UNIVERSIDADE FEDERAL DE SANTA CATARINA
DEPARTAMENTO DE INFORMÁTICA E ESTATÍSTICA

Willian Santos de Souza

INVESTIGAÇÃO DE MENÇÕES A ENTIDADES DE INTERESSE PARA *E-BUSINESS* EM TWEETS

Florianópolis

2019

Willian Santos de Souza

INVESTIGAÇÃO DE MENÇÕES A ENTIDADES DE INTERESSE PARA *E-BUSINESS* EM TWEETS

Trabalho de Conclusão de Curso submetido ao Curso de Ciências da Computação para a obtenção do Grau de Bacharel em Ciências da Computação.
Orientador: Prof. Dr. Renato Fileto

Florianópolis

2019

Willian Santos de Souza

INVESTIGAÇÃO DE MENÇÕES A ENTIDADES DE INTERESSE PARA *E-BUSINESS* EM TWEETS

Este Trabalho de Conclusão de Curso foi julgado aprovado para a obtenção do Título de "Bacharel em Ciências da Computação", e aprovado em sua forma final pelo Curso de Ciências da Computação.

Florianópolis, 03 de Julho 2019.

Prof. José Francisco D. de G. C. Fletes
Coordenador do Curso

Banca Examinadora:

Prof. Dr. Renato Fileto
Orientador

Prof. Dr. Elder Rizzon Santos

Prof. Dr. Ronaldo Dos Santos Mello

Dedico este trabalho primeiramente Deus, à minha esposa, Maieli, por sempre apoiar minhas decisões, ao meu irmão, Beto, e minha mãe, Nilza, que sempre me incentivaram a buscar e seguir meus sonhos e nunca desistir.

AGRADECIMENTOS

Agradeço o meu orientador, professor Dr. Renato Fileto, por me orientar e dedicar esforço para repassar seus conhecimentos.

Aos Brothers da Computação por pelos momentos de descontração.

Aos meus gestores Lucas Zago, Wagner e Fábio Moreira, pelo apoio durante as ausências, dando flexibilidade necessária no trabalho durante esse longo período.

Ao Vilmar, pela ajuda e discussões que contribuíram para a construção deste trabalho.

Aos colegas de LISA, que mesmo em momentos de tenção, nunca se deixaram abater e sempre para manter o clima do laboratório harmônico e divertido.

Ao Doutorando Ítalo, que contribuiu de forma direta na reta final deste trabalho, que me ajudou dando conselhos em momentos que eu me encontrava sem um direcionamento, seja através de revisão do trabalho escrito, tirando dúvidas, ou simplesmente batendo um papo.

E aos colegas da Universidade de Leipzig, Matthias, Olaf e Rainer pela colaboração na construção das pontes para o domínio de negócio usadas neste trabalho.

A mente que se abre a uma nova ideia,
jamais voltará ao seu tamanho original.

Albert Einstein

RESUMO

O crescimento de mídias sociais como meio de comunicação trouxe a possibilidade de usar dados abertos que trafegam por essas mídias em uma miríade de aplicações, desde detecção de eventos e desastres naturais, até marketing e recomendação de produtos e serviços. Textos de postagens de usuários nessas mídias podem conter menções a empresas, produtos, locais, etc. Ferramentas de anotação semântica permitem identificar tais menções e ligá-las a recursos que as descrevem com semântica bem definida. Este trabalho apresenta um estudo investigativo de menções a entidades relacionadas ao domínio de *e-Business* em postagens de mídias sociais e propõe um processo para construção de hierarquias semânticas dispondo de uma ontologia de domínio para construção de pontes entre os recursos semanticamente anotados com *Linked Open Data (LOD)* utilizando ferramentas de anotação semântica para identificar tais menções. Experimentos demonstram que anotações relacionadas à *e-Business* estão entre as mais comuns em postagens de mídias sociais enviadas do território brasileiro no período de execução deste trabalho. Isto sugere o potencial de uso de tais dados, anotados semanticamente para sistemas de *Business Intelligence* (BI).

Palavras-chave: web semântica, mídias sociais, dados abertos ligados, anotação semântica, e-business, data warehouse, dimensões de análise, hierarquia semântica.

ABSTRACT

The growth of social media as one of the mainstream communication forms has brought the possibility of using the open data that travels through these media in a myriad of applications, from the detection of events and natural disasters to marketing and recommendation of products and services. Texts from user postings in these media may contain mentions to companies, products, locations, etc. The semantic annotation tools allow to identify such mentions and link them to resources that describe them with well-defined semantics. This work presents an investigative study of mentions of interest to entities related to the e-Business domain in social media posts and proposes a process for constructing semantic hierarchies with a domain ontology for building bridges between semantically annotated resources with Linked Open Data (LOD) using semantic annotation tools to identify such references. Experiments show that annotations related to e-Business are among the most common in social media postings sent from Brazilian territory during the execution period of this work. This suggests the potential use of such data, annotated semantically for Business Intelligence (BI).

Keywords: semantic web, social media, linked open data, semantic annotation, e-business, data warehouse, semantic hierarchy, analysis dimensions.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE CÓDIGOS

LISTA DE ABREVIATURAS E SIGLAS

SUMÁRIO

1 INTRODUÇÃO

O surgimento de mídias sociais teve um impacto grande nos estudos sobre o comportamento humano (TUFEKCI, 2014). A ascensão de tais plataformas modificou a forma como as pessoas interagem (HEATH; SINGH; GANESH, 2014). Usando mídias sociais, pessoas podem criar conteúdo, compartilhá-lo e indicar seus sentimentos em relação a eles (e.g., marcar se gostou ou não) (ASUR; HUBERMAN, 2010), entre outras possibilidades, contribuindo significativamente para a geração de uma grande quantidade de dados.

Todavia, dados textuais de postagens em mídias sociais são considerados não-estruturados para fins de processamento computacional. Além disso, textos de postagens em mídias sociais são usualmente curtos (o que resulta em pouca informação de contexto) e sujeitos a ruídos (e.g., erros de ortografia e gramática, acrônimos, abreviações, gírias, etc.) e ambiguidade devido a fenômenos linguísticos (e.g., homonímia, sinonímia). Por exemplo, postagens na rede social Twitter[1] (denominados de *tweets* no restante deste trabalho) têm natureza informal, tamanho limitado, semântica pouco definida e frequentemente muitos ruídos (FILETO et al., 2016). Um dos principais desafios é lidar com a ambiguidade, i.e., uma palavra ou expressão pode ser usada em diferentes contextos, denotando coisas diferentes. Esses fatores dificultam a análise das postagens e demonstram a necessidade de um tratamento prévio das informações para obter êxito na análise. Assim, para fazer uso do vasto volume de dados disponíveis em mídias sociais é necessário extrair automaticamente dos textos de postagens informações mais estruturadas e semanticamente precisas.

A *Extração de Informação* (EI) é a área da Computação que visa extrair informação (semi-)estruturada de dados não estruturados, tais como textos livres. Um sistema de EI analisa um texto escrito em linguagem natural para extrair informações sobre diferentes tipos de eventos, entidades ou relacionamentos. A informação coletada pode ser armazenada em uma Base de Conhecimento que mantém descrições semânticas dos objetos (e.g., locais, épocas, pessoas, organizações, ações, eventos) e suas relações de modo formal, para que possam ser processadas por máquinas, compartilhadas, consultadas e utilizadas (HABIB; KEULEN, 2014).

Atualmente, a *Wikipedia*[2] é uma das principais enciclopédias di-

[1] https://twitter.com/
[2] https://www.wikipedia.org/

gitais e é mantida por diversos contribuidores. *DBpedia*[3] (AUER et al., 2007) é uma base de conhecimentos extraídos da *Wikipedia* e disponibilizada na Web sob a forma de dados interligados abertos (do inglês *Linked Open Data - LOD*). Os padrões e diretrizes de publicação e interconexão de LOD promovem a sua integração e reuso para diversas finalidades, incluindo a anotação semântica, i.e., associação de porções de conteúdo (possivelmente não estruturados) a recursos com semântica bem definida descritos em uma base de conhecimento (e.g., *DBpedia*, *Yago*[4], *Freebase*[5]).

As seguintes ferramentas gratuitas de anotação semântica de textos estão entre as mais proeminentes atualmente: *DBpedia-Spotlight*[6] (MENDES et al., 2011), *Federated knOwledge eXtraction Framework* (FOX)[7] (SPECK; NGOMO, 2014b) (SPECK; NGOMO, 2014a) e *Babelfy*[8] (MORO; RAGANATO; NAVIGLI, 2014). As duas primeiras ligam menções encontradas nos textos a recursos da DBpedia, enquanto que a última liga as menções a recursos da *BabelNet* (NAVIGLI; PONZETTO, 2010). Entretanto, o uso de tais ferramentas ainda precisa ser investigado em domínios específicos, tais como *e-Business*, anotando semanticamente dados de natureza complexa (não estruturados), tais como postagens em mídias sociais.

Este trabalho investiga menções a entidades relacionadas a *e-Business* em postagens de mídias sociais, mais especificamente *tweets*, usando a *DBpedia-Spotlight* como ferramenta de anotação e gerando hierarquias semânticas com instâncias e classes da *DBpedia*, ligadas através das propriedades `rdf:type` e `rdfs:subClassOf`, filtradas através de pontes construídas usando uma ontologia para o domínio de interesse.

1.1 OBJETIVOS

O objetivo geral deste trabalho é verificar a incidência de menções relacionadas com *e-Business* (e.g., empresas, produtos, serviços, promoções) em *tweets*. É utilizado a *DBpedia-Spotlight* para efetuar anotações semânticas, seleção de anotações de interesse com base em

[3] https://wiki.dbpedia.org/

[4] https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/

[5] https://developers.google.com/freebase/

[6] https://www.dbpedia-spotlight.org/

[7] http://fox-demo.aksw.org/#!/home

[8] http://babelfy.org/

pontes para uma ontologia de alto nível específica do domínio considerado e uso de software desenvolvido pelo autor para automatizar a verificação das incidências.

Os objetivos específicos deste trabalho são:

1. Estudar o estado da arte na área de anotação semântica e investigar menções a objetos e entidades relacionadas à *e-Business* em *tweets*.

2. Efetuar anotação semântica de *tweets* usando ferramentas de anotação (e.g., *DBpedia-Spotlight*) e bases de conhecimento (e.g., DBpedia) atualmente disponíveis;

3. Filtrar anotações de interesse para *e-Business* usando pontes entre classes dos recursos anotados no Objetivo Específico 2 e classes da ontologia de alto nível Good Relations (HEPP, 2008);

4. Analisar a incidência de menções a classes e instâncias relacionadas à *e-Business* em *tweets*;

    4.1. Desenvolvimento de software para automatização deste processo.

5. Prospectar possibilidades de aplicação dos dados anotados para Análise de Informação em *Data Warehouse* e em sistemas de recomendação.

Durante o desenvolvimento do trabalho, foi identificado que uma sequência de passos ocorre para gerar hierarquias com informações para o domínio de interesse. O trabalho se limita a fazer anotações semânticas em amostras de *tweets* previamente coletados e armazenados em um banco de dados do *Laboratório de Integração de Sistemas e Aplicações (LISA*[9]*)* (*dataset BR-2015*) e realizar a análise da incidência de entidades e classes relacionadas a *e-Business* utilizando as hierarquias filtradas.

1.2 MÉTODO DE PESQUISA

Este trabalho foi realizado no LISA, utilizando conhecimentos relacionados a área de anotações semânticas e *datasets* de *tweets* obtidos em trabalhos anteriores. Os *tweets* armazenados nos bancos de dados

[9]http://lisa.inf.ufsc.br

do LISA foram coletados usando a ferramenta *SeMovGet*, desenvolvida no trabalho (KLEIN, 2015). A metodologia de pesquisa usada se dividiu em 4 etapas:

1. Levantamento bibliográfico. O trabalho inicia com a síntese e análise da fundamentação teórica do processo e de ferramentas existentes para a anotação semântica automática de textos, particularmente de postagens em mídias sociais, visando analisar menções relacionadas à *e-Business*;

2. Preparação das bases de dados e conhecimento e das ferramentas de anotação para experimentos;

3. Realização de experimentos de anotação semântica com a ferramenta *DBpedia-Spotlight*;

4. Filtragem das anotações de interesse para *e-Business* usando pontes das classes dos recursos usados nos valores das anotações produzidas para classes da ontologia de alto nível GoodRelations[10];

5. Análise dos resultados obtidos nos experimentos;

6. Prospecção de possibilidades de uso para os resultados gerados pelas ferramentas de anotação em *Data Warehouses*.

1.3 ORGANIZAÇÃO DO TRABALHO

O restante deste documento está organizado como indicado a seguir. O Capítulo 2 apresenta os fundamentos necessário para o entendimento do trabalho. O Capítulo 3 descreve os passos do processo de filtragem de menções de interesse, passando por técnicas de pré-processamento de dados, anotação semântica, seleção de anotações confiáveis e a construção de hierarquias usando como base medidas, tais como, *hits* diretos e indiretos. O Capítulo 4 apresenta os experimentos aplicados em um estudo de caso de *e-Business*. O Capítulo 5 apresenta os trabalhos correlatos encontrados no levantamento bibliográfico. Por fim, o Capítulo 6 apresenta as conclusões e as perspectivas de trabalhos futuros.

[10] http://www.heppnetz.de/projects/goodrelations/

2 FUNDAMENTOS

Este capítulo descreve os fundamentos necessários para a compreensão deste trabalho, incluindo definições, conceitos-chave e exemplos para os seguintes tópicos da Web Semântica: Representação de Conhecimento; Dados Ligados; Anotação Semântica; e Modelagem Dimensional.

2.1 WEB SEMÂNTICA

A *World Wide Web (WWW)* impactou de forma significativa a maneira com que as pessoas interagem e comercializam objetos na Web. Contudo, a WWW em sua versão original possui algumas limitações, como, por exemplo, não possuir suporte a recursos que permitam descrever com semânticas bem definidas dados em páginas na Web. Tal fato dificulta que sistemas (e.g. motor de busca) interpretem as informações contidas em páginas Web. Além disso, as informações presentes em páginas Web são organizadas de forma não estruturada, sendo construídas exclusivamente para a compreensão humana (BERNERS-LEE; HENDLER; LASSILA, 2001).

A segunda versão da Web (Web 2.0), também chamada de Web Social por causa do seu contraste com a Web 1.0, permite que seu conteúdo seja publicado de forma mais rápida por seus usuários e a inteligência coletiva dos usuários tem encorajado o uso mais democrático dessa tecnologia. Originalmente a WWW era usada para compartilhar ideias e promover discussões em comunidades científicas. A Web 2.0 propôs modificações na maneira que a WWW estava sendo utilizada e aplicou em áreas como saúde e educação, melhorando a interação entre os usuários (BOULOS; WHEELER, 2007).

A Web Semântica ou Web 3.0(MARKOFFNOV, 2006), cujo termo surgiu pela primeira vez em (BERNERS-LEE; HENDLER; LASSILA, 2001), é um complemento a Web atual, em que a informação é dada com um significado bem definido, i.e., as informações são apresentadas de maneira estruturada ou semi-estruturada, facilitando a interpretação e interação de máquinas com a Web. Desse modo, a Web semântica permite que ambos humanos e máquinas trabalhem em cooperação (BERNERS-LEE; HENDLER; LASSILA, 2001).

Em (HARMELEN, 2004), os autores apresentam duas motivações para a adoção da Web Semântica: (i) integração dos dados, fator que

quando ausente limita o desempenho de diversas aplicações; (ii) suporte mais inteligente e personalizado para o usuário final, permitindo, por exemplo, que aplicações utilizem informações previamente obtidas para recomendar novas páginas, forneçam uma experiência mais precisa e completa na busca de informações, personalização dos resultados de busca, entre outras melhorias.

2.1.1 Representação de Informação na Web Semântica

Uma das contribuições mais relevantes da área de gerenciamento de inteligência da informação é a exploração da Web como uma plataforma de dados e informações integradas, principalmente para buscas (NGOMO et al., 2014). Para atingir esse objetivo diversas tecnologias foram propostas, de forma a representar as informações existentes em páginas da Web de maneira estruturada e processável por máquinas.

Em (BERNERS-LEE; HENDLER; LASSILA, 2001), o autor cita um importante padrão para a Web Semântica: o *Resource Description Framework* (RDF). De acordo com a *World-Wide Web Consortium*[1], principal organização de padronização da Web, RDF é um *framework* usado para representação de informação na Web (WOOD; LANTHALER; CYGANIAK, 2014). O RDF codifica a informação em formato de triplas e cada tripla é representada no formato *(sujeito-predicado-objeto)*, no exemplo ilustrado na Figura 1 a relação entre `http://example.com/hary` e `foaf:Person` é feita por meio da propriedade `rdf:type`; neste cenário, `http://example.com/hary` representa o sujeito, `rdf:type` é o predicado e o objeto é `foaf:Person`.

Na Web Semântica, as estruturas que definem os conceitos e relacionamentos usados para descrever e representar um determinado interesse são chamadas de vocabulários (W3C..., 2015b). Vocabulários são usados para classificar termos que podem ser usados em uma aplicação/domínio, caracterizar relacionamentos e definir restrições de uso desses termos. Portanto, vocabulários são a base para a construção de técnicas para inferência na Web Semântica.

Segundo a (W3C..., 2015b), há uma divisão bem clara entre o que é referenciado com "vocabulário"e "ontologia". Popularmente o termo "ontologia"vem sendo para descrever vocabulários complexos e com descrição formal dos termos, já o termo "vocabulário"tem sido utilizado quando o formalismo não necessário.

Para facilitar a integração e compartilhamento de dados com

[1] https://www.w3.org

semânticas bem definidas, diversas ontologias e vocabulários são utilizados em conjunto com o RDF. Um exemplo de vocabulário RDF é o projeto *Friend of a Friend (FOAF)* (BRICKLEY; MILLER, 2007), que descreve pessoas, suas atividades e seus relacionamentos com outras pessoas ou objetos usando o formato RDF. A Figura 1, foi baseada no exemplo de RDF utilizando o dicionário FOAF retirado do site *www.xml.com*[2], ilustra um modelo RDF, usando o vocabulário FOAF para definir uma amostra de rede com relacionamentos baseado na relação de "conhecimento"(`foaf:knows`). Assim, observamos que a entidade *Peter Parker* conhece as entidades *Harry Osborn* e *Aunt May*, e ainda, a entidade *Harry Osborn* possui ligação a um outro documento RDF com maiores informações sobre a própria, cuja semântica é expressa através do relacionamento `foaf:seeAlso`. Esta propriedade liga um significado bem definido entre a base disponível com mais detalhes em <http://www.osborn.com/harry.rdf> e o documento local sobre *Harry Osborn*.

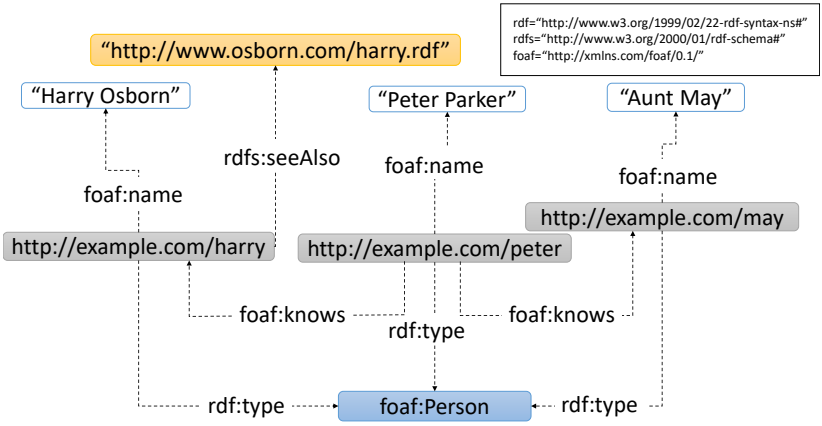


Figura 1 – Exemplo de grafo RDF, com dados de fontes distintas ligados via predicados de diferentes vocabulários (rdf, rdfs e foaf).

