessoas, organizações, locais), expressões temporais (hora e data) e alguns tipos de expressão numérica (percentual e valor monetário) escritos em um texto. Segundo Kozareva (2006), a tarefa de identificar as entidades consiste em determinar suas fronteiras, ou seja, qual o seu início e seu fim. Isso é importante para entidades compostas de mais de uma palavra, como, por exemplo, “Universidade Federal de Santa Catarina”.

A técnica de extração de entidades pode ser vista como um problema de classificação em que as palavras são assinadas para uma ou mais classes semânticas. Quando a entidade encontrada não pode ser assinada para uma classe específica, é atribuída a uma classe “geral” (KONCHADY, 2006, p. 156-157).

Como afirmado acima, uma entidade pode fazer parte de uma ou mais classes. A seção abaixo fala sobre a resolução de ambiguidade das

entidades reconhecidas, já que este é um problema encontrado durante o processo do reconhecimento de entidades.

2.3.4.1 Resolução de ambiguidade

É bastante comum encontrar entidades que pertencem a mais de uma classe, como, por exemplo, a palavra Bahia, a qual pode estar relacionada a uma classe "Lugar" ou ainda "Time de futebol". Konchady (2006) explica que, para o processo de "classificação" ou assinatura das entidades, são utilizadas listas de termos. Cada uma delas possui uma quantidade de entidades que compõem a classe em questão. Uma entidade pode estar presente em mais de uma lista, dificuldade que os sistemas de reconhecimento de entidades enfrentam. Pode-se construir algoritmos para melhorar a classificação das entidades, o que irá auxiliar no caso de ambiguidade.

Na grande maioria das vezes, os algoritmos criados para auxiliar na desambiguação utilizam os termos próximos para identificar a melhor classificação como, por exemplo: "o estado da Bahia" demonstra que a entidade Bahia pertence à classe "Lugar" (CECI et al., 2010).

Segundo Weiss et al. (2005), para auxiliar na desambiguação também são utilizados alguns métodos estatísticos que permitem verificar a frequência da entidade e quais termos estão acompanhando.

2.3.4.2 Utilização de sistemas NER

O emprego de sistemas NER traz uma série de vantagens para outros sistemas e áreas, como, por exemplo:

- auxiliar no processo de recuperação de informação: o sistema NER identifica as entidades do texto antes do processo de indexação, fazendo com que seja indexada a entidade (que pode ser composta de vários termos) em vez de apenas os termos;
- detecção de eventos: por meio das datas encontradas nos textos, pode-se fazer uma relação com os termos próximos e verificar a evolução destes;
- manutenção em ontologias: através das entidades levantadas pelo sistema NER, pode-se verificar qual delas é uma possível classe da ontologia em questão e quais termos estão relacionados com a classe a fim de atualizar essa ontologia (GIULIANO, 2009, p. 201-209).

Uma grande vantagem na utilização de sistemas NER é que, quando uma entidade é reconhecida, o sistema identifica uma possível classe para ela. O problema encontrado em utilizar listas de termos (ou tabelas léxicas) para classificação e reconhecimento das entidades é que isso torna o sistema sensível à linguagem das entidades levantadas, já que cada classe do sistema terá uma lista de termos relacionados nos idiomas previamente selecionados.

Uma solução para o reconhecimento de entidades de forma que estas não precisem ser previamente cadastradas numa lista de termos é utilizar a técnica de *clusterização*, que será apresentada na seção a seguir.

2.3.5 Clusterização

Segundo Konchady (2006), a *clusterização* (ou agrupamento) é classificada como aprendizado não supervisionado, já que não tem nenhum treinamento de dados que permita criar a classificação aprendida para agrupar os documentos.

O uso de *clusterização* em documentos é muito importante para a área de recuperação de informação, pois possibilita a visualização dos resultados das buscas na forma de conjuntos de documentos relacionados pelo seu conteúdo. Outra possibilidade da *clusterização* em documento é a descoberta não supervisionada de temas e principais tópicos da coleção de documentos (ALJABER et al., 2009).

Segundo Sculley (2010), o uso de *clusterização* não supervisionada é uma tarefa importante para aplicações baseadas na web e que são utilizadas para agrupar resultados de busca para a identificação de resultados duplicados e para a agregação de conteúdos similares.

O processo de *clusterização* básico funciona da seguinte maneira: primeiramente são levantados alguns termos e/ou palavras-chave em um vetor que identifiquem o documento. O próximo passo é determinar a proximidade entre os documentos com base nos vetores levantados na etapa anterior. Desse processo, é gerado um valor que irá definir se os documentos devem ou não ser agrupados (EBECKEN; LOPES; COSTA, 2003, p. 358-360).

O modelo proposto por este trabalho adota técnicas de *clusterização* para extrair as entidades da base de documentos. Existem várias outras pesquisas que utilizam *clusterização* para extrair informações e conhecimento, como, por exemplo, o trabalho de Ahmad, Alahakoon e Chau (2009), que utilizam a técnica numa base biomédica

para classificação de sequência de proteínas. Um trabalho bastante semelhante a este modelo é o de Fortuna, Lavrac e Velardi (2008), em que também foi utilizada a *clusterização* para encontrar possíveis instâncias de uma ontologia.

2.3.5.1 Algoritmo STC

O algoritmo *Suffix Tree Clustering* (STC) não considera um documento apenas como um conjunto de palavras, e sim como uma sequência de palavras utilizando as informações de proximidade entre essas palavras. O STC usa uma árvore de sufixo para identificar de forma eficiente os conjuntos de documentos em que as partes de frases são comuns e usa essas informações para criar os *clusters* (ZAMIR; ETZIONI, 1998).

Basicamente o algoritmo STC possui quatro fases: a primeira é a "limpeza do documento", que consiste em tirar os espaços maiores que 1 entre as palavras e aplicar *stemming*. Na segunda fase, são identificadas as frases comuns. Depois, na terceira fase, é calculado um valor para cada frase levando em consideração sua relevância para o documento. Na última fase, é feito o *merging* entre os *clusters* (WEI et al., 2008, p. 501-504).

Segundo Wang e Li (2008), o algoritmo STC tenta relacionar esses fragmentos de frase em todos os documentos da coleção, podendo gerar uma grande combinação de frases e documentos que acabam atrapalhando o processo de visualização e escolha dos *clusters* resultantes.

2.3.5.2 Algoritmo Lingo

Este algoritmo é adequado para resolver problemas de agrupamento de resultados de pesquisas. Diferentemente de outros algoritmos, ele tenta primeiramente descobrir o nome descritivo para os grupos (*clusters*) com os documentos correspondentes (OSINSKI; WEISS, 2004).

Segundo Carpineto et al (2009), o Lingo possui quatro fases de entrada: (1) pré-processamento de trechos, (2) extração de frases frequentes, (3) inclusão dos rótulos nos *clusters* e (4) alocação de conteúdo. O que diferencia este algoritmo das demais abordagens é que ele elege os melhores trechos para servirem como rótulo para os *clusters*. Para Osinski (2003), este algoritmo está dividido em quatro fases: (1) filtragem do texto, que tem como função limpar os caracteres

e rótulos não válidos, por exemplo, *tags* HTML; (2) identificação da linguagem do texto, que se constitui na etapa fundamental para as duas últimas, pois, caso o idioma não seja identificado, elas não são executadas. As duas últimas fases são (3) *stemming* e (4) remoção de *stopwords*.

A grande vantagem da utilização deste algoritmo é a sua preocupação com os nomes dos *clusters*. Como o presente trabalho utiliza esses nomes como possíveis entidades encontradas, este algoritmo apresenta-se como uma boa opção para a descoberta de entidades sem uma base de termos previamente levantada.

2.3.6 Uso de dicionário ou tesauros

Para Ebecken, Lopes e Costa (2005, p. 349), um dicionário ou *thesaurus* "pode ser definido como um vocabulário controlado que representa sinônimos, hierarquias e relacionamentos associativos entre termos para ajudar os usuários a encontrar a informação de que eles precisam".

Um bom tesauro tem de possibilitar o relacionamento das palavras principais do seu contexto com seus sinônimos tanto no processo de armazenamento das informações como na busca (KORFHAGE, 1997, p. 138-139).

Segundo Gonzales e Lima (2003), os tesauros podem desempenhar as seguintes funções:

- auxiliar na classificação de documentos bem como na caracterização de seus conceitos;
- auxiliar na produção e na tradução de textos;
- auxiliar no processo decisório na classificação de assuntos; e
- apoiar a recuperação de informação.

Este trabalho utiliza como base de conhecimento colaborativa a Wikipédia, que, segundo Mihalcea (2007), é uma enciclopédia on-line livre que representa o resultado do esforço contínuo de colaboração de voluntários de todo o mundo, estando disponível em mais de 200 idiomas distintos. A seção abaixo apresenta com mais detalhes bases de dados/conhecimento colaborativas.

2.4 Bases de conhecimento colaborativas

A internet possibilitou uma série de ferramentas e recursos, e com o advento da chamada Web 2.0, o leque de ferramentas para a construção de informação e de conhecimento de maneira colaborativa é ainda maior. Segundo Coutinho e Junior (2007), a primeira geração da internet possibilitava o acesso a uma grande quantidade de informação e conhecimento. Na Web 2.0, surgiram ferramentas em que qualquer pessoa podia publicar conteúdo sem precisar conhecer uma linguagem de marcação. Pode-se listar ferramentas como blogs, wikis, podcasts, entre outras.

Segundo Zesch, Muller e Gurevych (2008), bases de conhecimento colaborativas armazenam conhecimentos construídos colaborativamente por voluntários especialistas ou não na web através do uso da inteligência coletiva. Como instâncias dessas bases, podemos citar a Wikipédia. A seguir, é apresentado o conceito de inteligência coletiva.

2.4.1 Inteligência coletiva

O termo "inteligência coletiva" vem sendo usado por décadas, e com o advento das novas tecnologias está tornando-se cada vez mais popular. Sua definição pode ser uma combinação de comportamentos, preferências ou concepções de um grupo de pessoas, cominações essas que podem ser usadas para criar novas ideias. Embora métodos para a inteligência coletiva existissem antes da internet, a capacidade de coleta de informação na rede de milhares de pessoas tem aberto muitas novas possibilidades (SEGARAN, 2008, p. 2-3).

Um exemplo de base de conhecimento que utiliza a inteligência coletiva é a Wikipédia, que, como definido anteriormente, é uma enciclopédia livre na qual as pessoas podem descrever um determinado conceito de maneira colaborativa. Na seção abaixo é apresentada a construção social do conhecimento.

2.4.2 Construção social do conhecimento

Ramalho e Tsunoda (2007) afirmam que os wikis são espaços de aprendizagem em rede com um grau de complexidade muito mais elevado do que em espaços tradicionais na construção social do conhecimento. Essa construção nasce da interação do conhecimento (tácito ou explícito) dos indivíduos. O compartilhamento e a

comunicação de informações em grupo pretendem expor o conhecimento tácito, interno ao indivíduo. Esse processo é denominado "socialização do conhecimento" (NONAKA; TAKEUCHI, 2003).

Coutinho e Junior (2007) concluem que a Web 2.0 é uma forma de utilização colaborativa da internet, em que o conhecimento é compartilhado de maneira coletiva e descentralizado de autoridade para utilizá-lo e reeditá-lo.

Na próxima seção, são apresentadas mais informações sobre a Wikipédia e qual a sua representação e utilização neste trabalho.

2.4.3 O uso da Wikipédia

A principal função da Wikipédia no presente trabalho é identificar quais entidades reconhecidas são válidas (existentes na sua base de conhecimento) e, através da sua descrição, apresentar uma possível classificação.

Em etapas de processamento de linguagem natural não é a primeira vez que a Wikipédia é usada. Pode-se citar o trabalho de Mihalcea (2007), que a utiliza para auxiliar na fase de desambiguação entre as palavras. Outro trabalho que pode ser mencionado é o de Pehcevski et al. (2010), em que foram utilizadas as classes e os *link*s da Wikipédia para auxiliar no *ranking* das entidades recuperadas para responder a perguntas. Como mais um exemplo de utilização da Wikipédia tem-se a classificação de textos (WANG et al., 2008, p. 265-281).

A Wikipédia possui uma utilização importante também na área de extração de conceitos de bases textuais. Como seu conteúdo é cuidadosamente selecionado pelos seus editores, ela passa a ser uma fonte confiável para sistemas de classificação (PARAMESWARAN; GARCIA-MOLINA; RAJARAMAN, 2010, p. 266-277).

No trabalho de Bohn e Norvag (2010), os autores fazem uso da Wikipédia para gerar os termos que irão compor os *gazetteers*, que são utilizados pelos algoritmos de reconhecimento de entidades nomeadas (técnica essa utilizada em conjunto com a clusterização no modelo proposto pelo presente trabalho), além de auxiliar na extração de sinônimos dos termos levantados.

2.5 Modelos correlatos

Nesta seção, são detalhados alguns modelos/*frameworks* correlatos objetivando estabelecer uma base comparativa com o modelo proposto neste trabalho.

O critério utilizado de modo a viabilizar a comparação entre os modelos foi a escolha de trabalhos que possuam como objetivo a manutenção ou população de ontologias de maneira semiautomática ou automática.

2.5.1 "A Flexible Framework to Experiment with Ontology Learning Techniques" segundo Gacitua, Sawyer e Rayson (2007)

No trabalho realizado por Gacitua, Sawyer e Rayson (2007), os autores propõem a disponibilização de várias técnicas de processamento de linguagem natural e aprendizagem de máquina para que o engenheiro de ontologias possa combiná-las e extrair delas o maior número de informações possível para compor uma ontologia.

Este *framework* caracteriza-se como semiautomático, já que o usuário deve fazer a combinação de técnicas para ver qual a mais indicada para aplicar ao seu domínio. O nome dado a este *framework* é OntoLancs, e ele divide-se em quatro fases, como descrito a seguir.

1) Fase 1 (Anotação semântica ao *corpus*): os termos sofrem o processo de *tagging*, ou seja, anotação semântica. A aplicação assina a categoria semântica para cada palavra. Nesta etapa, é utilizado um *framework* sensível à linguagem que categoriza os termos de maneira automática.
2) Fase 2 (Extração de conceitos): são extraídas as terminologias do domínio a partir do *corpus* submetido à fase 1 para identificar a lista de termos candidatos. Nesta fase, o sistema disponibiliza um conjunto de técnicas de processamento de linguagem natural e de aprendizagem de máquina para que o engenheiro de ontologias possa combinar e identificar os conceitos candidatos.
3) Fase 3 (Construção da ontologia de domínio): nesta fase, o domínio léxico é construído, e as definições de cada conceito são extraídas de fontes on-line de forma automática, utilizando-se para tal tesauros de domínio. Na construção da ontologia de domínio, a classificação hierárquica dos termos é estruturada.

4) Fase 4 (Edição da ontologia de domínio): nesta fase final, os dados são disponibilizados em OWL.

A Figura 4 a seguir apresenta as fases principais para a utilização do OntoLancs:

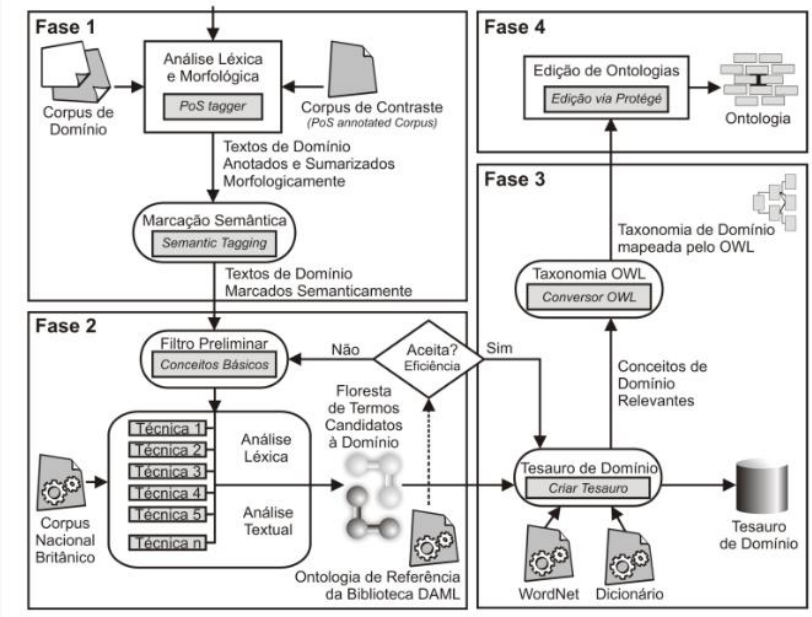


Figura 4 - Framework OntoLancs
Fonte: adaptado de Gacitua, Sawyer e Rayson (2007)

A abordagem apresentada neste *framework* é interessante, já que possibilita a flexibilidade na escolha dos algoritmos para a extração das informações da base textual utilizada como entrada. Entretanto, a proposta obriga necessariamente que o engenheiro de ontologia tenha domínio sobre as técnicas disponíveis na primeira fase. Outro ponto crítico é a necessidade de se criar um *corpus* anotado semanticamente para que o processo de extração de entidades possa ser realizado. Esse processo de anotação semântica está previsto na primeira fase deste *framework*, mas, dependendo do tamanho do *corpus*, essa tarefa pode dispendiosa. Para auxiliar na classificação dos termos encontrados, são utilizadas ontologias já disponíveis na web em conjunto com técnicas de processamento de linguagem natural que, se não forem combinadas de maneira adequada, podem não apresentar um resultado satisfatório para o usuário.

2.5.2 “A Hybrid Approach for Taxonomy Learning from Text” segundo El Sayed e Hacid (2008)

O trabalho em questão apresenta um *framework* para aprendizagem de ontologias chamado de OLea, o qual possui uma proposta híbrida para esse processo que utiliza combinações baseadas em padrões (linguísticos) e abordagens estatísticas.

A arquitetura deste *framework* é dividida em três principais estágios, como mostrado nas descrições e na Figura 5 a seguir.

1) Estágio 1: estima-se uma “taxa de confiança” para um conjunto de relações semânticas com base na coocorrência de termos encontrados no *corpus*. Para se chegar a essa relação semântica, primeiramente é calculado o grau de relação utilizando-se a distância entre dois termos que podem ser encontrados no dicionário WordNet.
2) Estágio 2: as relações semânticas são usadas como entradas para um algoritmo de aprendizagem de conceitos que agrupa os termos levando em conta o seu sentido encontrado no Wordnet. A partir disso, com base no WorkNet é criada uma hierarquia de conceitos.
3) Estágio 3: é utilizada a interação humana para validar os resultados, verificando as palavras encontradas e a forma como elas estão relacionadas hierarquicamente compondo uma taxonomia.

Cita-se ainda como características do OLea a sua capacidade de lidar com a natureza esparsa do texto, oferecendo reconhecimento mais flexível para as relações semânticas entre os termos. A partir dessas relações, são construídos os agrupamentos a fim de popular a taxonomia em questão, a qual posteriormente auxiliará na descoberta das relações dos temos na próxima iteração, proporcionando um ambiente de aprendizagem supervisionada.

É interessante mencionar que esse modelo possui dependência de uma estrutura formal de conhecimento (WordNet).

