Towards a Semi-Automatic Approach for Ontology Maintenance

Flávio Ceci (Universidade Federal de Santa Catarina, SC, Brasil) – flavio@stela.org.br
Dhiogo Cardoso da Silva (Universidade Federal de Santa Catarina, SC, Brasil) – dhiogo@stela.org.br
Denilson Sell (Universidade Federal de Santa Catarina, SC, Brasil) – denilson@stela.org.br
Alexandre Leopoldo Gonçalves (Universidade Federal de Santa Catarina, SC, Brasil) – a.l.goncalves@stela.org.br

Abstract: Due to the large volume of documents that retain valuable and implicit employees' knowledge about business activities, it becomes imperative for organizations to have means to extract and organize that knowledge in order to support knowledge management initiatives. This paper presents a strategy based on the recognition and co-relation extraction of named entities from textual documents. The extracted entities and relations are used for ontology creation and maintenance in a semi-automatic approach. A prototype is presented in order to demonstrate the applicability of the proposed strategy. A case study is described in which direct and indirect relations are extracted from academic and professional activities registered in a résumé database.

Keywords: Ontology Maintenance, Named Entity Recognition, Knowledge Engineering, Knowledge Management.

1. Introdução

As informações produzidas nas organizações nos mais variados meios e formatos podem ser insumos ao processo de gestão e tomada de decisão. Frequentemente, milhares de e-mails são trocados, conteúdos textuais variados são publicados na web e em intranets, atas de reuniões e documentos de projetos são criados. Por conta disso, diversas práticas da Gestão do Conhecimento como a Gestão por Competências (PRAHALAD; HAMEL, 1996; STAAB et al., 2001) e a Inteligência Competitiva (DISHMAN et al., 1999, 2003; KAHANER, 1996) demandam estratégias adequadas para coletar e aproveitar o conhecimento das fontes de dados textuais da organização. A dificuldade de gerir a grande quantidade de dados, dispostos interna ou externamente ou ainda, de maneira estruturada ou não estruturada, demonstra a premência do estudo, o desenvolvimento e a combinação de técnicas objetivando facilitar a busca de informações e a obtenção de conhecimento. Considerando esse cenário, a gerência organizacional possui grandes desafios relacionados à obtenção, ao uso e à gestão do conhecimento e, dessa forma, o papel da Engenharia do Conhecimento como meio auxiliador faz-se necessário.

A Engenharia do Conhecimento nasceu de um ramo da inteligência artificial e, como disciplina, estuda técnicas e métodos para a extração, a manipulação e a classificação do conhecimento, dando suporte à construção de sistemas de conhecimento para atender às demandas da Gestão do Conhecimento (STUDER;

BENJAMINS; FENSEL, 1998). Os sistemas baseados em conhecimento podem ser úteis para auxiliar a coleta de dados e informações a fim de inferir e explicitar o conhecimento dos mais variados tipos de fontes de dados das organizações. Dessa forma, a Engenharia do Conhecimento, na qual preconiza que o conhecimento pode ser explicitado, representado e codificado, tem emergido junto às demandas da gestão organizacional.

Um grande problema enfrentado pela Engenharia do Conhecimento refere-se a como codificar e representar o conhecimento contido nos repositórios de dados e documentos das organizações, de forma que ele possa ser armazenado, disseminado e inserido no processo de gestão. Ontologias têm sido aplicadas em abordagens automáticas ou semiautomáticas de extração do conhecimento com vistas a apoiar ações da Gestão do Conhecimento. O desafio para que os sistemas computacionais sejam capazes de raciocinar e colaborar em conjunto com as pessoas é uma das principais motivações para o desenvolvimento de ontologias (BERNEERS-LEE et al., 2001). As ontologias têm como função a representação formal do conhecimento consensuado e utilizado por um grupo de pessoas ou organização (GRUBER, 1998).

Como o conhecimento normalmente não é estático, as ontologias precisam de constantes atualizações para estar de acordo com o conhecimento que representam. Logo, elas também sofrem manutenções ao longo de seu ciclo de vida. Dessa forma, o processo de evolução de classes, relacionamentos e instâncias da ontologia deve ser revisado à medida que o conhecimento muda e o repositório de dados da organização cresce. Dado que esse processo de manutenção e evolução de ontologias pode ser custoso, torna-se indispensável o uso de ferramentas e técnicas apropriadas.

Este trabalho tem como objetivo apresentar uma abordagem para auxiliar o processo de manutenção de ontologias com base em fontes de dados não estruturadas, fontes estas que registram boa parte do conhecimento implícito e histórico das organizações. Nessa proposta, instâncias das classes da ontologia bem como os relacionamentos com as demais instâncias podem ser identificados a partir de documentos textuais. A metodologia de trabalho é baseada nos métodos e técnicas das áreas de Extração de Informação e Descoberta de Conhecimento em Bases Textuais. Para tal, uma amostra dos *résumés* curriculares da Plataforma Lattes dos pesquisadores vinculados à UFSC foi selecionada como fonte de dados para um estudo de caso ilustrativo. Um protótipo foi desenvolvido baseado em técnicas de Mineração de Textos, visando demonstrar como um conjunto de instâncias e classes de uma ontologia e seus possíveis relacionamentos podem ser identificados e obtidos a partir de textos. O *framework* de reconhecimento de entidades BALIE (Baseline Information Extraction) foi utilizado na implementação do protótipo. Como resultado, essa pesquisa obteve um conjunto de redes de relacionamentos entre acadêmicos da Plataforma Lattes, demonstrando as intersecções entre as áreas de conhecimento e as organizações em que estes atuam.

O trabalho está organizado em seis seções principais. Após a introdução, a segunda seção trata das técnicas clássicas da literatura relacionadas à Descoberta de Conhecimento em bases textuais, com enfoque às técnicas associadas aos objetivos deste trabalho. A terceira seção descreve o uso de ontologias para a Engenharia do Conhecimento e relata seus benefícios e a importância de sua manutenção. No quarto tópico, o *framework* BALIE utilizado

no desenvolvimento do protótipo é apresentado juntamente com as características do processo de reconhecimento de entidades para a manutenção de ontologias. Na quinta seção, descreve-se a solução proposta e a sua validação através de um estudo de caso para a manutenção de uma ontologia relacionada ao contexto acadêmico, em que um conjunto de *résumés* curriculares da Plataforma Lattes da UFSC foi utilizado como fonte de dados. Por fim, na última seção, são apresentadas as conclusões e as discussões sobre trabalhos futuros

2. Processos de Descoberta de Conhecimento em Textos

A manutenção de ontologias pode ser facilitada pelo uso de um conjunto de técnicas de Descoberta de Conhecimento em Textos (ou em inglês *KDT – Knowledge Discovery in Text*). De acordo com Feldman e Hirsh (1997), o processo KDT pode ser definido como a extração não trivial de informações implícitas, previamente desconhecidas e potencialmente úteis de grandes bases textuais. Nahm e Mooney (2002) declaram que é um processo para encontrar padrões interessantes e úteis, modelos, direções, tendências ou regras a partir de textos não estruturados. Ambas as definições resumidamente fazem menção ao processo tradicional de Descoberta de Conhecimento (*KDD – Knowledge Discovery in Databases*), porém aplicado em textos.

As etapas do processo tradicional de descoberta de conhecimento em bases textuais são apresentadas abaixo.

2.1. Etapas do processo Descoberta de Conhecimento

As etapas do processo KDT, bem como as etapas do processo KDD, são interativas, pois requerem análises por meio de um especialista, que estabelece a relevância do resultado de cada subetapa, e também análises iterativas, porque a seleção dos dados e a obtenção de um modelo de representação do problema nem sempre são imediatas, e pode o resultado de uma etapa não ser satisfatório, sendo necessário retornar a etapas anteriores. A Figura 1 abaixo mostra as etapas de todo o processo, desde o texto puro até a interpretação dos resultados e o conhecimento obtido.