[2]https://www.xml.com/pub/a/2004/02/04/foaf.html, visitado em Abril de 2017.

2.1.2 Dados Ligados

O termo Dado Ligado (*Linked Data*) faz referência a um conjunto de boas práticas para publicar e conectar dados estruturados na Web. Tais práticas permitem que dados de diferentes fontes na Web sejam interligados de forma padronizada, utilizando ontologias e vocabulários conhecidos. Segundo (W3C..., 2015a), para criar e navegar entre Dados Ligados, os dados deverão estar disponíveis em um formato comum, por exemplo RDF, que se encontra descrito na subseção 2.1.1.

Tim Berners-Lee (BERNERS-LEE et al., 2006) cita um conjunto de regras para publicar dados na Web, de maneira que os dados publicados façam parte de um mesmo espaço global (BERNERS-LEE, 2006). As regras são apresentadas a seguir:

1. Utilizar *Universal Resource Identifier (URI)* para identificação de recursos (e.g., pessoas, loja, objetos);
2. Usar HTTP URIs de forma que seja possível encontrar os recursos referenciados na Web;
3. Ao ser acessados através da URI, devem estar disponíveis diversas informações do recurso acessado, utilizando padrões já estabelecidos (e.g., RDF, OWL, SPARQL);
4. Incluir ligações, quando possível, para outras URIs, para que seja possível descobrir mais informações.

A *Wikidata* (VRANDEcIć; KRöTZSCH, 2014) e a *DBpedia* (AUER et al., 2007) são casos de repositórios de Dados Ligados que usam RDF em conjunto com *Ontology Web Language*[3] (OWL) para publicação de dados na Web. Na *DBpedia* cada dado possui diversos relacionamentos, sendo um deles uma URI que aponta para definições dos recursos em páginas da *Wikipedia*.

2.1.3 Anotação Semântica

Anotações semânticas são documentos que ligam dados, ou partes de dados, à recursos semânticos em bases de conhecimento, sendo perfeitamente processáveis por máquinas (UREN et al., 2006).

O processo de anotação semântica associa recursos com semânticas bem definidas (e.g., conceitos e instâncias de conceitos de uma base

[3] https://www.w3.org/OWL/

de conhecimento) a um dado qualquer a ser anotado (e.g., documentos de textos, imagens, áudios, vídeos). Quando um recurso é semanticamente anotado, ele é enriquecido com informações que podem ser interpretadas, combinadas e usadas por diversas aplicações.

Na Figura 2 é ilustrado como exemplo, um *tweet* semanticamente anotado com recursos da base de conhecimento *DBpedia*. No texto do *tweet* foram identificados as menções "`laptop`"e "`Dell #xps13`"que se referem, respectivamente, ao conceito "`Laptop`[4]"e a entidade "`Dell XPS`[5]". Uma das possíveis formas de identificar e desambiguar menções presentes em textos é através da comparação da palavra com o nome de superfície (*surface name*) dos recursos. O nome de superfície compreende os possíveis nomes e apelidos que podem ser usados para identificar um recurso. O reconhecimento e desambiguação de menções encontradas no texto não serão discutidas profundamente neste trabalho, para fins de delimitação de escopo.

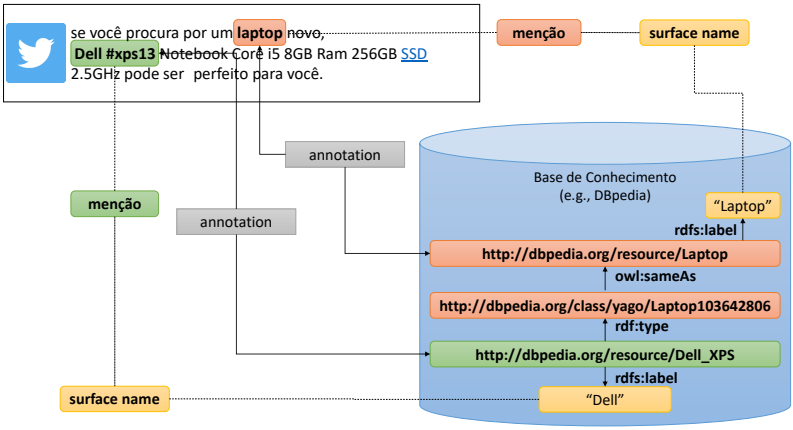


Figura 2 – Exemplo de anotações semânticas de menções de interesse para e-business encontradas em um tweet.

Uma anotação semântica pode ser feita de forma manual ou através de ferramentas, que efetuam de modo automatizado ou semi-automatizado (baseada em sugestões). Na anotação manual, o usuário explicitamente anota um determinado dado, especificando qual recurso

[4]http://dbpdia.org/page/Laptop
[5]http://dpbedia.org/page/Dell_XPS

em uma base de conhecimento melhor descreve o dado (ou parte dele) a ser anotado. O processo de anotação semi-automatizado tem como objetivo o "aprendizado de regras", automatizando a criação de padrões de extração de documentos previamente marcados ou semiestruturados (NOBATA; SEKINE, 1999). Por fim, o processo de anotação semântica automatizado, pode ser feito usando técnicas de aprendizado de máquina (TANG et al., 2012). Atualmente, uma das ferramentas mais proeminentes para anotação semântica de dados textuais é a *DBpedia-Spotlight*.

2.2 MODELAGEM DIMENSIONAL

A técnica de Modelagem Dimensional é utilizada em arquiteturas para Armazém de Dados, mais conhecido como *Data Warehouses (DW)*, que provê suporte a processos de apoio à tomada de decisão e gestão. Segundo (KIMBALL; ROSS, 2013), a modelagem dimensional tem sido usada como técnica dominante para apresentação de DW. Isto ocorre devido que a apresentação dos dados deve ser fundamentada principalmente na simplicidade, sendo esta uma característica da modelagem dimensional.

Modelos dimensionais, os quais representam dados de um determinado *DW*, podem ser implementados em sistemas de gerenciamento de banco de dados (SGBDs) relacionais, sendo, nestes casos, chamados de esquemas em estrela (KIMBALL; ROSS, 2013). Um esquema em estrela é ilustrado na Figura 3.

O esquema em estrela é formado por uma tabela chamada **fato** e várias tabelas **dimensionais**. A tabela fato armazena as medidas de desempenho resultantes dos eventos do processo de negócio, i.e., armazena dados quantitativos. As tabelas dimensionais complementam à tabela fato, possuindo o contexto textual associado a um evento de medição do processo de negócios, i.e., armazena os dados qualitativos.

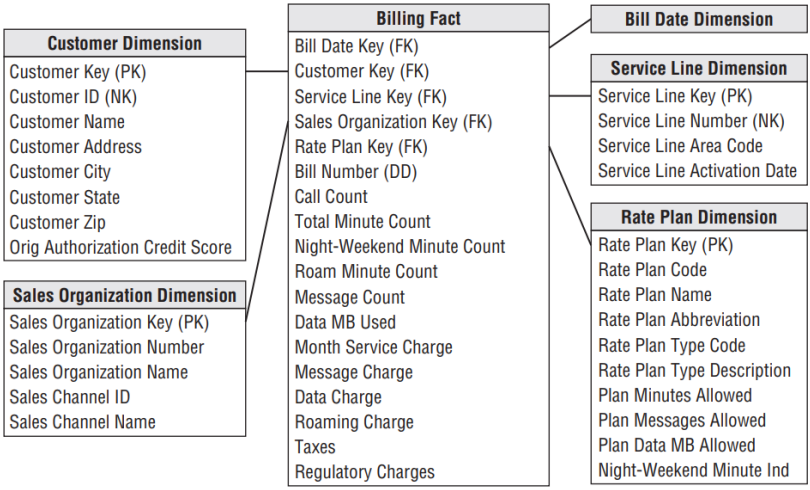


Figura 3 – Exemplo de um esquema estrela com medidas de faturamento. **Fonte:** (KIMBALL; ROSS, 2013)

3 PROCESSO PARA FILTRAR MENÇÕES DE INTERESSE

Para fazer uso do conjunto de dados disponíveis em mídias sociais é necessário extrair de textos as informações de forma estruturada e resolver problemas semânticos, tais como ambiguidade. Este trabalho apresenta uma proposta para a construção de um processo para filtragem de menções à *e-Business* em *tweets*.

Durante o desenvolvimento do trabalho, foi identificado que uma sequência de passos ocorre para gerar hierarquias com informações para o domínio de interesse. Este processo propõe-se a buscar, em textos não estruturados, menções a objetos que representem recursos (e.g. classes ou instâncias) relacionadas a entidades de *e-Business* e construir hierarquias usando os recursos anotados.

A Figura 4 ilustra um diagrama de alto nível do processo de enriquecimento semântico proposto. No processo são usadas como entradas um conjunto de *tweets*, um Grafo de Conhecimentos (e.g. *DBPedia*) e uma ontologia de domínio de alto nível (e.g. *Good Relations*). A saída é uma hierarquia filtrada apenas por instâncias e conceitos definidos na Ontologia de Domínio Alto Nível.

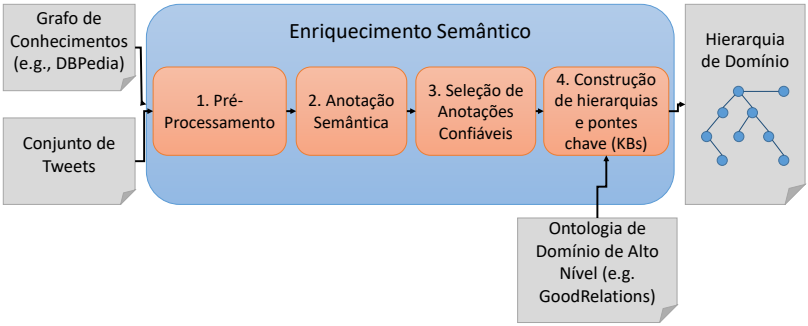


Figura 4 – Processo de Enriquecimento Semântico.

3.1 DEFINIÇÕES BÁSICAS

Definição 1 *Siglas que serão usadas para próximas definições*

t	*tweet*
idT	*identificador único de t*
sa	*anotação semântica de uma menção em t*
m	*menção de t que é alvo (target) da sa*
KG	*Grafo de Conhecimentos (Knowledge Graph - KG)*

Para a descrição do processo são necessárias as definições abaixo apresentadas, extraídas do trabalho de (JÚNIOR et al., 2018).

Um Grafo de Conhecimentos $KG(V, E)$ é definido pelo par de conjuntos V e E onde cada vértice $r \in V$ representa um recurso do KG e cada aresta $(r, r', \rho) \in E \subseteq V \times V \times R$ representa uma ligação em KG de r e r', indicando que ele está conectado pela relação $\rho \in R$.

Definição 2 *Dado um grafo de conhecimentos $KG(V, E)$ e um tweet t, uma* ***anotação semântica*** *de t, é uma tupla*

$$sa = (idT, m, r)$$

, onde:

r $\in V$ *é o recurso de KG que é o resultado da sa*

Toda anotação semântica usada neste trabalho faz a ligação entre uma menção marcada em texto de *tweet* para um recurso $r \in V$ da KG, como especificado na Definição 2. Contudo, o recurso r pode ser usado para semanticamente descrever a menção.

As menções de interesse em *tweet* podem ser mensuradas de acordo com suas anotações semânticas, assim como as classes dessas anotações. Chamamos de *hit* direto para um recurso r no KG, uma anotação semântica que tem como resultado o próprio r conforme a Definição 3. É possível visualizar essa informação na menção para `laptop` anotada semanticamente no *tweet* localizado na parte inferior esquerda da Figura 7.

Definição 3 *Dado um grafo de conhecimento $KG(V, E)$ e um tweet t, um* ***hit direto*** *de uma anotação semântica em t, é o par*

$$(sa, r)$$

, onde:

r *é recurso resultante da sa e* $r \in V$

Uma ferramenta de marcação, como *DBpedia-Spotlight*, gera anotações semânticas que podem trazer como resultados classes ou instâncias, porque seu processo de anotação faz correspondência de menções

marcadas em textos com o nome de superfície de classes e instâncias (MENDES et al., 2011) e, portanto, *hits* diretos dessas anotações podem ser classes ou instâncias. Além do *hit* direto (associação direta de menções em *tweets* com recursos contidos no KG, via anotação semântica), dois tipos de *hits* indiretos são relevantes para a análise das menções semanticamente anotadas em *tweets*:

(i) ***hit* indireto via `rdf:type`**, é o recurso no qual a menção no texto de *tweet* semanticamente anotada, faz ligação através da propriedade *rdf:type*. A Figura 7 permite a visualização de um ***hit* indireto via `rdf:type`**, através da menção #**dell** no *tweet* anotado semanticamente, dessa forma, podemos dizer que o recurso `Dell` é uma instância da classe `Company`, ou seja, a menção `Dell` anotada no *tweet* gera um ***hit* indireto via `rdf:type`** para o recurso `Company`, conforme especificado na Definição 4.

(ii) ***hit* indireto via `rdfs:subClassOf`**, ocorre quando uma classe $c\prime$ do KG é superclasse da classe c com um *hit* direto ou indireto via `rdf:type`, de acordo com a especificação da Definição 5. Na Figura 7 é possível observar que a menção ao recurso `laptop`, semanticamente anotada no *tweet*, gera um ***hit* indireto via `rdfs:subClassOf`** para o recurso `dbo:Device` ($c\prime$), através do *hit* indireto via `rdf:type` do recurso `Information Appliance` (c).

Definição 4 *Dado a grafo de conhecimento $KG(V, E)$, um tweet t e um hit direto (sa, r) de uma anotação semântica em t, um* ***hit indireto via `rdf:type`*** *de t em uma classe $r' \in V$ é uma tripla*

$$(sa, r, r')$$

, onde:

$r \in V$; $\boldsymbol{(r, r', \texttt{rdf:type}) \in E}$ *e r é uma instância da classe r' no KG*

Definição 5 *Dado um grafo de conhecimento $KG(V, E)$ e um tweet t, existe um* ***hit indireto via `rdfs:subClassOf`*** *se e somente se $\exists(r', r'', \texttt{rdfs:subClassOf}) \in E$, com r' e $r'' \in V$, tal que r' é uma subclasse de r'' e existe um hit direto do recurso r' anotado semanticamente em t ou um hit indireto via `rdf:type`* *de t em uma classe r'.*

3.2 ENRIQUECIMENTO SEMÂNTICO

O processo de construção da hierarquia semântica é realizado por meio de 4 etapas de processamento. São estas: pré-processamento (descrito na subseção 3.2.1), anotação semântica e seleção de anotações confiáveis (subseção 3.2.2), e construção de hierarquias e criação de pontes para ontologia de domínio de alto nível (subseção 3.2.3).

3.2.1 Pré-processamento

No contexto de processamento de dados, o pré-processamento é a etapa inicial e fundamental para garantir a acuracidade dos resultados. A omissão deste passo incorrerá em resultados errôneos (e.g., combinação impossível de dados, dados incompletos, etc.). Dados reais frequentemente apresentam os seguintes tipos de problemas: dados incompletos (e.g., abreviações), dados com ruídos (e.g., erros de ortografia e gramática, gírias, etc.) e inconsistência de dados (KOTSIANTIS; KANELLOPOULOS; PINTELAS, 2006).

Existem diversas técnicas para resolver os problemas mais recorrentes, dentre elas as mais utilizadas se baseiam na limpeza de dados, remoção de pontuação e caracteres especiais; redução dos dados, minimizar a quantidade de atributos relevantes e/ou discrepantes para os experimentos; integração de dados, utilizar dados de diferentes fontes para maior confiabilidade (KOTSIANTIS; KANELLOPOULOS; PINTELAS, 2006).

Posto que este trabalho utiliza informações textuais oriundas do *Twitter* (*tweets*) com dados não estruturados com palavras de natureza informal e possivelmente com diversos caracteres especiais, tais como, "|"(pipe), *emoticons*[1], esta etapa exigiu o uso de algumas técnicas de pré-processamento, tais como: técnica de limpeza de dados e redução para remover os *emoticons*, *emojis*[2], URLs, *ReTweets (RT)* e caracteres especiais.

3.2.2 Anotação Semântica e Seleção de Anotações Confiáveis

Na etapa de anotação semântica (etapa 2 do processo) é usado o serviço Web da *DBpedia-Spotlight*, que provê uma interface programá-

[1] https://pt.wikipedia.org/wiki/Emoticon
[2] https://pt.wikipedia.org/wiki/Emoji

tica para a fase de *spotting* (fase de reconhecimento dos recursos que deverão ser anotados) e desambiguação (DAIBER et al., 2013), através de um servidor Web REST. Neste processo são feitas múltiplas requisições ao servidor *DBpedia-Spotlight*, que retorna os recursos anotados em um *tweet*, conforme ilustrado na Figura 2.

Na área de Ciência da Computação e Análise de Dados, a técnica de deduplicação de dados é usada para eliminar cópia e dados repetidos. Essa técnica contribui para a melhoria na utilização de mecanismos de armazenamento, reduzindo a quantidade de *bytes* que necessitam ser transferidos em uma rede de computadores e em análise de dados, especificamente, é muito usado para evitar dados duplicados e aumentar a precisão dos resultados (DUTCH, 2008). Para a etapa de Seleção de Anotações Confiáveis, foi utilizada a técnica de deduplicação para prover um resultado mais confiável e evitar falsos positivos e/ou negativos, tais como, classes com *nomes de superfície* semelhantes, mas não equivalentes, ou até mesmo, classes com *nomes de superfície* bem distintos, porém equivalentes.

A lógica de deduplicação dos recursos anotados na etapa 3.2.2 foi feita usando as propriedades ontológicas `owl:sameAs`, que determina se dois indivíduos ou instâncias são iguais e `owl:equivalentClass`, semelhante à propriedade `owl:sameAs`, porém é usada para determinar se existe uma relação de equivalência entre classes (DEAN; SCHREIBER, 2004). Ainda nesse passo, foi necessário definir uma prioridade na escolha da base de dados que será escolhida na deduplicação os recursos anotados. Uma vez que *DBpedia* e a *schema.org* atualmente são as bases de dados de maior relevância para este estudo, a *DBpedia* foi definida como sendo mais prioritária, seguida pela *schema.org*.

A Figura 5 ilustra o relacionamento entre as classes `Organisation` da *DBpedia* e `Organization` da *schema.org* através da propriedade `owl:equivalentClass`. Neste caso, após o processo de deduplicação, o recurso `Organisation` da *DBpedia* é escolhido e o recurso equivalente `Organization` da *schema.org* é descartado, da mesma forma ocorre com a Instância `Dell` e `Q30873` ligadas pela propriedade `owl:sameAs`. Empregando esse procedimento nos resultados da Anotação Semântica, é considerado apenas um dentre os recursos conectados através dessas propriedades.