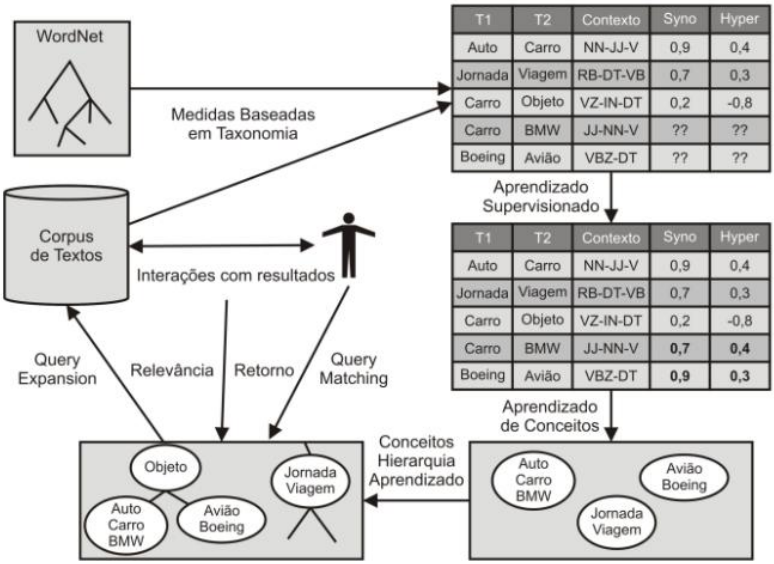


Figura 5 - *Framework* OLea
Fonte: adaptado de El Sayed e Hacid (2008)

2.5.3 "Advancing Topic Ontology Learning through Term Extraction" segundo Fortuna, Lavrac e Velardi (2008)

No trabalho de Fortuna, Lavrac e Velardi (2008), é apresentada a metodologia OntoTermExtraction, a qual baseia-se no *framework* OntoGen. Caracteriza-se como uma proposição semiautomática para a construção e edição de uma ontologia de tópicos. Numa ontologia de tópico, cada nó é um grupo de documento representado por uma palavra-chave, sendo os nós ligados por uma relação. Além de dispor de técnicas de mineração de texto, possui uma interface gráfica amigável para reduzir a complexidade da etapa de construção da ontologia.

Os termos encontrados pelo processo de agrupamento do OntoGen são termos simples, compostos de uma única palavra. Para solucionar essa deficiência, esses termos são aplicados a mais uma ferramenta, chamada TermExtractor, a qual obtém palavras compostas de uma coleção de documentos. Para visualizar o resultado na forma de árvore de termos hierárquicos, é utilizado o algoritmo K-Means.

A aplicação possibilita a conexão com ferramentas de busca como o Google para permitir a descoberta de novos termos a partir do

resultado da busca, fazendo com que a base cresça e evolua a partir dos novos termos armazenados.

O trabalho apresentado emprega técnica similar à do modelo proposto para a extração de entidades, visto que parte da abordagem de geração de termos dá-se a partir dos rótulos atribuídos aos agrupamentos de documentos. Além disso, ressalta-se que o trabalho apresenta uma forma bastante interessante para a visualização da ontologia, na forma de árvores hiperbólicas, o que facilita o entendimento do usuário.

A principal diferença entre as duas abordagens é que no corrente trabalho pode-se utilizar uma base de conhecimento colaborativa para auxiliar na descoberta, validação e classificação das entidades reconhecidas, facilitando o trabalho do engenheiro de ontologias. Vale mencionar que no presente trabalho o processo de extração também ocorre por meio da utilização da técnica tradicional de reconhecimento de entidades.

2.5.4 "Automated Ontology Learning and Validation Using Hypothesis Testing" segundo Granitzer et al. (2007)

O trabalho de Granitzer et a. (2007) apresenta a utilização de testes de hipóteses para auxiliar no processo de aprendizagem de ontologia, denominado de AVALON (*Acquisition and VALidation of ONtologies*). Para tal, emprega bases de documentos textuais para a extração de elementos/entidades visando à composição da ontologia. A fase de validação ocorre por meio indicadores do mundo real, como, por exemplo, dados da web.

Segundo os autores, esse projeto faz uso do estado da arte da área de extração de conhecimento para a composição da ontologia, e o seu diferencial está na utilização de hipóteses para validar o resultado obtido do processo de extração (por meio das técnicas clássicas da área) com indicadores existentes na web para refinar o resultado encontrado. A Figura 6 a seguir apresenta o seu modelo conceitual:

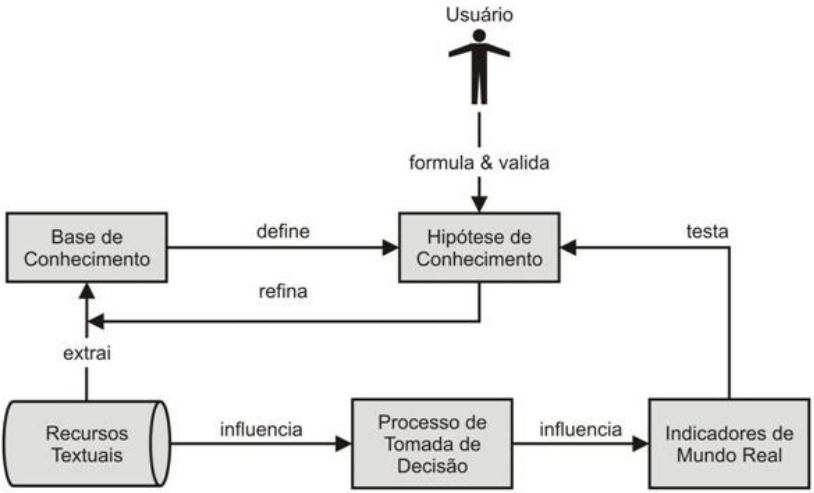


Figura 6 - Modelo conceitual do AVALON
Fonte: adaptado de Granitzer et al. (2007)

A formulação de hipóteses de base ontológica merece atenção especial. A granularidade da hipótese relaciona-se diretamente com a granularidade da ontologia. AVALON possui três pilares levando em consideração o ponto de vista do seu algoritmo, como mostrado a seguir.

1) Determinação da estrutura de domínio via aprendizagem de ontologia a partir de textos não estruturados: pode ser utilizada qualquer ferramenta para população de ontologias (o autor cita Text2Onto ou Kim Platform).
2) População da base de conhecimento a partir da extração de informação: a lista de termos (*gazetteers*) é definida com base em instâncias identificadas via extração de informação.
3) Seleção de hipótese a partir de mineração gráfica: é o diferencial apresentado nesta proposta, que foca na classificação dos resultados encontrados.

Como mencionado por Granitzer et al. (2007), tem-se a utilização de dados do "mundo real", pois as principais contribuições ao processo de aprendizagem de ontologias se baseiam na web e na formulação de hipóteses.

2.5.5 "Evaluation of OntoLearn, a methodology for automatic learning of domain ontologies" segundo Velardi et al. (2003)

O trabalho de Velardi et al. (2003) apresenta a ferramenta automática para aprendizagem de ontologias de domínio, a qual é chamada de OntoLearn. Como entrada de dados, é utilizada uma coleção de documentos referentes ao domínio em questão. A Figura 7 abaixo apresenta o funcionamento dessa ferramenta.

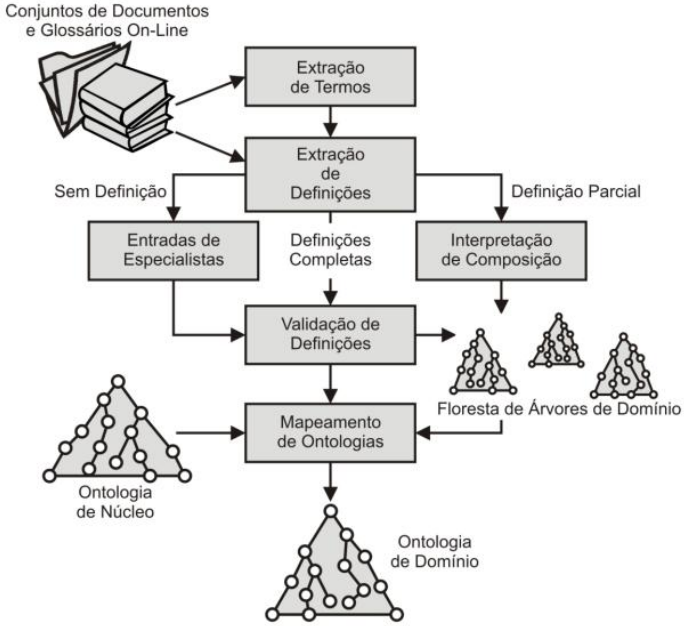


Figura 7 - Estrutura do OntoLearn
Fonte: adaptado de Velardi et al. (2003)

Os autores dividem o processo de aprendizagem de ontologias em dois problemas: (1) o primeiro está relacionado à extração das informações que irão compor a ontologia e (2) o segundo diz respeito à construção da ontologia em seu conjunto. As principais etapas são as seguintes:

1) extração dos termos: são utilizadas ferramentas baseadas em análises estatísticas ou de processamento de linguagem natural;

2) extração de definições em linguagem natural: esta etapa consiste em buscar na web definições para o termo encontrado via glossários, utilizando expressões regulares e parses semânticos;
3) separação das definições em linguagem natural: como o processo anterior pode retornar mais de uma definição, são separadas todas as definições encontradas para a devida classificação em um processo posterior;
4) resolução de ambiguidade semântica: para esta etapa é utilizado o algoritmo de ambiguidade semântica SSI, que, segundo os autores, é o núcleo do OntoLearn;
5) identificação das relações semânticas: identifica as relações semânticas utilizando bases como Euro Wordnet, DOLCE, FramNet, entre outras. Esta etapa vincula o termo ao seu contexto (classificação).

Os autores afirmam que a qualidade da ontologia gerada está diretamente ligada à execução de cada etapa listada acima.

O trabalho de Velardi et al. (2003) apresenta uma arquitetura bastante completa para a etapa de aprendizagem de ontologias. Contudo, pelo fato de ser uma abordagem automática, existe uma taxa de erro diretamente relacionada à qualidade da base de documento de entrada. Outro ponto a considerar é a dependência da língua (neste caso o inglês) por parte dos algoritmos de processamento de linguagem natural.

2.5.6 "Text2Onto - A Framework for Ontology Learning and Data-Driven Change Discovery" segundo Cimiano e Volker (2005)

O *framework* Text2Onto foi desenvolvido a partir de outro trabalho proposto por Maedche e Staab (2000), e introduz dois paradigmas adicionais para aprendizagem de ontologias a partir de textos, sendo: (i) Modelos de Ontologias Probabilísticas (*Probabilistic Ontology Models, POM*), em que as interconexões entre classes, instâncias/conceitos e relações encontradas recebem um grau probabilístico visando auxiliar o especialista no entendimento de determinado domínio; e (ii) identificação de mudanças nos dados, as quais são responsáveis pela detecção de mudanças no *corpus* a partir da variação do delta da ontologia probabilística calculada anteriormente. A seguir, na Figura 8, é apresentada a arquitetura do Text2Onto.

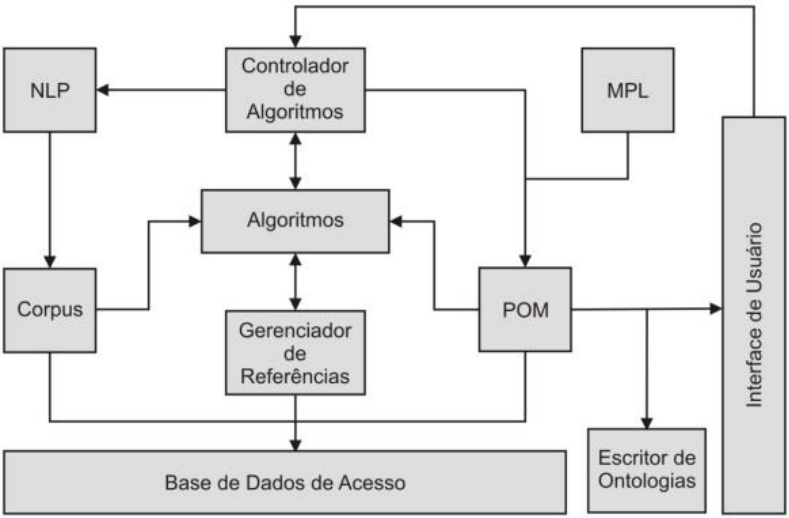


Figura 8 - Arquitetura do Text2Onto
Fonte: adaptado de Cimiano e Volker (2005)

Este *framework* combina algoritmos de análise linguística e técnicas de aprendizagem de máquinas para extrair classes, instâncias/conceitos e relações. Todas as etapas relacionadas ao processamento de linguagem natural são realizadas utilizando-se o *framework* GATE.

Um dos grandes diferenciais dessa proposta está na interface gráfica disponível para o usuário interagir com os resultados encontrados pelos algoritmos propostos. O Text2Onto possui uma interface gráfica que se acopla ao ambiente de desenvolvimento Eclipse.

Como características importantes, citam-se a não necessidade de uma ontologia pré-construída para a sua utilização e a geração de índices (deltas) que rotulam os relacionamentos entre os resultados (classes, instância/conceitos, relações) encontrados. A partir desses índices, é possível calcular uma variação que identifique mudanças na fonte de dados, podendo assim manter a ontologia atualizada.

2.6 Considerações finais

Neste capítulo, foram mostrados os principais temas para apoiar as etapas utilizadas no modelo de solução proposto.

Primeiramente, apresentou-se a gestão do conhecimento, e foi situado o problema de pesquisa nesse tema: tratou-se da etapa de extração e aquisição do conhecimento e, por fim, das etapas de representação e armazenamento de conhecimento.

Em seguida, foram abordadas as ontologias como a forma utilizada por este trabalho para fazer a representação e o armazenamento do conhecimento. Algumas das áreas da Engenharia do Conhecimento são apresentadas para auxiliar no processo de construção e manutenção dessas ontologias.

As técnicas abordadas foram as seguintes:

- recuperação de informação;
- reconhecimento de entidades nomeadas; e
- clusterização.

Apresentou-se uma seção falando sobre as bases de conhecimento colaborativas que são utilizadas nas etapas de validação e classificação do modelo proposto, o qual é tratado no capítulo a seguir.

Por fim apresentou-se uma seção sobre os modelos correlatos ao modelo proposto neste trabalho, com o objetivo de promover subsídios à análise comparativa com o modelo proposto descrita na seção 4.4

3 MODELO PROPOSTO

3.1 Introdução

A literatura relacionada com a área da engenharia de ontologias afirma que, pelo fato de o conhecimento de modo geral não ser estático, isso faz com que as ontologias de domínio de uma organização estejam em constante evolução (DAVIES; FENSEL; VAN HARMELEN, 2003; VASCONCELOS; ROCHA; KIMBLE, 2003). O processo de manutenção e construção de ontologias é bastante custoso e, na maioria das vezes, necessita da participação de um especialista para o levantamento das informações. Conta com o mapeamento e levantamento de todas as classes e instâncias que compõem a ontologia de domínio e como elas interagem entre si.

Segundo Giuliano (2009), o processo de população de ontologias que se utilizam da técnica de reconhecimento de entidades nomeadas em bases textuais não estruturadas é uma tarefa crucial para a web semântica e para sistemas de Gestão do Conhecimento. Muitos pesquisadores vêm trabalhando para auxiliar o processo de manutenção de ontologias, apresentando modelos automáticos e semiautomáticos.

O modelo proposto neste trabalho emprega (a) técnicas de *clusterização* e reconhecimento de entidades nomeadas na identificação das entidades a partir de uma base textual e (b) técnicas de recuperação de informação para validação e classificação. Para a etapa de classificação, também são utilizadas a identificação de padrões a partir de expressões regulares, a presença de um usuário especialista e bases de conhecimento colaborativas ou uma base de folksonomia (no presente trabalho, foi considerada a Wikipédia) para as fases de classificação e validação dos termos encontrados. Segundo Tang et al. (2009), uma base colaborativa ou de folksonomia refere-se a uma coleção de *tags* contendo sua descrição definida por usuários e publicada na web. A seção a seguir apresenta de forma mais detalhada o modelo proposto.

3.2 Descrição do modelo

Como mencionado anteriormente, muito do conhecimento existente nas organizações está concentrado no formato textual e digital, como, por exemplo, textos livres, contratos, logs de bate-papo, e-mails, etc. (CHAVES, 2007). Por esse motivo, tem-se como fonte de informação primária do modelo uma coleção de documentos textuais

que seja representativa para determinado domínio de interesse de uma organização. O modelo proposto é dividido em três importantes etapas:

1) reconhecimento das entidades da coleção de documentos;
2) validação das entidades recuperadas; e
3) classificação das entidades e adição das instâncias encontradas na ontologia de domínio da organização.

Essas três etapas estão conectadas por uma estrutura formal de representação de conhecimento. Para as duas últimas etapas (validação e classificação), é necessária a presença de um usuário caracterizando o modelo como semiautomático. Em geral, essa abordagem promove resultados interessantes, pois minimiza o erro e acelera os processos de construção e manutenção da ontologia.

Abaixo é apresentada a Figura 9, que representa a arquitetura lógica do modelo proposto em que se pode observar como cada etapa está conectada.

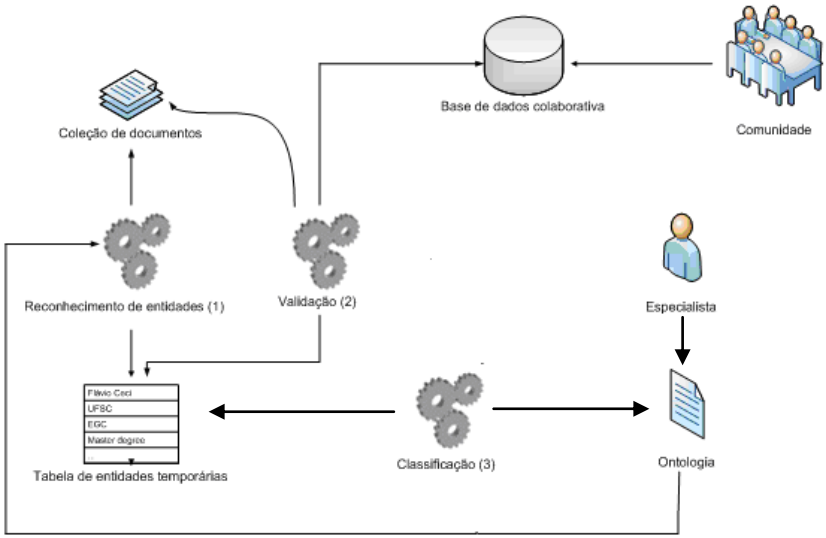


Figura 9 - Arquitetura lógica do modelo proposto

As próximas subseções apresentam um detalhamento das três principais etapas deste modelo.

3.2.1 Reconhecimento de entidades (1)

O sistema tem como entrada uma coleção de documentos representativos para o domínio da organização. Esses documentos são

convertidos em um índice textual que facilita o acesso às informações contidas nos documentos da coleção.

Para essa etapa de reconhecimento de entidades, são utilizadas duas técnicas da área da extração de informação: (1) agrupamento e (2) reconhecimento de entidades nomeadas. Quando a organização não possui uma ontologia de domínio bem formada, o processo de agrupamento é utilizado a partir dos seus documentos. Para dar suporte a essa etapa, emprega-se o algoritmo de agrupamento de documentos Lingo.

O agrupamento de documentos é iniciado a partir de uma consulta textual, em que o resultado é apresentado em forma de conjuntos de documentos agrupados (*cluster*) pelo seu conteúdo. Além disso, é gerado um rótulo (*label*) para o agrupamento, que corresponde a uma sequência de termos que o caracterizam. Esses títulos representam sequências de palavras que ocorrem muitas vezes na coleção de documentos e são possíveis entidades (instâncias da ontologia).

Os títulos resultantes são utilizados de maneira iterativa para gerar novas consultas, objetivando produzir novos agrupamentos até que nenhum novo termo seja encontrado. Abaixo é apresentado o fluxo dessa etapa por meio de um algoritmo:

- passo 1: informa o termo inicial (uma semente);
- passo 2: armazena o termo no vetor de agrupamento;
- passo 3: extrai (e elimina) o primeiro termo do vetor de agrupamento;
- passo 4: gera grupos a partir do termo retirado do vetor;
- passo 5: se gerou grupos, vai para o passo 6; se não, passo 11;
- passo 6: extrai o título dos agrupamentos e o armazena num vetor de títulos;
- passo 7: retira o primeiro termo do vetor de títulos e verifica se ele existe na base de conhecimento;
- passo 8: se o termo existe na base de conhecimento, vá para o passo 10; se não, passo 9;
- passo 9: o termo é armazenado na tabela de entidades temporárias e no vetor de agrupamento;
- passo 10: se o vetor de títulos possui termos, vá para o passo 7; se não, passo 11;
- passo 11: se o vetor de agrupamentos possuir termos, vá para o passo 3; se não, passo 12; e

- passo 12: finalização do processo de reconhecimento de entidades.