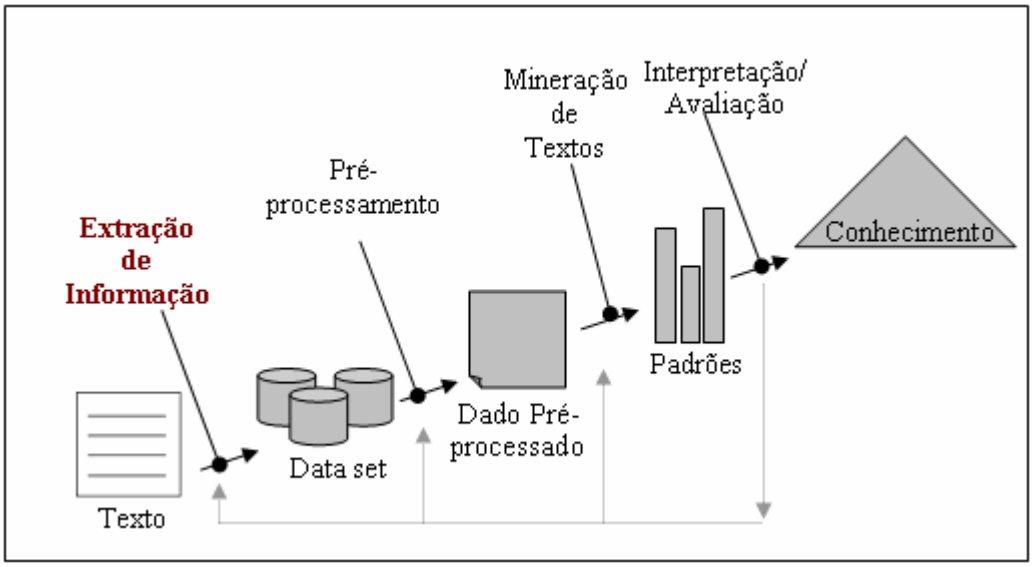

Figura 1 - Etapas do processo KDT
(Fonte: adaptado de MOONEY; NAHM, 2005)

Dados os objetivos que se deseja alcançar com o processo, o primeiro passo é eleger o conjunto de textos que será utilizado. A partir desse conjunto de documentos, inicia-se o processo de pré-processamento dos dados. O propósito do pré-processamento é eliminação de ruídos, termos não relevantes (stop-words), redução das palavras aos seus radicais (stemming), correções ortográficas e outros aspectos morfológicos e também sintáticos que as expressões textuais possuem. Após o pré-processamento, a etapa de transformação do texto é iniciada. Nessa etapa, ocorre a normalização do texto e sua transformação e representação no formato de vetor, tabela, matrizes, etc. As próximas etapas são a seleção e a projeção dos dados, em que há uma redução da dimensionalidade do modelo criado no passo anterior, e a escolha das palavras relevantes. Os textos têm a característica de possuírem alta dimensionalidade, visto que cada palavra pode ser uma dimensão do vetor ou matriz. Portanto, reduzir a dimensionalidade é importante para que o resultado seja encontrado com maior eficiência e desempenho. Dado o modelo estabelecido, as técnicas de KDT podem ser usadas no passo seguinte, com a escolha do algoritmo. Conforme o objetivo do problema, existem muitos métodos de descoberta de padrões em textos, com aprendizado supervisionado e não supervisionado, analogamente aos métodos de descoberta em banco de dados. Por fim, o último passo do processo KDT constitui a interpretação dos resultados obtidos e a obtenção do conhecimento.

Há duas grandes áreas relacionadas à Descoberta de Conhecimento em Textos: (1) Recuperação de Informação e (2) Extração de Informação.

2.2. Extração e Recuperação de Informação

A Recuperação de Informação, consoante Korfhage (1997), é a área que estuda as atividades e os aspectos de descrição de informação e sua especificação para busca, além de qualquer técnica, sistema ou máquina empregada para realizar ou auxiliar essas tarefas. Sua finalidade é ajudar na localização dos documentos relevantes, dado um conjunto de documentos. Os sistemas de Recuperação de Informação têm como entrada a consulta realizada pelo usuário do sistema, e então realizam uma pesquisa inter-documentos para obter os mais correlacionados com os termos da consulta. Já a Extração de Informação é a área que tem como objetivo a obtenção ou a extração de fatos relevantes pertencentes a determinados documentos (RILOFF; LEHNERT, 1994). A Extração de Informação tem objetivos mais específicos, e as pesquisas são executas intra-documento para encontrar e classificar elementos textuais (GRISHMAN, 2007).

O foco do presente trabalho está na área de Extração de Informação, a qual será apresentada em mais detalhes a seguir.

A Extração de Informação pode ser dividida em outras cinco subáreas: (1) Extração de Terminologias; (2) Reconhecimento de entidades; (3) Resolução de correferências; (4) Construção de elementos e relacionamentos; e (5) Extração de modelos de cenários. Entre as cinco áreas citadas, a área Reconhecimento de

Entidades está associada ao foco da pesquisa. Logo abaixo, cada área é explanada.

2.2.1. Extração de terminologias

A extração de termos compreende um dos primeiros passos para se obter uma extração de informação adequada. Essa subárea da Extração de Informação é a base para as demais, pois abrange diversas tarefas:

- Parsing: leitura das fontes de dados textuais, que podem estar em diferentes formatos, como arquivos CSV, XML, etc.;
- Part-of-speech tagging (POS-tagging): consiste em extrair as classes gramáticas dos termos de uma sentença (substantivo, verbos, etc.);
- Eliminação de stop-words: retira do texto caracteres com muitas ocorrências, como preposições, artigos, etc. e que não são úteis para as tarefas de extração e recuperação de informação;
- Técnicas de stemming: consiste em reduzir as palavras ao seu radical, útil para diminuir a dimensionalidade do documento; e
- Morfologia e sintaxe do idioma: o idioma torna sensível o desenvolvimento de sistemas de extração em virtude das especificidades de cada idioma.

2.2.2. Reconhecimento de entidades

Para Negri e Magnini (2004), o reconhecimento de entidades aplica-se para identificar e classificar os elementos textuais em categorias predeterminadas, tais como pessoas, organizações, locais, valores monetários, datas, etc.

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

Há muitas técnicas automáticas que podem ser utilizadas para realizar tal função, tais como: a aplicação de expressões regulares (muitas usadas para identificar datas, e-mails, URI, nomes seguidos de abreviações, etc.); o uso de dicionários (thesaurus); os modelos estatísticos; as heurísticas, regras conforme o padrão léxico e sintático do idioma; e também o uso de ontologias.

Segundo Kozareva (2006), a tarefa de identificar as entidades consiste em determinar as fronteiras das entidades, ou seja, qual o seu início e seu fim. Isso é importante para entidades compostas por mais de uma palavra, como, por exemplo, “Universidade Federal de Santa Catarina”. Outro problema está ligado à resolução de ambiguidades, pois há entidades que, dependendo do contexto, podem pertencer a mais de uma classe.

2.2.3. Resolução de correferências

Após o reconhecimento de entidades, a próxima área de estudo da Extração de Informação é a identificação e a resolução de correferências (*Coreference Resolution* - COR). Correferências são as terminologias que ocorrem duas ou mais vezes no texto e que representam a mesma entidade no mundo (SOON et al., 2001). Por exemplo, dada a frase: “O presidente viaja para

Florianópolis. Lula ficará por dois dias na ilha". Os termos "Florianópolis" e "ilha" nessa sentença representam a mesma entidade, assim como os termos "presidente" e "Lula". Algumas medidas da Recuperação de Informação podem ser úteis para o desenvolvimento dos sistemas de resolução de correferências, citando-se como exemplo as análises de coocorrência em determinados contextos do texto, a distância entre as entidades e a sua força de relacionamento.