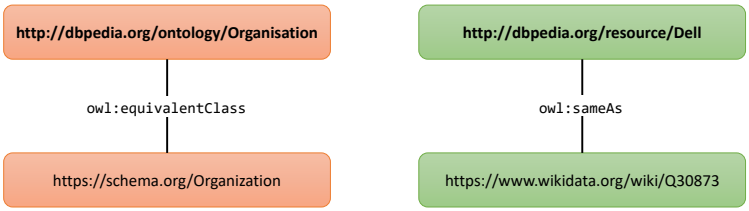


Figura 5 – Relacionamento entre classes usando `owl:equivalentClass` e entre instâncias usando `owl:sameAs`.

3.2.3 Construção de Hierarquias Semânticas - Semantic Hierarchy (SH)

A etapa de Construção de Hierarquias Semânticas é uma adaptação do modelo proposto por (SACENTI et al., 2015) e tem como objetivo compor hierarquias de recursos conectados aos recursos eleitos através do passo anterior (subseção 3.2.2).

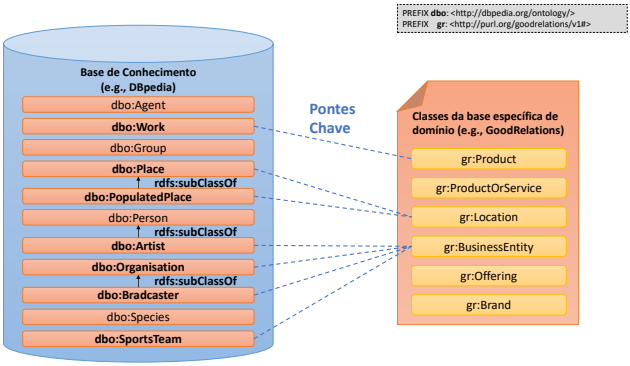


Figura 6 – Exemplo de pontes entre classes da *DBpedia* e classes da ontologia de alto nível.

Os *hits* são usados para identificar os recursos de maior interesse entre as anotações efetuadas. Os *hits* diretos, são mapeamento di-

reto (i) e induzem acesso por meio cadeias de mapeamento indireto (ii) (i.e., são extraídas por meio de conexões de exploração de cadeias) à suas classes e superclasses. Neste trabalho é considerado apenas o impacto indireto por um tipo de instância para sua classe na *DBpedia*. Usando sua classe, é feita a navegação usando o relacionamento *rdfs:subClassOf* e acumulando os *hits* indiretos por *rdfs:subClassOf*. Dessa forma, são usados os recursos anotados para navegar na *DBpedia* usando a propriedade `rdf:type` para encontrar a classe correspondente à instância anotada e consequentemente achar suas superclasses navegando pela propriedade `rdfs:subClassOf` até sua raiz *owl:Thing*. Ainda nesta etapa, durante a navegação entre os recursos da *DBpedia* é efetuada a contabilização dos *hits* diretos e indiretos para compor a Hierarquia final, com os recursos alcançados.

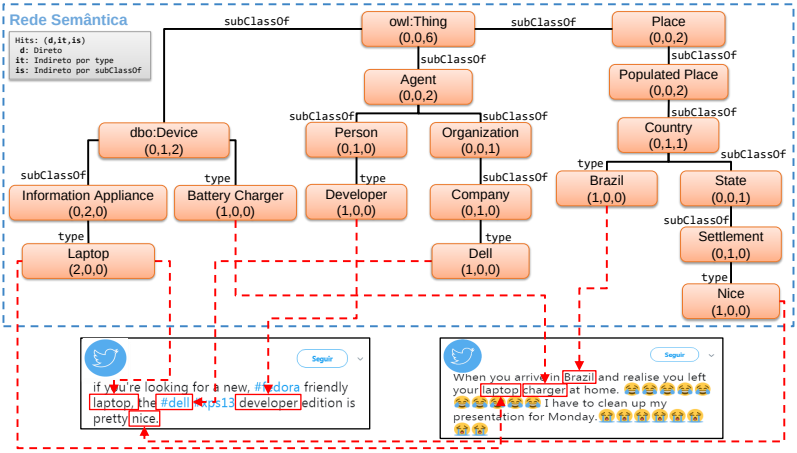


Figura 7 – Hits de classes da *DBpedia* por menções semanticamente anotadas em *tweets* e organizadas hierarquicamente

Para cada recurso de anotação confiável é construída uma cadeia de recursos que os liga por meio de conexões de mapeamento direto (i) à cadeias de mapeamento indireto (ii), que são extraídas por meio de conexões de exploração de cadeias.

(i) **mapeamento direto**, são recursos semanticamente anotados em textos de *tweets*, ou seja, possuem conexão direta para um Grafo de Conhecimento (*Knowledge Graph (KG)*)

(ii) **mapeamento indireto**, são recursos alcançados por meio de conexões de exploração de cadeias usando um relacionamento de ordem, tais como, `rdf:type` e `rdfs:subClassOf`

Durante este processo também são filtradas as hierarquias pertencentes às classes pontes para a ontologia de alto nível. A Figura 6 ilustra um exemplo de mapeamento usado para filtrar as classes de LOD para ontologias de alto nível.

A Figura 7 ilustra um fragmento do Grafo gerado após a conclusão da etapa descrita na subseção 3.2.3. Nesse grafo há uma organização hierárquica em forma de árvore apresentando, nas folhas, as menções de texto anotadas nos *tweets*. Nos vértices do grafo estão *Label* (ou nome de superfície) seguido por uma tupla que identifica os *hits* entre os parênteses, conforme apresentado na legenda: o primeiro elemento identifica a quantidade de *hits* diretos, o segundo a quantidade de *hits* indiretos por `rdf:type` e o terceiro elemento identifica a quantidade de *hits* indiretos por `rdfs:subClassOf`.

4 EXPERIMENTOS

Este capítulo apresenta os experimentos executados para a realização da investigação proposta neste trabalho. Na seção 4.1 são descritos os materiais e métodos utilizados para realização dos experimentos. A seção 4.2 apresenta os resultados da pesquisa e a na seção 4.3 são discutidos os resultados dos experimentos.

4.1 MATERIAIS E MÉTODOS UTILIZADOS

O *dataset* utilizado nos experimentos foi disponibilizado pelo laboratório LISA (*dataset BR-2015*), com aproximadamente 100.000 *tweets* postados entre 30 de Novembro de 2015 e 15 de Dezembro de 2015, com seu conteúdo em português. A pesquisa contou com a colaboração de três pesquisadores do Departamento de Informática Empresarial da Universidade de Leipzig, com experiência em *Customer Relationship Management* (CRM) Social que construíram, em consenso, 30 pontes para os experimentos. A construção das pontes foi feita de forma ordenada com prioridade pelas classes com um maior número de *hits*.

4.1.1 Ferramentas para Anotação Semântica de Dados

Existem diferentes tipos de ferramentas para anotação semântica de dados e com diferentes abordagens para processamento de linguagem natural. As três ferramentas mais proeminentes são:

(i) FOX, um *framework* para extração de conhecimento que combina uma coleção de *frameworks* NER (SPECK; NGOMO, 2014b) e utiliza técnicas de *ensemble learning* (SPECK; NGOMO, 2014a; DIETTERICH, 2002). O *framework* pode ser usado via *RESTful*[1] *Web Service* ou via *binds* em linguagens de programação (e.g., java e python);

(ii) Babelfy, que tem em seu processo dois importantes passos: *Entity Link (EL)* e *Word Sense Disambiguation (WSD)* (MORO; RAGANATO; NAVIGLI, 2014). A WSD é a tarefa que escolhe o sentido correto de uma palavra dentro de um determinado contexto. O EL descobre menções de entidades dentro de um texto e os liga

[1] https://en.wikipedia.org/wiki/Representational_state_transfer

à entrada mais adequada em uma base de conhecimento de referência.

(iii) *DBpedia-Spotlight*, uma ferramenta de código fonte aberto, usada para automatizar a anotação semântica de menções a recursos da *DBpedia* em textos não estruturados e usa princípios LOD para fazer a ligação entre os dados da *DBpedia*. A ferramenta disponibiliza três formas de acesso: (i) por meio de uma aplicação WEB, onde é possível inserir o texto a ser anotado e as configurações (e.g. nível de confiança, linguagem da anotação) para executar a anotação; (ii) por meio de API's (*Application Programming Interfaces*) escritas em linguagens de programação *Java* e *Scala* e; (iii) por um *Web Service REST*, que suporta saídas nos formatos *text/html* e *application/json*.

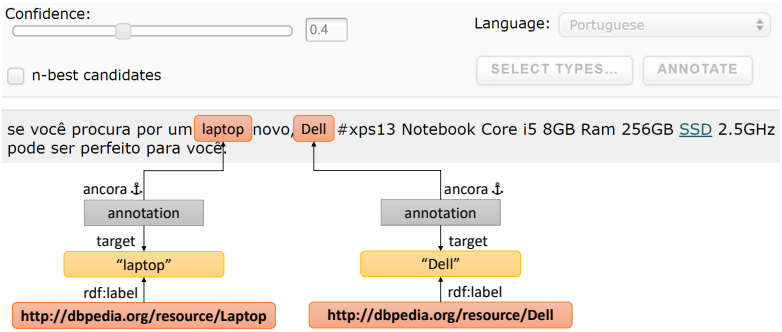


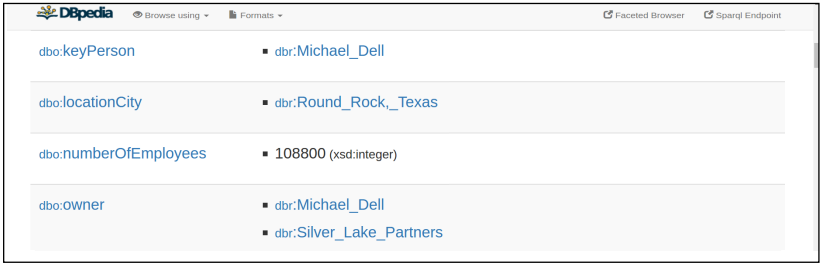
Figura 8 – Anotação de menções em tweet usando a ferramenta DBpedia-Spotlight.

A ferramenta escolhida para a execução dos experimentos deste trabalho foi a *DBpedia-Spotlight*. Tal escolha está embasada nos seguintes critérios: (i) apresentar bons resultados: dado um *tweet* é possível avaliar as menções anotadas de maneira rápida usando a interface WEB; (ii) possuir flexibilidade: permite que se configure um nível de confiança para o reconhecimento das entidades no texto; (iii) facilidade de uso: possui um serviço REST que permite automatizar o processo

de anotação semântica dos dados; (iv) sustentar um serviço estável. A Figura 8 ilustra a anotação de um *tweet* usando a interface WEB, que tem como resultado dois recursos anotados e seus respectivos links que apontam para as definições na *DBpedia*. Ainda é possível notar algumas configurações usadas para executar as anotações, tais como o nível de confiança no canto superior esquerdo da imagem e a linguagem utilizada como base para anotação. Neste caso, foram usados nível de confiança de 40% e a linguagem Português como base para a anotação semântica.

A *DBpedia* também possui uma interface WEB que permite visualizar as definições de um determinado recurso. Neste caso, a representação WEB de um documento no formato RDF com a descrição do recurso, suas propriedade e relacionamentos. Este documento possui informações adicionais ao recurso anotado, como por exemplo, o tipo de entidade, ilustrado na Figura 9.a, e algumas propriedades (*namespace* `dbo`), ligando este recurso a outros recursos (*namespace* `dbr`), ilustrado pela Figura 9.b. Dessa forma, podemos visualizar que a propriedade `dbo:owner` liga a instância `Dell` à instância `dbr:Michael_Dell`, que possui informações sobre o presidente e fundador da empresa Dell e muitas outras informações sobre o recurso **Dell**.

(a) Informações do tipo de entidade e descrição do dado anotado



(b) Propriedades e relacionamentos da anotação "Dell"

Figura 9 – Algumas propriedades do recurso *Dell*.

4.1.2 Filtragem e Concepção das Hierarquias

Os primeiros experimentos foram feitos de forma manual, realizando consultas na plataforma web *DBpedia-Spotlight* conforme ilustrado na Figura 8, inserindo textos, coletando as anotações e comparando os resultados das entidades anotadas com as classes da *Good Relations Ontology* (GRO), ontologia que descreve entidades e relacionamentos para negócios (*e-Business*) (HEPP, 2008). De acordo com o site do projeto[2], essa ontologia tem sido usada por grandes empresas, tais como *Google*, *Yahoo*, *BestBuy*, *Sears*, *Rakuten*, entre muitas outras. As classes mais relevantes da GRO consideradas pelos especialistas de negócio para análise da informação foram:

- `gr:BusinessEntity`: representa um agente, empresa ou indivíduo de negócio
- `gr:Offering`: representa uma venda ou oferta
- `gr:ProductOrService`: identifica produtos ou serviços
- `gr:Location`: identifica o local de uma loja ou oferta disponível

Com os resultados das anotações manuais foi possível traçar uma forma de identificar *tweets* que fazem menção a alguma entidade de *e-Business*. Essa abordagem foi essencial para a compreensão e organização das etapas subsequentes e permitiu aos especialistas de negócio mapear as menções anotadas à classes equivalentes na GRO. A análise iniciou-se com as classes de *LOD* mais mencionadas diretas e indiretamente através da ligação com classes por meio das propriedades `rdf:type` e `rdfs:subClassOf` nas anotações semânticas.

A Tabela 1 exibe uma amostra das classes da *DBpedia* que possuem maior incidência nas anotações, mapeadas pelos especialistas de domínio *(Key Bridges)*.

Os experimentos preliminares permitiram compreender o problema, conduzir e preparar a base de dados e conhecimento e as ferramentas de anotação para os experimentos. Para a realização dos experimentos em uma base de dados maior (*dataset* BR-2015), foi necessário criar um software para automatizar o pré-processamento (seção 3.2.1) dos textos e a anotação semântica dos dados (seção 3.2.1). O software foi construído utilizando a linguagem de programação *Python*[3] junto

[2] http://www.heppnetz.de/projects/goodrelations/
[3] https://docs.python.org/3.7/

com *RDFLib*[4], que implementa suporte para manipulação de documentos em RDF, e *SPARQLWrapper*[5] que fornece um cliente para efetuar buscas em *SPARQL*. A abordagem é executada em duas etapas: a primeira recebe como entrada um arquivo contendo um documento texto com um conjuntos de *tweets* organizado com um único *tweet* por linha, aplica a etapa 1, o pré-processamento (subseção 3.2.1), e, para cada *tweet*, faz chamadas a API REST da *DBpedia-Spotlight* para obter as anotações com os seguinte parâmetros:

- `confidence`: 0.5
- `language`: pt
- `text`: texto do tweet

Ainda, o software armazena o resultado original retornado pela da API em um banco de dados *MongoDB*[6].

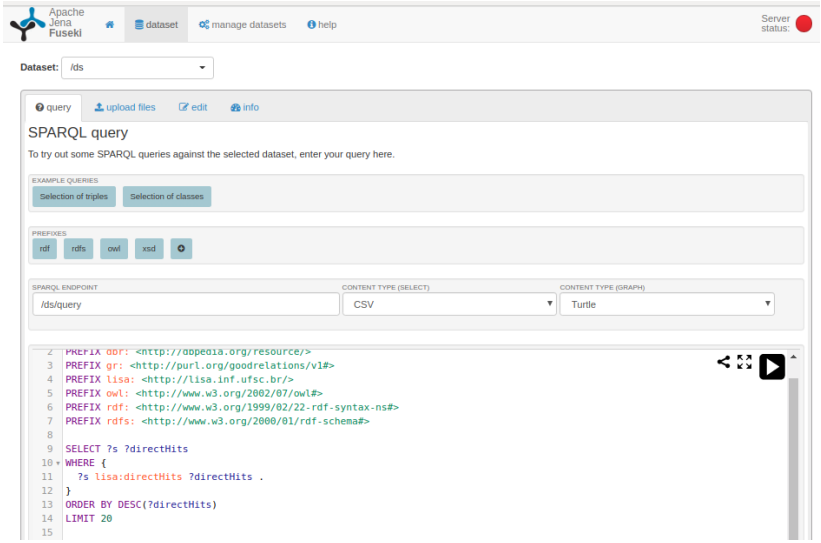


Figura 10 – Servidor *Apache Jena Fuseki* usado para auxiliar na filtragem usando as pontes para ontologia de específica de domínio.

[4]http://rdflib.readthedocs.io
[5]https://pypi.org/project/SPARQLWrapper/
[6]https://www.mongodb.com/

Na segunda etapa, o software desenvolvido faz uso das anotações armazenadas no *MongoDB* (etapa anterior) para construção das hierarquias, percorrendo o KG e contabilizando os *hits* diretos e indiretos, usando as relações `rdf:type` e `rdfs:subClassOf`. Conforme é percorrido o Gráfico de Conhecimento (Knowledge Graph - *KG*), as classes e recursos alcançados, assim como suas relações (e.g., `rdf:type`, `rdfs:subClassOf`), são armazenadas em um *Triple Data Base (TDB)*.

Ao final da execução destes passos do software, foi inicializado um servidor *Apache Jena Fuseki*[7] que faz uso dos dados armazenados no banco TDB para análise e filtragem das instâncias e classes, usando SPARQL.

Os experimentos rodaram em máquinas com dois perfis diferentes de configuração:

- Intel(R) Core(TM) i5-4570 CPU @ 3.20GHz, com 7846 MiB de memória e sistema operacional Manjaro Linux.
- Intel(R) Xeon(R) Platinum 8175M CPU @ 2.50GHz, com 3979 MiB de memória e sistema operacional Ubuntu 18.04.2 LTS (GNU/-Linux 4.15.0-1039-aws x86/64)

[7] https://jena.apache.org/documentation/fuseki2/

Tabela 1 – Pontes-chave (*KB*) elaboradas por especialistas de domínio para as 20 classes da DBpedia mais mencionadas nos *tweets* semanticamente anotados

#	**Classe da DBpedia**	**Possível Classe da GRO**
1	`dbo:Agent`	Mais geral que as classes da GRO
2	`dbo:Work`	`gr:Product`
3	`dbo:Group`	Não pertence à GRO; Em geral relevante para *e-Business*
4	`dbo:Place`	`gr:Location`
5	`dbo:Person`	Não pertence à GRO; Em geral relevante para *e-Business*
6	`dbo:PopulatedPlace`	`gr:Location`
7	`dbo:Artist`	Não é uma classe da GRO (Assim como `dbo:Person`)
8	`dbo:Organisation`	`gr:BusinessEntity`
9	`dbo:Broadcaster`	`gr:BusinessEntity`
10	`dbo:Species`	Não é uma classe da GRO; Relevância depende do domínio
11	`dbo:SportsTeam`	`gr:BusinessEntity`
12	`dbo:Eukaryote`	Não é uma classe da GRO; Relevância depende do domínio
13	`dbo:Politician`	Não é uma classe da GRO
14	`dbo:Website`	Não é classe da GRO; Em geral relevante para *e-Business*
15	`dbo:Settlement`	`gr:Location`
16	`dbo:Genre`	Não é uma classe da GRO
17	`dbo:Topical`	Não é uma classe da GRO
18	`dbo:MusicalWork`	`gr:ProductOrService`
19	`dbo:Animal`	`gr:ProductOrService`
20	`dbo:Plant`	Não é uma classe da GRO
21	`dbo:Band`	`gr:BusinessEntity`

4.2 RESULTADOS

Esta seção apresenta os resultados experimentais frutos desta pesquisa. Todos os resultados abaixo apresentados foram realizados utilizando os procedimentos apontados na seção 4.1 e usando como entrada de dados o *dataset BR-2015*.