Para facilitar a visualização dos passos do algoritmo apresentado anteriormente, foi modelado um fluxograma que é apresentado na Figura 10 abaixo.

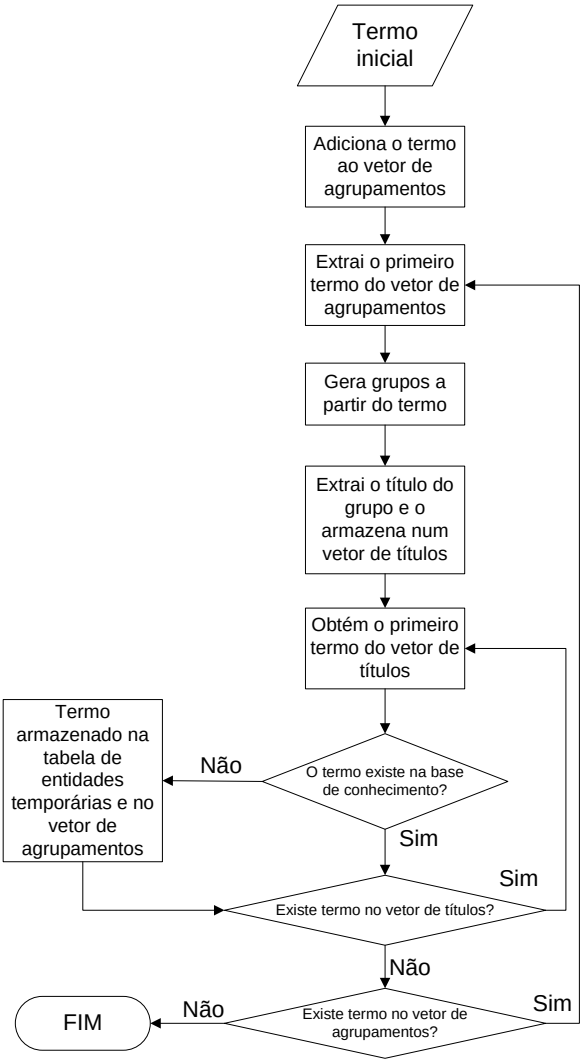


Figura 10 - Fluxograma do algoritmo de reconhecimento de entidades

A abordagem apresentada anteriormente é mais viável para descobrir padrões de texto quando não se tem uma base de conhecimento já constituída. Isso pode ser visto como o primeiro passo de um processo clássico de reconhecimento de entidades que precisa obrigatoriamente de uma lista de termos já classificados. O resultado desse processo é persistido em uma área de estagiamento, que serve como ponto de partida para a etapa de validação, a qual será apresentada na próxima seção.

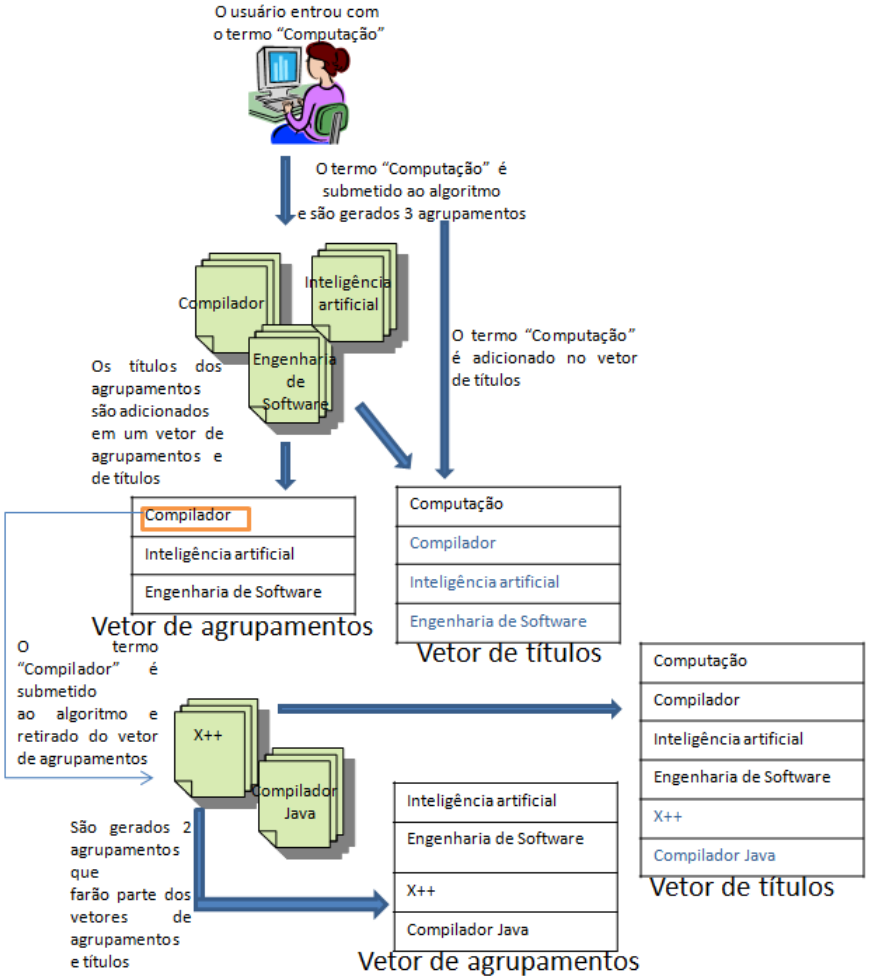


Figura 11 - Exemplo do algoritmo de reconhecimento de entidades

A Figura 11 tem como objetivo exemplificar a utilização do algoritmo de reconhecimento de entidades utilizando a técnica de *clusterização* (agrupamento). Primeiramente o usuário da aplicação entra com um termo para iniciar o processo. No caso do exemplo, o termo escolhido foi "Computação". O processo de agrupamento gerou três grupos de documentos com os seguintes rótulos: "Compilador", "Inteligência artificial" e "Engenharia de Software".

Os rótulos dos grupos são adicionados ao vetor de agrupamento, o qual possui todos os termos que serão submetidos ao algoritmo de reconhecimento de entidades. A figura não ilustra, mas é verificado antes de adicionar ao vetor se o termo já existe na base de conhecimento. Caso ele exista, é desconsiderado e não será adicionado ao vetor.

Os três rótulos dos agrupamentos também são adicionados ao vetor de títulos, que possui todos os termos candidatos a entidades e que serão persistidos ao final do processo.

O próximo passo é obter o primeiro termo do vetor de agrupamentos para submeter ao processo de *clusterização*. No caso do exemplo acima, foi o termo "Compilador". Quando um termo é submetido ao algoritmo, ele é retirado do vetor de agrupamentos. O resultado da submissão do termo "Compilador" gera mais dois agrupamentos cujos rótulos são "X++" e "Compilador Java". Verifica-se se esses termos existem na base de conhecimento. Caso não existam, são adicionados aos vetores de agrupamento e de títulos. Esse processo se repete até que não haja mais nenhum novo termo no vetor de agrupamentos.

Ao final do processo, os termos contidos no vetor de títulos são persistidos em uma tabela temporária para serem submetidos à próxima etapa – a etapa de validação.

Além da geração de termos a partir do processo de agrupamento, essa etapa também conta com a técnica de reconhecimento de entidades nomeadas clássicas. Essa técnica se utiliza de informações (termos) para compor as listas de termos (*gazetteer*) que representam as classes de entidades que se deseja reconhecer. Por exemplo, menciona-se como fonte para a geração da classe "Pessoa" nomes e sobrenomes que em uma organização podem ser obtidos nas tabelas de clientes e/ou funcionários.

Na técnica de reconhecimento de entidades nomeadas é realizada uma marcação dos termos do documento que existam no *gazetteer* (geralmente termos simples). O reconhecimento de termos compostos é

possibilitado apenas por meio da combinação de termos já conhecidos (presentes no *gazetteer*). Essa combinação de termos pode permitir o encontro de novos termos ou apenas explicitar a presença desses termos no texto já com a sua possível classificação.

Os termos encontrados no documento pela técnica de reconhecimento de entidades nomeadas são adicionados aos termos descobertos pelo processo de agrupamento (*clusterização*) com a particularidade de que estes já possuem uma classificação disponível. Todos os termos descobertos são submetidos à etapa de validação, apresentada na seção a seguir.

3.2.2 Validação (2)

O processo de validação proposto por este modelo é semiautomático e possui como ponto de entrada os termos/entidades pré-classificados adicionados à lista inicial de entidades pelo processo anterior.

Como o próprio nome sugere, este processo tem a função de verificar se determinada entidade reconhecida é válida ou não antes de ser disponibilizada para a avaliação do especialista. Esse processo de validação necessita de uma base de termos já validados (neste trabalho, foi utilizada a Wikipédia como base de conhecimento ou base de folksonomia) para identificar se o termo é válido ou não.

Obteve-se a Wikipédia como base de validação a partir de um sítio de domínio público e da posterior indexação dos termos e conceitos. Para uma entidade ser classificada como válida, é necessário que ela conste no índice da Wikipédia. Caso o termo não exista, é feito um levantamento dos termos com grafias similares contidas no índice. Os dez termos mais similares são submetidos a uma busca no índice da coleção de documentos. Se o termo existir, é adicionado à listagem de termos válidos, se não existir, vai para a lista de termos não válidos. Abaixo é apresentado um algoritmo para demonstrar com mais detalhes o funcionamento desta etapa com os seguintes passos:

- passo 1: recupera as entidades da tabela temporária e as armazena num vetor de entidades;
- passo 2: obtém (e remove) o primeiro termo do vetor de entidades e verifica se ele existe na Wikipédia;
- passo 3: se a entidade existe, vá para o passo 4; se não, passo 5;

- passo 4: altera o estado da entidade para entidade válida e vai para o passo 11;
- passo 5: verifica as entidades similares na Wikipédia e carrega um vetor de termos similares;
- passo 6: obtém (e remove) o primeiro termo do vetor de termos similares e verifica se ele existe no índice de documentos (*corpus*);
- passo 7: se existe, vá para o passo 8; se não, passo 9;
- passo 8: persiste a entidade com o estado de válida na tabela de entidades temporárias;
- passo 9: verifica se existe mais termos no vetor de termos similares;
- passo 10: se existe, vá para o passo 6; se não existe, vá para o passo 11;
- passo 11: verifica se existe mais alguma entidade no vetor de entidades;
- passo 12: se existe, vá para o passo 2; se não existe, passo 13; e
- passo 13: finaliza o processo de validação.

Para facilitar o entendimento do algoritmo acima, na Figura 12 é apresentado um fluxograma com todos os passos utilizados pelo algoritmo de validação e seus possíveis caminhos.

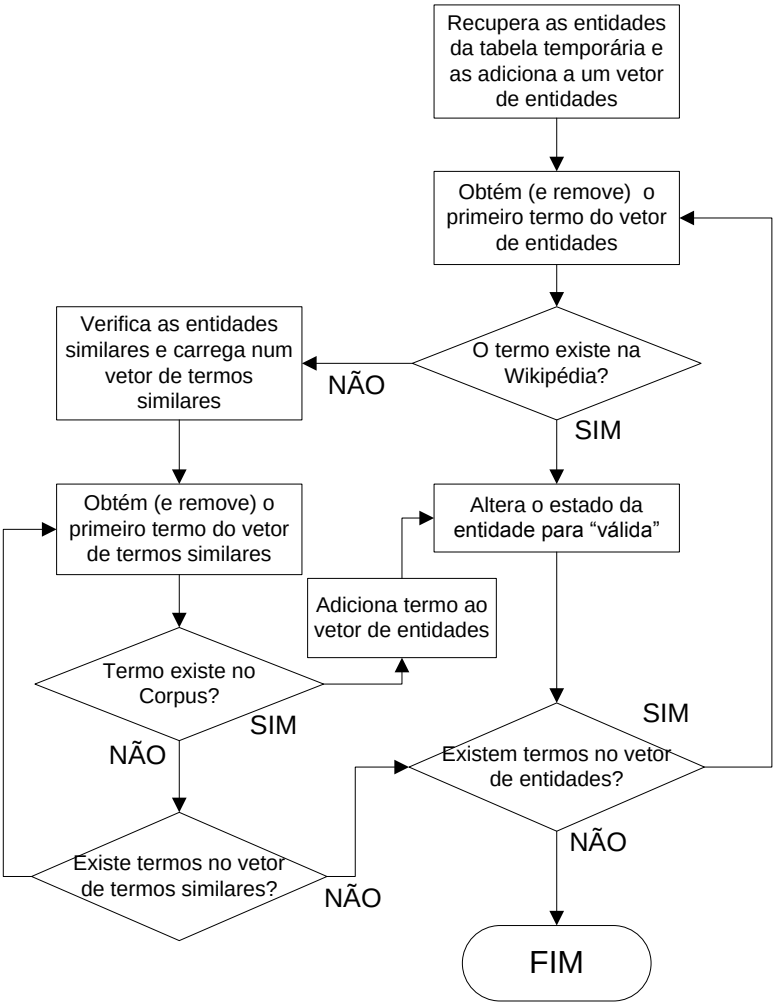


Figura 12 - Fluxograma do algoritmo de validação

A Figura 13 a seguir apresenta um exemplo do funcionamento do algoritmo exposto acima.

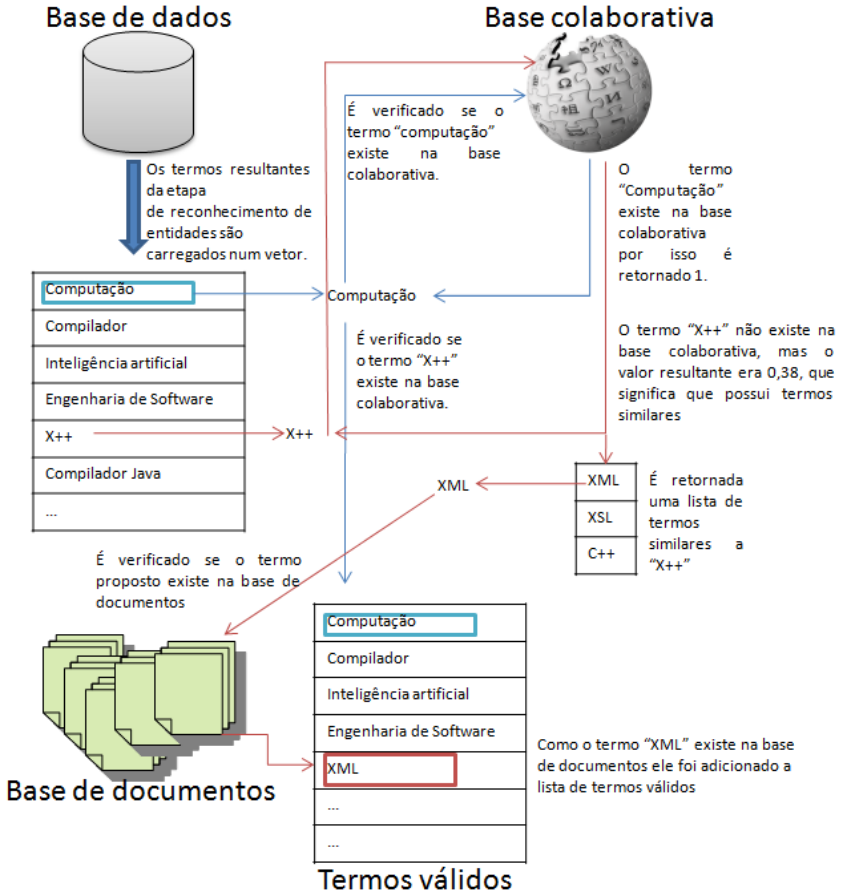


Figura 13 - Exemplo do algoritmo de validação de entidades

A Figura 13 apresenta um exemplo para facilitar o entendimento do algoritmo de validação de entidades. Inicialmente obtêm-se as entidades persistidas em uma tabela da base de dados operacional, as quais são copiadas para um vetor. No exemplo apresentado acima (Figura 13), considera-se inicialmente o termo "Computação", verificando-se em seguida se ele existe na base colaborativa (no caso do protótipo foi utilizada a Wikipédia). Como ele existe, é retornado para a aplicação o número 1 (caso ele não existisse, o valor seria 0, e se ele tivesse termos aproximados, o valor estaria entre 0 e 1).

Como o valor retornado pela consulta à base colaborativa foi 1, esse termo é considerado como termo válido. Se o valor retornado pela

consulta à base colaborativa fosse 0 (zero), o termo seria marcado como não válido e estaria disponível para a validação manual do especialista.

A figura também exemplifica quando a consulta à base colaborativa retorna um valor entre 0 e 1. É o caso do termo "X++", em que a consulta retornou o valor 0,38 (valor relacionado ao termo inicial buscado na Wikipédia, que neste caso foi X++) juntamente com uma lista de termos identificados como similares.

Retirou-se o primeiro termo da lista, "XML" no caso do exemplo apresentado, realizando-se uma consulta ao índice da base de documentos para verificar se o termo existe ou não. Nesse exemplo, ele existia e foi marcado como válido, sendo adicionado à lista de termos válidos; caso ele não existisse, seria descartado.

Ao final desse processo, as entidades pertencentes à lista de entidades, além de possuírem o nome (algumas terão a classe devido ao processo tradicional de reconhecimento de entidades), também terão a informação se são válidas ou não. Tal informação é relevante para a próxima etapa (classificação). Como mencionado anteriormente, essa etapa é semiautomática, pois o usuário pode interagir com os resultados da lista de entidades, principalmente daquelas rotuladas como inválidas, visando aprimorar o resultado do processo.

Na seção a seguir é apresentada a etapa de classificação e de população de ontologias, considerada a etapa final do processo em sua totalidade.

3.2.3 Classificação e população da ontologia (3)

A etapa de classificação, assim como a de validação, é semiautomática e possui como entrada os registros contidos na lista de entidades válidas. A descrição (quando disponível) do termo contido na base de conhecimento, nesse caso a Wikipédia, também é utilizada como entrada de dados para esse processo.

Esta etapa possui três métodos computacionais que auxiliam a classificação das entidades reconhecidas, sendo: (1) métodos estatísticos; (2) expressão regular; e (3) reconhecimento de entidades nomeadas.

Esses métodos têm como objetivo apresentar ao usuário uma possível classificação para as entidades encontradas, visto que essa classificação está diretamente ligada à classe de uma ontologia.

O primeiro método utilizado pelo presente trabalho é o NER, que obtém as instâncias contidas na lista de entidades e procura essas instâncias no seu dicionário léxico. Se a instância for localizada, a ela

são atribuídas a classe correspondente e a informação persistida na base de conhecimento.

O segundo método utiliza a descrição (significado) da entidade contida na Wikipédia. Para tal, é necessário que o usuário anteriormente informe algumas palavras-chave que auxiliem na caracterização de determinada classe, ou seja, é criado um contexto. Abaixo é apresentado um exemplo.

Considere "Steve Wozniak" como uma das entidades encontradas que utilizam o modelo proposto. O passo seguinte será a realização de uma consulta ao índice da Wikipédia a fim de encontrar o significado dessa entidade. Como resultado, tem-se o seguinte (texto retirado da Wikipédia):

> Stephen Gary Wozniak (São José, 11 de agosto de 1950), conhecido como Woz ou Wizard of Woz (em alusão ao filme The Wizard of Oz), é um engenheiro de computação, cofundador da Apple Computers, agora a Apple, Inc., junto com Steve Jobs. Foi pioneiro na iniciativa de colocar computadores disponíveis para o consumidor comum. Apesar de sua contribuição ter sido uma compilação de poucas bem conhecidas ideias que coincidiram perfeitamente com o surgimento da tecnologia necessária para produção em massa de computadores, a engenhosidade de Stephen Wozniak, sua persistência e criatividade deram-lhe o crédito por iniciar a revolução do computador pessoal.