2.2.4. Construção de elementos e relacionamentos

Um dos avanços das áreas NER e COR é a construção ou produção de elementos e relacionamentos entre as entidades do texto (*Template Element and Relation Construction* – TERC). Essa tarefa pode adicionar informação aos resultados da tarefa NER usando também as tarefas da área COR. Além de localizar e classificar uma entidade (pessoa, organização, data, etc.), associa informações descritivas às entidades. Por exemplo, levando-se em conta a mesma frase do exemplo da seção anterior, poder-se-ia ter como resultado da construção de elementos e relacionamentos a seguinte sentença: O presidente viaja para a *cidade de* Florianópolis. Lula ficará por dois dias na ilha *de Santa Catarina*. Note que os termos "*cidade de"* e "de Santa Catarina" foram acrescentados ao texto original. Portanto, esse tipo de construção requer dos sistemas de conhecimento a capacidade de raciocínio e um maior poder de representatividade da informação.

2.2.5. Produção de modelos de cenários

A produção de modelos de cenários (*Scenario Template Production*) é a última e mais complexa área de estudo da Extração de Informação descrita neste trabalho. A extração de modelos de cenários utiliza os resultados de NER e TERC para encontrar acontecimentos, fatos e eventos sobre as entidades reconhecidas. Novas informações são acrescentadas, associadas ou inferidas com o objetivo de dar maior riqueza às entidades-fim.

No exemplo da seção anterior, poder-se-ia extrair modelos de cenários para a entidade "Florianópolis" e varrer o documento a fim de achar outros dados relevantes sobre essa entidade, tais como dados da sua geografia, habitantes e clima.

Considerando-se a riqueza de informações que se pode obter das fontes de dados da organização pelo uso das técnicas de Extração de Informação supracitadas, há muitas possibilidades de aplicação na Engenharia do Conhecimento. Entre as aplicações possíveis, foco deste trabalho, cita-se a criação e a manutenção de ontologias para auxiliar a organização a representar seu conhecimento. As seções a seguir descrevem a importância do uso de ontologias para a Engenharia do Conhecimento bem como o seu processo de manutenção.

3. Uso e manutenção de ontologias

Ontologias são especificações formais de alto nível de um domínio de conhecimento (GRUBER, 1993). Uma ontologia define as regras que guiam a

combinação entre os termos e as relações em um domínio do conhecimento, sendo essas geralmente desenvolvidas por especialistas. De acordo com Pérez e Benjamins (1999), as ontologias constituem-se em:

- 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 – utilizados para modelar sentenças verdadeiras;
- instâncias – representam elementos específicos, os próprios dados das ontologias (geralmente ligados a uma classe, como instância de uma classe).
- Funções – eventos que podem ocorrer no contexto da ontologia

Uma ontologia é constituída por construtos semelhantes aos do paradigma de desenvolvimento de software orientado a objeto. As ontologias descrevem conceitos sobre um dado domínio através de classes e subclasses, as propriedades dos conceitos são representadas por meio dos atributos (slots), as restrições sobre as propriedades são demonstradas através dos tipos (cardinalidade) e as relações entre os conceitos através das igualdades e disjunções (NOY et al, 2001).

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

3.1. Benefícios de ontologias para a Engenharia do Conhecimento

Sabe-se que as ontologias no contexto da Engenharia do Conhecimento têm, entre outros objetivos, a representação do conhecimento, a qual pode auxiliar a construção de aplicações e de sistemas de conhecimento. Sobre os benefícios do uso das ontologias, Freitas (2003) faz as seguintes afirmações:

- a possibilidade de reusar as ontologias e as bases de conhecimento pelos desenvolvedores mesmo com adaptações e extensões. Já que a fase de construção de bases de conhecimento é a etapa mais custosa e demorada;
- a possibilidade de tradução entre diferentes linguagens e formalismos de representação de conhecimento;
- vasta quantidade de ontologias disponíveis em bases de conhecimento na web para reúso e possibilitando um vocabulário uniforme.
- Mapeamento entre formalismo de representação de conhecimento inspirado no componente de conectividade para sistemas gerenciadores de banco de dados.

As ontologias auxiliam em várias áreas do conhecimento. Algumas dessas áreas são vistas abaixo, conforme declara Morais (2006):

- Recuperação de informação – permite a reutilização de ontologias na web semântica, provendo estrutura de buscas em banco de dados de ontologias e outros documentos semânticos na internet;
- Processamento de linguagem natural – pode auxiliar em processo 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 – ontologias são aplicadas para agregar expressividade e sentido semântico sobre conteúdos e serviços na web.

3.2. Manutenção de ontologias

O conhecimento não é estático, ele pode evoluir ou mesmo tornar-se obsoleto com o tempo. Por esse motivo, as ontologias necessitam de constantes atualizações, objetivo do processo de manutenção de ontologias.

Para Navigli e Velardi (2004), o processo de criação das ontologias é algo que demanda tempo e envolve especialistas de vários campos. Com o processo de manutenção, não é diferente. Uma das grandes dificuldades para a manutenção é levantar as novas classes ou subclasses bem como os novos relacionamentos entre elas. Além disso, esse processo também trata da identificação de instâncias dessas classes e quais estão relacionadas entre si.

Este trabalho propõe uma solução que utiliza a técnica oriunda da área de Extração de Informação denominada Reconhecimento de Entidades Nomeadas, já explanada anteriormente. No contexto deste trabalho, os novos relacionamentos entre as classes podem ser extraídos a partir de documentos textuais não estruturados por meio de um protótipo desenvolvido.

As seções adiante mostram como o protótipo foi desenvolvido juntamente com o framework de reconhecimento de entidades chamado BALIE.

4. Reconhecimento de Entidades com BALIE

Objetivando a realização do processo de reconhecimento de entidades, utilizou-se neste trabalho a ferramenta BALIE (*Baseline Information Extraction*) versão 1.8 desenvolvida em Java. De acordo com Nadeau (2005), BALIE consiste em um sistema de extração de informação textual multilíngue. Com ele, é possível extrair e classificar os elementos textuais para qualquer idioma. Os idiomas suportados pela versão 1.8 são: inglês, romeno, francês, espanhol, italiano e alemão. Visto que não há suporte para o idioma português (do Brasil), uma extensão teve de ser desenvolvida. Para compreender as modificações realizadas, adiante é descrito como o BALIE está organizado e como funciona.

BALIE possui dois módulos principais. O primeiro módulo é utilizado para a criação de *Gazetters* (dicionário ou listas com os termos pertencentes a cada classe). Já o segundo módulo, usa heurísticas simples para identificar e classificar as entidades conforme o contexto, ou seja, resolução de ambiguidades de entidades. Esse último módulo utiliza os algoritmos de classificação da biblioteca de algoritmos de mineração de dados denominada WEKA (WITTEN; FRANK, 2005).

4.1. Geração de Gazetters

A geração automática do dicionário de entidades nomeadas (gazetters) é o primeiro passo para o reconhecimento de entidades. Nadeau propõe uma forma

de gerar *gazetters* por meio de um algoritmo de recuperação de informação na web, ou seja, um WebCrawler.

Toda geração de gazetters baseia-se no retorno de páginas web por meio de quatro palavras usadas como semente da busca. Por exemplo, dadas quatro cidades – Florianópolis, Curitiba, Porto Alegre e São Paulo –, a ideia é buscar diversas páginas com nomes de cidades para que novas sejam recuperadas (NADEU, 2005).

Com as entidades recuperadas e classificadas em cada lista separada, o aplicativo BALIE localiza e compara cada termo (token) de um determinado texto nesse dicionário. Trata-se de uma estratégia simples e bastante utilizada por sistemas de reconhecimento de entidades, porém apresenta diversos problemas, tais como ambiguidade entre entidades. Por exemplo, o termo 'manga', dependendo do contexto, pode ter o sentido de uma fruta ou uma parte da camisa. Esses problemas de ambiguidade são tratados pelo segundo módulo da ferramenta e estão descritos na seção seguinte.

