

**UNIVERSIDADE FEDERAL DE SANTA CATARINA  
DEPARTAMENTO DE INFORMÁTICA E ESTATÍSTICA**

Vilmar César Pereira Júnior

**UMA ABORDAGEM SEMÂNTICA PARA ANALISAR  
MENÇÕES DE INTERESSE EM UM DOMÍNIO EM  
CLIPES TEXTUAIS**

Florianópolis

2018



Vilmar César Pereira Júnior

**UMA ABORDAGEM SEMÂNTICA PARA ANALISAR  
MENÇÕES DE INTERESSE EM UM DOMÍNIO EM  
CLIPES TEXTUAIS**

Dissertação submetida ao Programa  
de Pós-Graduação em Ciência da Com-  
putação para a obtenção do Grau de  
Mestre em Ciência da Computação.  
Orientador: Prof. Renato Fileto, Dr.

Florianópolis

2018

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

Pereira Júnior, Vilmar César  
Uma Abordagem Semântica para Analisar Menções de  
Interesse em Um Domínio em Clipes Textuais / Vilmar  
César Pereira Júnior ; orientador, Renato Fileto,  
2018.

86 p.

Dissertação (mestrado) - Universidade Federal de  
Santa Catarina, Centro Tecnológico, Programa de Pós  
Graduação em Ciência da Computação, Florianópolis,  
2018.

Inclui referências.

1. Ciência da Computação. 2. Web Semântica. 3.  
Anotações Semânticas. 4. Data warehouses. 5. Dados  
Ligados Abertos (LOD). I. Fileto, Renato. II.  
Universidade Federal de Santa Catarina. Programa de  
Pós-Graduação em Ciência da Computação. III. Título.

Vilmar César Pereira Júnior

**UMA ABORDAGEM SEMÂNTICA PARA ANALISAR  
MENÇÕES DE INTERESSE EM UM DOMÍNIO EM  
CLIPES TEXTUAIS**

Esta Dissertação foi julgada aprovada para a obtenção do Título de “Mestre em Ciência da Computação”, e aprovada em sua forma final pelo Programa de Pós-Graduação em Ciência da Computação.

Florianópolis, 30 de Novembro 2018.



---

Prof. José Luís Almada Güntzel, Dr.  
Coordenador do Curso

**Banca Examinadora:**



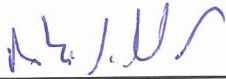
---

Prof. Renato Fileto, Dr.  
Orientador



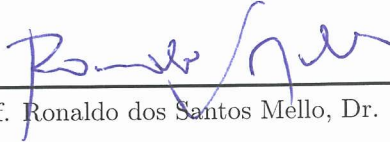
---

Prof. Denilson Sell, Dr.



---

Prof. Roberto Willrich, Dr.



---

Prof. Ronaldo dos Santos Mello, Dr.

Dedico este trabalho a Araci e Gabriela.





## AGRADECIMENTOS

Agradeço primeiramente aos meus pais, Vilmar e Araci, pelo apoio incondicional.

À minha mulher Gabriela, companheira em todos os momentos, pela cumplicidade, amor e paciência.

Ao meu orientador, professor Fileto, por ter aceito o meu convite em mais um ciclo de orientação, conselhos e ajuda no desenvolvimento deste trabalho.

Ao meu irmão Adriano, pelos conselhos e incentivo nesta jornada.

Aos meus amigos Franklin, Samuel, Josimar, Lizandro, Marcelo, Anderson, Michel, Lucas e Wagner pelos momentos de descontração ao longo deste período.

Aos meus amigos e colegas de trabalho Hallan, Daniel, Kátyra, Ernani, Roberto, Adriano e Luciano, pela ajuda durante todo o curso, conselhos, incentivos, cafés tomados e também na confecção do trabalho.

Ao Willian, pela ajuda na construção dos experimentos e contribuições nas inúmeras reuniões ao longo destes dois anos.

Aos colegas de LISA, Danielly, Fábio e Ítalo pela ajuda com os dados coletados em trabalhos anteriores, que muito auxiliaram nos experimentos do estudo de caso.

Ao Juarez, pelos conselhos acerca do mestrado e também pela contribuição dele com o seu trabalho anterior, que muito inspirou esta dissertação.

Aos colaboradores da Universidade de Leipzig, Matthias, Olaf e Rainer pela ajuda na construção do estudo de caso deste trabalho.