No texto acima, pode-se identificar palavras como "engenheiro", "cofundador", "consumidor", etc., as quais podem ser usadas para classificar como uma instância da classe "Pessoa". No caso de existirem palavras cadastradas de tipos distintos, é verificada a quantidade de vezes que cada tipo (classe) ocorre, sendo apresentada a que tiver maior frequência. Em caso de possuírem o mesmo número de ocorrências, é apresentada mais de uma sugestão de classificação.

O último método computacional disponível para esta etapa é baseado em expressões regulares, em que o usuário vincula uma expressão regular a uma classe. Essa expressão pode ser aplicada tanto na entidade quando da descrição dela. Um exemplo são palavras sucedidas por "Ltda.", que representam organizações, ou palavras antecedidas por Dr., M.Sc., Eng., que são representações de pessoas.

O processo de classificação está diretamente ligado ao processo de população (ou manutenção) de ontologias, já que o resultado desse processo é um conjunto de instâncias já classificadas segundo uma base de conhecimento que compõe uma ontologia de domínio.

O usuário pode validar a classificação realizada de maneira automática bem como sugerir novas classificações e instâncias para compor a ontologia. É importante observar que os processos de reconhecimento de entidades, de validação e de classificação são incrementais. Vale lembrar ainda que os termos submetidos a essas etapas e posteriormente anexados como instâncias da ontologia são insumos para futuras operações de manutenção de ontologia. Desse modo, pode-se dizer que o sistema "aprende" ao longo do seu uso.

3.3 Considerações finais

Este capítulo teve como objetivo apresentar o modelo proposto para construção e manutenção de ontologias a partir de bases textuais não estruturadas.

O modelo pode ser dividido em três etapas, sendo a primeira responsável pelo reconhecimento das entidades. Essa etapa pode utilizar duas técnicas distintas – (1) a técnica de reconhecimento de entidades nomeadas clássica ou (2) a extração de rótulos de agrupamento de documentos –, sendo esta última mais eficiente para reconhecer entidades quando não existe uma lista previamente construída de termos válidos.

A segunda etapa é a de validação, que, como o próprio nome sugere, é responsável por verificar se uma entidade encontrada é válida para o domínio em questão ou não. Nessa etapa, utilizam-se bases de conhecimento colaborativas como insumo para esse processo, que é semiautomático e necessita da interação do usuário.

A terceira e última etapa é responsável pela classificação das entidades válidas. Esse processo, que também é semiautomático, procura primeiramente realizar a classificação utilizando expressões regulares que o usuário deve cadastrar (por exemplo, para reconhecer datas). Outra possibilidade é o cadastro de algumas palavras-chave como contexto de determinada classe que auxiliam na definição da entidade em bases de conhecimento colaborativas. Nessa última etapa, o usuário interage com o sistema para manipular as entidades encontradas vinculando-as junto às classes da ontologia.

O próximo capítulo analisa a viabilidade do modelo, apresentando o protótipo desenvolvido e alguns comparativos com trabalhos já publicados nessa área.

4 DEMONSTRAÇÃO DE VIABILIDADE E ANÁLISE COMPARATIVA

Este capítulo demonstra a aplicação do modelo proposto descrito no capítulo anterior. Primeiramente, apresenta-se um protótipo baseado no modelo e, em seguida, um estudo de caso visando demonstrar a sua viabilidade. Por fim, é realizado um comparativo entre modelos similares disponíveis publicamente e o modelo proposto neste trabalho.

4.1 Arquitetura física da solução

Nesta seção, apresentam-se quais *frameworks* e aplicações são utilizados nesse modelo e como eles se comunicam entre si formando a solução proposta. A Figura 14 abaixo exibe esta arquitetura.

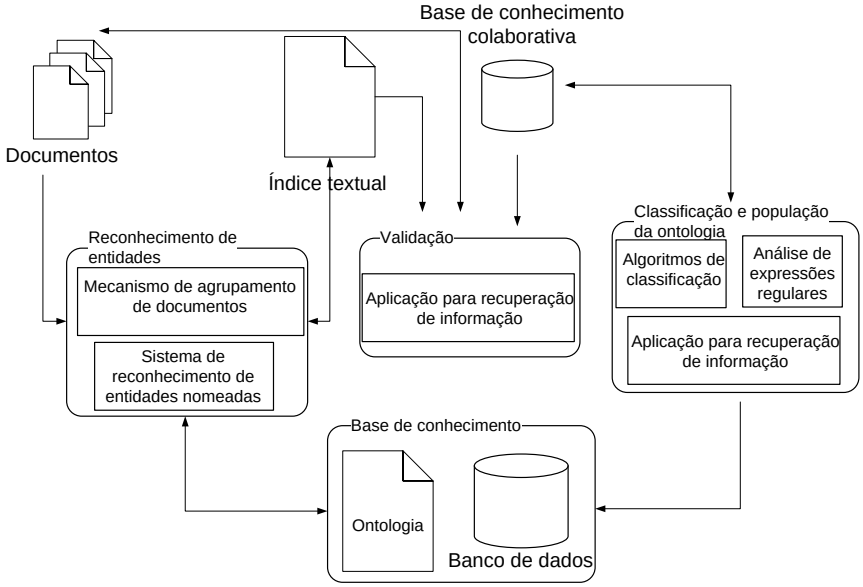


Figura 14 - Arquitetura lógica do modelo proposto

A etapa de reconhecimento de entidades foi desenvolvida com o uso de técnicas computacionais, as quais são relacionadas a seguir.

- Sistema de reconhecimento de entidades nomeadas: é empregado para fazer o reconhecimento de entidades de

maneira clássica, utilizando uma lista de palavras já classificadas e identificando-as na coleção de documentos.

- Aplicação para recuperação de informação: *framework* para indexação e busca de informação, é utilizado nesta etapa para interpretar as informações do índice gerado para o algoritmo de agrupamento de documentos.
- Mecanismo de agrupamento de documentos: é um *framework* para recuperação de informação que possibilita a apresentação do resultado em forma de *clusters* (agrupamentos). É por meio desse algoritmo de *clusterização* que se extraem as entidades (que são os rótulos dos grupos encontrados).

A etapa de validação das entidades reconhecidas utiliza apenas a aplicação de recuperação de informação, já que se adotou a Wikipédia como base de conhecimento colaborativa, sendo transformada em um índice textual. Nessa etapa, basicamente são feitas buscas nos índices. Uma entidade é válida quando consta na Wikipédia, ou, caso se encontre um termo relacionado na Wikipédia que exista na coleção de documentos (busca no índice de coleção de documentos), a entidade é adicionada à lista e marcada como válida.

A base de conhecimento da organização é formada por uma ontologia de domínio (que pode já estar construída e ser apenas atualizada, ou estar em branco). A ontologia representa o final do fluxo deste modelo, mas também é utilizada por todas as etapas. Dessa maneira, o modelo vai "aprendendo" à medida que for utilizado.

O sistema de reconhecimento de entidades também é usado para a etapa de classificação das entidades encontradas. Ele utiliza as suas tabelas léxicas para identificar uma possível classificação para as entidades em questão. As outras duas estratégias disponibilizadas por esse modelo foram inteiramente desenvolvidas em Java e sem o emprego de bibliotecas de terceiros.

Ainda na etapa de classificação e população das ontologias, pode-se encontrar a presença do servidor web, que é o responsável por disponibilizar uma interface gráfica ao usuário via navegador para as etapas de validação e classificação das entidades, bem como toda a manipulação da ontologia de domínio da organização.

A próxima seção detalha o protótipo desenvolvido para demonstrar a viabilidade do modelo, assim como um comparativo do modelo com outras ferramentas e soluções para manutenção de ontologias a partir de uma coleção de documentos não estruturados.

4.2 Apresentação do protótipo

Para atestar a viabilidade do modelo proposto, optou-se pelo desenvolvimento de um protótipo dividido basicamente em duas partes: (1) um sistema de reconhecimento de entidade e (2) uma aplicação web para o usuário interagir com os resultados e manipular a ontologia em questão. A Figura 15 abaixo apresenta a arquitetura física da solução proposta, dividindo os seus componentes entre a aplicação para reconhecimento de entidades (em azul) e a aplicação web (em amarelo).

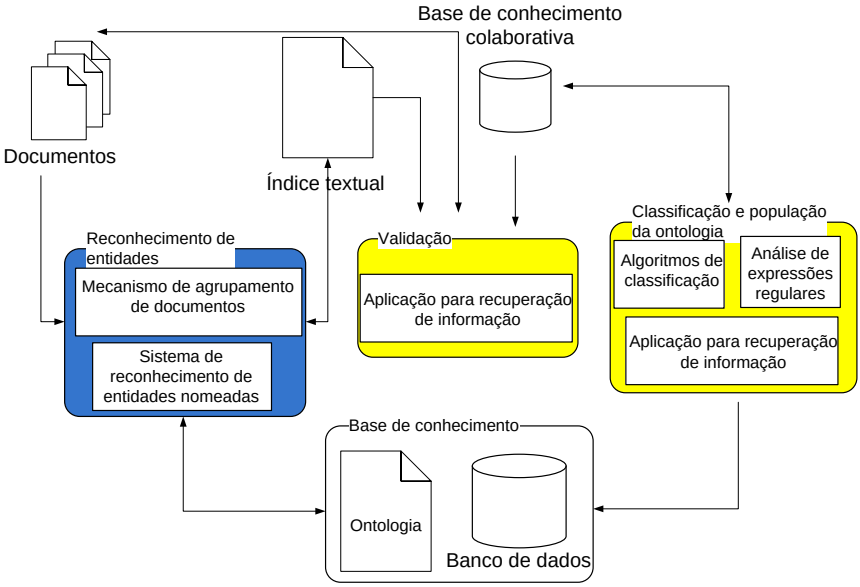


Figura 15 - Componentes das aplicações *back-end* e web

As próximas seções apresentam em mais detalhes a aplicação de reconhecimento de entidades e a aplicação web.

4.2.1 Aplicação para reconhecimento de entidades

Como foi descrito na seção 3.2.1, a entrada dos dados é uma coleção de documentos, a qual deve estar organizada em uma pasta que o usuário deverá informar ao sistema.

Quando o usuário inicia a aplicação de reconhecimento de entidades, esta irá varrer os locais (diretórios) previamente configurados a fim de recuperar todos os arquivos que serão submetidos aos algoritmos desenvolvidos. O usuário deverá também informar ao sistema qual a semente de busca inicial para que a aplicação inicie o processo. Após a configuração e execução inicial, o processo se repete de tempos em tempos, independentemente da intervenção do usuário.

Quando o sistema é iniciado, verifica-se se existe um índice que reflita a coleção de documentos informados no arquivo de configuração. Caso não exista, é feita a geração desse índice. Se existir um índice, examina-se se há algum documento novo na coleção; caso conste, o índice é atualizado.

Como apresentado na Figura 15, a aplicação de *back-end* é responsável por toda a etapa de reconhecimento de entidades. Toda a parte computacional é feita por essa aplicação, que terá como saída uma lista de termos pré-validados que deverão ser manipulados pelo usuário.

A aplicação de reconhecimento de entidades irá persistir os dados em uma tabela de entidades temporária, identificando se essas instâncias são válidas ou não. Isso auxilia a etapa de validação por parte do usuário, que é feita por meio da aplicação web.

A próxima seção apresenta com mais detalhes a aplicação web.

4.2.2 Aplicação web

Para facilitar a interação do engenheiro de ontologias (usuário) com as aplicações de *back-end*, foi desenvolvida uma aplicação web que permite interagir com todas as três etapas do modelo (reconhecimento de entidades, validação e classificação), facilitando o entendimento do modelo e auxiliando no processo de manutenção de determinada ontologia. A Figura 16 a seguir demonstra a interface da aplicação.

Figura 16 - Página inicial da aplicação web

Como se pode observar na figura, na parte superior existe um menu que apresenta as principais etapas do modelo proposto. Além dessas etapas, há um *link* "Ontologia" que possibilita a visualização da ontologia de forma gráfica. Outro *link* disponível é "Configurações". Por meio dele, o usuário tem a possibilidade de informar os dados de configuração, como o local onde a coleção de documentos de entrada está no servidor, qual o caminho do índice da base de conhecimento utilizado para a validação e a classificação e o local onde o arquivo da ontologia (em formato OWL[2]), que irá receber os dados extraídos pelo modelo, encontra-se no servidor. A Figura 17 apresenta o menu superior da aplicação.

Figura 17 - Menu da aplicação web

Após o usuário ter configurado os caminhos da coleção dos documentos e solicitado a geração do índice (caso este não exista) pela página de configuração, ele deve configurar também o caminho do índice da base de conhecimento e, por fim, o arquivo OWL da ontologia que será carregada (populada).

[2] OWL (*Ontology Web Language*) – linguagem utilizada para definir e instanciar ontologias.

Com as configurações iniciais realizadas, o próximo passo (primeira etapa descrita no modelo proposto) é o reconhecimento das entidades nomeadas. Para tal, o usuário deve acessar o *link* "Reconhecimento Entidades". Após isso, uma página como a apresentada na Figura 18 é aberta, e nela deve-se entrar com um termo de busca (semente inicial), uma vez que a abordagem de agrupamento implementada possui essa necessidade. Uma maneira de contornar essa limitação seria apresentar ao usuário uma lista (10 elementos, por exemplo) para escolha dos termos mais significativos da coleção de documentos indexada.

Figura 18 - Página para iniciar o reconhecimento das entidades

Entretanto, na abordagem atual é extremamente recomendado que o usuário utilize uma palavra-chave relacionada com o domínio representado na coleção de documentos, pois isso irá ajudar na montagem e separação dos grupos por parte do algoritmo de agrupamento utilizado.

Após informar o termo de busca, o usuário deve pressionar o botão "Processar" para iniciar o processo de reconhecimento de entidade e, por fim, o botão "Salvar" para concluir essa etapa e poder manipular o resultado na etapa seguinte.

A etapa de reconhecimento de entidades produz como resultado uma lista de termos encontrados na coleção de documentos de entrada, em que cada item dessa lista de termos é submetido ao algoritmo de validação, o qual irá verificar se o termo existe na base colaborativa. Caso o termo exista, este vai para uma lista de termos válidos; se não

existir, é feita uma busca na base colaborativa por termos similares. Os dez primeiros termos similares retornados pela busca são submetidos a nova outra busca, mas dessa vez na coleção de documentos. Se eles existirem, vão para a lista de termos válidos.

Para o usuário visualizar as entidades reconhecidas, deve clicar no *link* "Validação" do menu superior. Será aberta uma página como a demonstrada na Figura 19. Nessa página, está disponível a lista de entidades reconhecidas divididas em duas colunas, uma com as entidades válidas e outra com as entidades não válidas.

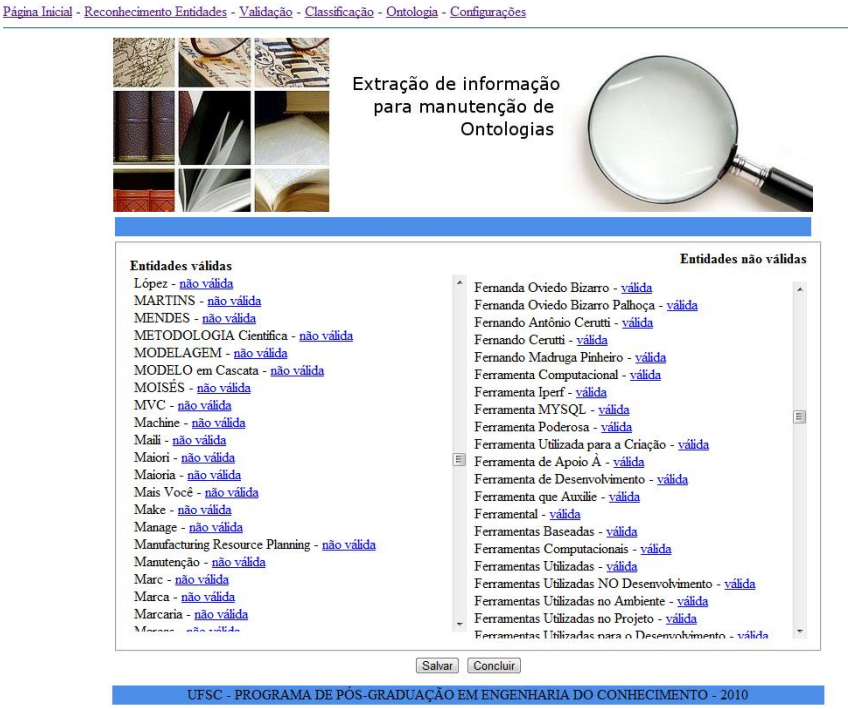


Figura 19 - Página de validação das entidades reconhecidas

Cada entidade listada possui um *link* ao lado que possibilita alterar o seu estado de "entidade válida" para "entidade não válida", ou vice-versa. É por meio dessas operações que o usuário interage com o processo de validação de entidades.

Existem mais duas operações disponíveis nesta página: uma por meio do botão "Salvar", que grava as operações de mudança de estado entre as entidades, mas não disponibiliza o resultado para a próxima

etapa. A outra operação disponível é através do botão “Concluir”, que disponibiliza as entidades válidas para a próxima etapa – a etapa de classificação. Vale mencionar que as entidades não válidas são armazenadas para que, em um processo de extração futuro, não sejam mais levadas em consideração.

Ao final desta etapa, os termos que estão com o estado de entidade válida ficam à disposição da próxima etapa, que é a de classificação.

A página inicial do módulo de classificação, apresentada na Figura 20, disponibiliza um ambiente para o especialista cadastrar as classes já conhecidas da ontologia. Caso o usuário tenha informado no módulo de configurações uma ontologia (que será submetida ao processo de manutenção), as classes nela cadastradas ficam disponíveis automaticamente.

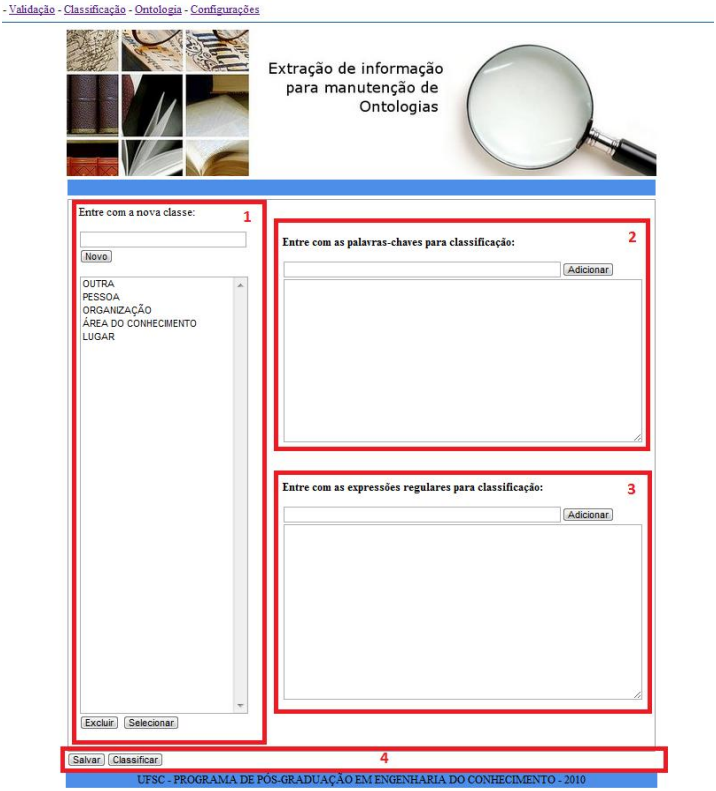


Figura 20 - Página inicial do módulo de classificação

As classes da ontologia são disponibilizadas na região 1 da Figura 20, em que se pode observar a existência de um campo que permite ao usuário cadastrar manualmente uma nova classe. As classes ficam visíveis na lista logo abaixo do campo de cadastro.