4.2. Reconhecimento de entidades e resolução de ambiguidades

O reconhecimento de entidades realizado com a estratégia de busca em dicionários oferece três problemas de ambiguidade, conforme declara Nadeau (2005), sendo: 1) erro de ambiguidade entre substantivos e entidades; 2) erro de detecção de limite de entidades; e 3) erro de ambiguidade entre entidades.

O erro de ambiguidade entre substantivos e entidades ocorre quando ambos são homógrafos. Assim, o termo "Jobs" no idioma inglês pode ora referenciar uma entidade da classe Pessoa (sendo usado como sobrenome) ou, ainda, pode significar o substantivo *trabalho*. Para resolver tal ambiguidade, BALIE usa heurísticas simples. Dado um documento, assume-se que o termo é uma entidade nomeada caso apareça capitalizado (primeira letra em maiúscula), exceto quando:

- há ocorrências no documento sem capitalização;
- somente aparece no início de uma frase ou entre aspas;
- aparece dentro de sentenças nas quais todas as palavras com mais de três caracteres iniciam com maiúsculo, como, por exemplo, em títulos ou seções.

O problema de detecção de limites de entidade refere-se à estratégia de reconhecer onde uma entidade inicia e onde ela termina no texto. Isso acontece sempre com entidades que são formadas por duas ou mais palavras, como São Paulo (cidade) Folha de São Paulo (jornal), por exemplo. BALIE trata esses casos realizando uma união das entidades consecutivas quando estas são do mesmo tipo, ou ainda, quando a união resultante é reconhecida no dicionário.

O terceiro e último problema está associado à ambiguidade entre entidades. Esse caso ocorre quando as mesmas entidades (mesmos tokens) pertencem a mais de um tipo de classe, isto é, quando tecnicamente se encontram em duas listas de termos no dicionário. Um algoritmo específico foi criado para solucionar esse tipo de erro. Quando uma entidade ambígua é encontrada, seus termos podem ser usados de duas formas. Primeiramente, se uma palavra que define uma entidade não é ambígua, esta pode auxiliar na desambiguação. Por exemplo, "Oceano Atlântico" pode significar um lugar ou

localização, porém “Atlântico” isoladamente pode ter sentido de localização ou até mesmo de uma organização. Se ambas as palavras pertencem ao conjunto de termos que formam a entidade, então se assume que o conjunto inteiro é do tipo localização. A segunda forma de resolver a ambiuidade seria incluir palavras próximas aos termos da entidade para usar um contexto maior para classificá-las (NADEAU, 2005).

5. Solução proposta

A solução proposta neste trabalho tem como objetivo reconhecer entidades em documentos textuais não estruturados. Essas entidades são possíveis instâncias e possuem tipos/classes utilizadas para compor uma ontologia de domínio. Por meio do cálculo de coocorrência, é possível estabelecer a força da relação entre as entidades.

Abaixo são apresentados os passos aplicados pela abordagem proposta, para que, partindo-se de uma coleção de documentos textuais, seja possível inferir as entidades e suas relações e registrá-las numa ontologia:

1. **Reconhecimento de entidades**: nesta etapa, a técnica de reconhecimento de entidades na base de documentos textuais é aplicada, e as entidades textuais são classificadas;
2. **Frequência das entidades**: identificação da frequência de ocorrências das entidades na coleção de documentos;
3. **Validação de entidades**: validação das entidades retornadas pelo algoritmo de reconhecimento de entidades;
4. **Cálculo da relação entre entidades**: o grau de correlação entre as entidades é estabelecido por meio do cálculo de janelas (número de palavras entre duas entidades) e frequência de ocorrências;
5. **Apresentação do resultado**: etapa em que um especialista do domínio é envolvido. Nesta etapa é feita a apresentação dos resultados para o usuário, seja adicionando-se as instâncias à ontologia de domínio ou utilizando-se um grafo para facilitar o usuário na construção da nova ontologia.

5.1. Arquitetura da solução

O sistema tem como entrada um documento textual (ou uma coleção de documentos) não estruturado. A partir deste documento, são extraídos os elementos textuais candidatos a entidades reconhecidas. Esses elementos textuais são submetidos a uma pré-classificação realizada por meio do framework BALIE, o qual disponibiliza informações como entidades, suas possíveis classes, a posição e a sentença (frase) em que cada entidade é encontrada. Essas informações são importantes para posteriormente efetuar o cálculo de correlação entre as instâncias. Cada entidade *E* pode ser representada assim: E={nome, classe, posições, número frase}, em que o *nome* é a própria entidade (ex.: Universidade Federal de Santa Catarina), *classe* é a classe da ontologia à que a entidade pertence (ex.: instituição), *posições* é a lista de posições em que a

entidade é encontrada na sentença e o *número frase* é a posição da sentença no texto.

Essas entidades devem ser submetidas para a validação de um usuário que irá excluir as entidades não relevantes para o domínio. O vetor resultante do processo de validação é submetido ao algoritmo de correlação, que irá encontrar o peso das relações entre as entidades e as classes (mais informações sobre o algoritmo na seção 5.1.2.). A Figura 2 abaixo representa a arquitetura da solução proposta:

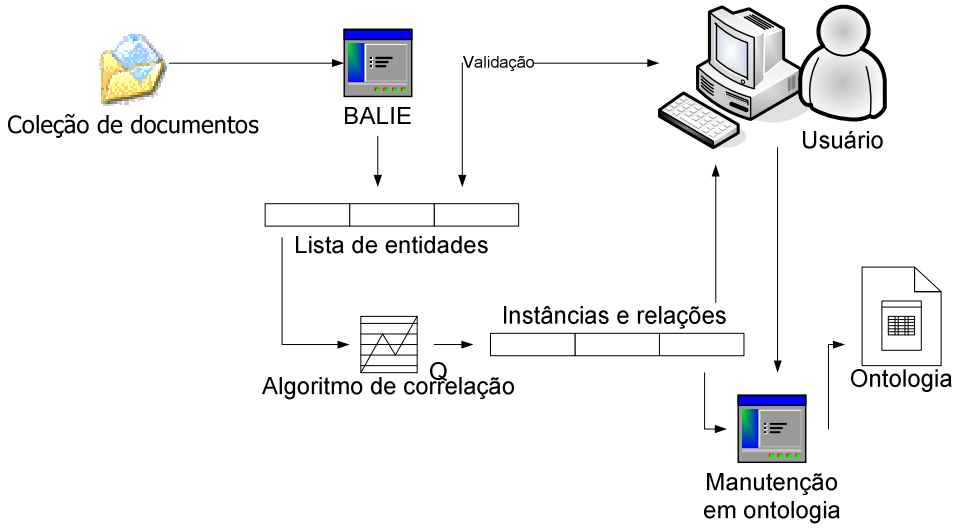

Figura 2 - Arquitetura da solução

O usuário seleciona os documentos de entrada, e o sistema retorna as instâncias e suas relações. Essa informação serve de insumo para a adição de novas instâncias às classes da ontologia. O usuário interage validando a lista de entidades geradas a partir de cada documento. No próximo passo, é criada uma matriz que irá apresentar as instâncias e o grau de correlação entre elas. Por fim, a matriz é utilizada para identificar novas instâncias de uma ontologia de domínio e, também, os relacionamentos entre essas instâncias, os quais são extraídos a partir do grau de correlação obtido em cada documento.

A arquitetura proposta auxilia tanto a manutenção de uma ontologia de domínio existente quanto a criação de uma ontologia. Por meio dessa abordagem, caso a ontologia ainda não exista, o usuário pode perceber classes, instâncias ou ainda relacionamentos através das entidades e das relações obtidas pelo algoritmo de correlação. Uma vez definidas, tais entidades e relações podem ser visualizadas por meio de um grafo que irá auxiliar na construção da ontologia de domínio da organização.