*Me dê o zero e o um que eu tiro o tudo e  
o nada lá de dentro.*

Júlio Felipe Szeremeta



## RESUMO

Clipes textuais (*Textual Clips - TCs*), tais como postagens em mídias sociais e uma variedade de outros textos livres possivelmente georreferenciados, podem conter muitas informações. No entanto, a análise adequada de um TC requer a captura da semântica do que é mencionado em seu conteúdo, filtrando o que é de interesse para determinados domínios de aplicação e estruturando as informações extraídas de maneira adequada para análise. Este trabalho propõe uma abordagem para analisar incidências de menções de interesse para domínios de aplicação específicos em TCs, a qual combina as tecnologias de Web Semântica e *Business Intelligence (BI)*. Esta abordagem é suportada por um processo de Extração, Transformação e Carga (*ETL*) de dados que anota semanticamente *TCs* com dados abertos ligados (*Linked Open Data - LOD*), filtra recursos de interesse nas anotações usando pontes definidas neste trabalho entre classes *LOD* e uma ontologia de domínio de alto nível, aprimora essas pontes e adapta hierarquias existentes de classes e instâncias de *LOD* para servir como dimensões para análise de informação. A abordagem foi validada em um estudo de caso que analisa menções a itens de interesse para negócios (*business*) em tweets. Resultados experimentais mostram que uma quantidade considerável de tweets recentemente enviados do Brasil têm ao menos uma menção de interesse para o domínio de negócios e que a abordagem proposta permite análises que não são suportadas pelos atuais sistemas de BI para dados de mídias sociais.

**Palavras-chave:** Web Semântica. Anotações Semânticas. Mídias Sociais. *Data warehouses*. Dimensões de Análise. Dados Ligados Abertos (*LOD*). *E-Business*. *Social CRM*.



## ABSTRACT

Textual clips, such as social media posts and a variety of other time-stamped and sometimes georeferenced free text, can carry lots of information. However, their proper analysis requires capturing the semantics of what is mentioned in their contents, filtering what is of interest for particular application domains, and structuring the extracted information in a suitable way for analysis. In this work, we propose an approach to analyze the incidences of mentions of interest for particular application domains in textual clips, by combining Semantic Web and Business Intelligence (BI) technologies. This approach is supported by a data Extraction, Transformation and Loading (ETL) process that semantically annotates textual clips with Linked Open Data (LOD), filters *LOD* resources of interest in the annotations using bridges between *LOD* classes and a high-level domain ontology, improves those bridges, and adapts existing hierarchies of *LOD* classes and instances accordingly to serve as information analysis dimensions. This approach is validated in a case study that analyzes mentions to things of interest for business in tweets. Experimental results show that a considerable amount of the tweets recently sent from Brazil have at least a mention to something of interest for business and that the proposed approach enables analyses that are not supported by current BI systems for analyzing social media data.

**Keywords:** Semantic Web. Semantic Annotation. Social media. Data warehouses. Analysis dimensions. Linked Open Data. E-Business. Social CRM.





## LISTA DE FIGURAS

Figura 1	Exemplo de tweets com menções de interesse para a área de negócios ( <i>business</i> ).....	26
Figura 2	Exemplo de anotação semântica em um Clipe Textual .	36
Figura 3	Exemplo de Entidades Nomeadas anotadas em um Clipe Textual .....	38
Figura 4	Exemplo de anotação da ferramenta DBpedia-Spotlight	40
Figura 5	Exemplo de Clipe Textual anotado com o FOX.....	41
Figura 6	Exemplo de modelo dimensional.....	43
Figura 7	O processo de <i>ETL</i> semanticamente estendido proposto.	45
Figura 8	Classes da DBpedia com <i>hits</i> de menções semanticamente anotadas em <i>tweets</i> , organizadas em uma Hierarquia de Recursos com <i>Hits</i> (HHR) .....	50
Figura 9	Exemplos de pontes entre classes de <i>LOD</i> (à esquerda) e específicas de domínio (à direita): (1) pontes-chave ( <i>KB</i> ), criadas por especialistas de domínio e representadas por linhas contínuas; (2) pontes novas ( <i>NB</i> ), inferidas pelo algoritmo <i>CheckComplete</i> e representadas por linhas tracejadas; (3) única ponte inconsistente ( <i>IB</i> ) neste exemplo, detectada pelo algoritmo <i>CheckComplete</i> e representada por uma linha pontilhada. ....	57
Figura 10	Hierarquias de dimensões delimitadas em uma <i>HHR</i> ...	58
Figura 11	Diagrama ER da base de dados gerada para análises ...	60
Figura 12	Principais conceitos da GRO (HEPP, 2008).....	63
Figura 13	Classes mais mencionadas nas anotações do <i>dataset</i> BR-2015 .....	68
Figura 14	Resultados do algoritmo <i>CheckComplete</i> categorizados por instâncias.....	69
Figura 15	Resultados do algoritmo <i>CheckComplete</i> categorizados por classes .....	69
Figura 16	Subclasses de <i>owl:Thing</i> com mais <i>hits</i> .....	70
Figura 17	Hierarquia de subclasses de <i>dbo:Organization</i> .....	71