Tabela 2 – Top-10 instâncias com maior número de menções diretas à *DBpedia*, antes da realização dos filtros, ordenadas de forma descendente pelo número de *hits* diretos

Instância	Hits Diretos
`dbr:Deus`	1037
`dbr:One_Direction`	843
`dbr:Dilma_Rousseff`	752
`dbr:Brasil`	658
`dbr:Lista_de_personagens_de_Kingdom_Hearts`	547
`dbr:Twitter`	516
`dbr:Impeachment`	503
`dbr:Saudade`	417
`dbr:Ku_Klux_Klan`	412
`dbr:YouTube`	410

Para a apresentação de alguns resultados foram omitidas as classes mais genéricas como `dbo:Agent`, por não possuir uma ponte para a ontologia específica de domínio, segundo a análise feita pelos especialistas de domínio (GRO).

A Tabela 2 aponta os recursos com maiores incidências entre os recursos anotados, organizados em ordem, do recurso com maior incidência para o recurso com menor incidência na tabela. De maneira semelhante, a Tabela 3 apresenta os recursos com maiores incidências, após a realização dos filtros usando a ontologia específica para o domínio de negócio (GRO).

Ao todo, foram avaliados 35795 recursos mencionados diretamente nos *tweets* contidos no *dataset BR-2015*. Desses 35795 recursos, 22726 possuem pontes para ontologia de domínio de negócio (GRO) de forma indireta, por meio de ligações indiretas via `rdf:type` e `rdf:subClassOf`, com classes da DBpedia Ontology (DBO).

O processo de contabilização dos *hits* praticado segue uma ordem ascendente, partindo dos recursos mais específicos (recursos semanticamente anotados) para os recursos mais genéricos (e.g. `dbo:Agent`,

Tabela 3 – Top-10 instâncias com maior número de menções diretas à *DBpedia*, após a realização dos filtros, ordenadas de forma descendente pelo número de *hits* diretos

Ponte (GRO)	**Instância**	**#Hits Diretos**
gr:BusinessEntity	dbr:Esporte_Interativo	5
gr:ProductOrService	dbr:Diário_Oficial_da_União	5
gr:Product	dbr:Death_Note	4
gr:Product	dbr:Naruto	4
gr:Product	dbr:One_Piece	4
gr:Product	dbr:Tokyo_Ghoul	4
gr:BusinessEntity	dbr:Air_France	3
gr:BusinessEntity	dbr:Avianca_Holdings	3
gr:BusinessEntity	dbr:Bayer_04_Leverkusen	3
gr:BusinessEntity	dbr:BTG_Pactual	3

`owl:Thing`).

A Figura 11 apresenta um *ranking* das 20 classes com maior número de *hits* nas anotações semânticas geradas pela *DBpedia-Spotlight* para o *dataset* BR-2015. A classificação é ordenada em número decrescente de ocorrências acumuladas. Na figura é possível notar classes como `dbo:Organisation` ou `dbo:SportsTeam`, que possuem pontes para `gr:BusinessEntity`, indicando a possibilidade de haver algum interesse de negócio, i.e., algum artigo esportivo de um clube para a venda ou até mesmo um torcedor interessado na compra.

O resultado do processo de Construção das Hierarquias Semânticas (subseção 3.2.3) apresentada na Figura 12 aponta um resultado satisfatório para um experimento inicial. O gráfico mostra a proporção de classes e instâncias da *DBpedia* anotadas e que possuem pontes para GRO. Cerca de 61.3% das instâncias anotadas possuem ponte para ontologia de domínio, enquanto a cobertura das classes coletadas da *DBpedia*, 39.2% possuem pontes. Se fizermos um comparativo com a quantidade de classes que a *DBpedia* possui (685 classes[8]), neste experimento foi possível avaliar 186 classes da *DBpedia*, o que equivale a aproximadamente 27% de classes da DBO.

Por outro lado, o gráfico apresentado na Figura 13 demonstra a distribuição da proporção entre as instância/recursos anotados e a ontologia de domínio, e sugere que aproximadamente 40% das menções em *tweets* do conjunto de dados avaliado refletem algum tipo de

[8]https://wiki.dbpedia.org/services-resources/ontology

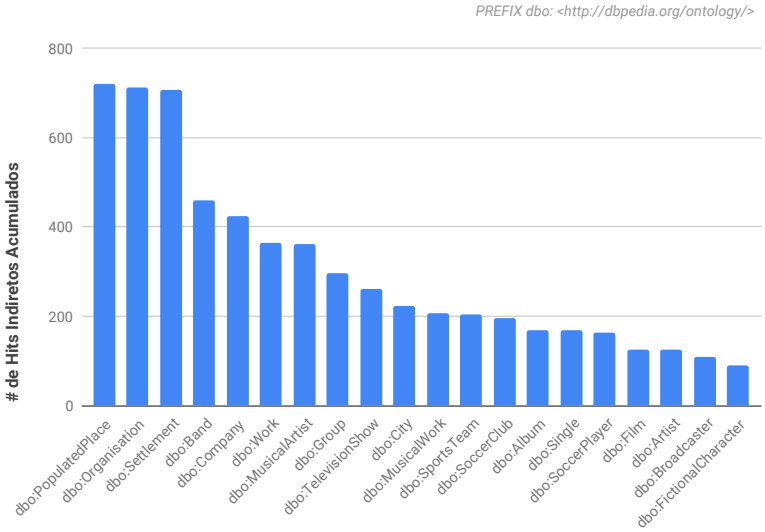


Figura 11 – Top 20 classes mais mencionadas nas anotações do *dataset* BR-2015, após o filtro

interesse em produtos e/ou serviços, pouco mais de 39% refletem interesse de negócio, uma menção a alguma entidade de negócio, tais como, emissoras de TV (`dbr:MTV`), canais de TV (`dbr:SportTV`), entre outros. Apesar de `gr:BusinessEntity` não representar um interesse específico e/ou direto para alguma entidade de negócio, pode ser usado em CRM Social para estreitar relação entre consumidor e empresa, capturar *leads*, pessoas que podem se tornar oportunidades de negócios reais através de marketing direcionado, entre outras oportunidades, assim como `gr:Location`, que pode corresponder a localização de uma empresa, loja, assim como o interesse de uma pessoa em uma viagem para uma localização específica.

Os dados apresentados na Tabela 4 apresentam os resultados da mesma busca efetuada para gerar a Figura 11, porém os resultados contidos na tabela foram gerados após o processo de Filtragem e Concepção das Hierarquias (descritos na subseção 4.1.2). A tabela está ordenada de forma descendente com as classes que possuem um maior número de *hits* acumulados (mais mencionadas) no topo da tabela. Além disso, é possível identificar as pontes usadas como para a ontologia de do-

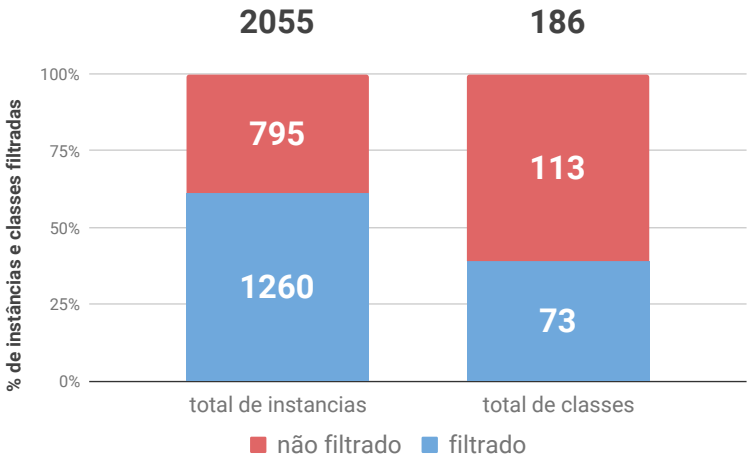


Figura 12 – Quantidade de instâncias e classes filtradas usando pontes para ontologia de alto nível

mínio de negócio (GRO), assim como a presença de novas classes da *DBpedia* no gráfico, tais como `dbo:MusicalArtist` que possui ponte para `gr:BusinessEntity` e como o próprio nome sugere, trata-se de menções a cantores, intérpretes, etc.

A Figura 14 ilustra um esquema dimensional para análise de menções de interesse em textos apresentado por (JÚNIOR et al., 2018) e demonstra uma possibilidade de aplicação para *Data Warehouse Semânticos (DWS)*. No esquema é apresentado uma tabela **fato** *tweetMeasures* com medidas mensuráveis para *#tweets*, *#users* e *#mentions*. O modelo segue a notação de (GOLFARELLI; MAIO; RIZZI, 1998) e apresenta dimensões espaço-temporais comuns em *Data Warehouse (DW)*, `Time` e `Place` e que podem ser preenchidas com os resultados das pontes para `gr:Location` ou por metadados oriundos de *tweets*, tais como data e hora, e geolocalização. As dimensões `Product or Service` e e `Business Entity` podem ser carregadas com as informações obtidas pelo processo de construção da SH. `Product or Service` pode agregar os resultados das pontes `gr:Products` e `gr:ProductsOrServices`, como por exemplo `dbo:Single` e `dbo:Album` que, de acordo com a Tabela 4 está ligado às duas pontes. E por sua vez, a dimensão `Business Entity` pode ser carregada os recursos filtrados usando a ponte para

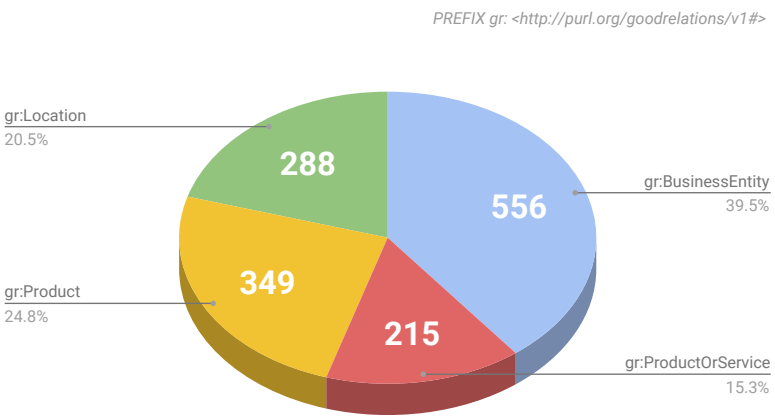


Figura 13 – Distribuição da quantidade de instâncias mencionadas e filtradas com as pontes para a ontologia de alto nível

`gr:BusinessEntity`, *e.g.,* `dbo:MusicalArtist`, `dbo:SoccerClub`, `dbo:TelevisionStation`, entre outros.

Não foi possível apresentar um estudo que incorporasse sistemas de recomendação. Isso ocorre devido ao conjunto de dados utilizados neste trabalho. O conjunto de dados possui apenas textos de *tweets*, sem quaisquer informações dos usuários que postaram os *tweets*, como um id ou localização, os quais são essenciais para sistemas de recomendação. No entanto, o autor deste trabalho acredita que o uso do modelo em sistemas de recomendação seja factível, uma vez que se tenha a identificação dos usuários.

Tabela 4 – Top 20 classes com maior número de *Hits* (diretos + indiretos), filtradas usando as pontes para ontologia de alto nível e ordenados de forma descendente pelo número de *hits*

Pontes (GRO)	**Classes (DBO)**	**Hits**
gr:Location	dbo:PopulatedPlace	480
gr:Location	dbo:Settlement	470
gr:BusinessEntity	dbo:MusicalArtist	181
gr:Location	dbo:City	148
gr:BusinessEntity	dbo:Company	141
gr:Product	dbo:TelevisionShow	130
gr:BusinessEntity	dbo:Band	115
gr:Product	dbo:MusicalWork	103
gr:BusinessEntity	dbo:Group	99
gr:BusinessEntity	dbo:SoccerClub	98
gr:BusinessEntity	dbo:SportsTeam	68
gr:Product	dbo:Film	63
gr:Product	dbo:Album	56
gr:Product	dbo:Single	56
gr:ProductOrService	dbo:Album	56
gr:ProductOrService	dbo:Single	56
gr:Product	dbo:WrittenWork	43
gr:BusinessEntity	dbo:Broadcaster	36
gr:Location	dbo:Country	36
gr:BusinessEntity	dbo:TelevisionStation	32

4.3 DISCUSSÃO

Durante o desenvolvimento do trabalho, foi identificado que uma sequência de passos ocorre e propor um processo denominado Enriquecimento Semântico, apresentado no capítulo 3. Já na fase de Análise dos Resultados foi possível constatar que `dbr:Deus` é a instância com maior número de menções, seguidas por `dbr:One_Direction`, banda britânica, `dbr:Dilma_Rousseff` e `dbr:Brasil`. Na lista também consta entre os dez recursos mais anotados, o recurso `dbr:Impeachment`, os textos explorados na rede social relata o momento político vivido no Brasil no período entre 30/11/2015 e 11/12/2015 e evidenciado nos `dbr:tweets`. Nos resultados do conjunto de dados explorado também é possível ver que existe uma predominância das classes de negócio

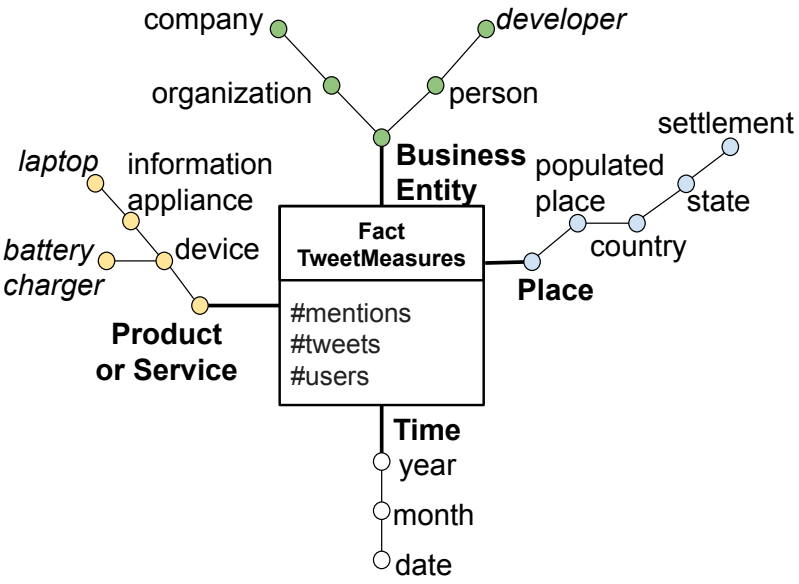


Figura 14 – Modelagem Semântica Dimensional proposta por Júnior *et al.* (JÚNIOR et al., 2018)

`dbo:MusicalArtist`, `dbo:Company` e `dbo:TelevisionShow` analisadas em tweets enviados do Brasil.

Abaixo são sintetizados alguns dados observados durante o experimento:

- Ao total foram registrados 35795 menções diretas a um recurso/instância da *DBpedia*, das 35795 menções anotadas, 22726 possuem pontes para GRO, ou seja, em torno de 63% dos recursos/instâncias que foram mencionadas nos *tweets* foram filtradas usando as pontes para ontologia de alto nível
- Foram anotadas 2055 Instâncias diferentes da DBpedia (dbr), das quais 1260 foram filtradas usando as pontes para a ontologia de domínio
- A *DBpedia Ontology (DBO)* possui 685 classes definidas em sua ontologia, neste experimento foi possível alcançar, por meio da navegação por `rdf:type` e `rdfs:subClassOf`, 186 classes da DBO,

ou 27% de cobertura da ontologia, e ainda, de 186 classes foi factível criar 73 associações para ontologia de domínio de negócio (*e-Business*).

Por fim, os resultados ilustram que é possível criar um DWS através de menções feitas em *tweets* e semanticamente anotadas e filtrados usando uma ontologia de alto nível que descreve semanticamente os interesses que se pretende filtrar, tal como a GRO, usada neste trabalho. E ainda, que em mídias sociais podem existir inúmeras possibilidades de uso, pode ser usado para análise estatística para identificação de oferta e demanda de produtos, tendências e padrões comportamentais e/ou tópicos emocionais.

5 TRABALHOS RELACIONADOS

Este capítulo apresenta os trabalhos correlatos à nossa proposta, selecionados a partir de uma revisão bibliográfica sistemática baseada nas diretrizes de (KITCHENHAM; CHARTERS, 2007). Esta revisão bibliográfica foi realizada em diversas fontes usando a seguinte expressão de consulta:

(semantic annotation OR annotation)
AND (social media OR tweets OR posts)
AND (analysis OR filtering)

Foram encontrados 40 artigos relacionados à expressão de consulta. Após a primeira filtragem, através da leitura e análise dos resumos e estruturas dos trabalhos, foram considerados apenas 7 artigos. Os artigos resultaram no conjunto de trabalhos relacionados a este, descritos e comparados no restante deste capítulo, seguindo os seguintes critérios:

- Postagens de mídias sociais *(Social Media Posts - SMP)*: origem do conjunto de dados. *Datasets* de textos originados de mídias sociais (*e.g. tweets*) são vastos e possuem *APIs* públicas para coleta. Os valores possíveis para os trabalhos correlatos são: usa (OK) ou não usa (NOK);
- Anotação Semântica *(Semantic Annotation - SA)*: abordagem faz uso (OK) ou não (NOK) técnicas de anotação semântica para enriquecimento dos textos de mídias sociais.
- Método de Anotação *(Annotation Method - AM)*: técnica utilizada para anotação (manual: apenas com conhecimento de especialistas; semiautomático: regras e ontologias de especialistas; automático: somente ontologias e base de conhecimento);
- Customizável *(Customizable - C)*: permite personalização para um domínio específico;
- Construção Automática de Hierarquias *(Automatic Build Hierarchies - ABH)*: a abordagem gera dimensões com hierarquias de propriedades de ordem parcial disponíveis em ontologias e *LODs*.

O trabalho de Abrahams (ABRAHAMS et al., 2012) propõe um processo que aplica técnicas de mineração de textos *(text mining)* em

postagens de fóruns de discussão on-line, procurando características de veículos para alimentar sistemas de recomendação. Em uma abordagem semelhante (VILLANUEVA et al., 2016), anota semanticamente os *tweets* para extrair informações para fins de recomendação.

Outras propostas constroem e preenchem um modelo dimensional para analisar de textos. Nebot *et al.* (NEBOT; BERLANGA, 2012) analisa registros de pacientes e médicos anotados manualmente usando o modelo dimensional. Nebot (NEBOT; BERLANGA, 2012) propõe um processo para preencher um DW a partir de dados semânticos, mas não é personalizável e não usa texto de postagens em mídias sociais.