5.1.1. BALIE

Essa seção descreve as atividades técnicas usadas com o framework BALIE para a realização do reconhecimento de entidades. Segundo Nadeau (2005), o BALIE é um sistema multi-idiomas para a extração de informação em documentos textuais utilizado mais precisamente para o reconhecimento de entidades. A versão 1.8 não possui suporte para o idioma português do Brasil, que é imprescindível para o correto reconhecimento dos textos da Plataforma Lattes. Além disso, o reconhecimento de entidades requer que as áreas de

conhecimento, nomes ou siglas das organizações e os nomes das pessoas estejam mapeados como gazetters. Portanto, dois passos adicionais precisaram ser desenvolvidos, como se segue:

- extensão das APIs Java do BALIE;
- geração de gazetters para as áreas de conhecimento, organizações e pessoas da Plataforma Lattes;

Para atender ao requisito de identificação do idioma português do Brasil, BALIE provê classes Java exclusivas para serem estendidas. Basicamente a classe *ca.uottawa.balie.LanguageSpecific* deve ser herdada, e consequentemente dois métodos devem ser implementados: o método *GetAbbreviations,* usado para retornar um conjunto de abreviações empregadas nas heurísticas; e, o método *Decompound*, empregado para quebrar o documento em tokens ou palavras. A decomposição do documento em tokens, optou-se pelo caractere de espaço em branco como único separador. A classe *LanguageSpecificPortuguese* foi criada para cumprir esse requisito.

Tecnicamente, BALIE possui em seu *classpath* uma pasta denominada *lexicon* com os arquivos no formato de texto (extensão *txt*), onde estão localizados os termos do dicionário. Para funcionar corretamente, todos os termos devem estar em minúsculo e sem quaisquer caracteres especiais, tais quais acentos e cedilha. BALIE já possui uma lista de termos classificados, como cidades, tempo (meses, feriados, dias de semana, etc.), pessoas e organizações, às quais se adicionaram os termos da Plataforma Lattes. Apenas um novo arquivo-texto contendo as áreas de conhecimento foi adicionado ao BALIE. Para conduzir a leitura do dicionário, é necessário modificar as seguintes classes Java: *LexiconOnDisk*, adicionando à nova lista de termos, no caso a lista de áreas de conhecimento; e *NamedEntityTypeEnum*, enumerando o novo tipo criado.

Após efetuar essas alterações no código-fonte, a ferramenta já se encontra preparada para extrair as entidades a partir do texto.

5.1.2. Algoritmo de correlação

Dado o *résumé* curricular de uma pessoa no formato de texto, o protótipo juntamente com o BALIE realizam o processo de reconhecimento de entidades. O próximo passo é a geração de um vetor contendo as entidades encontradas, a classe a que elas pertencem, qual a sua posição no texto e em qual sentença (frase) elas estão contidas. A partir deste vetor, são extraídos os termos distintos que são os índices da matriz. A matriz gerada é do tipo Entidade X Entidade, sendo que as células armazenam o valor da correlação entre elas. A tabela 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 Gonçalves et al., 2006. A abordagem utilizada é bastante simplificada e leva em consideração 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 calcular a correlação entre dois termos, onde *freq* é igual à frequência 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", a janela entre as entidades "Flávio Ceci" e "Ciência da Computação" é 4, pois existem 4 palavras entre as duas entidades.

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

A janela média ($\bar{j}$) é calculada pela fórmula abaixo:

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

Ainda na frase apresentada acima, podemos verificar que a frequência das entidades é igual a 1, já que os termos "Flávio Ceci" e "Ciência da Computação" só coocorrem uma única vez. Vamos aplicar este exemplo na fórmula anteriormente apresentada:

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

Com a matriz gerada, as entidades mais relevantes e o seu grau de correlação são apresentados ao usuário. Esse resultado auxilia no processo de manutenção, pois apresentam novas possíveis instâncias para as classes e quais as suas relações.

5.1.3. Estudo de caso: Résumés curriculares da Plataforma Lattes

Com o intuito de ilustrar a aplicação da abordagem em um cenário real, tomou-se como fonte de dados a Plataforma Lattes Institucional da UFSC, por meio da qual é possível realizar análises sobre as informações curriculares do corpo discente e docente. Através do *résumé* acadêmico, texto livre no qual a pessoa descreve suas atividades profissionais, áreas de conhecimento e instituições de atuação, aplicou-se o reconhecimento de entidades. O objetivo de

interesse 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 resumos como entrada de dados, e o sistema deve apontar as instâncias e seus relacionamentos. No fim do processo, são apresentados os resultados na forma de uma rede.

Para o BALIE reconhecer uma entidade, ela deve estar anteriormente cadastrada numa lista de termos, como mencionado na seção 5.1.1. O BALIE não suporta a linha portuguesa, e por esse motivo carregou-se a lista de pessoas, organizações e locais com os termos em português. Esses dados foram extraídos por meio de consultas realizadas na base operacional da Plataforma Lattes Institucional da UFSC.

Para facilitar o entendimento do processo, apresenta-se abaixo um extrato dos *résumés* dos currículos Lattes dos autores deste trabalho 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 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 co-autorias 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 co-orientou 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 co-autorias 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 base nos *résumés* fornecidos como insumo, o processo retorna uma lista de entidades reconhecidas. Essas entidades foram submetidas a um processo de validação por um usuário que retirou os termos que não fazem parte deste domínio ou não eram relevantes. Em seguida, as entidades restantes foram submetidas ao algoritmo de correlação. A Tabela 1 abaixo exibe um exemplo (resumido) do resultado obtido, considerando apenas o extrato de um *résumé* ilustrado anteriormente.

Tabela 1 - Exemplo de grau de correlação entre entidades

Entidades / ***Entidades***	*alexandre (pessoa)*
ciencias da computacao (*área*)	0,125
engenharia de producao (*área*)	0,026315789
engenharia do conhecimento (*área*)	0,028571429
universidade regional de blumenau (*organização*)	0,083333333
instituto stela (*organização*)	0,1

Na Tabela 1, as entidades textuais identificadas encontram-se em caixa baixa e sem acentuação devido ao pré-processamento do texto realizado anteriormente. As classes de cada entidade são mostradas entre parênteses. No exemplo, a entidade "alexandre" possui um grau de correlação de "0,125" com a entidade "ciencia da computacao". A relevância do grau de correlação pode ser ajustada conforme a necessidade do usuário ou contexto de aplicação, e assim poder-se-ia, por exemplo, definir um valor mínimo de aceitação para que uma relação fosse considerada.

Para facilitar a visualização dos resultados e melhor demonstrar ao usuário os relacionamentos entre as instâncias encontradas para auxiliar o processo de criação da ontologia, as instâncias e os seus relacionamentos são apresentadas ao usuário a partir de uma ferramenta incorporada à arquitetura da solução que gera gráficos em forma de redes. A Figura 3 a seguir mostra o gráfico de uma rede de relacionamento gerada com base nos *résumés*.