Na parte inferior correspondente à região 1, encontram-se dois botões, “Excluir” e “Selecionar”. Para utilizá-los, o usuário deve primeiramente selecionar uma das classes listadas. O botão “Excluir”, como o próprio nome sugere, apaga a classe selecionada; já o botão “Selecionar” possibilita que o usuário entre com os dados que irão auxiliar na etapa de classificação.

Esse processo de cadastro das classes e informações referentes a elas (adicionar palavras-chave ou expressões regulares) pode ser feito a qualquer momento, independentemente da etapa. Ele só é pré-requisito para a etapa de classificação, que utiliza essas informações como fonte.

Para a execução das operações identificadas nas regiões 2 e 3, é necessário que o usuário anteriormente tenha escolhido uma classe na região 1 por meio do botão “Selecionar”. Na região 2, é possível cadastrar as palavras-chave utilizadas na etapa de classificação. Para efetuar o cadastro, o usuário deve informar a palavra-chave no campo de texto e clicar no botão “Adicionar”. A palavra-chave então será apresentada na área de texto logo abaixo.

Como esse protótipo tem a função apenas de materializar a interação com o modelo proposto, não foram implementadas as operações de exclusão e edição das palavras-chave cadastradas em uma classe.

A região 3 apresenta as mesmas operações e limitações da região 2, mas as informações que são cadastradas na região 3 correspondem às expressões regulares que também auxiliam na classificação dos termos encontrados.

Por fim, a região 4 possui dois botões, “Salvar” e “Classificar”. No primeiro, todas as informações cadastradas e interações feitas são persistidas na base de dados do protótipo.

O botão “Classificar” remete à página que possibilita iniciar o processo de classificação das entidades reconhecidas e marcadas como válidas nos processos anteriores, como apresentado na Figura 21.

Figura 21 - Tela para início da etapa de classificação

A página ilustrada na Figura 21 possui o botão "Iniciar classificação". É por meio dele que a aplicação analisa as entidades reconhecidas e marcadas como válidas e assim aplica o algoritmo de classificação, levando em consideração as palavras-chave e as expressões regulares cadastradas na etapa anterior (Figura 20).

O resultado dessa operação é uma lista de entidades com a sua possível classe entre parênteses, como apresentado na Figura 22.

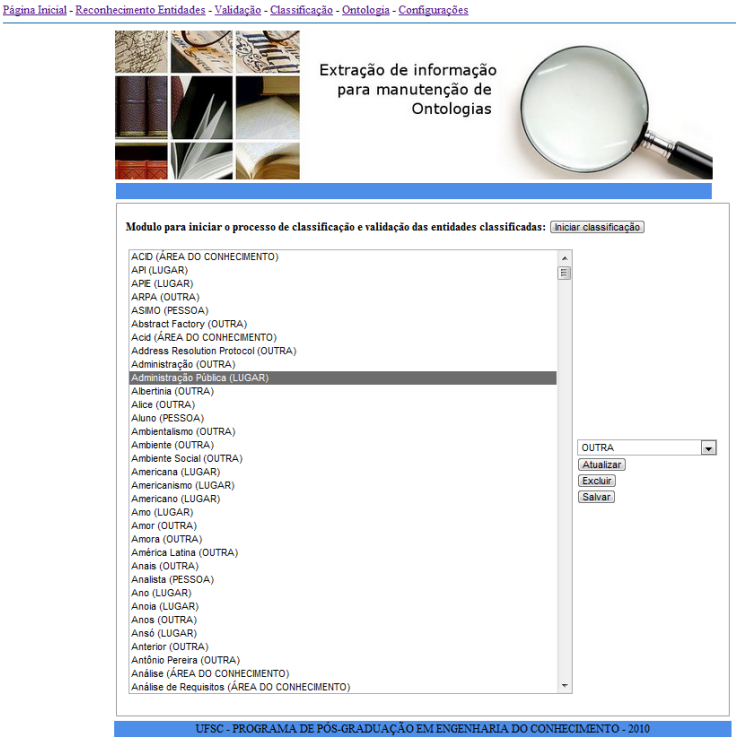


Figura 22 - Tela para validação do resultado da classificação

Cada entidade classificada terá ao seu lado uma classe associada como resultado do processo. O usuário possui ainda a possibilidade de alterar essa classificação, selecionando a entidade na lista de resultados e depois escolhendo a nova classe na região apresentada à direita da tela. Depois de escolhida a nova classe, basta pressionar o botão "Atualizar".

O usuário pode ainda excluir as entidades e sua classificação, selecionando a entidade na lista de resultados e pressionando o botão "Excluir". Para gravar todas as operações feitas, basta que o usuário pressione o botão "Salvar".

Após clicar em "Salvar", o usuário terá um arquivo no formato OWL. Essa ontologia pode ser editada e manipulada com qualquer ferramenta que interprete arquivos nesse formato. Para facilitar a visualização, utilizou-se a ferramenta Protègè, que permite visualizar a ontologia gerada a partir do protótipo criado. Vale lembrar que foram selecionados poucos resultados para facilitar a visualização (Figura 23).

A Figura 23 demonstra como está estruturado o resultado obtido pelo protótipo, onde se encontram as instâncias ligadas às classes sem explicitar o tipo de relacionamento ou a cardinalidade.

Este exemplo mostrou a criação de uma ontologia sem nenhuma referência a ontologias já construídas, utilizando apenas os dados do *corpus* de documentos (nesse caso, trabalhos de conclusão de curso dos cursos de Ciência da Computação e Sistemas de Informação). Utilizou-se a técnica de agrupamento para identificar os novos termos. Na seção 4.3, é apresentado um estudo de caso que emprega a técnica de reconhecimento de entidades nomeadas para ilustrar a outra forma que o modelo proposto disponibiliza para a etapa de extração de entidades.

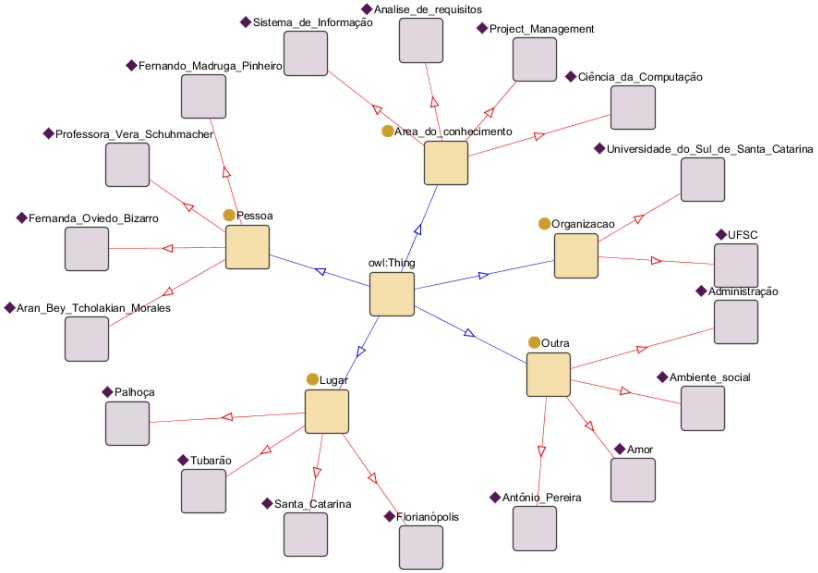


Figura 23 - Visualização da ontologia gerada pelo Protègè

4.2.3 Ferramental de apoio ao modelo

Nesta seção, são apresentados os *frameworks* e as aplicações utilizados no protótipo e como eles se comunicam entre si objetivando materializar o modelo proposto. A Figura 24 ilustra esta arquitetura:

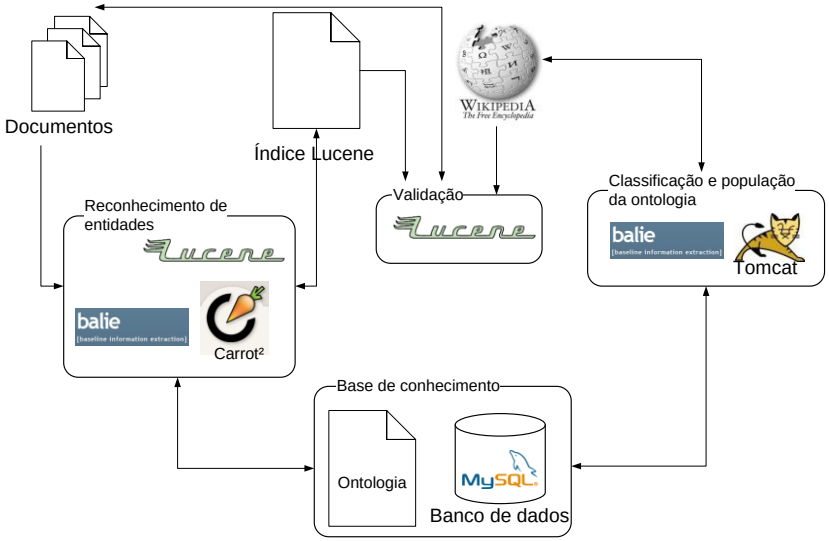


Figura 24 - Ferramentas de apoio ao modelo

A etapa de reconhecimento de entidades foi desenvolvida inteiramente em Java e utiliza algumas bibliotecas de terceiros para facilitar a sua implementação. Abaixo apresenta-se cada uma delas.

- Balie[3] (*baseline information extraction*) – é utilizada para fazer o reconhecimento de entidades de maneira clássica por meio de uma lista de palavras previamente classificadas.
- LUCENE – *framework* para indexação e recuperação de informação. Segundo Hatcher e Gospodnetic (2005), o LUCENE é a biblioteca de recuperação de informação mais popular entre as existentes e funciona como um motor para a indexação de termos de busca em textos completos. Permite que todas as regras de negócio sejam tratadas pelas aplicações que serão desenvolvidas a partir dele. A Figura 25 a seguir exibe a integração do LUCENE com uma aplicação.

[3] Disponível em: <http://balie.sourceforge.net/>. Acesso em: 20 jul. 2010.

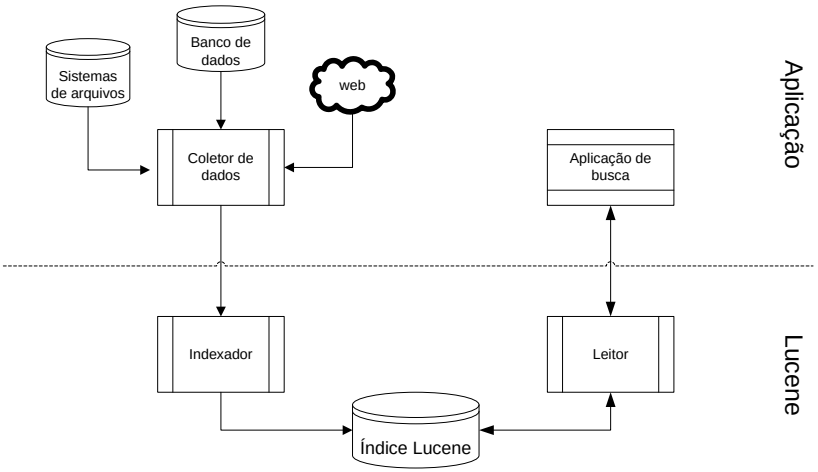


Figura 25 - Integração de uma aplicação com o LUCENE
Fonte: adaptado de Hatcher e Gospodnetic (2005, p. 8).

Hatcher e Gospodnetic (2005) afirmam ainda que o LUCENE pode efetuar buscas em qualquer tipo de arquivo desde que o seu conteúdo seja texto. Vale lembrar que este trabalho irá contemplar o uso do LUCENE apenas na etapa de indexação.

No LUCENE, os modelos booleano e espaço vetorial, além da possibilidade de utilização de pesos nos termos da consulta, são combinados para estabelecer a relevância de determinado documento em relação a uma consulta do usuário. Para a etapa de indexação, o LUCENE emprega a técnica de arquivo invertido. É usado nesta etapa para interpretar as informações do índice gerado para o Carrot².

- Carrot²[4] – é um *framework* para recuperação de informação (baseado no LUCENE) que possibilita a apresentação do resultado em forma de agrupamentos. É através desse algoritmo de clusterização que as entidades são extraídas (representadas pelos rótulos dos agrupamentos). Para iniciar o processo, é necessária uma "semente de busca", uma vez que o Carrot² constitui-se de uma solução completa de recuperação de informação em documentos textuais.

[4] Disponível em: <http://project.carrot2.org/> Acesso em: 10 out. 2010.

A etapa de validação das entidades reconhecidas utiliza o LUCENE, visto que a Wikipédia (base de conhecimento colaborativa empregada no processo) foi indexada permitindo assim consultas textuais. Uma entidade é válida se existir na Wikipédia. Se encontrar um termo relacionado à entidade obtida a partir de uma consulta na Wikipédia que conste na coleção de documentos (busca no índice de coleção de documentos), a entidade é adicionada à lista e marcada como válida.

A base de conhecimento da organização é formada por um banco de dados (no caso do protótipo, utilizou-se o MySQL) e por uma ontologia de domínio (que pode ou não possuir instâncias previamente cadastradas). A ontologia representa o final do fluxo deste modelo, mas também é aplicada em todas as etapas. Dessa maneira, o modelo vai "aprendendo" à medida que é utilizado.

O Balie também é usado na etapa de classificação das entidades encontradas por meio de suas tabelas léxicas, de modo a identificar uma possível classe para uma determinada entidade em análise. As outras duas estratégias disponibilizadas por esse modelo foram inteiramente desenvolvidas em Java e sem o emprego de bibliotecas de terceiros.

Ainda na etapa de classificação e população das ontologias, pode-se encontrar a presença do servidor web Tomcat, responsável por disponibilizar uma interface gráfica ao usuário via navegador web para as etapas de validação e classificação das entidades, bem como toda a manipulação da ontologia de domínio da organização.

4.3 Estudo de caso

Nesta seção, aborda-se um estudo de caso aplicando-se os dados gerados pelo modelo proposto para produzir uma rede de relacionamento entre pessoas, organizações e área do conhecimento a fim de demonstrar a viabilidade do modelo.

Como o protótipo apresentado na seção anterior é voltado à extração de entidades sem a utilização de uma base de conhecimento previamente construída, neste estudo de caso optou-se por empregar a abordagem de reconhecimento de entidades nomeada *clássica*, com vistas a validar essa alternativa que também é possibilitada pelo modelo.

A seção abaixo apresenta como foi realizada a preparação dos dados de entrada para o estudo de caso.

4.3.1 Preparação dos dados

Para este estudo voltado ao reconhecimento de entidades, tomou-se como fonte de dados a Plataforma Lattes do CNPq[5]. Para tal, foram extraídos *résumés* acadêmicos de alguns pesquisadores, conteúdo esse disponível publicamente no site do CNPq. Um *résumé* acadêmico constitui-se em um texto livre no qual a pessoa descreve suas atividades profissionais, áreas de conhecimento e instituições de atuação profissional.

O objetivo do estudo é explicitar os relacionamentos entre as entidades (ou instâncias) contidas no texto de modo a apoiar a manutenção de ontologias. Assim, quando novos relacionamentos entre instâncias de classes são detectados, o engenheiro de ontologias pode rever a ontologia para realizar possíveis alterações. Dessa forma, a manutenção de ontologias pode ser feita a partir das fontes de dados da própria organização com o auxílio da ferramenta desenvolvida.

Neste estudo de caso, não foi utilizada nenhuma ontologia para ser atualizada. Partiu-se apenas dos *résumés* como entrada de dados, em que, ao final, o sistema deve apontar as instâncias (entidades) e os relacionamentos entre elas, apresentando os resultados na forma de uma rede.

A base de conhecimento necessária ao processo de classificação/nomeação de entidades foi gerada a partir do currículo do autor, capturando-se via CV-Lattes (disponível on-line) a relação de áreas do conhecimento e de cursos para compor a tabela léxica de áreas do conhecimento. As instituições encontradas nos *résumés* utilizados no estudo foram coletadas e adicionadas à tabela léxica de organizações, enquanto que os nomes e os sobrenomes dos pesquisadores foram adicionados à tabela léxica de pessoas. Desse modo, as três classes – KNOWLEDGE_AREA (área do conhecimento), PERSON (pessoa) e ORGANIZATION (organização) – necessárias para este estudo foram compostas.

Para facilitar o entendimento do processo, apresenta-se abaixo um extrato dos *résumés* dos currículos Lattes de alguns pesquisadores fornecido como parte da entrada para o estudo de caso:

> Flávio Ceci concluiu a graduação em Ciência da Computação pela Universidade do Sul de Santa Catarina em 2007. Flávio é mestrando do curso de

[5] Disponível em: <http://lattes.cnpq.br>. Acesso em: 10 out. 2010.

Engenharia e Gestão do Conhecimento pela Universidade Federal de Santa Catarina, Atualmente é desenvolvedor do Instituto Stela. Possui 6 softwares e outro 1 item de produção técnica. Entre 2004 e 2007 participou de 4 projetos de pesquisa. Atualmente participa de 3 projetos de pesquisa. Flávio atua na área de Ciência da Computação, com ênfase em Sistemas de Informação, atuando principalmente nos seguintes temas: reconhecimento de entidades; técnicas de inteligência artificial aplicada à engenharia do conhecimento; população de ontologias; descoberta de conhecimento em bases textuais e recuperação de informação. Em suas atividades profissionais interagiu com 13 colaboradores em coautorias de trabalhos científicos.

Alexandre Leopoldo Gonçalves possui graduação em Bacharel em Ciências da Computação pela Fundação Universidade Regional de Blumenau (1997), mestrado em Engenharia de Produção pela Universidade Federal de Santa Catarina (2000) e doutorado em Engenharia de Produção pela Universidade Federal de Santa Catarina (2006). Atualmente Alexandre é colaborador e líder da Unidade de Produto do Instituto Stela. Alexandre tem experiência na área de Ciência da Computação, com ênfase em Engenharia do Conhecimento, atuando principalmente nos seguintes temas: extração e recuperação de informação, mineração de textos e extração e engenharia do conhecimento. Possui trabalhos publicados em periódicos especializados e em eventos nacionais e internacionais em diversos países, assim como softwares com e sem registro. Desde 2001 participa tanto na atuação quanto na coordenação de projetos de pesquisa no Brasil e no exterior.

Denilson Sell concluiu o doutorado em Engenharia de Produção pela Universidade Federal de Santa Catarina em 2007. Atualmente Denilson é Professor da Universidade Federal de Santa Catarina, Analista de Sistemas do Instituto

> Stela e Professor da Universidade do Estado de Santa Catarina. Publicou 1 artigo em periódico especializado e 16 trabalhos em anais de eventos. Possui 16 softwares, sendo 1 com registro e outros 11 itens de produção técnica. Participou de 3 eventos no exterior e 6 no Brasil. Denilson coorientou 5 dissertações de mestrado, além de ter orientado 2 trabalhos de conclusão de curso nas áreas de Ciência da Computação e Administração. Recebeu 2 prêmios e/ou homenagens. Entre 1997 e 2005, participou de 11 projetos de pesquisa. Atualmente participa de 5 projetos de pesquisa, sendo que coordena 2 destes. Atua na área de Ciência da Computação, com ênfase em Sistemas de Informação. Em suas atividades profissionais interagiu com 55 colaboradores em coautorias de trabalhos científicos.
>
> Dhiogo Cardoso da Silva possui graduação em Bacharelado em Sistemas de Informação pela Universidade Federal de Santa Catarina (2007), e no momento é mestrando de Engenharia do Conhecimento da Universidade Federal de Santa Catarina. Atualmente Dhiogo é colaborador do Instituto Stela. Dhiogo tem experiência na área de Ciência da Computação, com ênfase em Sistemas de Informação, atuando principalmente nos seguintes temas: Business Intelligence, Web Semântica, Data Warehousing e Text Mining.