O trabalho de Fileto *et al.* (FILETO et al., 2015) propõe um modelo ontológico e um processo para estruturar e enriquecer semanticamente dados de movimento em vários níveis de abstração, para apoiar a análise dos esquemas dimensionais enriquecidos como proposto em (FILETO et al., 2014). Já o trabalho de Sacenti *et al.* (SACENTI et al., 2015) propõe um método para construir dimensões para analisar postagens de mídias sociais semanticamente enriquecidos com LOD. Eles fazem isso adaptando hierarquias de classes e instâncias existentes em coleções de LOD de acordo com suas incidências como valores de anotação de determinados conjuntos de dados. Essas duas obras são semelhantes à nossa abordagem, entretanto não constroem as dimensões de análise para domínios de aplicação específicos que podem ser alterados no início do processo. Para isso, a abordagem proposta neste trabalho usa o conceito de pontes (já apresentado anteriormente). Inicialmente, os conceitos mais mencionados da Hierarquia gerada são verificados por grupo pequeno de especialistas de domínio, que selecionam as classes ou instâncias de fato relevantes e as associam uma a uma com conceitos de uma ontologia de domínio de alto nível escolhida, formando um conjunto inicial de pontes-chave KBs. O processo segue com a construção da SH associando as pontes chaves através da relação de equivalência (`owl:equivalentClass`). Tais pontes orientam a seleção de anotações relevantes e a adaptação de hierarquias de LOD usados nas anotações e podem servir como dimensões de análise de dados. O resultado final é uma hierarquia semântica construída especialmente para o domínio determinado pelos especialistas. Esta etapa de construção de pontes com a SH diferencia este trabalho dos dois trabalhos analisados ((FILETO et al., 2015), (SACENTI et al., 2015)).

Finalmente, a abordagem interativa EXODuS proposta em (CHOUDER; RIZZI; CHALAL, 2019) permite consultas OLAP exploratórias em bases NoSQL orientadas a documentos. Em tal abordagem, hierarquias são construídas mediante um método baseado em mineração de depen-

dências funcionais aproximadas entre elementos de documentos JSON. Tais hierarquias são montadas incrementalmente em porções envolvidas nas consultas multidimensionais à medida em que tais consultas são submetidas pelos usuários, de modo a conferir melhor desempenho. As consultas OLAP expressas sobre um modelo dimensional com hierarquias dinâmicas são traduzidas para a linguagem de consulta do MongoDB. A abordagem EXODuS é avaliada com dados da NBA (*National Basketball Association*), da DBLP (*DataBase systems and Logic Programming bibliography*) e *tweets*. Todavia, como as dimensões dos cubos de dados produzidos pela abordagem EXODuS são baseadas em elementos JSON (na maioria metadados), diferentemente do nosso trabalho, EXODuS não permite efetuar consultas de acordo com conceitos e instâncias mencionadas em textos, tais como os conteúdos textuais de *tweets* e artigos.

A Tabela 3 resume as características dos trabalhos relacionados selecionados, com base nos seis critérios de comparação descritos no início deste capítulo. A última linha da Tabela 3 refere-se à nossa proposta. Ela usa algumas ideias de trabalhos relacionados selecionados, como a exploração de anotações semânticas e hierarquias presentes em coleção de LOD. No entanto, este trabalho é o único que apresenta um processo geral que emprega e filtra as anotações semânticas de oriundas de mídias sociais (*e.g., tweet*) que sejam de interesse para um domínio de aplicação utilizando pontes para uma ontologia de alto nível para tal domínio.

Tabela 5 – Comparação de trabalhos correlatos

Trabalho	**SMP**	**SA**	**C**	**AM**	**ABH**	**Saída**
Abrahams(2012)	NOK	OK	NOK	Manual	NOK	Recomendações
Nebot(2012)	NOK	OK	NOK	Manual	OK	Cubo dimensional populado
Fileto(2015)	OK	OK	OK	Automático	NOK	Dados semânticos de movimento
Sacenti(2015)	OK	OK	OK	Automático	OK	Cubo dimensional não-populado
Villanueva(2016)	OK	OK	NOK	Semi-Automático	NOK	Recomendações
Chouder(2019)	OK	NOK	OK	NOK	OK	Cubo dimensional populado
Junior *et al.*(2018)	**OK**	**OK**	**OK**	**Automático**	**OK**	**Cubo dimensional populado**
Este trabalho	**OK**	**OK**	**OK**	**Automático**	**OK**	**Hierarquia Semântica**

SMP: Postagens de mídias sociais; **SA**: Anotação Semântica;
C: Customizável; **AM**: Método de Anotação; **ABH**: Construção Automática de Hierarquias

6 CONCLUSÕES E TRABALHOS FUTUROS

As estruturas e organizações de redes sociais tem variado de nome e formato, mas a maneira de como os dados são expostos segue possibilitando a exploração dos dados uso comercial e/ou fins científicos. Neste trabalho foi proposto um processo para detectar e avaliar menções a entidades de negócio (*business*) em textos de postagem de mídias sociais, tais como Twitter, utilizando ferramentas da Web Semântica para efetuar anotações semânticas em textos de classes de LOD e construir hierarquias utilizando pontes para conceitos de alto nível em uma ontologia de domínio. Para processo proposto é necessário que exista uma ontologia de domínio, neste trabalho foi usado a *Good Relations (GRO)* que define um vocabulário para o *e-commerce* e uma base de dados ligados (LOD), neste caso a DBpedia. Porém, esse processo permite que seja substituído qualquer uma das ferramentas sejam substituídas por outras que executem de maneira semelhante o mesmo passo que se deseja substituir, por exemplo, neste trabalho foi usada *DBpedia-Spotlight* como ferramenta de anotação semântica de dados para anotar semanticamente os recursos (*e.g.* conceitos e instâncias) de Grafos de Conhecimento *Knowledge Graph*.

Experimentos com *tweets* semanticamente enriquecidos com recursos da *DBpedia* usando a ferramenta *DBpedia-Spotlight* revelam que algumas pontes-chaves de classes da *DBpedia* para classes de alto nível da *Good Relations Ontology* são suficientes para verificar a consistência dessas pontes e derivar um número considerável de novas pontes consistentes. Essas pontes permitiram a detecção de menções de interesse para o domínio de negócios (*business*) em geral, determinando uma das principais contribuições deste trabalho.

Os experimentos feitos com *tweets* semanticamente anotados usando a ferramenta *DBpedia-Spotlight* mostraram que as 32 classes da *DBpedia* mapeadas para as classes da ontologia de domínio *Good Relations Ontology (GRO)*, foram suficientes para avaliar a consistência das pontes e obter os resultados esperados.

A principal dificuldade enfrentada neste trabalho foi a geração das anotações semânticas dos *tweets*, a ferramenta escolhida para a geração das anotações semânticas dos textos de *tweets* foi a *DBpedia-Spotlight*, que inicialmente apresentou um ótimo serviço e resultados satisfatórios. Contudo, a ferramenta começou a ficar intermitente e por diversos dias o serviço não funcionava o que retardou alguns experimentos e não permitiu a geração de novos.

Este trabalhos tem como principais contribuições:

1. Elaboração do processo para análise de menções a entidades de negócio (*business*) em *tweets*, montagem e filtragem das hierarquias com base em pontes para ontologias de domínio de *e-Business*

2. Criação do software (*semantic-hierarchy*[1]) para automação do processo de anotação semântica de textos (*tweets*) e construção das hierarquias

3. Apresentação dos resultados da pesquisa bem como dos experimentos efetuados. A proporção do mapeamento entre recursos semanticamente anotados de LOD da *DBpedia* e as classes da ontologia de domínio GRO

Não era o foco desta pesquisa, mas acabou se fazendo necessário a utilização de um servidor que armazena as informações geradas pelo processo na construção das hierarquias em um *Triple Data Base (TDB)* e disponibiliza uma interface WEB para análise/consultas à base de dados usando a linguagem *SPARQL*.

Dentre os principais trabalhos futuros vale a pena mencionar:

1. Repetir os experimentos com outras ferramentas de anotação semântica, tais como Babelfy (MORO; RAGANATO; NAVIGLI, 2014) e *Federated Knowledge Extraction Framework (FOX)* (SPECK; NGOMO, 2014b) (SPECK; NGOMO, 2014a)

2. A *Good Relations Ontology (GRO)*(HEPP, 2008) está atualmente sendo integrada à *Schema.org*[2], a qual está presente em boa parte do comércio eletrônico mundial e é mais aderente a este domínio que a DBpedia. Assim, seguindo sugestões do acadêmico autor deste trabalho, pretendemos efetuar estudos e novos experimentos com classes de valores de anotações semânticas referentes à *Schema.org* no lugar das classes da *DBepedia*.

[1] https://gitlab.com/souzawillian/semantic-hierarchy
[2] http://wiki.goodrelations-vocabulary.org/Cookbook/Schema.org

REFERENCIAS

ABRAHAMS, A. S. et al. Vehicle defect discovery from social media. *Decision Support Systems*, Elsevier, v. 54, n. 1, p. 87–97, 2012.

ASUR, S.; HUBERMAN, B. A. Predicting the future with social media. In: *IEEE/WIC/ACM: International Conference on Web Intelligence and Intelligent Agent Technology*. [S.l.: s.n.], 2010.

AUER, S. et al. Dbpedia: A nucleus for a web of open data. In: *The semantic web*. [S.l.]: Springer, 2007. p. 722–735.

BERNERS-LEE, T. *Linked Data*. 2006. <https://www.w3.org/DesignIssues/LinkedData.html>. Acessado em 22-06-2017.

BERNERS-LEE, T.; HENDLER, J.; LASSILA, O. The semantic web: A new form of web content that is meaningful to computers will unleash a revolution of new possibilities. *Scientific American*, p. 34–43, 2001.

BERNERS-LEE, T. et al. Self-describing delegation networks for the web. In: *Policies for Distributed Systems and Networks. Seventh IEEE International Workshop*. London, Ont., Canada: [s.n.], 2006.

BOULOS, M. N. K.; WHEELER, S. The emerging web 2.0 social software: an enabling suite of sociable technologies in health and health care education 1. *Health Information & Libraries Journal*, Wiley Online Library, v. 24, n. 1, p. 2–23, 2007.

BRICKLEY, D.; MILLER, L. *The Friend Of A Friend (FOAF) Vocabulary Specification*. November 2007. <http://xmlns.com/foaf/spec/>.

CHOUDER, M. L.; RIZZI, S.; CHALAL, R. Exodus: Exploratory olap over document stores. *Information Systems*, v. 79, p. 44 – 57, 2019. ISSN 0306-4379. Special issue on DOLAP 2017: Design, Optimization, Languages and Analytical Processing of Big Data. <http://www.sciencedirect.com/science/article/pii/S0306437917304507>.

DAIBER, J. et al. Improving efficiency and accuracy in multilingual entity extraction. In: *Proceedings of the 9th International Conference on Semantic Systems (I-Semantics)*. [S.l.: s.n.], 2013.

DEAN, M.; SCHREIBER, G. *OWL Web Ontology Language Reference*. [S.l.], fev. 2004. Http://www.w3.org/TR/2004/REC-owl-ref-20040210/.

DIETTERICH, T. G. Ensemble learning. *The handbook of brain theory and neural networks*, MIT Press: Cambridge, MA, v. 2, p. 110–125, 2002.

DUTCH, M. Understanding data deduplication ratios. In: *SNIA Data Management Forum*. [S.l.: s.n.], 2008. p. 7.

FILETO, R. et al. The baquara 2 knowledge-based framework for semantic enrichment and analysis of movement data. *Data & Knowledge Engineering*, Elsevier, v. 98, p. 104–122, 2015.

FILETO, R. et al. A semantic model for movement data warehouses. In: *Proceedings of the 17th International Workshop on Data Warehousing and OLAP, DOLAP 2014, Shanghai, China, November 3-7, 2014*. [s.n.], 2014. p. 47–56. <http://doi.acm.org/10.1145/2666158.2666180>.

FILETO, R. et al. Análise de métodos e ferramentas para reconhecimento de palavras relevantes em microblogs. In: *XII Brazilian Symposium on Information Systems*. Florianópolis, SC, Brasil: [s.n.], 2016. p. 345–352.

GOLFARELLI, M.; MAIO, D.; RIZZI, S. The dimensional fact model: A conceptual model for data warehouses. *International Journal of Cooperative Information Systems*, World Scientific, v. 7, n. 02n03, p. 215–247, 1998.

HABIB, M. B.; KEULEN, M. V. Information extraction for social media. In: *In Proceedings of the Third Workshop on Semantic Web and Information Extraction (SWAIE)*. Dublin, Ireland: [s.n.], 2014. W14-62, p. 9–16.

HARMELEN, F. van. A numerical model for thin airfoils in unsteady motion. *IEEE Distributed Systems Online 1541-4922*, v. 5, n. 3, 2004.

HEATH, D.; SINGH, R.; GANESH, J. Social media at sociosystems inc.: A socio-technical systems analysis of strategic action. In: *47th Hawaii International Conference on System Science*. Waikoloa, HI, USA: [s.n.], 2014. p. 584–593.

HEPP, M. Goodrelations: An ontology for describing products and services offers on the web. *Knowledge Engineering: Practice and Patterns*, Springer, p. 329–346, 2008.

JÚNIOR, V. C. P. et al. A semantic BI process for detecting and analyzing mentions of interest for a domain in tweets. In: *Proceedings of the 24th Brazilian Symposium on Multimedia and the Web, WebMedia 2018, Salvador-BA, Brazil, October 16-19, 2018*. [s.n.], 2018. p. 197–204. <http://doi.acm.org/10.1145/3243082.3243100>.

KIMBALL, R.; ROSS, M. *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling*. 3rd. ed. [S.l.]: Wiley Publishing, 2013. ISBN 1118530802, 9781118530801.

KITCHENHAM, B.; CHARTERS, S. *Guidelines for performing Systematic Literature Reviews in Software Engineering*. [S.l.], 2007. <http://www.dur.ac.uk/ebse/resources/Systematic-reviews-5-8.pdf>.

KLEIN, D. Estudo de técnicas e ferramentas aplicáveis a mídias sociais para reconhecimento e desambiguação de entidades nomeadas. Universidade Federal de Santa Catarina, p. 47–48, 2015.

KOTSIANTIS, S.; KANELLOPOULOS, D.; PINTELAS, P. Data preprocessing for supervised leaning. *International Journal of Computer Science*, Citeseer, v. 1, n. 2, p. 111–117, 2006.

MARKOFFNOV, J. *Entrepreneurs See a Web Guided by Common Sense*. November 2006. Http://www.nytimes.com/2006/11/12/business/12web.html. Accessed: 19-09-2017.

MENDES, P. N. et al. Dbpedia spotlight: Shedding light on the web of documents. In: *Proceedings of the 7th International Conference on Semantic Systems (I-Semantics)*. Graz, Austria: [s.n.], 2011.

MORO, A.; RAGANATO, A.; NAVIGLI, R. Entity linking meets word sense disambiguation: a unified approach. *Transactions of the Association for Computational Linguistics*, v. 2, p. 231–244, 2014.

NAVIGLI, R.; PONZETTO, S. P. Babelnet: Building a very large multilingual semantic network. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the 48th annual meeting of the association for computational linguistics*. [S.l.], 2010. p. 216–225.

NEBOT, V.; BERLANGA, R. Building data warehouses with semantic web data. *Decision Support Systems*, Elsevier, v. 52, n. 4, p. 853–868, 2012.

NGOMO, A.-C. N. et al. Introduction to linked data and its lifecycle on the web. In: _____. *Reasoning Web. Reasoning on the Web in the Big Data Era: 10th International Summer School 2014, Athens, Greece, September 8-13, 2014. Proceedings*. Cham: Springer International Publishing, 2014. p. 1–99. ISBN 978-3-319-10587-1.

NOBATA, C.; SEKINE, S. Towards automatic acquisition of patterns for information extraction. In: *International Conference of Computer Processing of Oriental Languages*. [S.l.: s.n.], 1999. v. 108.

SACENTI, J. A. et al. Automatically tailoring semantics-enabled dimensions for movement data warehouses. In: SPRINGER. *International Conference on Big Data Analytics and Knowledge Discovery*. [S.l.], 2015. p. 205–216.

SPECK, R.; NGOMO, A.-C. N. Ensemble learning for named entity recognition. In: SPRINGER. *International semantic web conference*. [S.l.], 2014. p. 519–534.

SPECK, R.; NGOMO, A.-C. N. Named entity recognition using fox. In: CEUR-WS. ORG. *Proceedings of the 2014 International Conference on Posters & Demonstrations Track-Volume 1272*. [S.l.], 2014. p. 85–88.

TANG, J. et al. Automatic semantic annotation using machine learning. In: *Machine Learning: Concepts, Methodologies, Tools and Applications*. [S.l.]: IGI Global, 2012. p. 535–578.

TUFEKCI, K. Big questions for social media big data: Representativeness, validity and other methodological pitfalls. In: *Proceedings of the 8th International AAAI Conference on Weblogs and Social Media (ICWSM)*. [S.l.: s.n.], 2014.

UREN, V. et al. Semantic annotation for knowledge management: Requirements and a survey of the state of the art. *Web Semantics: Science, Services and Agents on the World Wide Web*, v. 4, n. 1, p. 14 – 28, 2006. ISSN 1570-8268. <http://www.sciencedirect.com/science/article/pii/S1570826805000338>.

VILLANUEVA, D. et al. Smore: Towards a semantic modeling for knowledge representation on social media. *Science of Computer Programming*, Elsevier, v. 121, p. 16–33, 2016.

VRANDEcIć, D.; KRöTZSCH, M. Wikidata: A free collaborative knowledgebase. *Commun. ACM*, ACM, New York, NY, USA, v. 57, n. 10, p. 78–85, set. 2014. ISSN 0001-0782. <http://doi.acm.org/10.1145/2629489>.

W3C - LINKED DATA. [S.l.], 2015. Https://www.w3.org/standards/semanticweb/data. Acessado em 22-06-2017.

W3C - VOCABULARIES. [S.l.], 2015. Https://www.w3.org/standards/semanticweb/ontology.html. Acessado em 22-06-2017.

WOOD, D.; LANTHALER, M.; CYGANIAK, R. *RDF 1.1 Concepts and Abstract Syntax*. [S.l.], fev. 2014. Http://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/.

APÊNDICE A – Artigo no Formato SBC

Investigação de Menções a Entidades de Interesse para *e-Business* em Tweets

Willian S. de Souza[1]

[1]Departamento de Informática e Estatística
Universidade Federal de Santa Catarina (UFSC)
Florianópolis – Santa Catarina – SC – Brasil

willianstosouza@gmail.com

Abstract. *The growth of social media as one of the mainstream communication forms has brought the possibility of using the open data that travels through these media in a myriad of applications, e.g., marketing and recommendation. Texts from postings in media may contain mentions to companies, products, etc. The semantic annotation tools allow you to identify such mentions and link them to resources that describe them with well-defined semantics. This work presents an investigative study of mentions of interest to entities related to the e-Business domain in social media posts and proposes a process for constructing semantic hierarchies with a domain ontology for building bridges between semantically annotated resources with Linked Open Data.*

Resumo. *O crescimento de mídias sociais como meio de comunicação trouxe a possibilidade de usar dados abertos que trafegam por essas mídias em uma miríade de aplicações, e.g., marketing e recomendação. Textos de postagens em mídias podem conter menções a empresas, produtos, etc. Ferramentas de anotação semântica permitem identificar tais menções e ligá-las a recursos que as descrevem com semântica bem definida. Este trabalho apresenta um estudo investigativo de menções a entidades relacionadas ao domínio de e-Business em postagens de mídias sociais e propõe um processo para construção de hierarquias semânticas dispondo de uma ontologia de domínio para construção de pontes entre os recursos semanticamente anotados com Linked Open Data.*

1. Introdução

O surgimento de mídias sociais teve um impacto grande nos estudos sobre o comportamento humano [Tufekci 2014]. A ascensão de tais plataformas modificou a forma como as pessoas interagem [Heath et al. 2014]. Usando mídias sociais, pessoas podem criar conteúdo, compartilhá-lo e indicar seus sentimentos em relação a eles (e.g., marcar se gostou ou não) [Asur and Huberman 2010], entre outras possibilidades, contribuindo significativamente para a geração de uma grande quantidade de dados.