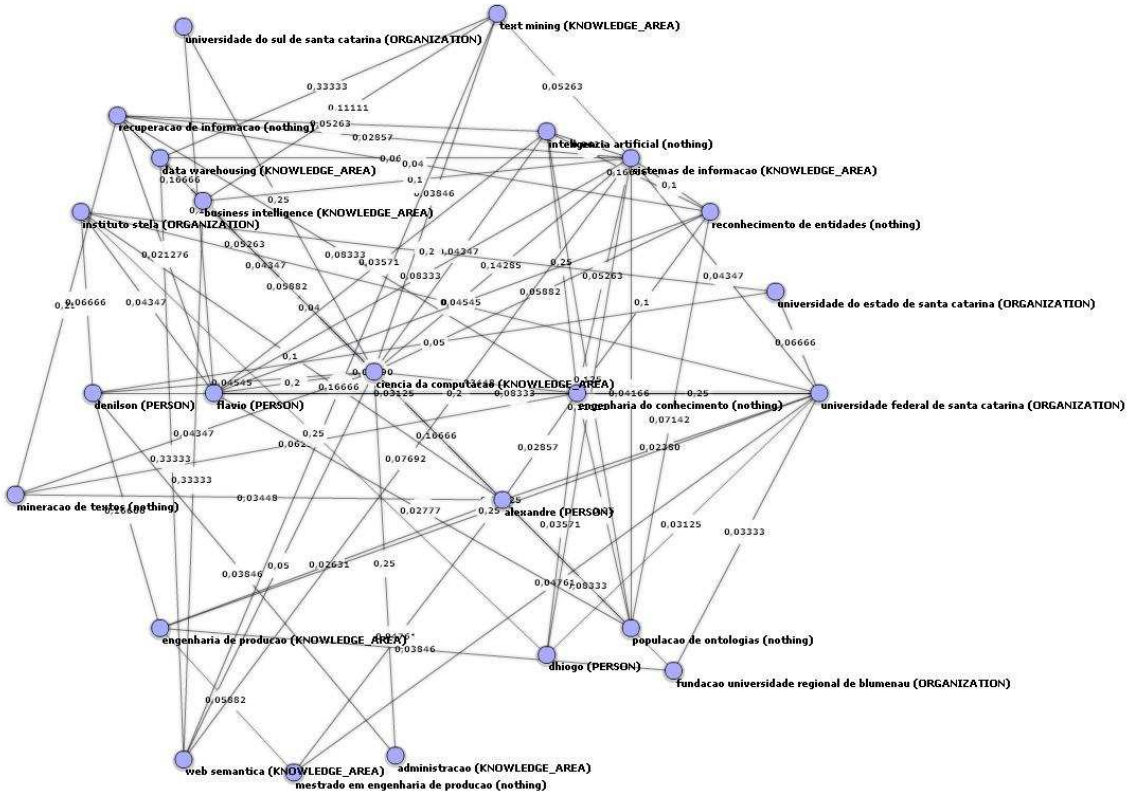

Figura 3 - Entidades e seus relacionamentos do resultado

Como não existia uma ontologia para se atualizar, optou-se por apresentar os resultados em forma de rede. Essa visualização auxilia o usuário que irá construir a nova ontologia, podendo ver as entidades e as suas classes e como elas estão relacionadas.

Na Figura 3, são apresentadas todas as instâncias encontradas com a classe a que ela pertence entre parênteses, e por meio das arestas é possível identificar o relacionamento e o grau de correlação entre as instâncias (entidades).

Como esse grafo resultante exibido na Figura 3 traz muitas informações, uma análise mais detalhada das instâncias torna-se importante. É por esse motivo que a aplicação permite que o usuário selecione as instâncias que gostaria de analisar, apresentando apenas as instâncias e os relacionamentos que estão ligados a ela. A Figura 4 mostra como a instância “ciência da computação” está ligada às outras instâncias, o que permite uma análise mais detalhada.

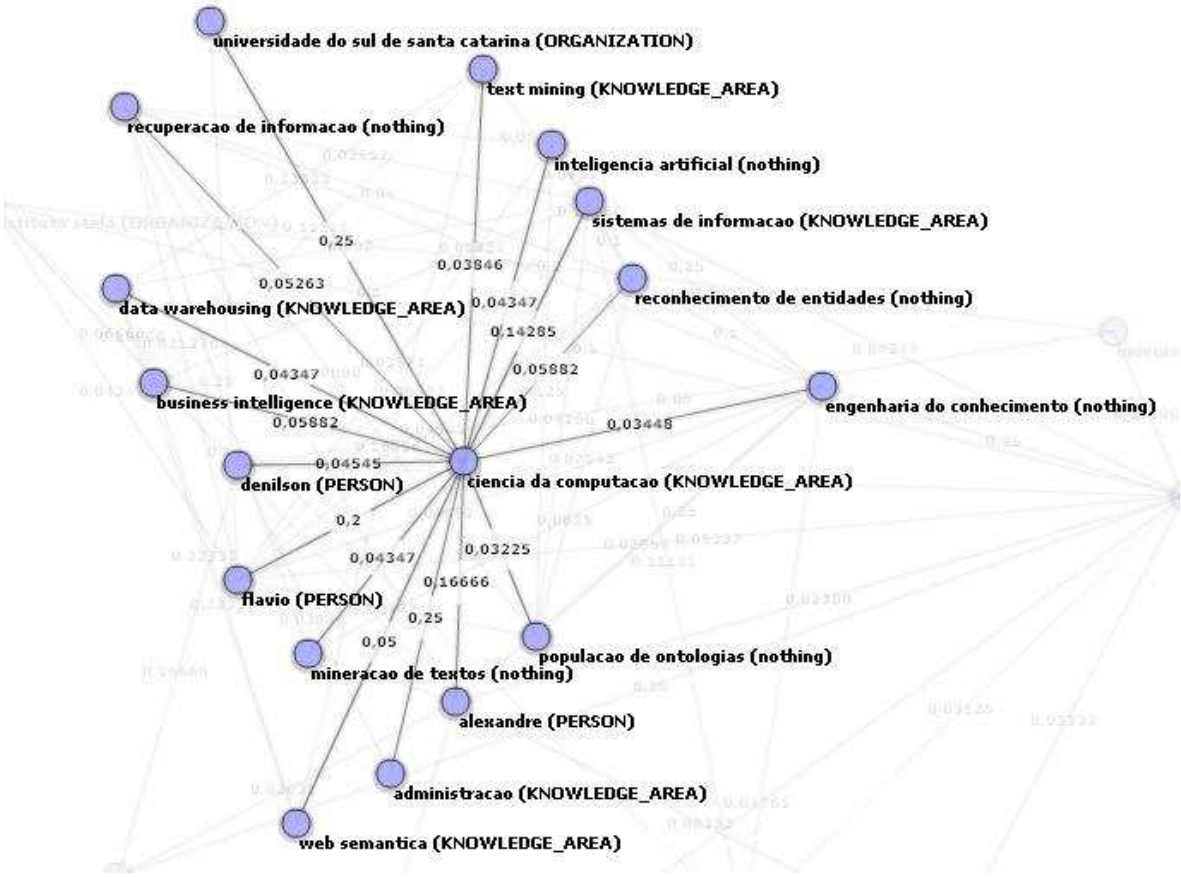

Figura 4 - Instâncias ligadas à instância “ciencia da computacao”

A aplicação classificou a instância “ciencia da computação” como “KNOWLEDGE_AREA”, ou seja, da classe área do conhecimento. A figura ilustra 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 em Ciência da Computação.

A instância “ciencia da computação” também está ligada a “sistemas de informacao" (também área de conhecimento), cursos com formação similar. Pode-se observar outras instâncias do tipo área do conhecimento que estão ligadas a “ciencia da computacao”, como, por exemplo: “inteligencia artificial”, “recuperacao de informacao”, “mineracao de texto”, “web semantica”, entre outras. Vale lembrar que essas relações só puderam ser identificadas dada a riqueza do conteúdo dos documentos selecionados. Outros documentos ou outro cenário de aplicação podem não encontrar resultados satisfatórios.

Logo abaixo, a Figura 5 demonstra os relacionamentos da instância “flavio" pertencente à classe “pessoa” com as demais instâncias identificadas no *résumé* da Plataforma Lattes.