## LISTA DE TABELAS

Tabela 1	Pontes-chave ( <i>KB</i> ) elaboradas por especialistas de domínio nas 20 classes mais mencionadas da DBpedia.....	65
Tabela 2	Lista das 10 instâncias com mais <i>hits</i> diretos .....	68
Tabela 3	Comparação de trabalhos correlatos .....	77



## LISTA DE ABREVIATURAS E SIGLAS

CRM	Customer Relationship Management . . . . .	25
UGC	User Generated Content . . . . .	25
CRM	Customer Relationship Management . . . . .	25
TC	Textual Clip . . . . .	25
OLAP	On-Line Analytical Processing . . . . .	26
ETL	Extract, Transform, Load . . . . .	26
LOD	Linked Open Data . . . . .	26
BI	Business Intelligence . . . . .	27
KG	Knowledge Graph . . . . .	29
SDC	Semantic Data Cube . . . . .	29
BI	Business Intelligence . . . . .	30
DW	Data Warehouse . . . . .	30
ETL	Extraction, Transformation and Loading . . . . .	30
LOD	Linked Open Data . . . . .	30
NLP	Natural Language Processing . . . . .	35
NED	Named Entity Disambiguation . . . . .	35
NER	Named Entity Recognition . . . . .	35
URI	Uniform Resource Identifier . . . . .	37
NERD	Named Entity Recognition and Disambiguation . . . . .	38
DFM	Dimensional Fact Model . . . . .	42
BI	Business Intelligence . . . . .	42
HHR	Hierarchy of Hit Resources . . . . .	49



## SUMÁRIO

<b>1 INTRODUÇÃO</b> .....	25
1.1 PROBLEMA .....	29
1.2 PERGUNTA DE PESQUISA .....	31
1.3 OBJETIVOS .....	31
1.4 MATERIAIS E MÉTODOS .....	32
1.5 ESTRUTURA DO TRABALHO .....	32
<b>2 FUNDAMENTOS</b> .....	35
2.1 ONTOLOGIAS, DADOS ABERTOS LIGADOS ( <i>LOD</i> ) E GRAFOS DE CONHECIMENTO ( <i>KGS</i> ) .....	35
2.2 ANOTAÇÕES SEMÂNTICAS .....	37
2.3 FERRAMENTAS PARA ANOTAÇÃO SEMÂNTICA AU- TOMÁTICA .....	39
2.4 MODELO DIMENSIONAL .....	42
2.5 CONSIDERAÇÕES FINAIS .....	43
<b>3 PROCESSO DE BI SEMÂNTICO PROPOSTO</b> .....	45
3.1 DEFINIÇÕES BÁSICAS .....	47
3.2 ENRIQUECIMENTO SEMÂNTICO E CONTAGEM .....	51
3.2.1 Pré-processamento .....	51
3.2.2 Anotação Semântica .....	52
3.2.3 Seleção de anotações confiáveis .....	52
3.2.4 Contagem de <i>Hits</i> e Construção da Hierarquia de Recursos com <i>Hits</i> ( <i>HHR</i> ) .....	53
3.3 CONSTRUÇÃO DO CUBO DE DADOS SEMÂNTICO ( <i>SDC</i> )	53
3.3.1 Construção de Pontes-Chave ( <i>KBs</i> ) para Conceitos de Alto Nível .....	54
3.3.2 Verificar e Completar Pontes .....	55
3.3.3 Geração de Dimensões de Análise .....	58
3.3.4 Cálculo de Medidas .....	58
3.4 ANÁLISE DA INFORMAÇÃO .....	59
<b>4 EXPERIMENTOS</b> .....	63
4.1 ESTUDO DE CASO: <i>BUSINESS</i> .....	63
4.2 CONFIGURAÇÕES DOS EXPERIMENTOS .....	66
4.3 RESULTADOS EXPERIMENTAIS .....	67
4.3.1 Contagem de <i>Hits</i> e Construção da Hierarquia de Recursos com <i>Hits</i> ( <i>HHR</i> ) .....	67
4.3.2 Verificar e Completar Pontes .....	69
4.3.3 Geração de Dimensões de Análise .....	70