Todavia, dados textuais de postagens em mídias sociais são considerados não-estruturados para fins de processamento computacional. Além disso, textos de postagens em mídias sociais são usualmente curtos (o que resulta em pouca informação de contexto) e sujeitos a ruídos (e.g., erros de ortografia e gramática, acrônimos, abreviações, gírias, etc.) e ambiguidade devido a fenômenos linguísticos (e.g., homonímia, sinonímia).

Por exemplo, postagens na rede social Twitter[1] (denominados de *tweets* no restante deste trabalho) têm natureza informal, tamanho limitado, semântica pouco definida e frequentemente muitos ruídos [Fileto et al. 2016]. Um dos principais desafios é lidar com a ambiguidade, i.e., uma palavra ou expressão pode ser usada em diferentes contextos, denotando coisas diferentes. Esses fatores dificultam a análise das postagens e demonstram a necessidade de um tratamento prévio das informações para obter êxito na análise. Assim, para fazer uso do vasto volume de dados disponíveis em mídias sociais é necessário extrair automaticamente dos textos de postagens informações mais estruturadas e semanticamente precisas.

Atualmente, a *Wikipedia*[2] é uma das principais enciclopédias digitais e é mantida por diversos contribuidores. *DBpedia*[3] [Auer et al. 2007] é uma base de conhecimentos extraídos da *Wikipedia* e disponibilizada na Web sob a forma de dados interligados abertos (do inglês *Linked Open Data - LOD*). Os padrões e diretrizes de publicação e interconexão de LOD promovem a sua integração e reuso para diversas finalidades, incluindo a anotação semântica, i.e., associação de porções de conteúdo (possivelmente não estruturados) a recursos com semântica bem definida descritos em uma base de conhecimento (e.g., *DBpedia*, *Yago*[4], *Freebase*[5]).

Este trabalho investiga menções a entidades relacionadas a *e-Business* em postagens de mídias sociais, mais especificamente *tweets*, usando a *DBpedia-Spotlight* como ferramenta de anotação e gerando hierarquias semânticas com instâncias e classes da *DBpedia*, ligadas através das propriedades `rdf:type` e `rdfs:subClassOf`, filtradas através de pontes construídas usando uma ontologia para o domínio de interesse.

1.1. Objetivos

O objetivo geral deste trabalho é verificar a incidência de menções relacionadas com *e-Business* (e.g., empresas, produtos, serviços, promoções) em *tweets*. É utilizado a *DBpedia-Spotlight* para efetuar anotações semânticas, seleção de anotações de interesse com base em pontes para uma ontologia de alto nível específica do domínio considerado e uso de software desenvolvido pelo autor para automatizar a verificação das incidências.

2. Definições Básicas

Definição 1 *Siglas que serão usadas para próximas definições*

t	*tweet*
idT	*identificador único de* t
sa	*anotação semântica de uma menção em* t
m	*menção de* t *que é alvo (target) da* sa
KG	*Grafo de Conhecimentos (Knowledge Graph - KG)*

Para a descrição do processo são necessárias as definições abaixo apresentadas, extraídas do trabalho de [Júnior et al. 2018].

[1] https://twitter.com/

[2] https://www.wikipedia.org/

[3] https://wiki.dbpedia.org/

[4] https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/

[5] https://developers.google.com/freebase/

Um Grafo de Conhecimentos $KG(V, E)$ é definido pelo par de conjuntos V e E onde cada vértice $r \in V$ representa um recurso do KG e cada aresta $(r, r', \rho) \in E \subseteq V \times V \times R$ representa uma ligação em KG de r e r', indicando que ele está conectado pela relação $\rho \in R$.

Definição 2 *Dado um grafo de conhecimentos $KG(V, E)$ e um tweet t, uma **anotação semântica** de t, é uma tupla*

$$sa = (idT, m, r)$$

, onde:

r $\in V$ *é o recurso de KG que é o resultado da sa*

Toda anotação semântica usada neste trabalho faz a ligação entre uma menção marcada em texto de *tweet* para um recurso $r \in V$ da KG, como especificado na Definição 2. Contudo, o recurso r pode ser usado para semanticamente descrever a menção.

As menções de interesse em *tweet* podem ser mensuradas de acordo com suas anotações semânticas, assim como as classes dessas anotações. Chamamos de *hit* direto para um recurso r no KG, uma anotação semântica que tem como resultado o próprio r conforme a Definição 3. É possível visualizar essa informação na menção para `laptop` anotada semanticamente no *tweet* localizado na parte inferior esquerda da Figura 3.

Definição 3 *Dado um grafo de conhecimento $KG(V, E)$ e um tweet t, um **hit direto** de uma anotação semântica em t, é o par*

$$(sa, r)$$

, onde:

r *é recurso resultante da sa e $r \in V$*

Uma ferramenta de marcação, como *DBpedia-Spotlight*, gera anotações semânticas que podem trazer como resultados classes ou instâncias, porque seu processo de anotação faz correspondência de menções marcadas em textos com o nome de superfície de classes e instâncias [Mendes et al. 2011] e, portanto, *hits* diretos dessas anotações podem ser classes ou instâncias. Além do *hit* direto (associação direta de menções em *tweets* com recursos contidos no KG, via anotação semântica), dois tipos de *hits* indiretos são relevantes para a análise das menções semanticamente anotadas em *tweets*:

(i) ***hit*** **indireto via `rdf:type`**, é o recurso no qual a menção no texto de *tweet* semanticamente anotada, faz ligação através da propriedade *rdf:type*. A Figura 3 permite a visualização de um ***hit*** **indireto via `rdf:type`**, através da menção **#dell** no *tweet* anotado semanticamente, dessa forma, podemos dizer que o recurso `Dell` é uma instância da classe `Company`, ou seja, a menção `Dell` anotada no *tweet* gera um ***hit*** **indireto via `rdf:type`** para o recurso `Company`, conforme especificado na Definição 4.

(ii) ***hit*** **indireto via `rdfs:subClassOf`**, ocorre quando uma classe $c\prime$ do KG é superclasse da classe c com um *hit* direto ou indireto via `rdf:type`, de acordo com a especificação da Definição 5. Na Figura 3 é possível observar que a menção

ao recurso `laptop`, semanticamente anotada no *tweet*, gera um ***hit* indireto via `rdfs:subClassOf`** para o recurso `dbo:Device` ($c\prime$), através do *hit* indireto via `rdf:type` do recurso `Information Appliance` (c).

Definição 4 *Dado a grafo de conhecimento $KG(V, E)$, um tweet t e um hit direto (sa, r) de uma anotação semântica em t, um* ***hit* indireto via `rdf:type`** *de t em uma classe $r' \in V$ é uma tripla*

$$(sa, r, r')$$

, onde:

$r \in V$; $(r, r', \texttt{rdf:type}) \in E$ *e r é uma instância da classe r' no KG*

Definição 5 *Dado um grafo de conhecimento $KG(V, E)$ e um tweet t, existe um* ***hit* indireto via `rdfs:subClassOf`** *se e somente se*
$\exists(r', r'', \texttt{rdfs:subClassOf}) \in E$, *com r' e $r'' \in V$, tal que r' é uma subclasse de r'' e existe um hit direto do recurso r' anotado semanticamente em t ou um hit indireto via `rdf:type` de t em uma classe r'.*

3. Processo para filtrar menções de interesse

Durante o desenvolvimento do trabalho, foi identificado que uma sequência de passos ocorre para gerar hierarquias com informações para o domínio de interesse. Este processo propõe-se a buscar, em textos não estruturados, menções a objetos que representem recursos (e.g. classes ou instâncias) relacionadas a entidades de *e-Business* e construir hierarquias usando os recursos anotados.

A Figura 1 ilustra um diagrama de alto nível do processo de enriquecimento semântico proposto. No processo são usadas como entradas um conjunto de *tweets*, um Grafo de Conhecimentos (e.g. *DBPedia*) e uma ontologia de domínio de alto nível (e.g. *Good Relations*). A saída é uma hierarquia filtrada apenas por instâncias e conceitos definidos na Ontologia de Domínio Alto Nível.

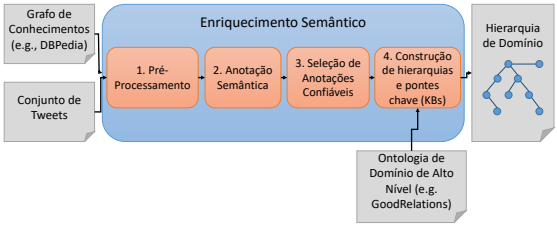


Figure 1. Processo de Enriquecimento Semântico.

O processo de construção da hierarquia semântica é realizado por meio de 4 etapas de processamento. São estas: pré-processamento (descrito na subseção 3.1), anotação semântica e seleção de anotações confiáveis (subseção 3.2), e construção de hierarquias e criação de pontes para ontologia de domínio de alto nível (subseção 3.3).

3.1. Pré-processamento

Posto que este trabalho utiliza informações textuais oriundas do *Twitter* (*tweets*) com dados não estruturados com palavras de natureza informal e possivelmente com diversos caracteres especiais, tais como, "|" (pipe), *emoticons*[6], esta etapa exigiu o uso de algumas técnicas de pré-processamento, tais como: técnica de limpeza de dados e redução para remover os *emoticons*, *emojis*[7], URLs, *ReTweets (RT)* e caracteres especiais.

3.2. Anotação Semântica e Seleção de Anotações Confiáveis

Na etapa de anotação semântica (etapa 2 do processo) é usado o serviço Web da *DBpedia-Spotlight*, que provê uma interface programática para a fase de *spotting* (fase de reconhecimento dos recursos que deverão ser anotados) e desambiguação [Daiber et al. 2013], através de um servidor Web REST. Neste processo são feitas múltiplas requisições ao servidor *DBpedia-Spotlight*, que retorna os recursos anotados em um *tweet*.

Na área de Ciência da Computação e Análise de Dados, a técnica de deduplicação de dados é usada para eliminar cópia e dados repetidos. Essa técnica contribui na redução da quantidade de *bytes* que necessitam ser transferidos em uma rede de computadores e em análise de dados, especificamente, é muito usado para evitar dados duplicados e aumentar a precisão dos resultados [Dutch 2008]. Na etapa de Seleção de Anotações Confiáveis, foi utilizada a técnica de deduplicação para prover um resultado mais confiável e evitar falsos positivos e/ou negativos. A lógica de deduplicação dos recursos anotados na etapa 3.2 foi feita usando as propriedades ontológicas `owl:sameAs`, que determina se dois indivíduos ou instâncias são iguais e `owl:equivalentClass`, semelhante à propriedade `owl:sameAs`, porém é usada para determinar se existe uma relação de equivalência entre classes [Dean and Schreiber 2004]. Ainda nesse passo, foi necessário definir uma prioridade na escolha da base de dados que será escolhida na deduplicação os recursos anotados. Uma vez que *DBpedia* e a *schema.org* atualmente são as bases de dados de maior relevância para este estudo, a *DBpedia* foi definida como sendo mais prioritária, seguida pela *schema.org*.

3.3. Construção de Hierarquias Semânticas - Semantic Hierarchy (SH)

A etapa de Construção de Hierarquias Semânticas é uma adaptação do modelo proposto por [Sacenti et al. 2015] e tem como objetivo compor hierarquias de recursos conectados aos recursos eleitos através do passo anterior (subseção 3.2).

Os *hits* são usados para identificar os recursos de maior interesse entre as anotações efetuadas. Os *hits* diretos, são mapeamento direto (i) e induzem acesso por meio cadeias de mapeamento indireto (ii) (i.e., são extraídas por meio de conexões de exploração de cadeias) à suas classes e superclasses. Neste trabalho é considerado apenas o impacto indireto por um tipo de instância para sua classe na *DBpedia*. Usando sua classe, é feita a navegação usando o relacionamento *rdfs:subClassOf* e acumulando os *hits* indiretos por *rdfs:subClassOf*. Dessa forma, são usados os recursos anotados para navegar na *DBpedia* usando a propriedade `rdf:type` para encontrar a classe correspondente à instância anotada e consequentemente achar suas superclasses navegando pela propriedade `rdfs:subClassOf` até sua raiz *owl:Thing*. Ainda nesta etapa, durante a

[6]https://pt.wikipedia.org/wiki/Emoticon
[7]https://pt.wikipedia.org/wiki/Emoji

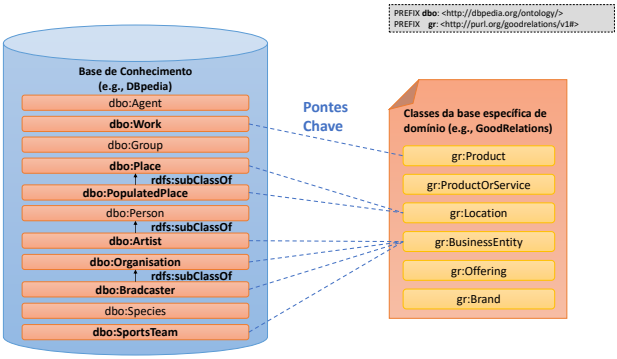


Figure 2. Exemplo de pontes entre classes da *DBpedia* e classes da ontologia de alto nível.

navegação entre os recursos da *DBpedia* é efetuada a contabilização dos *hits* diretos e indiretos para compor a Hierarquia final, com os recursos alcançados.

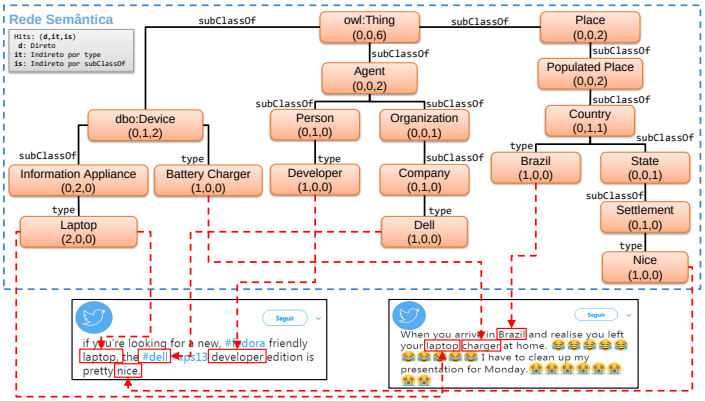


Figure 3. Hits de classes da *DBpedia* por menções semanticamente anotadas em *tweets* e organizadas hierarquicamente

Para cada recurso de anotação confiável é construída uma cadeia de recursos que os liga por meio de conexões de mapeamento direto (i) à cadeias de mapeamento indireto (ii), que são extraídas por meio de conexões de exploração de cadeias.

(i) **mapeamento direto**, são recursos semanticamente anotados em textos de *tweets*, ou seja, possuem conexão direta para um Grafo de Conhecimento (*Knowledge Graph (KG)*)

(ii) **mapeamento indireto**, são recursos alcançados por meio de conexões de exploração de cadeias usando um relacionamento de ordem, tais como,

`rdf:type` e `rdfs:subClassOf`

Durante este processo também são filtradas as hierarquias pertencentes às classes pontes para a ontologia de alto nível. A Figura 2 ilustra um exemplo de mapeamento usado para filtrar as classes de LOD para ontologias de alto nível.

A Figura 3 ilustra um fragmento do Grafo gerado após a conclusão da etapa descrita na subseção 3.3. Nesse grafo há uma organização hierárquica em forma de árvore apresentando, nas folhas, as menções de texto anotadas nos *tweets*. Nos vértices do grafo estão *Label* (ou nome de superfície) seguido por uma tupla que identifica os *hits* entre os parênteses, conforme apresentado na legenda: o primeiro elemento identifica a quantidade de *hits* diretos, o segundo a quantidade de *hits* indiretos por `rdf:type` e o terceiro elemento identifica a quantidade de *hits* indiretos por `rdfs:subClassOf`.

4. Experimento

O *dataset* utilizado nos experimentos foi disponibilizado pelo laboratório LISA (*dataset BR-2015*), com aproximadamente 100.000 *tweets* postados entre 30 de Novembro de 2015 e 15 de Dezembro de 2015, com seu conteúdo em português. A pesquisa contou com a colaboração de três pesquisadores do Departamento de Informática Empresarial da Universidade de Leipzig, com experiência em *Customer Relationship Management* (CRM) Social que construíram, em consenso, 30 pontes para os experimentos. A construção das pontes foi feita manualmente, analisando as 100 classes com maior número de *hits*.

4.1. Ferramentas para Anotação Semântica de Dados

A ferramenta escolhida para a execução dos experimentos deste trabalho foi a *DBpedia-Spotlight*. A *DBpedia-Spotlight*, uma ferramenta usada para automatizar a anotação semântica de menções a recursos da *DBpedia* em textos não estruturados e usa princípios LOD para fazer a ligação entre os dados da *DBpedia*. A ferramenta disponibiliza três formas de acesso: (i) por meio de uma aplicação WEB, onde é possível inserir o texto a ser anotado e as configurações (e.g. nível de confiança, linguagem da anotação) para executar a anotação; (ii) por meio de API's (*Application Programming Interfaces*) escritas em linguagens de programação *Java* e *Scala* e; (iii) por um *Web Service REST*, que suporta saídas nos formatos $text/html$ e $application/json$.

4.2. Filtragem e Concepção das Hierarquias

De acordo com o site do projeto[8], *Good Relation Ontology (GRO)* tem sido usada por grandes empresas, tais como *Google*, *Yahoo*, *BestBuy*, entre outras. As classes mais relevantes da GRO consideradas pelos especialistas de negócio para análise da informação foram: (i) `gr:BusinessEntity`: representa um agente, empresa ou indivíduo de negócio; (ii) `gr:Offering`: representa uma venda ou oferta; (iii) `gr:ProductOrService`: identifica produtos ou serviços; e (iv) `gr:Location`: identifica o local de uma loja ou oferta disponível.

Com os resultados das anotações manuais foi possível traçar uma forma de identificar *tweets* que fazem menção a alguma entidade de *e-Business*. Essa abordagem foi essencial para a compreensão e organização das etapas subsequentes e permitiu aos especialistas de negócio mapear as menções anotadas à classes equivalentes na GRO.

[8]http://www.heppnetz.de/projects/goodrelations/

A Tabela 1 exibe uma amostra das classes da *DBpedia* que possuem maior incidência nas anotações, mapeadas pelos especialistas de domínio *(Key Bridges)*.