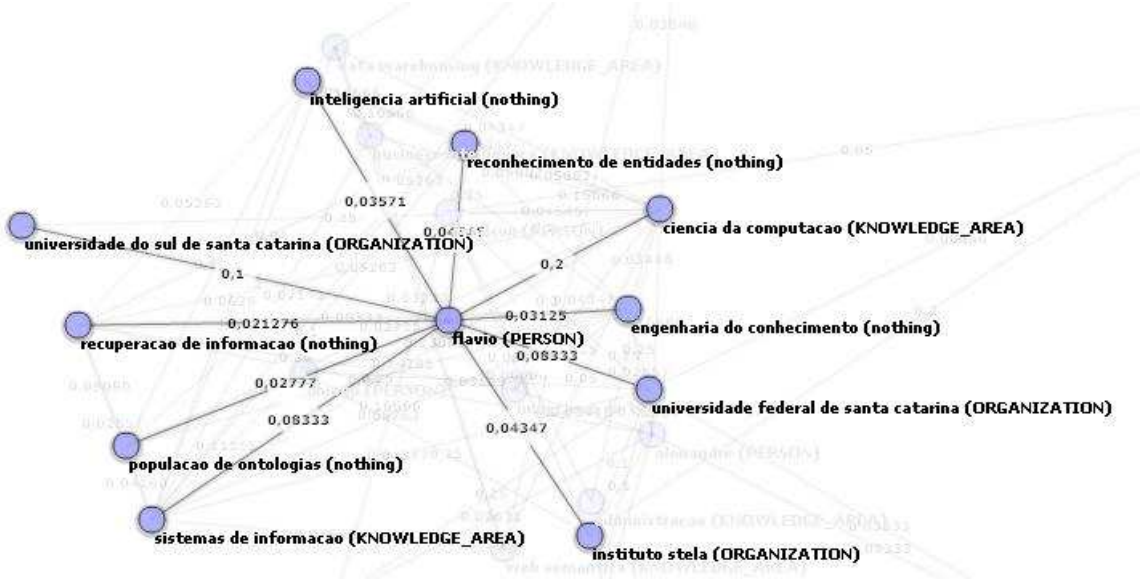

Figura 5 - Relacionamentos entre a pessoa “flavio” e demais instâncias

A aplicação classificou a instância “Flavio” como pessoa. Pode-se observar que o pesquisador possui relacionamentos com as instituições: Universidade do Sul de Santa Catariana, 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 5 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”, “reconhecimento de entidades”. Além disso, há outras instâncias em que as classes não foram identificadas no processo de reconhecimento de entidades. Por exemplo, a instância “populacao de ontologias”, apesar de possuir relação com a instância “flavio", não possui uma classe definida. Quando isso ocorre, cabe ao usuário definir qual a classe apropriada ou descartar a relação se não houver relevância.

A seguir, a Figura 6 exibe outra rede de relacionamento entre a instância “dhiogo” pertencente à classe *pessoa* com as demais instâncias reconhecidas no documento.

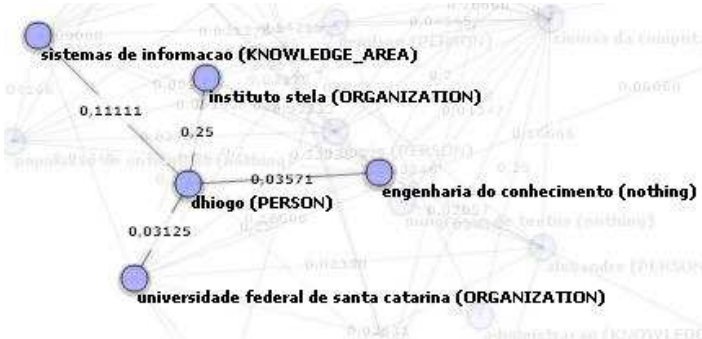

Figura 6 - Relacionamentos entre a pessoa “dhiogo” e demais instâncias

Pode-se identificar a relação entre a instância “dhiogo” classificada como Pessoa com as instâncias da classe *Organização*: “universidade federal de santa Catarina” e “instituto stela”. Nota-se que embora em seu *résumé* curricular o pesquisador Dhiogo tenha informado algumas áreas de conhecimento, elas não estão presentes nesta figura, pois nas sentenças em que o pesquisador se refere a elas, não é citado o seu nome em conjunto. Ou seja, elas não coocorrem na mesma sentença.

A última análise feita, utilizam-se as instâncias do tipo pessoa Denilson e Alexandre, conforme ilustra a Figura 7 a seguir.

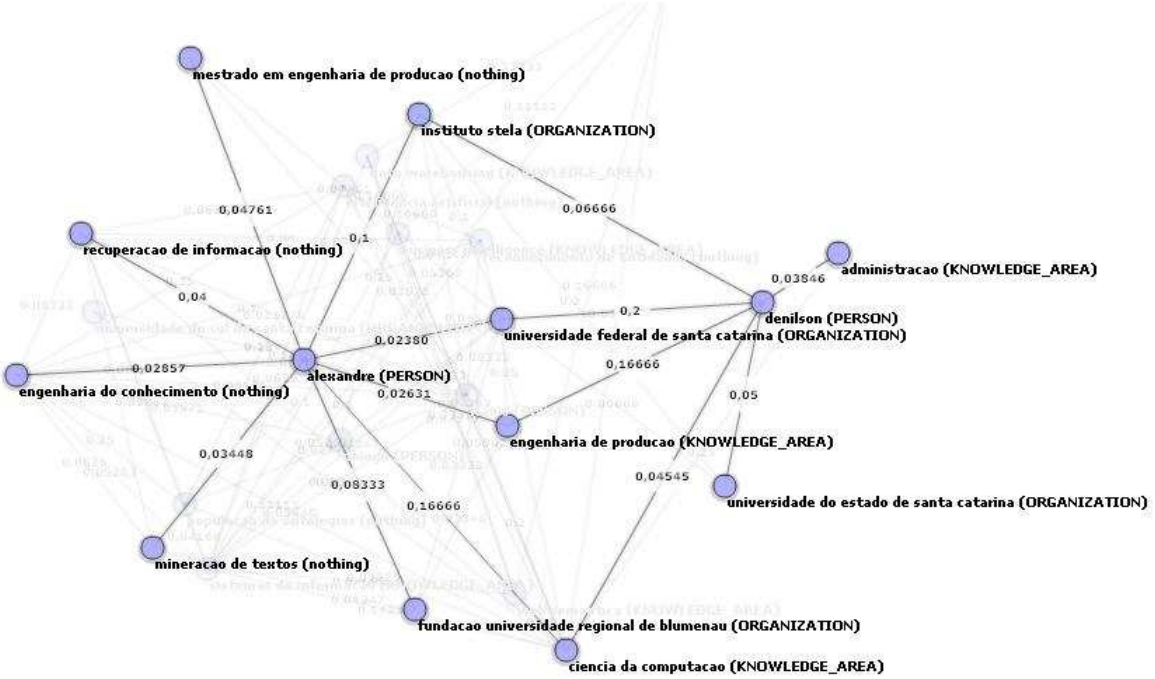

Figura 7 - Relacionamentos entre as instâncias "denilson" e "alexandre"

Observa-se que em nenhum momento nos *résumés* dos pesquisadores Denilson e Alexandre encontram-se essas instâncias coocorrendo. Nesse resultado, pode-se observar que essas instâncias 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 como Ciência da Computação e Engenharia de Produção. Um especialista poderia inferir que as duas pessoas trabalham nas mesmas organizações e que possuem interesses de pesquisas similares. Informações como essas podem ser utilizadas para apoiar ações como a formação de equipes de trabalho, 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.

6. Conclusão

O presente trabalho apresenta a utilização de um sistema de reconhecimento de entidades para auxiliar no processo de extração e representação de conhecimento, com vistas a apoiar a manutenção de ontologias. O estudo de caso ilustrou como o reconhecimento de entidades, subárea da Extração de Informação que trata da classificação de elementos textuais contidos em documentos, pode ser útil para explicitar relacionamentos entre instâncias de classes da ontologia.

Na solução atual da arquitetura, foi aplicado o framework BALIE para auxiliar o processo de reconhecimento de entidades. Para o estudo de caso apresentado, utilizou-se uma amostra de *résumés* da Plataforma Lattes como fonte primária de informação. Por meio desses extratos curriculares, realizou-se tanto o reconhecimento quanto o cálculo da aderência entre as entidades com a finalidade de explicitar a força dos relacionamentos entre elas e de apoiar a manutenção de ontologias.