Com os dados da entrada do processo definidos, o próximo passo foi a utilização desses dados no modelo proposto. As informações geradas foram persistidas em uma base de conhecimento no formato {entidade recuperada, frequência e posições por documento} para serem consumidas por qualquer sistema baseado em conhecimento. Na próxima seção, é apresentado o algoritmo de correlação que calcula o grau de relacionamento entre as entidades e permite, a partir desse resultado, a sua projeção na forma de rede. A projeção é realizada utilizando-se uma estrutura no formato GraphML por um componente chamado ISLinks[6].

[6] O ISLinks é de propriedade do Instituto Stela e foi gentilmente disponibilizado para a realização deste trabalho.

4.3.2 Algoritmo de correlação

Dado o *résumé* curricular de uma pessoa no formato de texto, o protótipo realiza o processo de reconhecimento de entidades. O próximo passo constitui-se na geração de um vetor contendo as entidades encontradas, a classe a que elas pertencem, quais são as suas posições no texto e em qual sentença (frase) elas estão contidas. A partir desse vetor, são extraídos os termos distintos que representam os índices da matriz. A matriz gerada é do tipo Entidade X Entidade, e as células armazenam o valor da correlação entre as entidades. A Tabela 1 a seguir exemplifica essa matriz:

Tabela 1 - Matriz de correlação entre entidades

	UFSC	EGC	Flávio Ceci	Instituto Stela
UFSC	-	2,7	0,9	0,012
EGC	2,7	-	1,2	0,88
Flávio Ceci	0,9	1,2	-	1,8
Instituto Stela	0,012	0,88	1,8	-

O sistema verifica a quantidade de entidades contidas no vetor do índice e gera uma matriz quadrada com o tamanho do vetor. Em seguida, são combinados todos os termos da matriz a fim de gerar a força da relação entre duas entidades quaisquer. O algoritmo para o cálculo de correlação foi inspirado em Zhu, Gonçalves e Uren (2005). A abordagem utilizada é bastante simplificada e considera a correlação em uma sentença (frase), e não em um documento.

A correlação entre as entidades é calculada utilizando-se as coocorrências divididas pela média das janelas entre as entidades. Abaixo é apresentada a equação usada para se calcular a correlação entre dois termos. Na equação, *freq* é igual à frequência com que as entidades coocorrem (frequência conjunta) na sentença, e $\bar{j}$ é a média das janelas, sendo uma janela definida como a quantidade de termos que existem entre as entidades na sentença.

Por exemplo, na sentença “Flávio Ceci concluiu a graduação em Ciência da Computação”, as entidades “Flávio Ceci” e “Ciência da Computação” possuem uma frequência conjunta de 1 e uma janela de 4, pois existem quatro palavras entre as duas entidades. A fórmula de correlação é a seguinte:

$$correlação = \sum_{i=1}^{n} \frac{freq}{\bar{j}}$$

onde, *freq* representa as coocorrências em um determinado texto/documento divididas pela janela média $\bar{j}$. Os dois parâmetros são considerados ao nível de sentença. A janela média ($\bar{j}$) é calculada pela seguinte fórmula:

$$\bar{j} = \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{1}{n} (x_1 + ... + x_n)$$

onde, *n* é o número de janelas existentes entre as entidades (ao nível de sentença) e x_i é a janela para a *i*th sentença do texto. Vale mencionar que em uma sentença pode existir mais de uma janela. Aplicando-se as informações acima, tem-se:

$$correlação = \sum_{1}^{n} \frac{1}{\left(\frac{1}{1} x 4\right)} = 0.25$$

O valor de 0,25 representa o grau de correlação entre as entidades "Flávio Ceci" e "Ciência da Computação". Esse valor será adicionado à intersecção dos valores na matriz de entidades.

Com a matriz gerada, as entidades mais relevantes e o seu grau de correlação são apresentados ao usuário. Esse grau de correlação pode ser bastante útil no processo de explicitação de conhecimento. Por meio da relação entre as entidades, é possível clarificar possíveis vínculos entre itens de classes distintas, como, por exemplo, de uma pessoa com uma organização ou área de conhecimento.

Na seção abaixo, são apresentados os dados dessa matriz no formato de rede para auxiliar na análise do conhecimento contido nos resumos.

4.3.3 Resultados do estudo de caso

Após submeter as instâncias encontradas pelo protótipo ao algoritmo de correlação apresentado na seção anterior, os dados foram transpostos para uma estrutura de redes de forma a facilitar a análise dos resultados encontrados.

Na Figura 26, pode-se verificar como todas as instâncias (nodos) se relacionam, bem como o grau de correlação (apresentado nas arestas) e a classificação entre parênteses.

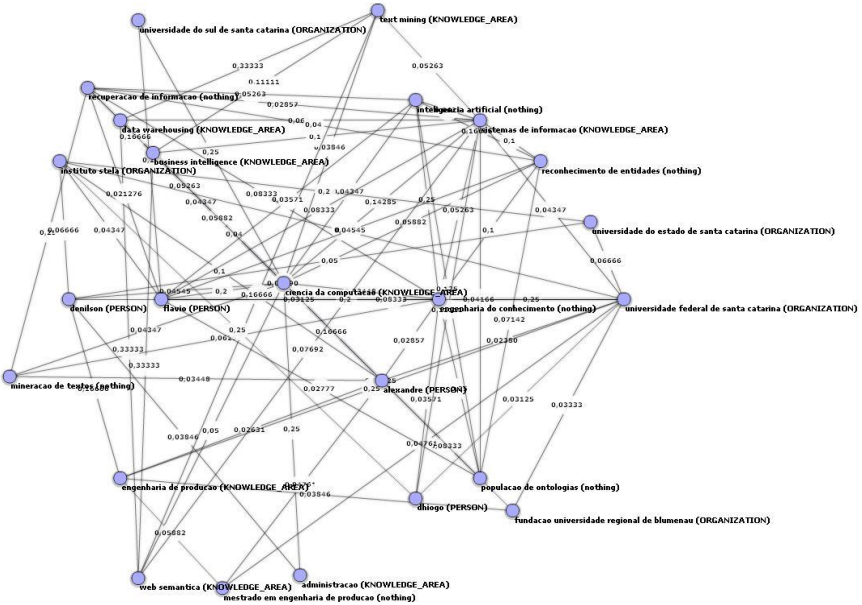


Figura 26 - Apresentação das entidades encontradas e seus relacionamentos

Mesmo com a representação dos resultados em forma de rede, pela quantidade de instâncias encontrada torna-se difícil uma análise mais detalhada. Por esse motivo, apresentam-se a seguir algumas imagens focadas em até duas instâncias.

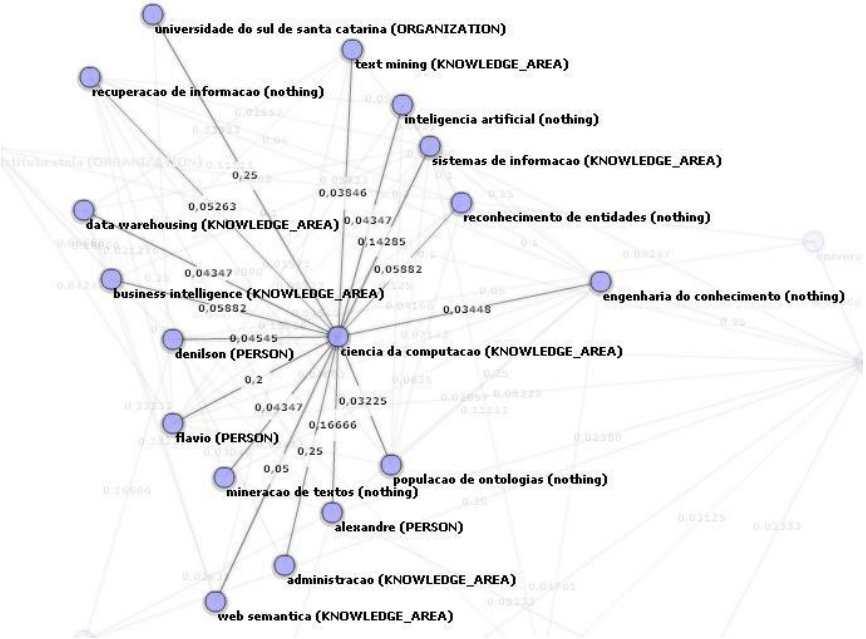


Figura 27 - Análise da instância “Ciência da computação”

A Figura 27 possibilita a realização de uma análise no termo “ciencia da computacao” e como ele está relacionado no domínio em questão.

A aplicação classificou a instância “Ciência da Computação” como “KNOWLEDGE_AREA”, ou seja, na classe *área do conhecimento*. Pode-se analisar ainda que as instâncias “flavio", “denilson" e “alexandre" estão ligadas a essa área, informação que poderia ser interpretada pelo especialista usuário da solução como se essas pessoas possuíssem graduação ou pós-graduação em Ciência da Computação.

A instância “ciencia da computacao” também está ligada a “sistemas de informacao" (também área de conhecimento), cursos cuja estrutura das disciplinas é bastante similar e possuem profissionais que competem pelas mesmas áreas de conhecimento, com formação análoga. Pode-se observar outras instâncias do tipo área do conhecimento que estão ligadas a “ciencia da computacao”, como, por exemplo, “recuperacao de informacao”, “mineracao de texto”, “web semantica”, entre outras.

A entidade “inteligencia artificial” também é uma área de conhecimento da Ciência da Computação, mas, como não existia referência a essa área na lista de termos classificados, ela acabou sendo reconhecida, porém não classificada.

Vale lembrar que essas relações só puderam ser identificadas dada a alta frequência com que esses termos se encontram nos documentos selecionados. Entretanto, dependendo da coleção de documentos e do domínio de análise, será necessário um processo incremental de manutenção de ontologia com múltiplas interações até que resultados satisfatórios sejam atingidos.

Após se fazer uma análise da instância “ciencia da computacao”, optou-se por analisar a instância “flavio” e verificar como ela está relacionada com as outras instâncias do domínio analisado, conforme apresentado na Figura 28.

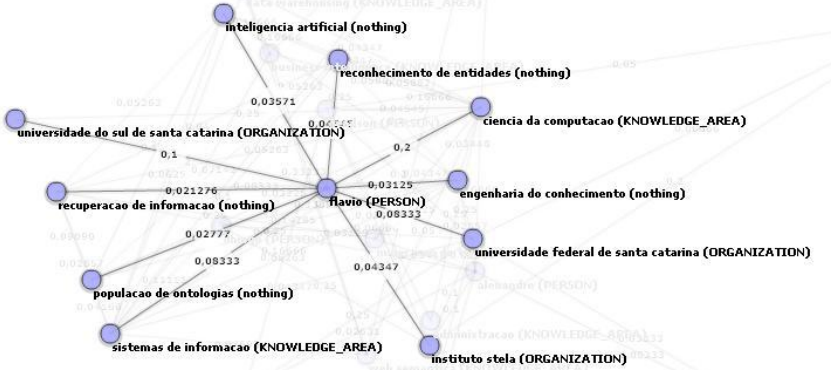


Figura 28 - Relacionamento entre a pessoa “flavio” e demais instâncias

A aplicação classificou a instância “flavio” como pessoa. Pode-se observar que o pesquisador possui relacionamentos com as instituições Universidade do Sul de Santa Catarina, Universidade Federal de Santa Catarina e Instituto Stela. Esses relacionamentos identificados devem ser nomeados pelo engenheiro de ontologias, visto que podem significar relacionamentos de vínculo profissional ou atuação acadêmica, por exemplo. Além disso, na Figura 28 é possível identificar, por meio dos relacionamentos explicitados, as áreas de conhecimento com as quais o pesquisador possui relação, como, por exemplo: “recuperacao de informação” e “reconhecimento de entidades”. Há também outras instâncias em que as classes não foram identificadas no processo de

reconhecimento de entidades. Por exemplo, a instância “populacao de ontologias”, apesar de ter relação com a instância “flavio", não possui uma classe definida. Quando isso ocorre, cabe ao usuário determinar qual a classe apropriada ou descartar a relação se não houver relevância.

A última análise foi feita entre as instâncias “denilson” e “alexandre”, que não possuem um relacionamento explícito, já que em nenhum momento são apresentadas as referências de uma instância a outra, ou seja, elas não coocorrem em um mesmo *résumé*.

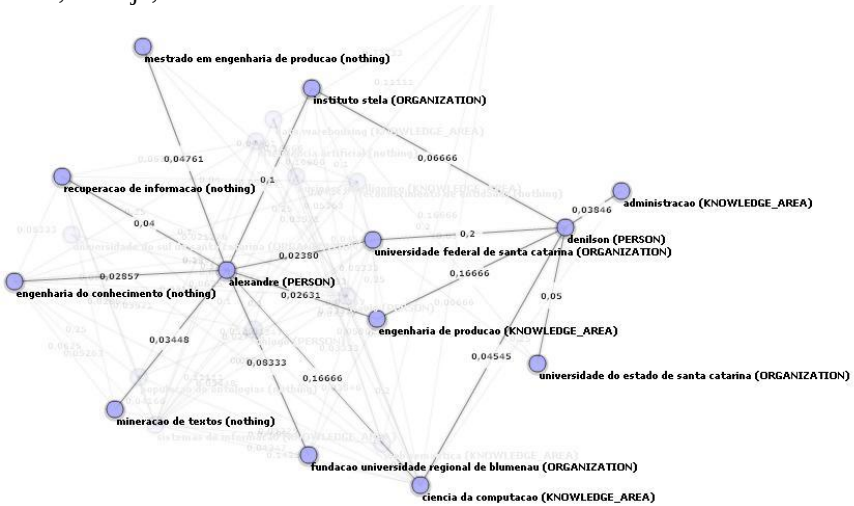


Figura 29 - Relacionamento entre as instâncias “denilson” e “alexandre”

No resultado apresentado na Figura 29, pode-se observar que as instâncias “denilson” e “alexandre” estão relacionadas indiretamente por meio de outras instâncias, como as organizações Instituto Stela e Universidade Federal de Santa Catarina, ou através das áreas de conhecimento, tais como Ciência da Computação e Engenharia de Produção. Um especialista poderia inferir que as duas pessoas trabalham ou trabalharam nas mesmas organizações e que possuem interesses de pesquisas similares ou formação nessas áreas. Informações desse tipo podem ser utilizadas para apoiar ações como a formação de equipes de trabalho em projetos, por exemplo.

As análises citadas acima são de suma importância para a manutenção ou a construção de ontologias. Depois que as instâncias já foram reconhecidas e classificadas, é possível identificar relações de pai e filho, entre outras informações, de maneira tal que seja possível

construir uma base de conhecimento que promova suporte às mais variadas aplicações de Gestão do Conhecimento.

4.3.4 Conclusão do estudo de caso

Este estudo de caso teve como objetivo a realização do processo tradicional de reconhecimento de entidades por meio da utilização de tabelas léxicas que identificam cada classe do domínio (também chamadas de base de conhecimento), assim como a aplicação de correlação, de modo que fosse possível projetar relacionamentos entre entidades na forma de rede.

Verificou-se que o uso de uma representação gráfica em forma de rede auxilia na explicitação do conhecimento contido nos *résumés* dos pesquisadores analisados. Também foi possível identificar relações indiretas por meio da análise das redes.

Este estudo de caso procurou demonstrar a viabilidade do modelo proposto e como os resultados gerados podem auxiliar tanto no processo de manutenção e construção de ontologias como na preparação dos dados para serem consumidos por sistemas baseados em conhecimento.

Os dados gerados a partir do estudo de caso foram salvos em OWL e visualizados em forma de árvore por uma ferramenta gráfica para manipulação de ontologias, como pode ser visto na Figura 30.

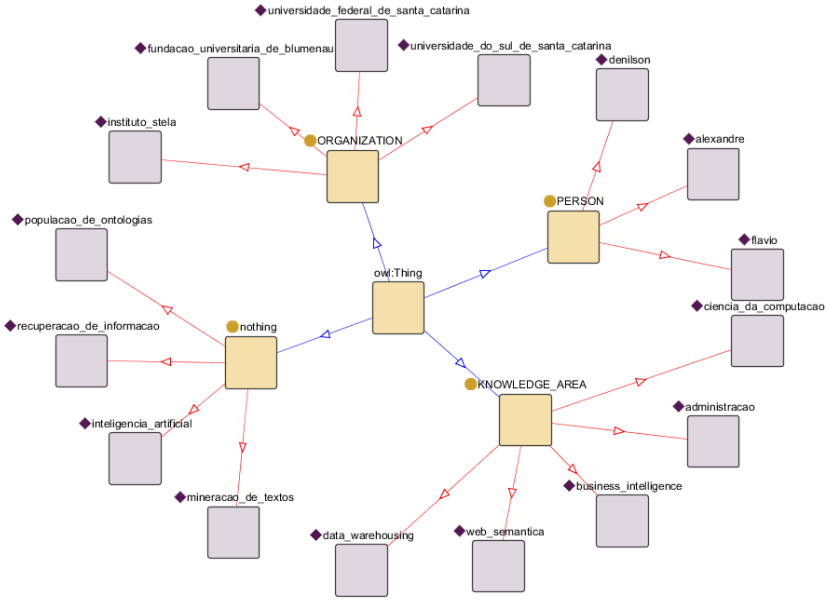


Figura 30 - Ontologia gerada a partir dos dados do estudo de caso

A próxima seção apresenta um estudo experimental que procura avaliar o modelo e o protótipo num contexto mais amplo.

4.3.5 Estudo experimental

A fim de avaliar a escalabilidade do modelo, o estudo de caso foi expandido para um número maior de *résumés*. Para tal, utilizou-se um grupo de currículos da Plataforma Lattes, os quais foram coletados a partir do Portal Inovação[7]. Especificamente, foram considerados os primeiros 1.000 currículos que mais citam a palavra "Biotecnologia". Para cada currículo, considerou-se o *résumé*, sendo este armazenado em um banco de dados, o qual constitui um *corpus* destinado à realização do estudo.

Para cada *résumé*, foram extraídas as entidades e as posições em que estas se encontravam no texto. Como resultado, obteve-se um vetor por *résumé*. O passo seguinte utiliza esses vetores para produzir uma

[7] Disponível em: <http://www.portalinovacao.mct.gov.br>. Acesso em: 11 nov. 2010.

matriz de correlação do tipo entidade-entidade. A partir da matriz de correlação, redes podem ser projetadas com base na escolha de uma entidade específica.

A Figura 31 apresenta as principais relações obtidas com base no termo "Biotecnologia". Entre esses termos, encontram-se a "Biologia", a "Biologia Molecular", a "Microbiologia", a "Bioquímica" e a "Genética". Vale lembrar que nessa representação foi utilizado, por questões de visualização, um máximo de cinco relacionamentos distintos em cada nível da rede e apenas dois níveis.

Caso se aumente o número de nodos por nível, outros importantes termos podem ser observados na rede, entre eles "Engenharia", "Medicina", "Química", "Biologia Celular" e "Ciência dos Alimentos". Tais termos são, portanto, instâncias na ontologia relacionadas à "Biotecnologia". Cada termo (instância na ontologia) está relacionado com vários outros termos. Desse modo, a projeção promove grafos densos com muitas conexões. A abordagem utilizada para a projeção de redes visa minimizar esse efeito em que, mesmo que um termo se conecte a vários outros em diferentes níveis, somente a conexão mais forte (maior correlação) é apresentada. Isso pode ser verificado para o termo "Química", que, apesar de estar relacionado com "Biotecnologia", possui maior conectividade com Bioquímica no segundo nível da rede.