Table 1. Pontes-chave (*KB*) elaboradas por especialistas de domínio para as 20 classes da DBpedia mais mencionadas nos *tweets* semanticamente anotados

#	Classe da DBpedia	Possível Classe da GRO
1	dbo:Agent	Mais geral que as classes da GRO
2	dbo:Work	gr:Product
3	dbo:Group	Não pertence à GRO; Em geral relevante para *e-Business*
4	dbo:Place	gr:Location
5	dbo:Person	Não pertence à GRO; Em geral relevante para *e-Business*
6	dbo:PopulatedPlace	gr:Location
7	dbo:Artist	Não é uma classe da GRO (Assim como dbo:Person)
8	dbo:Organisation	gr:BusinessEntity
9	dbo:Broadcaster	gr:BusinessEntity
10	dbo:Species	Não é uma classe da GRO; Relevância depende do domínio
11	dbo:SportsTeam	gr:BusinessEntity
12	dbo:Eukaryote	Não é uma classe da GRO; Relevância depende do domínio
13	dbo:Politician	Não é uma classe da GRO
14	dbo:Website	Não é classe da GRO; Em geral relevante para *e-Business*
15	dbo:Settlement	gr:Location
16	dbo:Genre	Não é uma classe da GRO
17	dbo:Topical	Não é uma classe da GRO
18	dbo:MusicalWork	gr:ProductOrService
19	dbo:Animal	gr:ProductOrService
20	dbo:Plant	Não é uma classe da GRO
21	dbo:Band	gr:BusinessEntity

Os experimentos preliminares permitiram compreender o problema, conduzir e preparar a base de dados e conhecimento e as ferramentas de anotação para os experimentos. Para a realização dos experimentos em uma base de dados maior (*dataset* BR-2015), foi necessário criar um software para automatizar o pré-processamento (seção 3.1) dos textos e a anotação semântica dos dados (seção 3.1). O software foi construído utilizando a linguagem de programação *Python*[9] junto com *RDFLib*[10], que implementa suporte para manipulação de documentos em RDF, e *SPARQLWrapper*[11] que fornece um cliente para

[9]https://docs.python.org/3.7/
[10]http://rdflib.readthedocs.io
[11]https://pypi.org/project/SPARQLWrapper/

efetuar buscas em *SPARQL*. A abordagem é executada em duas etapas: a primeira recebe como entrada um arquivo contendo um documento texto com um conjuntos de *tweets* organizado com um único *tweet* por linha, aplica a etapa 1, o pré-processamento (subseção 3.1), e, para cada *tweet*, faz chamadas a API REST da *DBpedia-Spotlight* para obter as anotações com os seguinte parâmetros:

- `confidence`: 0.5
- `language`: pt
- `text`: texto do tweet

Ainda, o software armazena o resultado original retornado pela da API em um banco de dados *MongoDB*[12].

Na segunda etapa, o software desenvolvido faz uso das anotações armazenadas no *MongoDB* (etapa anterior) para construção das hierarquias, percorrendo o KG e contabilizando os *hits* diretos e indiretos, usando as relações `rdf:type` e `rdfs:subClassOf`. Conforme é percorrido o Gráfico de Conhecimento (Knowledge Graph - *KG*), as classes e recursos alcançados, assim como suas relações (e.g., `rdf:type`, `rdfs:subClassOf`), são armazenadas em um *Triple Data Base (TDB)*.

5. Resultados

Todos os resultados abaixo apresentados foram realizados utilizando os procedimentos apontados na seção 4 e usando como entrada de dados o *dataset BR-2015*.

Table 2. Top-10 instâncias com maior número de menções diretas à *DBpedia*, antes da realização dos filtros, ordenadas de forma descendente pelo número de *hits* diretos

Instância	Hits Diretos
`dbr:Deus`	1037
`dbr:One_Direction`	843
`dbr:Dilma_Rousseff`	752
`dbr:Brasil`	658
`dbr:Lista_de_personagens_de_Kingdom_Hearts`	547
`dbr:Twitter`	516
`dbr:Impeachment`	503
`dbr:Saudade`	417
`dbr:Ku_Klux_Klan`	412
`dbr:YouTube`	410

Para a apresentação de alguns resultados foram omitidas as classes mais genéricas como `dbo:Agent`, por não possuir uma ponte para a ontologia específica de domínio, segundo a análise feita pelos especialistas de domínio (GRO).

A Tabela 2 aponta os recursos com maiores incidências entre os recursos anotados, organizados em ordem, do recurso com maior incidência para o recurso com menor incidência na tabela. De maneira semelhante, a Tabela 3 apresenta os recursos com maiores

[12]https://www.mongodb.com/

incidências, após a realização dos filtros usando a ontologia específica para o domíñio de negócio (GRO).

Ao todo, foi possível avaliar 35795 recursos mencionados diretamente nos *tweets* contidos no *dataset BR-2015*. Desses 35795 recursos mencionados diretamente, 22726 possuem pontes para ontologia de domínio de negócio (GRO) de forma indireta, por meio de ligações indiretas via `rdf:type` e `rdf:subClassOf`, com classes da DBpedia Ontology (DBO).

Table 3. Top-10 instâncias com maior número de menções diretas à *DBpedia*, após a realização dos filtros, ordenadas de forma descendente pelo número de *hits* diretos

Ponte (GRO)	Instância	#Hits Diretos
gr:BusinessEntity	dbr:Esporte_Interativo	5
gr:ProductOrService	dbr:Diário_Oficial_da_União	5
gr:Product	dbr:Death_Note	4
gr:Product	dbr:Naruto	4
gr:Product	dbr:One_Piece	4
gr:Product	dbr:Tokyo_Ghoul	4
gr:BusinessEntity	dbr:Air_France	3
gr:BusinessEntity	dbr:Avianca_Holdings	3
gr:BusinessEntity	dbr:Bayer_04_Leverkusen	3
gr:BusinessEntity	dbr:BTG_Pactual	3

O processo de contabilização dos *hits* praticado segue uma ordem ascendente, partindo dos recursos mais específicos (recursos semanticamente anotados) para os recursos mais genéricos (e.g. `dbo:Agent`, `owl:Thing`).

A Figura 4 apresenta um *ranking* das 20 classes com maior número de *hits* nas anotações semânticas geradas pela *DBpedia-Spotlight* para o *dataset* BR-2015. A classificação é ordenada em número decrescente de ocorrências acumuladas. Na figura é possível notar classes como `dbo:Organisation` ou `dbo:SportsTeam`, que possuem pontes para `gr:BusinessEntity`, indicando a possibilidade de haver algum interesse de negócio, i.e., algum artigo esportivo de um clube para a venda ou até mesmo um torcedor interessado na compra.

O resultado do processo de Construção das Hierarquias Semânticas (subseção 3.3) apresentada na Figura 5 aponta um resultado satisfatório para um experimento inicial. O gráfico mostra a proporção de classes e instâncias da *DBpedia* anotadas e que possuem pontes para GRO. Cerca de 61.3% das instâncias anotadas possuem ponte para ontologia de domínio, enquanto a cobertura das classes coletadas da *DBpedia*, 39.2% possuem pontes. Se fizermos um comparativo com a quantidade de classes que a *DBpedia* possui (685 classes[13]), neste experimento foi possível avaliar 186 classes da *DBpedia*, o que equivale a aproximadamente 27% de classes da DBO.

Por outro lado, o gráfico apresentado na Figura 6 demonstra a distribuição da proporção entre as instância/recursos anotados e a ontologia de domínio, e sugere que

[13]https://wiki.dbpedia.org/services-resources/ontology

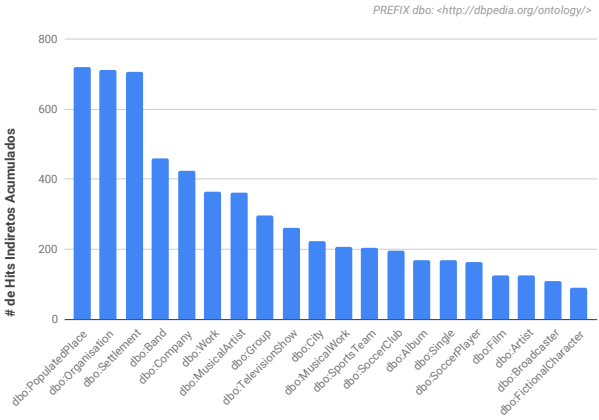


Figure 4. Top 20 classes mais mencionadas nas anotações do *dataset* BR-2015, após o filtro

aproximadamente 40% das menções em *tweets* do conjunto de dados avaliado refletem algum tipo de interesse em produtos e/ou serviços, pouco mais de 39% refletem interesse de negócio, uma menção a alguma entidade de negócio, tais como, emissoras de TV (`dbr:MTV`), canais de TV (`dbr:SportTV`), entre outros. Apesar de `gr:BusinessEntity` não representar um interesse específico e/ou direto para alguma entidade de negócio, pode ser usado em CRM Social para estreitar relação entre consumidor e empresa, capturar *leads*, pessoas que podem se tornar oportunidades de negócios reais através de marketing direcionado, entre outras oportunidades, assim como `gr:Location`, que pode corresponder a localização de uma empresa, loja, assim como o interesse de uma pessoa em uma viagem para uma localização específica.

Os dados apresentados na Tabela 4 apresentam os resultados da mesma busca efetuada para gerar a Figura 4, porém os resultados contidos na tabela foram gerados após o processo de Filtragem e Concepção das Hierarquias. A tabela está ordenada de forma descendente com as classes que possuem um maior número de *hits* acumulados (mais mencionadas) no topo da tabela. Além disso, é possível identificar as pontes usadas como para a ontologia de domínio de negócio (GRO), assim como a presença de novas classes da *DBpedia* no gráfico, tais como `dbo:MusicalArtist` que possui ponte para `gr:BusinessEntity` e como o próprio nome sugere, trata-se de menções a cantores, intérpretes, etc.

A Figura 7 ilustra um esquema dimensional para análise de menções de interesse em textos apresentado por [Júnior et al. 2018] e demonstra uma possibilidade de aplicação para *Data Warehouse Semânticos (DWS)*. No esquema é apresentado uma tabela **fato** *tweetMeasures* com medidas mensuráveis para *#tweets*, *#users* e *#mentions*. O modelo segue a notação de [Golfarelli et al. 1998] e apresenta dimensões espaço-temporais comuns em *Data Warehouse (DW)*, `Time` e `Place` e que podem ser preenchidas com

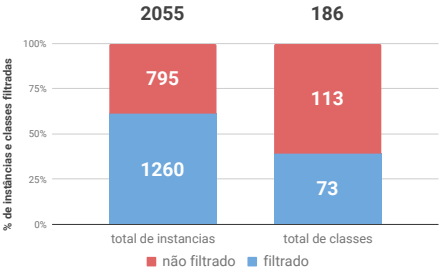


Figure 5. Quantidade de instâncias e classes filtradas usando pontes para ontologia de alto nível

os resultados das pontes para `gr:Location` ou por metadados oriundos de *tweets*, tais como data e hora, e geolocalização. As dimensões `Product or Service` e e `Business Entity` podem ser carregadas com as informações obtidas pelo processo de construção da SH. `Product or Service` pode agregar os resultados das pontes `gr:Products` e `gr:ProductsOrServices`, como por exemplo `dbo:Single` e `dbo:Album` que, de acordo com a Tabela 4 está ligado às duas pontes. E por sua vez, a dimensão `Business Entity` pode ser carregada os recursos filtrados usando a ponte para `gr:BusinessEntity`, *e.g.,* `dbo:MusicalArtist`, `dbo:SoccerClub`, `dbo:TelevisionStation`, entre outros.

Não foi possível apresentar um estudo que incorporasse sistemas de recomendação. Isso ocorre devido ao conjunto de dados utilizados neste trabalho. O conjunto de dados possui apenas textos de *tweets*, sem quaisquer informações dos usuários que postaram os *tweets*, como um id ou localização, os quais são essenciais para sistemas de recomendação. No entanto, o autor deste trabalho acredita que o uso do modelo em sistemas de recomendação seja factível, uma vez que se tenha a identificação dos usuários.

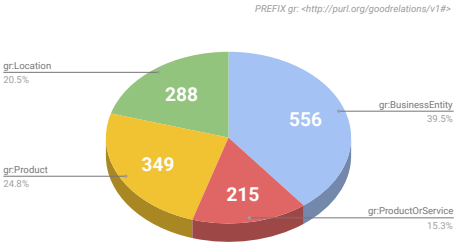


Figure 6. Distribuição da quantidade de instâncias mencionadas e filtradas com as pontes para a ontologia de alto nível

Table 4. Top 20 classes com maior número de *Hits* (diretos + indiretos), filtradas usando as pontes para ontologia de alto nível e ordenados de forma descendente pelo número de *hits*

Pontes (GRO)	Classes (DBO)	Hits
gr:Location	dbo:PopulatedPlace	480
gr:Location	dbo:Settlement	470
gr:BusinessEntity	dbo:MusicalArtist	181
gr:Location	dbo:City	148
gr:BusinessEntity	dbo:Company	141
gr:Product	dbo:TelevisionShow	130
gr:BusinessEntity	dbo:Band	115
gr:Product	dbo:MusicalWork	103
gr:BusinessEntity	dbo:Group	99
gr:BusinessEntity	dbo:SoccerClub	98
gr:BusinessEntity	dbo:SportsTeam	68
gr:Product	dbo:Film	63
gr:Product	dbo:Album	56
gr:Product	dbo:Single	56
gr:ProductOrService	dbo:Album	56
gr:ProductOrService	dbo:Single	56
gr:Product	dbo:WrittenWork	43
gr:BusinessEntity	dbo:Broadcaster	36
gr:Location	dbo:Country	36
gr:BusinessEntity	dbo:TelevisionStation	32

6. Discussão

Durante o desenvolvimento do trabalho, foi identificado que uma sequência de passos ocorre e propor um processo denominado Enriquecimento Semântico, apresentado no capítulo 3. Já na fase de Análise dos Resultados foi possível constatar que `dbr:Deus` é a instância com maior número de menções, seguidas por `dbr:One_Direction`, banda britânica, `dbr:Dilma_Rousseff` e `dbr:Brasil`. Na lista também consta entre os dez recursos mais anotados, o recurso `dbr:Impeachment`, os textos explorados na rede social relata o momento político vivido no Brasil no período entre 30/11/2015 e 11/12/2015 e evidenciado nos `dbr:tweets`. Nos resultados do conjunto de dados explorado também é possível ver que existe uma predominância das classes de negócio `dbo:MusicalArtist`, `dbo:Company` e `dbo:TelevisionShow` analisadas em tweets enviados do Brasil.

Abaixo são sintetizados alguns dados observados durante o experimento:

- Ao total foram registrados 35795 menções diretas a um recurso/instância da *DBpedia*, das 35795 menções anotadas, 22726 possuem pontes para GRO, ou seja, em torno de 63% dos recursos/instâncias que foram mencionadas nos *tweets* foram filtradas usando as pontes para ontologia de alto nível
- Foram anotadas 2055 Instâncias diferentes da DBpedia (dbr), das quais 1260 foram filtradas usando as pontes para a ontologia de domínio

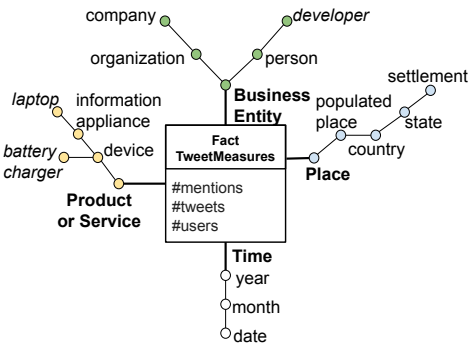


Figure 7. Modelagem Semântica Dimensional proposta por Júnior *et al.* [Júnior et al. 2018]

- A *DBpedia Ontology (DBO)* possui 685 classes definidas em sua ontologia, neste experimento foi possível alcançar, por meio da navegação por `rdf:type` e `rdfs:subClassOf`, 186 classes da DBO, ou 27% de cobertura da ontologia, e ainda, de 186 classes foi factível criar 73 associações para ontologia de domínio de negócio (*e-Business*).

Por fim, os resultados ilustram que é possível criar um DWS através de menções feitas em *tweets* e semanticamente anotadas e filtrados usando uma ontologia de alto nível que descreve semanticamente os interesses que se pretende filtrar, tal como a GRO, usada neste trabalho. E ainda, que em mídias sociais podem existir inúmeras possibilidades de uso, pode ser usado para análise estatística para identificação de oferta e demanda de produtos, tendências e padrões comportamentais e/ou tópicos emocionais.

7. Trabalhos Relacionados

Este capítulo apresenta os trabalhos correlatos à nossa proposta, selecionados a partir de uma revisão bibliográfica sistemática baseada nas diretrizes de [Kitchenham and Charters 2007]. Esta revisão bibliográfica foi realizada em diversas fontes usando a seguinte expressão de consulta:

(semantic annotation OR annotation)
AND (social media OR tweets OR posts)
AND (analysis OR filtering)

Foram encontrados 40 artigos relacionados à expressão de consulta. Após a primeira filtragem, através da leitura e análise dos resumos e estruturas dos trabalhos, foram considerados apenas 7 artigos. Os artigos resultaram no conjunto de trabalhos relacionados a este, descritos e comparados no restante deste capítulo, seguindo os seguintes critérios:

- Postagens de mídias sociais *(Social Media Posts - SMP)*: origem do conjunto de dados. *Datasets* de textos originados de mídias sociais (*e.g. tweets*) são vastos e possuem *APIs* públicas para coleta. Os valores possíveis para os trabalhos correlatos são: usa (OK) ou não usa (NOK);

- Anotação Semântica *(Semantic Annotation - SA)*: abordagem faz uso (OK) ou não (NOK) técnicas de anotação semântica para enriquecimento dos textos de mídias sociais.
- Método de Anotação *(Annotation Method - AM)*: técnica utilizada para anotação (manual: apenas com conhecimento de especialistas; semiautomático: regras e ontologias de especialistas; automático: somente ontologias e base de conhecimento);
- Customizável *(Customizable - C)*: permite personalização para um domínio específico;
- Construção Automática de Hierarquias *(Automatic Build Hierarchies - ABH)*: a abordagem gera dimensões com hierarquias de propriedades de ordem parcial disponíveis em ontologias e *LODs*.

O trabalho de Abrahams [Abrahams et al. 2012] propõe um processo que aplica técnicas de mineração de textos *(text mining)* em postagens de fóruns de discussão online, procurando características de veículos para alimentar sistemas de recomendação. Em uma abordagem semelhante [Villanueva et al. 2016], anota semanticamente os *tweets* para extrair informações para fins de recomendação.

Outras propostas constroem e preenchem um modelo dimensional para analisar de textos. Nebot *et al.* [Nebot and Berlanga 2012] analisa registros de pacientes e médicos anotados manualmente usando o modelo dimensional. Nebot [Nebot and Berlanga 2012] propõe um processo para preencher um DW a partir de dados semânticos, mas não é personalizável e não usa texto de postagens em mídias sociais.

O trabalho de Fileto *et al.* [Fileto et al. 2015] propõe um modelo ontológico e um processo para estruturar e enriquecer semanticamente dados de movimento em vários níveis de abstração, para apoiar a análise dos esquemas dimensionais enriquecidos como proposto em [Fileto et al. 2014]. Já o trabalho de Sacenti *et al.* [Sacenti et al. 2015] propõe um método para construir dimensões para analisar postagens de mídias sociais semanticamente enriquecidos com LOD. Eles fazem isso adaptando hierarquias de classes e instâncias existentes em coleções de LOD de acordo com suas incidências como valores de anotação de determinados conjuntos de dados. Essas duas obras são semelhantes à nossa abordagem, entretanto não constroem as dimensões de análise para domínios de aplicação específicos que podem ser alterados no início do processo. Para isso, a abordagem proposta neste trabalho usa o conceito de pontes (já apresentado anteriormente). Inicialmente, os conceitos mais mencionados da Hierarquia gerada são verificados por grupo pequeno de especialistas de domínio, que selecionam as classes ou instâncias de fato relevantes e as associam uma a uma com conceitos de uma ontologia de domínio de alto nível escolhida, formando um conjunto inicial de pontes-chave KBs. O processo segue com a construção da SH associando as pontes chaves através da relação de equivalência (`owl:equivalentClass`). Tais pontes orientam a seleção de anotações relevantes e a adaptação de hierarquias de LOD usados nas anotações e podem servir como dimensões de análise de dados. O resultado final é uma hierarquia semântica construída especialmente para o domínio determinado pelos especialistas. Esta etapa de construção de pontes com a SH diferencia este trabalho dos dois trabalhos analisados ([Fileto et al. 2015], [Sacenti et al. 2015]).