<b>4.3.4 Discussão</b> .....	71
<b>5 TRABALHOS RELACIONADOS</b> .....	73
<b>6 CONCLUSÕES E TRABALHOS FUTUROS</b> .....	79
<b>REFERÊNCIAS</b> .....	83



# 1 INTRODUÇÃO

Muitas pessoas e organizações frequentemente usam as mídias sociais para comunicar idéias, desejos, ofertas e o que acontece com elas e em seu ambiente (BONTCHEVA; ROUT, 2012; ATEFEH; KHREICH, 2015). Assim, as plataformas de mídias sociais podem transmitir informações relevantes e atualizadas, graças à sua popularidade e facilidade de uso. Este conteúdo é também conhecido como Conteúdo Gerado pelo Usuário (*User Generated Content - UGC*), que pode ser valioso em uma ampla variedade de domínios e aplicativos, incluindo:

- detecção de eventos (ATEFEH; KHREICH, 2015; VIEIRA et al., 2016);
- rastreamento e previsão de surtos de doenças (CHARLES-SMITH et al., 2015; SIGNORINI; SEGRE; POLGREEN, 2011; PAUL et al., 2016);
- monitoramento em tempo real de desastres naturais (SAKAKI; OKAZAKI; MATSUO, 2010; MIDDLETON; MIDDLETON; MODAFFERI, 2014; YIN et al., 2015);
- análise de tendências, situações e sentimentos (BONTCHEVA; ROUT, 2012).

Atualmente, a análise de dados de mídias sociais também tem sido usada para alavancar a área de negócios (ATEFEH; KHREICH, 2015; ZHAO et al., 2016; VIEIRA et al., 2016) e gerenciamento de relacionamento com o cliente *Customer Relationship Management (CRM)*, dando origem ao *social CRM* (LINDA, 2010; REINHOLD; ALT, 2012).

Chamamos de Clipe Textual (*Textual Clip - TC*) (GALLINUCCI; GOLFARELLI; RIZZI, 2015) qualquer texto livre acompanhado por um indicador de tempo (*timestamp*, geralmente indicando o momento do envio ou da criação do *TC*). Um *TC* também pode, adicionalmente, ser associado a coordenadas geográficas ou um indicador de lugar (geralmente denotando sua origem geográfica), além de outros metadados (*e.g.*, id do autor, indicação do idioma do texto). Alguns exemplos de *TCs* são postagens de usuários em mídias sociais e anotações em prontuários médicos. A análise de seu conteúdo pode revelar informações úteis sobre a incidência de doenças, tratamentos mais usados, etc. As postagens de mídias sociais, em particular, podem ser úteis para diversas aplicações, desde a análise de tendências e análise de sentimentos, até marketing, sistemas de recomendação e social *CRM*.

Os usuários das mídias sociais podem mencionar em suas postagens coisas de interesse para determinados domínios de aplicação (*e.g.*,

lugares, produtos, marcas e organizações de interesse para negócios). Por exemplo, um usuário do Twitter pode expressar o desejo de comprar um laptop com certas características, talvez relacionadas a uma marca ou a um determinado modelo.

A Figura 1 apresenta dois exemplos de tweets contendo tais tipos de menções. No primeiro tweet (a), o usuário recomenda um produto (laptop) de uma organização específica (Dell) e marca (XPS), para uso com o Sistema Operacional Fedora. No segundo tweet (b), o usuário lamenta ter esquecido o carregador do laptop ao viajar para o Brasil, fato que demonstra a intenção em adquirir um novo carregador.