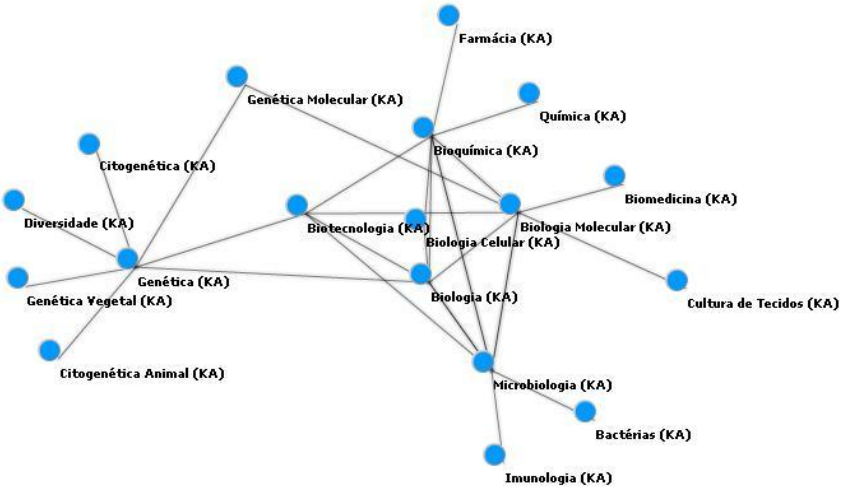


Figura 31 - Área de conhecimento (KA) relacionada à instância "Biotecnologia"

A Figura 32 apresenta uma rede que expande a projeção de "Biotecnologia" adicionando entidades do tipo *organização*. Entre as principais organizações extraídas, adicionadas à ontologia e que se relacionam ao termo (área) central, encontram-se USP, UFRGS, UFLA, UFBA e UFRJ. Por questões de visualização, somente são apresentados os nodos conectados diretamente à Biotecnologia (5 áreas e 5 organizações) e os nodos conectados em primeiro nível a essas 5 organizações. Tomando-se como exemplo a organização USP, é possível verificar, além do relacionamento com Biotecnologia, relacionamentos com Genética, Microbiologia, Bioquímica, Agronomia e Química. No caso da UFRGS, destacam-se Citogenética, Genética, Microbiologia, Biologia e Desenvolvimento de Medicamentos. Outra possibilidade de análise é a partir das áreas que conectam duas ou mais organizações. Esse é o caso, por exemplo, de "Genética", que possibilita a ligação indiretamente entre USP, UFRGS e UFLA.

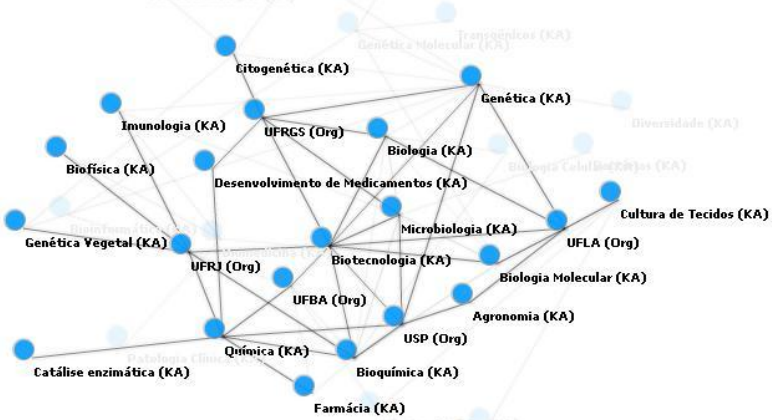


Figura 32 - Área de conhecimento (KA) e organizações (Org) relacionadas à instância "Biotecnologia"

Na próxima seção, são apresentados seis modelos/*frameworks* para manutenção de ontologias que utilizam como entrada uma fonte de dados textuais não estruturados. A partir dessa apresentação, é feita uma comparação com o modelo aqui proposto, levantando-se as suas vantagens e desvantagens.

4.4 Discussão da comparação entre os modelos/frameworks

Todos os seis modelos expostos nesta seção possuem pelo menos uma característica em comum entre eles. Para a fase de extração das entidades ou dos termos, os modelos não apresentaram grandes diferenças.

Para facilitar o entendimento, a seguir é relacionado um resumo dos modelos analisados, incluindo-se o modelo proposto neste trabalho. O resumo considera os seguintes critérios: a) extração de entidade; b) classificação das instâncias; c) aprendizagem; e d) ambiente para manutenção da ontologia:

a) Extração de entidade

Modelo/Framework	**Detalhamento**
OntoLancs	São utilizadas anotações semânticas para identificar termos.
OLea	Utiliza o dicionário WordNet e o conceito de relação semântica para localizar termos nos documentos.
OntoGen (OntoTermExtraction)	Extrai termos a partir dos agrupamentos de documentos usando o OntoGen. Como esse processo retorna apenas termos simples, é utilizada uma segunda ferramenta chamada TermExtractor para extrair termos compostos.
AVALON	Emprega técnica de reconhecimento de entidades nomeadas a partir de termos contidos nos *gazetteers* (tabelas léxicas).
OntoLearn	São utilizadas técnicas baseadas em análises estatísticas e de processamento de linguagem natural.
Text2Onto	Usa o *framework* GATE[8] para a extração das entidades da base textual.
Modelo proposto	Emprega a combinação da técnica clássica de reconhecimento de entidades nomeadas associada à

[8] Disponível em: <http://gate.ac.uk/>. Acesso em: 10 out. 2010.

	extração de rótulos a partir do processo de agrupamento (*clusterização*) de documentos de um determinado *corpus*.

b) Classificação de instância

Modelo/Framework	**Detalhamento**
OntoLancs	Ocorre por meio da utilização de tesauros de domínio combinados com técnicas de processamento de linguagem natural.
OLea	Identifica o termo ou seus sinônimos na taxonomia do WordNet e cria uma estrutura hierárquica de conceitos.
OntoGen (OntoTermExtraction)	A classificação dos termos encontrados é realizada utilizando-se o rótulo do agrupamento a que determinado termo pertence. O algoritmo KMeans é usado na fase de agrupamento.
AVALON	Cria hipóteses para classificar os termos encontrados e os valida por meio de informações disponíveis na web.
OntoLearn	Os termos encontrados são submetidos a buscas na web em sites previamente mapeados para identificação dos conceitos. A partir destes, é feita a classificação empregando-se expressões regulares. Para a retirada de ambiguidade, é utilizado o algoritmo SSI.
Text2Onto	Introduz-se o paradigma chamado modelo de ontologia probabilística (POM).
Modelo proposto	Antes do processo de classificação, ocorre a fase de validação das instâncias utilizando-se uma base de conhecimento colaborativa. As instâncias válidas são então classificadas por meio do processo

	clássico de nomeação de entidades, do contexto de determinada classe ou do uso de expressões regulares.

c) Aprendizagem

Modelo/Framework	**Detalhamento**
OntoLancs	Baseia-se no reúso da ontologia em que as informações descobertas fomentam o tesauro de domínio.
OLea	Caracterizado como um modelo supervisionado em que a validação do usuário gera insumo para as futuras iterações do *framework*.
OntoGen (OntoTermExtraction)	A aplicação permite a conexão com ferramentas de busca como o Google para possibilitar a descoberta de novos termos a partir do resultado da busca, fazendo com que a base cresça e evolua a partir dos novos termos armazenados.
AVALON	Identifica as mudanças nos termos via consultas à web e pode reconhecer evoluções e atualizações dos termos de maneira automática.
OntoLearn	Este trabalho usa bases de informação externas na web que possibilitam a evolução dos resultados encontrados. O processo apresentado é incremental e interativo, podendo sempre que utilizado permitir encontrar novos relacionamentos e instâncias.
Text2Onto	Verifica se com o passar do tempo ocorrem alterações no delta dos relacionamentos entre classes, instâncias e relações de modo a rever a necessidade de atualização.
Modelo proposto	Toda entidade reconhecida é submetida ao processo de validação, sendo armazenada como válida ou não. Essas informações são levadas em consideração em uma próxima iteração.

d) Ambiente para manutenção da ontologia

Modelo/Framework	Detalhamento
OntoLancs	Dados explicitados por meio do processo são disponibilizados para o engenheiro de ontologias via padrão OWL.
OLea	Não apresentado.
OntoGen (OntoTermExtraction)	Além da apresentação dos resultados em forma de árvore hiperbólica para auxiliar o engenheiro de ontologias, é disponibilizado um ambiente para fazer a manipulação da ontologia.
AVALON	Não apresentado.
OntoLearn	Não é apresentando um ambiente para manipulação de ontologias com a participação de um especialista. Contudo, os dados gerados podem ser adicionados à ontologia de maneira automática, permitindo a sua evolução. Pode-se utilizar uma ferramenta para manipulação de ontologias como o Protègè a partir do resultado gerado.
Text2Onto	Além de disponibilizar um mecanismo automático para identificar as mudanças no domínio para a atualização da ontologia, oferece também uma interface gráfica que permite manipular os resultados encontrados.
Modelo proposto	É disponibilizado através de um ambiente web para que o especialista possa interagir com os resultados obtidos em cada etapa do modelo. Possui como foco a classificação e o cadastro de novas classes ou instâncias, mas não promove suporte à edição do tipo de relacionamento entre elas.

De modo geral, pode-se verificar que, dos seis modelos analisados, os que foram propostos a partir de 2006 fazem uso de consultas à internet para melhorar os seus resultados. Verificando-se o critério de reconhecimento de entidades, todos os modelos analisados utilizam apenas uma técnica, a qual varia entre anotações semânticas em *corpus*, reconhecimento de entidades nomeadas *clássicas*, técnicas de agrupamento ou uso de tabelas léxicas. O modelo proposto emprega a combinação de duas técnicas: (1) extração de rótulos de agrupamento de textos a partir de uma busca, que demonstrou ser relevante para encontrar novos termos sem utilizar uma base previamente construída; e (2) técnica de reconhecimento de entidades nomeadas, que identifica e classifica novas entidades a partir da combinação dos termos contidos nos *gazetteers* (modelo clássico).

Pode-se citar a flexibilidade como vantagem dessa abordagem, uma vez que inicialmente se exige somente uma semente para gerar instâncias candidatas. Após o processo de validação e classificação, as instâncias armazenadas na ontologia retornam para formar/incrementar as tabelas léxicas (*gazetteers*) utilizadas no processo tradicional de extração de entidades.

Analisando-se o critério de classificação, verifica-se que é comum a utilização de outras ontologias ou bases de dados na web, com exceção do OntoTermExtraction, que utiliza a técnica de agrupamento para classificar os termos descobertos a partir do algoritmo KMeans. No modelo proposto, a fase de classificação é precedida por uma fase de validação que verifica quais das entidades reconhecidas são válidas para o domínio em questão ou não, realizando consultas a uma base de conhecimento colaborativa. Essa abordagem pode ser vantajosa, pois reduz o tempo de processamento da classificação (inclui-se aqui tanto o tempo da abordagem automática quanto o tempo de interação do especialista), uma vez que somente instâncias válidas serão classificadas. Por outro lado, a qualidade e a quantidade de instâncias válidas estão diretamente ligadas à abrangência/cobertura da base colaborativa.

A fase de classificação do modelo proposto baseia-se na lista de entidades marcadas como válidas e submete a descrição dessas entidades válidas (recuperadas da base de conhecimento colaborativa) aos algoritmos desenvolvidos, que se utilizam da análise de expressões regulares e da aderência do termo ou de sua definição perante o contexto das classes que pertencem à ontologia. Não é possível afirmar se essa abordagem é ou não mais eficiente em relação aos demais modelos

analisados, contudo nota-se que a combinação desses elementos tende a produzir resultados satisfatórios.

Analisando-se o critério de aprendizagem, percebe-se que os modelos utilizam desde abordagens mais simples, como no caso do OntoLancs, que se baseia apenas no reúso de uma ontologia, até abordagens mais elaboradas, como a proposta pelo TextToOnto, que gera um índice (delta) para possibilitar a verificação de mudanças e evoluções. O modelo proposto faz uso de uma abordagem interativa (necessita de um especialista), incremental e iterativa (o processo é realimentado pela própria ontologia). Como desvantagem, cita-se a dependência do especialista, uma vez que este tem importante papel na qualidade das informações. Entretanto, isso também pode ser constatado em outros modelos analisados.

Por fim, o último critério levantado foi o de manutenção, que verifica se os modelos provêem uma plataforma gráfica (interface de usuário) para visualização ou manipulação dos resultados encontrados. Verificou-se que dois modelos não suportam esse recurso: OntoLean e o AVALON. No caso do OntoLean, apesar de não possuir uma interface para o usuário interagir, ao final do processo permite que este gere uma ontologia que pode ser manipulada em qualquer editor OWL.

O modelo proposto disponibiliza um ambiente para que o usuário interaja com as fases desse modelo. É possível manipular elementos como classes e instâncias de uma ontologia por meio de uma interface mais simples baseada em listas e *combobox*. Também não promove suporte à inclusão de informações adicionais sobre o relacionamento entre instâncias e classes.

O melhor ambiente entre os modelos/*frameworks* avaliados é o do Text2Onto, que disponibiliza uma série de ferramentas para interagir com a ontologia de maneira gráfica, utilizando recursos do tipo arrasta e solta.

De maneira geral, o modelo proposto emprega a técnica de agrupamento em conjunto com a de reconhecimento de entidades nomeadas para a extração das entidades da base de documentos. Essa característica é entendida como um diferencial entre os modelos/*frameworkds* analisados.

Cita-se ainda como diferencial em relação aos demais modelos/*frameworks* a independência de linguagem no processo de extração, classificação e população da ontologia. Por fim, considera-se como um ponto relevante a utilização de bases de conhecimento colaborativas como meio de promover suporte às fases de validação e classificação, além de o sistema sugerir novas instâncias para a

ontologia. Desse modo, o processo em seu conjunto apresentado pelo modelo proposto neste trabalho é caracterizado como semiautomático, incremental e iterativo.

4.5 Considerações finais

Esta seção teve como objetivo mostrar a viabilidade do modelo proposto neste trabalho. Para atestar isso, primeiramente foi desenvolvido um protótipo dividido em duas partes: (1) *back-end*, que é responsável por todas as interações automáticas e computacionais; e (2) *front-end*, que permite ao usuário interagir com os resultados nas fases de validação e classificação.

Como esse protótipo foca na descoberta de novas instâncias (utilizando o processo de *clusterização*) e da população das ontologias, no estudo de caso optou-se pela utilização da técnica tradicional para o reconhecimento das entidades. Aplicou-se o resultado diretamente em um sistema baseado em conhecimento, atestando-se assim a qualidade dos dados recuperados. Com os resultados encontrados, foi possível explicitar visualmente relações diretas e indiretas entre as entidades extraídas.

Por fim, foram analisados seis trabalhos similares com foco na aprendizagem e na população das ontologias com o objetivo de demonstrar as contribuições promovidas pelo atual trabalho.

5 CONCLUSÕES

5.1 Conclusões

O presente trabalho apresentou um modelo para auxiliar no processo de extração e representação de conhecimento com vistas a apoiar a manutenção de ontologias a partir de bases de dados não estruturadas e que conta com a contribuição de bases de conhecimento colaborativas. Pressupõe-se ainda a participação de especialistas de domínio no processo de manutenção, caracterizando-se assim como um modelo semiautomático.

Para demonstrar a viabilidade do modelo proposto, desenvolveu-se um protótipo que foi aplicado em uma coleção de trabalhos de conclusão dos cursos de Ciência da Computação e Sistemas de Informação. Também foi elaborado um estudo de caso em que as entidades extraídas e validadas a partir do protótipo foram correlacionadas, objetivando-se demonstrar possíveis relacionamentos entre pessoas, organizações e áreas do conhecimento.

Destaca-se ainda a elaboração de uma análise comparativa com outros modelos correlatos visando destacar, segundo alguns critérios, as vantagens e as desvantagens da abordagem proposta por este trabalho.

Durante a elaboração do trabalho, observou-se que o emprego em conjunto das técnicas de reconhecimento de entidades nomeadas juntamente com a de identificação de rótulos de agrupamentos de documentos a partir de busca geram um bom resultado para a etapa de extração de entidades em bases não estruturadas.

A inclusão de uma fase de validação utilizando uma base de conteúdo colaborativa (neste caso a Wikipédia) possibilitou a separação de termos válidos e não válidos, promovendo, assim, informação relevante ao processo de manutenção de ontologias. Constatou-se também que a utilização de uma base colaborativa voltada à geração de insumos à fase de classificação contribui positivamente para o processo. Contudo, ressalta-se que ambas as fases necessitam da interação de um especialista para lapidar os resultados e com isso proporcionar um aprendizado por parte dos algoritmos. Desse modo, o processo em sua totalidade pode ser caracterizado como semiautomático, iterativo e incremental, uma vez que o refinamento da ontologia interfere em resultados futuros na fase de extração de entidades.

De maneira geral, este modelo possibilita a construção inicial de uma ontologia de domínio levando em consideração coleções de documentos (bases de dados não estruturadas) que auxiliam na etapa de

manutenção da ontologia. A projeção de entidades conectadas por meio do cálculo de coocorrências pode fornecer subsídios para a investigação de possíveis relacionamentos, mas para o estabelecimento dos tipos de relações entre as instâncias da ontologia torna-se necessária a intervenção de um engenheiro de ontologias.

5.2 Sugestões para trabalhos futuros

Como trabalhos futuros, vislumbram-se a evolução dos algoritmos de classificação das instâncias extraídas da base não estruturada. No modelo atual, é possível classificar as instâncias utilizando-se duas abordagens, sendo: (1) através de expressões regulares, e (2) a partir de análise estatística que considere os termos presentes na descrição de determinada entidade. A aplicação de técnicas de agrupamento ou até mesmo de análise semântica do contexto dos termos pode auxiliar nessa etapa.

Outro ponto com espaço para evolução é a construção de um ambiente voltado à visualização dos resultados gerados e persistidos na ontologia de maneira gráfica, de modo tal que pudesse ser um facilitador no entendimento e na manipulação dessa ontologia.

Por fim, pode-se identificar como trabalhos futuros a construção de mecanismos que auxiliem na definição dos tipos de relações entre as instâncias das classes da ontologia, analisando-se, por exemplo, o verbo que liga duas entidades quaisquer.

REFERÊNCIAS

AHMAD, Norashikin; ALAHAKOON, Damminda; CHAU, Rowena. Cluster identifications and separation in the growing self-organization map: application in protein sequence classification. **Neural Computing & Applications Journal**, London, 2009.

ALJABER B. et al. Document clustering of scientific texts using citation contexts. **Information Retrieval Journal**, Springer Science+Business Media, 2009.

ALMEIDA, Mauricio B. Roteiro para a construção de uma ontologia bibliográfica através de ferramentas automatizadas. **Perspect**. ciEnc. inf., Belo Horizonte, v. 8, n. 2, p. 164-179, 2003.

_____.; BAX, Marcello P. Uma visão geral sobre ontologias: pesquisa sobre definições, tipos, aplicações, métodos de avaliação e construção. **Ci. Inf**., Brasília, v. 32, n. 3, p. 7-20, 2003.

BAEZA-YATES, Ricardo A.; RIBEIRO-NETO, Berthier. **Modern information retrieval**. New York: ACM Press, 1999.

BOHN, Christian; NORVAG, Kjetil. Extraction named entities and synonyms from Wikipedia. In: IEEE INTERNATIONAL CONFERENCE, 24., 2010, Perth, Australia. **Proceedings…** Perth, Australia, 2010.

BRACHMAN, Ronald J.; LEVESQUE, Hector J. **Knowledge representation and reasoning**. Morgan Kaufmann Publishers, 2004.

CALHOUN, Mikelle A.; STARBUCK, William H. Barriers do creating knowledge. In: EASTERBY-SMITH, M.; LYLES, M. **Handbook of organizational learning and knowledge management**. Malden: Blackwell, 2005. p. 473-492.

CARDOSO, Olinda Nogueira Paes. Recuperação de informação. **Infocomp**, Lavras, v. 2, n. 1, p. 33-38, 2000.

CARPINETO, Claudio et al. A survey of web clustering engines. **ACM Computing Surveys**, New York, v. 41, n. 3, 2009.

CECI, Flávio et al. Towards a semi-automatic approach for ontology maintenance. In: CONTECSI INTERNATIONAL CONFERENCE ON INFORMATION SYSTEMS AND TECHNOLOGY MANAGEMENT, 7., 2010, São Paulo. **Anais…** São Paulo: USP, 2010.