Finalmente, a abordagem interativa EXODuS proposta em [Chouder et al. 2019]

permite consultas OLAP exploratórias em bases NoSQL orientadas a documentos. Em tal abordagem, hierarquias são construídas mediante um método baseado em mineração de dependências funcionais aproximadas entre elementos de documentos JSON. Tais hierarquias são montadas incrementalmente em porções envolvidas nas consultas multidimensionais à medida em que tais consultas são submetidas pelos usuários, de modo a conferir melhor desempenho. As consultas OLAP expressas sobre um modelo dimensional com hierarquias dinâmicas são traduzidas para a linguagem de consulta do MongoDB. A abordagem EXODuS é avaliada com dados da NBA (*National Basketball Association*), da DBLP (*DataBase systems and Logic Programming bibliography*) e *tweets*. Todavia, como as dimensões dos cubos de dados produzidos pela abordagem EXODuS são baseadas em elementos JSON (na maioria metadados), diferentemente do nosso trabalho, EXODuS não permite efetuar consultas de acordo com conceitos e instâncias mencionadas em textos, tais como os conteúdos textuais de *tweets* e artigos.

A Tabela 3 resume as características dos trabalhos relacionados selecionados, com base nos seis critérios de comparação descritos no início deste capítulo. A última linha da Tabela 3 refere-se à nossa proposta. Ela usa algumas ideias de trabalhos relacionados selecionados, como a exploração de anotações semânticas e hierarquias presentes em coleção de LOD. No entanto, este trabalho é o único que apresenta um processo geral que emprega e filtra as anotações semânticas de oriundas de mídias sociais (*e.g., tweet*) que sejam de interesse para um domínio de aplicação utilizando pontes para uma ontologia de alto nível para tal domínio.

Table 5. Comparação de trabalhos correlatos

Trabalho	SMP	SA	C	AM	ABH	Saída
Abrahams(2012)	NOK	OK	NOK	Manual	NOK	Recomendações
Nebot(2012)	NOK	OK	NOK	Manual	OK	Cubo dimensional populado
Fileto(2015)	OK	OK	OK	Automático	NOK	Dados semânticos de movimento
Sacenti(2015)	OK	OK	OK	Automático	OK	Cubo dimensional não-populado
Villanueva(2016)	OK	OK	NOK	Semi-Automático	NOK	Recomendações
Chouder(2019)	OK	NOK	OK	NOK	OK	Cubo dimensional populado
Junior *et al.*(2018)	**OK**	**OK**	**OK**	**Automático**	**OK**	**Cubo dimensional populado**
Este trabalho	**OK**	**OK**	**OK**	**Automático**	**OK**	**Hierarquia Semântica**

References

Abrahams, A. S., Jiao, J., Wang, G. A., and Fan, W. (2012). Vehicle defect discovery from social media. *Decision Support Systems*, 54(1):87–97.

Asur, S. and Huberman, B. A. (2010). Predicting the future with social media. In *IEEE/WIC/ACM: International Conference on Web Intelligence and Intelligent Agent Technology*.

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., and Ives, Z. (2007). Dbpedia: A nucleus for a web of open data. In *The semantic web*, pages 722–735. Springer.

Chouder, M. L., Rizzi, S., and Chalal, R. (2019). Exodus: Exploratory olap over document stores. *Information Systems*, 79:44 – 57. Special issue on DOLAP 2017: Design, Optimization, Languages and Analytical Processing of Big Data.

Daiber, J., Jakob, M., Hokamp, C., and Mendes, P. N. (2013). Improving efficiency and accuracy in multilingual entity extraction. In *Proceedings of the 9th International Conference on Semantic Systems (I-Semantics)*.

Dean, M. and Schreiber, G. (2004). OWL web ontology language reference. W3C recommendation, W3C. http://www.w3.org/TR/2004/REC-owl-ref-20040210/.

Dutch, M. (2008). Understanding data deduplication ratios. In *SNIA Data Management Forum*, page 7.

Fileto, R., May, C., Renso, C., Pelekis, N., Klein, D., and Theodoridis, Y. (2015). The baquara 2 knowledge-based framework for semantic enrichment and analysis of movement data. *Data & Knowledge Engineering*, 98:104–122.

Fileto, R., Raffaetà, A., Roncato, A., Sacenti, J. A. P., May, C., and Klein, D. (2014). A semantic model for movement data warehouses. In *Proceedings of the 17th International Workshop on Data Warehousing and OLAP, DOLAP 2014, Shanghai, China, November 3-7, 2014*, pages 47–56.

Fileto, R., Sorato, D., Goularte, F. B., and Nassar, S. M. (2016). Análise de métodos e ferramentas para reconhecimento de palavras relevantes em microblogs. In *XII Brazilian Symposium on Information Systems*, pages 345–352, Florianópolis, SC, Brasil.

Golfarelli, M., Maio, D., and Rizzi, S. (1998). The dimensional fact model: A conceptual model for data warehouses. *International Journal of Cooperative Information Systems*, 7(02n03):215–247.

Heath, D., Singh, R., and Ganesh, J. (2014). Social media at sociosystems inc.: A sociotechnical systems analysis of strategic action. In *47th Hawaii International Conference on System Science*, pages 584–593, Waikoloa, HI, USA.

Júnior, V. C. P., Fileto, R., de Souza, W. S., Wittwer, M., Reinhold, O., and Alt, R. (2018). A semantic BI process for detecting and analyzing mentions of interest for a domain in tweets. In *Proceedings of the 24th Brazilian Symposium on Multimedia and the Web, WebMedia 2018, Salvador-BA, Brazil, October 16-19, 2018*, pages 197–204.

Kitchenham, B. and Charters, S. (2007). Guidelines for performing systematic literature reviews in software engineering. Technical Report EBSE 2007-001, Keele University and Durham University Joint Report.

Mendes, P. N., Jakob, M., García-Silva, A., and Bizer, C. (2011). Dbpedia spotlight: Shedding light on the web of documents. In *Proceedings of the 7th International Conference on Semantic Systems (I-Semantics)*, Graz, Austria.

Nebot, V. and Berlanga, R. (2012). Building data warehouses with semantic web data. *Decision Support Systems*, 52(4):853–868.

Sacenti, J. A., Salvini, F., Fileto, R., Raffaetà, A., and Roncato, A. (2015). Automatically tailoring semantics-enabled dimensions for movement data warehouses. In *International Conference on Big Data Analytics and Knowledge Discovery*, pages 205–216. Springer.

Tufekci, K. (2014). Big questions for social media big data: Representativeness, validity and other methodological pitfalls. In *Proceedings of the 8th International AAAI Conference on Weblogs and Social Media (ICWSM)*.

Villanueva, D., González-Carrasco, I., López-Cuadrado, J. L., and Lado, N. (2016). Smore: Towards a semantic modeling for knowledge representation on social media. *Science of Computer Programming*, 121:16–33.

APÊNDICE B – Código fonte do software - SemanticHierarchy

```
import argparse
import re
import traceback

from utils import logger

from semantic_hierarchy import database
from semantic_hierarchy import spotlight

# annotation globals
_CONFIDENCE = '0.5'
_LANGUAGE = 'pt'


class FormatsAllowed:
    TXT = 'txt'


log = logger.get_logger(__name__)

parser = argparse.ArgumentParser(
    'semantic_hierarchy',
    description='Build a semantic hierarchy using input file'
)
parser.add_argument(
    'filename',
    type=argparse.FileType(bufsize=16 * 1024),
    help='file name used as input to build %(prog)s'
)
parser.add_argument(
    '-f', '--format',
    default='txt',
    choices=['txt'],
    help='file format used (default: %(default)s)'
)
parser.add_argument(
    '-d', '--dataset',
    # metavar='dataset_name',
    help='name of dataset used on experiment'
)

sp = spotlight.Spotlight(_LANGUAGE, _CONFIDENCE)
db = database.AnnotationDB()


def preprocessing(text):
    # https://stackoverflow.com/questions/33404752/removing-emojis-from-
        a-string-in-python
    # remove_emojis(inputString):
    txt = text.encode('ascii', 'ignore').decode('ascii')
    # remove_urls(inputString):
    txt = re.sub(r'https?://\S+', "", txt)
```

```
    # remove_mail(inputString):
    txt = re.sub(r'[\w\d_\-\.]+@[\w\d_\-\.]+(\.\w+)+', "", txt)
    return txt


def make_annotation(text, collection):
    # import pdb; pdb.set_trace()
    annot = sp.annotate(text)
    # db.insert(collection, annot)
    return annot


def build_from_text_plain(file, dataset_name):
    for i, line in enumerate(file):
        if i < 26706:
            continue
        # maybe should filter not RT
        try:
            text = preprocessing(line)
            make_annotation(text, dataset_name.replace(" ", "_"))
            log.info('success')

        except Exception as exp:
            traceback_str = traceback.format_exc()
            log.error(
                f'Line[{i}]: {line}\nAnnotation error: {repr(exp)}\n{
                    traceback_str}')


def main_annotation(args):
    arguments = parser.parse_args(args)
    if arguments.format == FormatsAllowed.TXT:
        build_from_text_plain(arguments.filename, arguments.dataset)
```

Código B.1 – semantic_annotation.py

```
from semantic_hierarchy import hierarchy_builder


if __name__ == "__main__":
    hb = hierarchy_builder.HierarchyBuilder()
    hb.exec(5)
    print()
```

Código B.2 – semantic_builder.py

```
from pymongo import MongoClient
DATABASE = 'research'
SEMANTIC_ANNOTATIONS = 'fabio_bif_2015'

class AnnotationDB:
    def __init__(self, host='localhost', port=27017):
```

```
        self.client = MongoClient(host, port)
        self.db = self.client[DATABASE]

    def insert(self, *annotations):
        self.db[SEMANTIC_ANNOTATIONS].insert_many(annotations)

    def find(self):
        return self.db[SEMANTIC_ANNOTATIONS].find(
            {'Resources': {'$exists': True, '$ne': []}},
            {'@text': 1, 'Resources': 1}
        )
```

Código B.3 – semantic_hierarchy/database.py

```
from string import Template
from backoff import on_exception, expo
from SPARQLWrapper import SPARQLWrapper


DBPEDIA_SPARQL_ENDPOINT = 'http://dbpedia.org/sparql'
RELIABLE_DATASETS = [
    'dbpedia.org/ontology',
    'schema.org'
]

_SPARQL_QUERY_TEMPLATE = Template(r'''
    SELECT DISTINCT ?elem
        WHERE {
            <$resource> $prop ?elem .
            FILTER(
                REGEX(?elem, 'https?://($re_realiable_dataset) ')
            )
            MINUS {
                ?otherClass <http://www.w3.org/2002/07/owl#
                     equivalentClass> ?elem
            }
        }
''')


class Sparql:
    def __init__(self):
        self.sparql = SPARQLWrapper(DBPEDIA_SPARQL_ENDPOINT)

    def rdf_types_from(self, resource: str):
        return self._rdf_find_property_from('a', resource)

    def rdf_parents_from(self, resource: str):
        return self._rdf_find_property_from(
            '<http://www.w3.org/2000/01/rdf-schema#subClassOf>',
            resource,
        )
```

```
38
39      def _rdf_find_property_from(self, prop_to_find: str, resource: str):
40          query = _SPARQL_QUERY_TEMPLATE.substitute(
41              prop=prop_to_find,
42              resource=resource,
43              re_realiable_dataset='|'.join(RELIABLE_DATASETS),
44          )
45          resp = self._sparql_query(query)
46          return {
47              # the 'elem' key must be the same used in sparql query
48              result['elem']['value'] for result in resp['results']['
                    bindings']
49          }
50
51      @on_exception(expo, Exception, max_tries=8)
52      def _sparql_query(self, query: str):
53          self.sparql.setReturnFormat('json')
54          self.sparql.setQuery(query)
55          return self.sparql.queryAndConvert()
```

Código B.4 – semantic_hierarchy/dbpedia.py

```
1  import traceback
2  from concurrent import futures
3
4  from semantic_hierarchy import database, dbpedia, tdb
5  from utils import logger
6
7  # parallel globals
8  _NUM_WORKERS = 4
9  DS_ENDPOINT = 'http://localhost:3030/ds'
10
11 log = logger.get_logger(__name__)
12
13
14 class HierarchyBuilder:
15     def __init__(self):
16         self.mongodb = database.AnnotationDB()
17         self.sparql = dbpedia.Sparql()
18         self.tdb = tdb.TripleDB(DS_ENDPOINT+'/query', DS_ENDPOINT+'/
                update')
19
20     def _walking_in_subclassof(self, resource: str):
21         # build subclassof hierarchy
22         parents = self.sparql.rdf_parents_from(resource)
23         for parent in parents:
24             tdb.increase_subclassof_hit(parent)
25             self.tdb.add_subclassof(resource, parent)
26             self._walking_in_subclassof(parent)
27
28     def _walking_in_type(self, resource: str):
29         # build type hierarchy
```

```
30          rdf_types = self.sparql.rdf_types_from(resource)
31          for tp_class in rdf_types:
32              tdb.increase_type_hit(tp_class)
33              self.tdb.add_type(resource, tp_class)
34              self._walking_in_subclassof(tp_class)
35
36      def build(self, annotations: dict):
37          # direct hits
38          for resource in annotations['Resources']:
39              r = resource['@URI'].replace('pt.', '')
40              # insert a relation from annotated resource to text from
                    tweet r --tweet--> text
41              self.tdb.add_tweet(r, annotations['@text'])
42              tdb.increase_direct_hit(r)
43              self._walking_in_type(r)
44
45      def exec(self, max_workers=_NUM_WORKERS):
46          annotations = self.mongodb.find().limit(50)
47          # after filter: 27389
48          print("# of annotated resources:", annotations.count())
49          with futures.ThreadPoolExecutor(max_workers=max_workers) as
                executor:
50              hierarchies = {
51                  executor.submit(self.build, ann): ann
52                  for ann in annotations
53              }
54              for future in futures.as_completed(hierarchies):
55                  ann = hierarchies[future]
56                  try:
57                      future.result()
58                  except Exception as exp:
59                      log.error(
60                          f'An error ocurred on building hierarchy step: \n'
61                          f'{{ "_id":{ann["_id"]}, "@text":{ann["@text"]}, "
                                Resources": {ann["Resources"]} }}\n'
62                          f'({repr(exp)}). {traceback.format_exc()}")'
63                      )
64                  else:
65                      log.info(f'id:{ann["_id"]} - success')
66          self.tdb.add_semantic_hits()
```

Código B.5 – semantic_hierarchy/hierarchy_builder.py

```
1  import aiohttp
2  import requests
3
4  from ratelimit import limits, RateLimitException
5  from backoff import on_exception, expo
6
7
8  # _SPOTLIGHT_URL_BASE = 'http://model.dbpedia-spotlight.org' # old
       endpoint
```

```
_SPOTLIGHT_URL_BASE = 'https://api.dbpedia-spotlight.org'
# _SPOTLIGHT_URL_BASE = 'http://localhost:8080/rest'


class Spotlight:
    MINUTE = 60

    def __init__(self, language, confidence):
        self.lang = language
        self.confidence = confidence
        self.session = aiohttp.ClientSession(
            timeout=aiohttp.ClientTimeout(total=60))

    @on_exception(expo, aiohttp.ClientError, max_tries=8)
    async def annotate(self, tweet):
        '''
        returns a list of pairs (URI, LABEL) annotateds
        '''
        url = f'{_SPOTLIGHT_URL_BASE}/{self.lang}/annotate'
        # url = f'{_SPOTLIGHT_URL_BASE}/annotate',
        async with self.session.post(url,
                                     data={
                                         'text': tweet,
                                         'confidence': self.confidence
                                     },
                                     headers={
                                         'accept': 'application/json',
                                         'content-type': 'application/x-www-
                                             form-urlencoded'
                                     }) as resp:
            # if resp.status_code != 200:
            # import pdb; pdb.set_trace()
            resp.raise_for_status()
            return await resp.json()
```

Código B.6 – semantic_hierarchy/spotlight.py

```
import collections
import threading

from backoff import on_exception, expo

import rdflib
from rdflib import namespace as ns
from rdflib.plugins.stores import sparqlstore

from utils import logger


SEMANTIC_HITS = {
    'directHits': collections.Counter(),
    'indirectHitsByType': collections.Counter(),
```

```
    'indirectHitsBySubClassOf': collections.Counter(),
}

LISA = ns.ClosedNamespace(
    uri=rdflib.URIRef("http://lisa.inf.ufsc.br/"),
    terms=["directHits", "indirectHitsByType",
           "indirectHitsBySubClassOf", "tweet"]
)

log = logger.get_logger(__name__)


class TripleDB:
    def __init__(self, query_endpoint: str, update_endpoint: str):
        store = sparqlstore.SPARQLUpdateStore()
        self.gs = rdflib.ConjunctiveGraph(store)
        self.gs.open((query_endpoint, update_endpoint))

    def add_semantic_hits(self):
        log.info('Inserting the semantic hits on TDB')
        for prop_name, resources in SEMANTIC_HITS.items():
            for r, n in resources.items():
                self._add_relation(
                    rdflib.URIRef(r),
                    LISA[prop_name],
                    rdflib.Literal(str(n))
                )
        log.info('Semantic hits inserted with success')

    @on_exception(expo, Exception, max_tries=8)
    def _add_relation(self, subject, predicate, obj):
        q = f'''
            INSERT DATA
            {{
                {subject.n3()} {predicate.n3()} {obj.n3()} .
            }}
        '''
        self.gs.update(q)

    def add_tweet(self, subject: str, text_tweet: str):
        self._add_relation(
            rdflib.URIRef(subject),
            LISA.tweet,
            rdflib.Literal(text_tweet)
        )

    def add_label(self, subject: str, label: str):
        self._add_relation(
            rdflib.URIRef(subject),
            ns.RDFS.label,
            rdflib.Literal(label)
        )
```

```
68
69      def add_type(self, subject: str, obj: str):
70          self._add_relation(
71              rdflib.URIRef(subject),
72              ns.RDF.type,
73              rdflib.URIRef(obj)
74          )
75
76      def add_subclassof(self, subject: str, obj: str):
77          self._add_relation(
78              rdflib.URIRef(subject),
79              ns.RDFS.subClassOf,
80              rdflib.URIRef(obj)
81          )
82
83
84  lock = threading.Lock()
85
86
87  def increase_subclassof_hit(resource):
88      _increase_hit('indirectHitsBySubClassOf', resource)
89
90
91  def increase_type_hit(resource):
92      _increase_hit('indirectHitsByType', resource)
93
94
95  def increase_direct_hit(resource):
96      _increase_hit('directHits', resource)
97
98
99  def _increase_hit(hit_name, resource):
100     with lock:
101         SEMANTIC_HITS[hit_name][resource] += 1
102
103
104 def to_dict():
105     return {
106         key: dict(value)
107         for (key, value) in SEMANTIC_HITS.items()
108     }
```

Código B.7 – semantic_hierarchy/tdb.py