Figura 1 – Exemplo de tweets com menções de interesse para a área de negócios (*business*)

Esse tipo de dado pode ser útil em uma variedade de aplicações, desde análises de tendências simples e *marketing* direcionado até análise de sentimentos, recomendação e avaliação do potencial do cliente em sistemas de *CRM* social. Analogamente, uma análise cuidadosa do conteúdo dos prontuários médicos pode revelar informações úteis sobre a incidência de doenças, tratamentos mais utilizados, etc. Este trabalho visa analisar menções de interesse encontradas em *TCs* com técnicas de Inteligência de Negócio (do inglês *Business Intelligence - BI*) (BROHMAN et al., 2000), especificamente com Processamento Analítico *On-Line* (do inglês *On-Line Analytical Processing - OLAP*) com as menções de interesse dos *TCs* organizadas em um modelo dimensional.

No entanto, a detecção automática de menções de interesse em *TCs* pode ser uma tarefa difícil. Uma abordagem manual para reconhecer menções e ligá-las a descrições semânticas precisas não é viável em muitas situações práticas, devido aos custos envolvidos. Essa tarefa apresenta problemas inerentes às linguagens naturais, como ambiguidades ou o contexto do clipe textual. No exemplo anterior, “nice” refere-se à qualidade do produto mencionando, e não à cidade francesa de Nice, e “Fedora” menciona o Sistema Operacional, não um tipo de chapéu. Assim, é essencial adotar um processo automatizado que use ferramen-

tas de estado-da-arte para identificar e desambiguar menções relevantes em *TCs*. Em seguida, organizar estas menções para realizar a análise sistemática das informações, por exemplo, utilizando um modelo dimensional para suportar OLAP. A organização da informação destas menções em um modelo dimensional é feita por meio de um processo de Extração, Transformação e Carga (*Extraction, Transformation, Loading - ETL*) (KIMBALL; ROSS, 2011; INMON, 1992) estendido com tarefas para pré-processamento de *TCs* - visando tratar ruídos nos seus conteúdos textuais -, anotação semântica de *TCs* com dados abertos ligados (do inglês *Linked Open Data (LOD)*), filtragem das anotações de interesse para o domínio e criação de dimensões de análise de dados baseadas em hierarquias de classes de *LOD* mencionadas nas anotações de interesse.

Este trabalho é motivado pela escassa literatura que combina análise dimensional de dados de mídias sociais com enriquecimento semântico. A análise de propostas correlatas ao problema proposto na literatura acadêmica encontrou abordagens que apresentam soluções parciais. Tais propostas foram agrupadas em:

1. sistemas de recomendação baseados em *TCs* anotados manualmente em fóruns *on-line* (ABRAHAMS et al., 2012) ou em mídias sociais (VILLANUEVA et al., 2016)
2. anotação semântica automática de *TCs* (FILETO et al., 2015; SACENTI et al., 2015; FRANCIA; GOLFARELLI; RIZZI, 2014);
3. construção e operação de bancos de dados dimensionais para processamento analítico *on-line* (OLAP) de *TCs* brutos (TAO et al., 2016) ou semanticamente anotados (NEBOT; BERLANGA, 2012; FRANCIA; GOLFARELLI; RIZZI, 2014; CUZZOCREA et al., 2015; FILETO et al., 2015; SACENTI et al., 2015; CHOUDER; RIZZI; CHALAL, 2019).

Assim, até onde sabemos, a nova abordagem automatizada para detecção e análise de menções de interesse em *TCs* proposta neste documento é a única que integra anotação semântica, filtragem de anotações de interesse baseada em pontes para uma ontologia de domínio e gera automaticamente dimensões de análise de informações em um processo de *ETL* para *BI* semântico em *TCs*. Uma ontologia pode ser definida como uma “conceitualização compartilhada de um universo de discurso” (GUARINO et al., 1998). Neste trabalho anotamos *TCs* com recursos de coleções de *LOD*, as quais têm ontologias subjacentes, que

funcionam como modelos conceituais para descrever os conceitos e relações semânticas cobertos por tais coleções (NGOMO et al., 2014). Uma ponte associa uma instância ou classe de uma ontologia geral de *LOD* usados em anotações semânticas a uma classe de uma ontologia de alto nível de domínio específico.

As principais contribuições deste trabalho são:

1. Definição de um processo geral para enriquecer, filtrar e analisar semanticamente menções de interesse em (*TCs*);
2. Criação de um algoritmo para verificar e completar pontes entre conceitos e dados abertos ligados - *LOD* (classes dos valores das anotações semânticas de *TCs*) a conceitos de uma ontologia de domínio de alto nível;
3. Adaptação de técnicas previamente propostas para construir automaticamente dimensões para analisar dados semanticamente anotados (SACENTI et al., 2015) para também filtrar menções de interesse com base nas pontes melhoradas.