CHAVES, Marcirio Silveira. **Uma metodologia para construção de ontologias e integração de conhecimento geográfico**. 2007. Proposta de Tese (Qualificação de Doutoramento) – Programa de Doutoramento em Informática da Universidade de Lisboa, Universidade de Lisboa, Portugal, 2007.

CIMIANO, Philipp; VOLKER, Johanna. Text2Onto: a framework for ontology learning and data-driven change discovery. In: INTERNATIONAL CONFERENCE ON APPLICATIONS OF NATURAL LANGUAGE TO INFORMATION SYSTEMS (NLDB), 10., 2005, Alicante, Spain. **Proceedings…** Alicante, Spain: Springer, 2005.

COUTINHO, Clara Pereira; JUNIO, João Batista Bottentuit. Blog e Wiki: os futuros professores e as ferramentas da web 2.0. In: SIMPÓSIO INTERNACIONAL DE INFORMÁTICA EDUCATIVA, 9., 2007, Portugal. **Actas…** Portugal, 2007. p. 199-204.

DAVENPORT, Thomas H.; PRUSAK, Laurence. **Working knowledge**: how organizations manage what they know. USA: Harvard Business School Press, 2000.

DAVIES, John; FENSEL, Dieter; VAN HARMELEN, Frank. **Towards the semantic web ontology-driven knowledge management**. England: John Wiley & Sons, 2003.

DRUMOND, Lucas; GIRARDI, Rosario. Extracting ontology concept hierarchies from text using markov logic. In: SYMPOSIUM ON APPLIED COMPUTING, 25., 2010, Switzerland. **Proceedings…** Switzerland, 2010.

EBECKEN, Nelson F. F.; LOPES, Maria Celia S.; COSTA, Myrian C. A. Mineração de texto. In: REZENDE, Solange O. (Coord.). **Sistemas inteligentes**: fundamentos e aplicações. São Paulo: Manole, 2005.

EL SAYED, Ahmad; HACID, Hakim. A hybrid approach for taxonomy learning from text. **COMPSTAT 2008**. p. 255-266, 2008.

FENSEL, Dieter. **Ontologies**: silver bullet for knowledge management and electronic commerce. Berlin: Springer-Verlag, 2001.

FERNEDA, Edberto. **Recuperação de informação**: análise sobre a contribuição da Ciência da Computação para a Ciência da Informação. 2003. Tese (Doutorado em Ciências da Comunicação) – Universidade de São Paulo, São Paulo, 2003.

FIALHO, Francisco Antônio Pereira et al. **Gestão do conhecimento e aprendizagem**: as estratégias competitivas da sociedade pós-industrial. Florianópolis: Visualbooks, 2006.

FINN, Aidan; KUSHMERICK, Nicholas. Multi-level boundary classification for information extraction. In: EUROPEAN CONFERENCE ON MACHINE LEARNING (ECML), 15., 2004, Pisa, Italy. **Proceedings…** Pisa, Italy, 2004. p. 111-122.

FORTUNA, Blaz; LAVRAC, Nada; VELARDI, Paola. Advancing topic ontology learning through term extraction. In: PACIFIC RIM INTERNATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE, 10., 2008, Heidelberg, Berlim. **Proceedings...** Heidelberg, Berlim: Springer-Verlag, 2008.

FREITAS, Frederico Luiz G. de. Ontologias e a web semântica. In: CONGRESSO DA SOCIEDADE BRASILEIRA DE COMPUTAÇÃO, 28., 2003, Campinas. **Anais...** Campinas: SBC, 2003. p. 1-52.

FREITAS, Jackeline S.; PEREIRA, Rachel C. Um modelo de representação fuzzy em um sistema de recuperação de informação. In: ENCONTRO NACIONAL DE INTELIGÊNCIA ARTIFICIAL, 5., 2005, Rio Grande do Sul. **Anais...** Rio Grande do Sul, 2005.

GACITUA, Ricardo; SAWYER, Pete; RAYSON, Paul. A flexible framework to experiment with ontology learning techniques. In: RESEARCH AND DEVELOPMENT IN INTELLIGENT SYSTEMS, 24., 2007, London . **Proceedings...** London: Springer, 2007. p. 153-166.

GARCIA, Ana Cristina B.; VAREJÃO, Flávio M.; FERRAZ, Inhaúma N. Aquisição de conhecimento. In: REZENDE, Solange O. (Coord.). **Sistemas inteligentes**: fundamentos e aplicações. São Paulo: Manole, 2005.

GIULIANO, Claudio. Fine-grained classification of named entity exploiting latent semantic kernels. In: CONFERENCE ON COMPUTATIONAL NATURAL LANGUAGE LEARNING (CoNLL), 13., 2009, Boulder, Colorado. **Proceedings…** Boulder, Colorado, 2009. p. 201-209.

GÓMEZ-PÉREZ, Asunción; FERNÁNDEZ-LÓPEZ, Mariano; CORCHO, Oscar. **Ontology Engineering – with examples from the areas of Knowledge Management, e-Commerce and the Semantic Web**. London: Springer-Verlag, 2004.

GONZALES, Marco; LIMA, Vera L. S. de. Recuperação de informação e processamento de linguagem natural. In: JORNADA DE MINI-CURSOS DE INTELIGÊNCIA ARTIFICIAL, 3., 2003, Campinas. **Anais...** Campinas: UNICAMP, 2003, p. 347-395.

GRANITZER, Michael et al. Automated ontology learning and validation using hypothesis testing. **Advances in Soft Computing**, Berlin, v. 43, p. 130-135, 2007.

GRUBER, T. R. Toward principles for the design of ontologies used for knowledge sharing. In: INTERNATIONAL WORKSHOP ON FORMAL ONTOLOGY, 1993, Padova, Italy. **Proceedings…** Padova, Italy, 1993.

GUARINO, Nicola. **Formal ontology and information systems**. FOIS´98. Amsterdam: IOS Press, 1998.

HATCHER, Erik; GOSPODNETIC, Otis. **Lucene in action**. Greenwich: Manning Publications, 2005.

KONCHADY, Manu. **Text mining application programming**. Massachusetts: Charles River Media, 2006.

KORFHAGE, Robert R. **Information Storage and Retrieval**. New York: John Wiley & Sons, Inc., 1997.

KOZAREVA, Zornitsa. Bootstrapping named entity recognition with automatically generated gazetteer lists. In: CONFERENCE OF THE EUROPEAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS (EACL), 11., 2006, Trento, Italy. **Proceedings…** Trento, Italy, April 2006.

KRUGLIANSKAS, Isak; TERRA, José Cláudio Cyrineu. **Gestão do conhecimento em pequenas e médias empresas**. Rio de Janeiro: Campus, 2003.

LASTRES, Helena; ALBAGLI, Sarita. **Informação e globalização na era do conhecimento**. Rio de Janeiro: Campus, 1999.

MAEDCHE, A.; STAAB, S. The text-to-onto ontology learning environment. In: INTERNATIONAL CONFERENCE ON CONCEPTUAL STRUCTURES, 8., 2000, Darmstadt, Germany. **Proceedings…** Darmstadt, Germany, 2000. p. 14-18.

MANNING, Christopher D.; RAGHAVAN, Prabhakar; SCHUTZE, Hinrich. **Introduction to information retrieval**. New York: Cambridge University Press, 2008.

_____.; SCHUTZE, Hinrich. **Foundations of Statistical Natural Language Processing**. Cambridge, Massachusetts: MIT Press, 1999.

MIHALCEA, Rada. Using Wikipedia for automatic word sense disambiguation. In: THE ANNUAL NORTH AMERICAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS (NAACL), 2007, Rochester, United States. **Proceedings…** Rochester, United States, 2007.

MORAIS, Edison Andrade Martins. O estado da arte no estudo das ontologias. In: SIMPÓSIO DE ESTUDOS E PESQUISAS: EDUCAÇÃO, CULTURA E PRODUÇÃO DO CONHECIMENTO DA FASAM, 2006, Goiânia. **Anais…** Goiânia, 2006.

NAVIGLI, Roberto; VELARDI, Paola. **Learning domain ontologies from document warehouses and dedicated web sites**. Universit`a di Roma “La Sapienza”, 2004.

NÉDELLEC, C.; NAZARENKO, A. Ontologies and information extraction. **LIPN Internal Report**, 2005.

NEGRI, Matteo; MAGNINI, Bernardo. Using Wordnet predicates for multilingual named entity recognition. In: GLOBAL WORDNET CONFERENCE, 2., 2004, Czech Republic. **Proceedings...** Czech Republic: Masaryk University, Brno, 2004. p. 169-174.

NONAKA, Ikujiro; TAKEUCHI, Hirotaka. **Criação de conhecimento na empresa**: como as empresas japonesas geram a dinâmica da inovação. Rio de Janeiro: Campus, 2003.

NOY, Natalya F.; MCGUINNESS, Deborah L. **Ontology development 101:** a guide to creating your first ontology. California, United States: Stanford University, 2001.

OSINSKI, Stanislaw. An algorithm for clustering of web search results. 2003. Tese (Doutorado em Ciência da Computação) – Poznan University of Technology, Poland. 2003.

_____.; WEISS, David. Conceptual clustering using Lingo algorithm: evaluation on open directory project data. In: INTERNATIONAL IIS: IIPWM´04 CONFERENCE, 2004, Zakopane, Poland. **Proceedings…** Zakopane, Poland, 2004. p. 369-378.

PARAMESWARAN, Aditya; GARCIA-MOLINA, Hector; RAJARAMAN, Anand. Towards the web of concepts: from large datasets. In: INTERNATIONAL CONFERENCE ON VERY LARGE DATA BASES (VLDB), 36., 2010, Singapore. **Proceedings…** Singapore, 2010. p. 566-577.

PEHCEVSKI, Jovan et al. Entity rankink in Wikipedia: utilizing categories, links and topic difficulty predication. **Information Retrieval Journal**, Springer Science+Business Media, 2010.

PINHEIRO, Carlos André Reis. **Inteligência analítica**: mineração de dados e descoberta de conhecimento. Rio de Janeiro: Ciência Moderna, 2008.

RAMALHO, Leiridiane; TSUNODA, Denise Fukumi. A construção colaborativa do conhecimento a partir do uso de ferramentas Wiki. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 8., 2007, Salvador. **Anais...** Salvador, 2007.

RUSSELL, Stuart; NORVIG, Peter. **Inteligência artificial**. São Paulo: Elsevier, 2004.

SALTON, Gerard. **Automatic information organization and retrieval**. New York: McGraw-Hill, 1968.

SCHREIBER, G. et al. **Knowledge engineering and management**: the commonKADS methodology. MIT Press: Cambridge, 2002.

SCULLEY, D. **Web-scale K-means clustering**. In: INTERNATIONAL CONFERENCE ON WORLD WIDE WEB, 19., 2010, Raleigh, North Carolina, USA. **Proceedings…** Raleigh, North Carolina, USA, 2010.

SEGARAN, Toby. **Programando a inteligência coletiva:** desenvolvendo aplicativos inteligentes Web 2.0. Rio de Janeiro: AltaBooks, 2008.

SILVA, E. L. da; MENEZES, E. M. **Metodologia da pesquisa e elaboração de dissertação**. 3. ed. rev. atual. Florianópolis: Laboratório de Ensino a Distância da UFSC, 2001.

SILVIA, Icléia; SPITZ, Rejane. A gestão do conhecimento organizacional e sua relação com a vantagem competitiva. Diseño en Palermo. In: ENCUENTRO LATINOAMERICANO DE DISEÑO, 1., 2006, Buenos Aires, Argentina. **Actas…** Buenos Aires, Argentina, 2006.

SOUZA, Luiz Cláudio Guarita. **Regras de raciocínio aplicadas a ontologias por meio de sistemas multiagente para a decisão organizacional**. 2003. Dissertação (Mestrado em Informática Aplicada) – Pontifícia Universidade Católica do Paraná, Curitiba, 2003.

STEIL, Andrea Valéria. **Competência e aprendizagem organizacional**. Como planejar programas de capacitação para que as competências individuais auxiliem a organização a aprender. Florianópolis: Stela, 2006.

_____. **Estado da arte das definições de gestão do conhecimento e seus subsistemas**. Florianópolis: Instituto Stela, 2007. Technical Report.

STUDER, Rudi; BENJAMINS, V. Richard; FENSEL, Dieter. Knowledge engineering**:** principles and methods. **IEEE Transactions on Data and Knowledge Engineering**, 1998.

SUN, Z; HAO, G. HSM: a hierarchical spiral model for knowledge management. In: INTERNATIONAL CONFERENCE ON INFORMATION MANAGEMENT AND BUSINESS (IMB2006), 2., 2006, Sydney, Australia. **Proceedings…** Sydney, Australia, 2006. p. 542-555.

TANG, Jie et al. Towards ontology learning from folksonomies. In: INTERNATIONAL JOINT CONFERENCE ON ARTIFICIAL INTELLIGENCE, 21., 2009, San Francisco, CA. **Proceedings…** San Francisco, CA: Morgan Kaufmann Publishers Inc., 2009.

VASCONCELOS, José Braga de; ROCHA, Álvoro; KIMBLE, Chris. Sistema de informação de memória organizacional: uma abordagem ontológica para a definição de competências de grupo. In: CONFERÊNCIA DA ASSOCIAÇÃO PORTUGUESA DE SISTEMAS DE INFORMAÇÃO, 4., 2003, Porto, Portugal. **Actas...** Porto, Portugal, 2003.

VELARDI, Paola et al. Evaluation of OntoLearn, a methodology for automatic learning of domain ontologies. In: BUITELAAR, Paul; CIMIANO. Philipp; MAGNINI, Bernardo (Eds.). **Ontology learning from text**: methods, applications and evaluation. Amsterdam: IOS Press, 2003.

WANG, Jianhua; LI, Ruixu. A new cluster merging algorithm of suffix tree clustering. In: SHI, Z.; SHIMOHARA, K.; FENG D. (Eds.). **Intelligent Information Processing III**. Boston: Springer, 2008. p. 197-203.

WANG, Pu et al**.** Using Wikipedia knowledge to improve text classification. **Knowledge and Information Systems Journal**, London, Spring 2008.

WEI, Zhang et al. A new method for clustering search result. **Wuhan University Journal of Natural Sciences**, 2008.

WEISS, Sholom M. et al. **Text mining predictive methods for analyzing unstructured information**. New York: Springer, 2005.

WILKS, Yorick; CATIZONE, Roberta. Can we make information extraction more adaptive? **LNAI**, Berlin, p. 1-16, 1999.

ZAMIR, Oren; ETZIONI, Oren. Web document clustering. In: ANNUAL INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 21., 1998, Melbourne, Australia. **Proceedings…** Melbourne, Australia, 1998.

ZESCH, Torsten; MULLER, Christof; GUREVYCH, Iryna. Extracting lexical semantic knowledge from Wikipedia and Wiktionary. In: CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION (LREC), 2008, Marrakech. **Proceedings…** Marrakech, 2008.

ZHANG, Yun-tao; GONG, Ling; WANG, Yong-cheng. An improved TF-IDF approach for text classification. **Journal of Zhejiang University Science**, China, 2005.

ZHU, Jianhan; GONÇALVES, Alexandre L.; UREN, Victoria. Adaptive named entity recognition for social network analysis and domain ontology maintenance. **Tech Report kmi-04-30**. Knowledge Media Institute, The Open University, UK, 2005.

GLOSSÁRIO

Agrupamento: é classificado como aprendizado não supervisionado que permite criar grupos de documentos a partir de uma busca.

Bases de conhecimento: são bases que armazenam o conhecimento de maneira estruturada e que servem de apoio a sistemas baseados em conhecimento.

Bases de conhecimento colaborativas: são bases que armazenam conhecimentos construídos colaborativamente por voluntários na Web, por exemplo, a Wikipédia (ZESCH; MULLER; GUREVYCH, 2008).

Bases de dados: são consideradas bases de dados estruturadas como, por exemplo, aquelas que utilizam o banco de dados relacionais de uma organização. São estruturas formais que facilitam o armazenamento e a recuperação dos dados.

Bases textuais: conjunto de documentos textuais cujo conteúdo não é estruturado, ou seja, está escrito de maneira livre sem marcações.

Conhecimento: é a combinação completa de informação, dados e relações que levam os indivíduos à tomada de decisão, ao desenvolvimento de novas informações ou conhecimentos e à realização de tarefas (FIALHO et al., 2006).
Clusterização: ver agrupamento.

Dado: são representações simbólicas para a descrição de atributos de qualquer nível. A camada de dados é responsável pela existência e pela operação dos sistemas transacionais, os quais têm como responsabilidade apoiar as operações de determinada organização (FIALHO et al., 2006; PINHEIRO, 2008).

Dados estruturados: dados que possuem uma estrutura formal que facilita a sua recuperação (por exemplo: arquivos XML e banco de dados relacionais).

Dados não estruturados: dados que não possuem uma estrutura formal para a sua recuperação. São dados escritos em linguagem natural e livre.

Engenharia do Conhecimento: promove o ferramental para sistematizar e apoiar processos da gestão que culminam na concepção de sistemas de conhecimento (SCHREIBER et al., 2002).

Entidades: termos simples ou compostos que foram submetidos a um processo prévio de validação e classificação. As entidades podem ser nome de pessoas, organizações, áreas de conhecimento, lugares, entre outros.

Especialista no domínio: indivíduo que conhece profundamente os processos e as características do ambiente do problema em questão.

Extração de informação: é uma área-filha da área de processamento de linguagem natural e tem como foco identificar informações importantes em bases textuais. A extração de informação é um processo de identificação de itens relevantes em documentos textuais que verifica as "fronteiras" entre os termos que formam as entidades (FINN; KUSHMERICK, 2004).

Gestão do Conhecimento: é um conjunto de processos que auxiliam na criação, aquisição, representação, no armazenamento, na manipulação, distribuição e utilização do conhecimento (KRUGLIANSKAS; TERRA, 2003).

Índice (ou índice textual): estrutura de dados que relaciona termos a documentos, visando facilitar a recuperação de informação.

Índice invertido: é uma estrutura de dados em que se relaciona cada palavra com todos os documentos que a contêm, armazenando ainda a frequência com que a palavra ocorre no documento, de modo tal que permita a ordenação de documentos pela sua relevância em relação a determinada consulta de usuário.

Informação: é o conjunto de dados que são devidamente processados e tornam-se compreensíveis, ou seja, a informação é a disposição dos dados de uma forma que apresentem um significado, criando padrões e acionando significados na mente dos indivíduos (FIALHO et al., 2006).

Lista de termos (tabelas léxicas ou *gazetteers*): são listas de termos válidos e relacionados a uma determinada classe, sendo que cada classe possui a sua lista.

Modelo booleano: é baseado nos conceitos da lógica ou álgebra booleana. Tais conceitos utilizam os operadores lógicos E, OU e NÃO para refinar as consultas (KORFHAGE, 1997).

Modelo vetorial: representa cada documento como um vetor ou uma lista ordenada de termos e um peso. Esse peso pode ser considerado como o grau de importância aplicado num espaço euclidiano de *n* dimensões, em que *n* é o número de termos (KORFHAGE, 1997).

Ontologias: é uma especificação formal e explícita de troca de conceitos e relações que existem em um domínio e que são compartilhados por uma comunidade (STUDER; BENJAMINS; FENSEL, 1998).

Reconhecimento de entidades: processo que identifica (reconhece) termos simples ou compostos (por exemplo: Universidade Federal de Santa Catarina) em meio a documentos não estruturados. É uma técnica da área de extração de informação (EI) que tem como função reconhecer entidades em textos de diferentes tipos e de diferentes domínios (ZHU; GONÇALVES; UREN, 2005).

Recuperação de informação: tem como objetivo identificar documentos a partir da busca de um ou mais termos. A recuperação de informação preocupa-se com os principais processos para lidar com a informação, como, por exemplo, estrutura, análise, organização, representação, recuperação e busca das informações (SALTON, 1968).