O experimento demonstra que a partir de pequenas porções de texto, como os *résumés* da Plataforma Lattes, é possível a geração de mapas que promovam uma ideia de como determinado conhecimento se distribuiu e se relaciona em um domínio de conhecimento. Isso fica evidente nas projeções (redes) utilizadas para explicitar as entidades e a força do relacionamento entre elas. Além disso, tal representação promove subsídios ao engenheiro de ontologias em seu trabalho de manutenção desse tipo de estrutura, caracterizando-se assim um processo semiautomático.

Trabalhos futuros residem em quatro pontos principais. Primeiro, pretende-se incrementar o modelo pela identificação de relacionamentos factuais, e não somente pela força do relacionamento. Em seguida, torna-se necessária a conexão com bases colaborativas (Wikipédia, por exemplo), visando facilitar a correta extração e classificação de entidades.

Pretende-se ainda avaliar a solução sobre grandes coleções de documentos de modo que seja possível medir a precisão e a facilidade de todo o processo. Por fim, vislumbra-se a utilização da abordagem para apoiar aplicações de Gestão do Conhecimento, tais como localização de especialistas, identificação de hiatos de competência e mesmo outras iniciativas no contexto da Inteligência Competitiva.

Referências

ALMEIDA, M. B.; BAX, M. 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.

ALMEIDA, M. B. **Um modelo baseado em ontologias para a representação da memória organizacional**. Tese apresentada ao curso de Doutorado do Programa de Pós Graduação da Escola de Ciência da Informação da Universidade Federal de Minas Gerais, como requisito parcial à obtenção do título de Doutor em Ciência da Informação. Belo Horizonte, 2006.

ANTONIOU, G.; HARMELEN, F. **A Semantic Web Primer**. The MIT Press Cambridge, Inglaterra, 2ª ed. 2008.

BERNERS-LEE, T.; HENDLER, J.; LASSILA, O. **The Semantic Web**. Scientific American, 2001.

CARDOSO, O. **Recuperação de Informação**. Universidade Federal de Lavras, Minas Gerais, 2007.

CORDEIRO; J. **Extração de Elementos Relevantes em Texto/Páginas da World Wide Web**. Faculdade de Ciências do Porto, 2003.

DISHMAN, P.; FLEISHER, C.S.; V. KNIP. **Chronological and Categorized Bibliography of Key Competitive Intelligence Scholarship**: Part 1, Journal of Competitive Intelligence and Management, pag. 16-78, 2003.

FELDMAN, R.; HIRSH, H. **Exploiting Background information in Knowledge discovery from text**. Journal of Intelligent Information System, 1997.

FREITAS, F. L. **Ontologias e a Web Semântica**. In: Renata Vieira; Fernando Osório. (Org.). Anais do XXIII Congresso da Sociedade Brasileira de Computação. Volume 8: Jornada de Mini-Cursos em Inteligência Artificial. Campinas: SBC, v. 8, p. 1-52, (2003).

GONÇALVES, Alexandre L.; ZHU, Jianhan; SONG, Dawei; UREN, Victoria; PACHECO, Roberto C S. **LRD: Latent Relation Discovery for Vector Space Expansion and Information Retrieval**. In: Seventh International Conference on Web-Age Information Management (WAIM 2006), 2006, Hong Kong. J.X. Yu, M. Kitsuregawa, and H.V. Leong (Eds.): WAIM 2006, Lecture Notes in Computer Science, 2006. v. 4016. p. 122-133.

GRISHMAN; R. **Information Extraction: Techniques and Challenge**. New York University, 2007.

GRUBER, T. **A translation approach to portable ontology specification. Knowledge Acquisition**, v. 5, n. 2, pag. 199-220, 1993.

KORFHAGE, R. R. **Information Retrieval and Storage**. New York: John Wiley & Sons, p. 349, 1997.

KOZAREVA, Z. **Bootstrapping named entity recognition with automatically generated gazetteer lists**. in Proceedings of EACL student session (EACL 2006) , Trento, Italy, 2006.

KAHANER, L. Competitive Intelligence: **How to Gather Analyze and Use Information to Move Your Business to the Top**. Simon & Schuster. 1996

MOONEY, R. J.; NAHM, U. Y; **Text Mining with Information Extraction**. In: INTERNATIONAL MIDP COLLOQUIUM DAELEMANS, 4., September 2003,

Bloemfontein, South Africa. W., du PLESSIS, T., SNYMAN, C. and TECK, L. (Eds.). Proceedings… Bloemfontein, South Africa: Van Schaik Pub., p.141-160. 2005.

MORAIS, E. A. M. **O estado da arte no estudo das ontologias. In: Simpósio de Estudos e Pesquisas: Educação, Cultura e Produção do Conhecimento da FASAM**, Goiânia. Anais do Simpósio. 2006.

NADEAU. D.; **BALIE – Base Line Information Extraction: Multilingual Information Extraction from Text with Machine Learning and Natural Languages Techniques**. University of Ottawa, Canadá, 2005.

NADEAU; D.; TURNEY P.; MATWIN S;. **Unsupervised Named-Entity Recognition: Generating Gazetteers and Resolving Ambiguity**. 19th Canadian Conference on Artificial Intelligence, 2006.

NAHM. U.; MOONEY. R. **Text Mining with Exformation Information**. University of Texas, 2002.

NAVIGLI, R; VELARDI, P. **Learning Domain Ontologies from DocumentWarehouses and Dedicated Web Sites.** Universit`a di Roma "La Sapienza", 2004.

NEGRI, M.; MAGNINI, B. **Using wordnet predicates for multilingual named entity recognition**. In Proceedings of The Second Global Wordnet Conference, pag. 169–174. 2004.

NOY, Natalya F.; MCGUINNESS, Deborah L. **Ontology Development 101: A Guide to Creating Your First Ontology**. Stanford University, California, United States, 2001

PÉREZ, G; BENJAMINS, V. **Overview of Knowledge Sharing and Reuse Components: Ontologies and Problem-Solving Methods**, 1999.

PRAHALAD C. K.; HAMEL, G. **The Core Competence of the Corporation**. Springer Berlin Heidelberg, 1996.

RILOFF, E. LEHNERT, W. **Information Extraction as a Basis for High-Precision Text Classification**. ACM Transactions on Information Systems,1994. Disponível na Internet em <http://citeseer.ist.psu.edu/old/21204.html>

SOON. W; LIM; D.; Ng.; H. **A Machine Learning Approach to Coreference Resolution of Noun Phrases**. DSO National Laboratories, 2001.

STUDER R.; BENJAMINS, R.; FENSEL, D. **Knowledge Engineering: Principles and Methods**. 1998. – Disponível na internet em <http://www.aifb.uni-karlsruhe.de/WBS/Publ/1998/dke98.html>. Acessado em Junho de 2009.

STAAB, S.; STUDER, R.; SCHNURR, H.P.; SURE, Y. **Knowledge processes and ontologies**. IEEE Intelligent systems. V.16, n.1, p. 26-34, 2001.

WITTEN I. H.; FRANK E. **Data Mining: Practical Machine Learning Tools and Techniques**. Morgan Kaufmann, 2ª ed. 2005.

WIVES, L. **Tecnologias de Descoberta de Conhecimento em Textos aplicadas à Inteligência Competitiva**. Universidade Federal do Rio Grande do SUL, 2002.

ZHU, Jianhan; UREN, Victoria; MOTTA, Enrico. **ESpotter: Adaptive Named Entity Recognition for Web Browsing**. Professional Knowledge Management Conference, Springer-Verlag LNAI p. 518-529, Alemanha, 2005.

ZOHAR; Y. **Introducting to Text Mining**. University of Illinois, 2002.