Apesar de utilizar abordagens análogas especialmente na etapa de anotação semântica dos *TCs*, este trabalho difere da abordagem de (SACENTI et al., 2015) na construção das dimensões de análise, visto que estas dimensões agora são construídas usando pontes entre conceitos de ontologias (de *LOD* e de domínio). Na abordagem proposta neste trabalho, um pequeno conjunto de tais pontes precisa ser gerado por especialistas de domínio, para então ser verificado e ampliado usando um algoritmo que projetamos, implementamos e testamos para tal finalidade.

A avaliação da proposta deste trabalho é realizada por meio de um estudo de caso na área de negócios (*business*). Resultados experimentais mostram a distribuição de classes e instâncias da DBpedia mencionadas em amostras do Twitter e as características de cada dimensão de análise gerada automaticamente a partir da hierarquia de *subsumption* da DBpedia. Estas dimensões permitem novas consultas analíticas úteis para o processo de tomada de decisão. São exemplos destas novas consultas:

- *Q1: “Quais são as organizações mais mencionadas nos tweets enviados do Brasil, durante um certo período de tempo?”*
- *Q2: “Quais são os produtos ou serviços mais mencionados nos tweets enviados na cidade de Florianópolis nos últimos 30 dias?”*

- Q3: “Quais são os produtos de informática mais mencionados em tweets no Estado de Santa Catarina no ano de 2018?”

## 1.1 PROBLEMA

Este trabalho objetiva investigar e avaliar as incidências de instâncias e classes de interesse para algum domínio (e.g. negócios) em *TCs* tais como postagens de mídias sociais. Tal investigação é feita por meio de um processo que primeiramente efetua anotação semântica do *TCs* usando ferramentas de anotação existentes e então contabiliza anotações cujos os valores são recursos de *LOD* tomados de uma coleção usada para anotação semântica (e.g. DBpedia). Posteriormente, o processo proposto emprega associações de equivalência e subclasse entre classes da ontologia subjacente a tal coleção de *LOD* que foram usadas direta ou indiretamente em anotações e classes de interesse de uma ontologia de domínio específico, tal como negócios, para filtrar as anotações de interesse.

Neste trabalho, estas associações entre classes (da ontologia de anotação e da ontologia de domínio) são chamadas de **pontes**. Umhas poucas dezenas de tais pontes são inicialmente criadas por especialistas de domínio, e em seguida verificadas e expandidas de forma automática por um algoritmo proposto. O conjunto de pontes verificado e expandido é usado para filtrar anotações de interesse e construir um *data warehouse* com dimensões específicas para análise de informação anotada e selecionada nos *TCs* conforme o domínio em questão. Assim, as análises *OLAP* no cubo de dados resultante são mais específicas e semanticamente ricas do que as realizadas em um *Data Warehouse* convencional para analisar os *TCs* somente de acordo com seus tempos de envio, localizações e palavras-chaves do seu conteúdo (TAO et al., 2016; CUZZOCREA et al., 2015; WITTEWER et al., 2016).

Usuários de mídias sociais podem ser fornecedores, vendedores e, acima de tudo, consumidores de produtos ou serviços. Entidades específicas de domínio (e.g., entidades comerciais: empresas, instituições, organizações) podem ser mencionadas em *Textual Clips - TCs* como postagens em mídias sociais (e.g. tweets). É comum os usuários expressarem intenções de compra ou interesses nessas postagens. Todavia, o alto volume de dados e textos livres com muitos ruídos tornam a tarefa de análise de dados de certos *TCs*, particularmente postagens em mídias sociais, uma tarefa não trivial. O trabalho de (RITTER et al., 2011) mostra que as postagens possuem características que impedem

sua compreensão automática. Assim, anotações semânticas são úteis para desvendar e analisar o conteúdo desses *TCs*.

As informações extraídas são expressas através de anotações semânticas, aderentes a padrões de dados semiestruturados, que permitem a representação de dados com semântica bem-definida. Essas anotações promovem o desenvolvimento de novas técnicas e ferramentas se apoiam em semântica para *data warehousing*, análise de informações, análise de sentimentos, *data mining* e sistemas de suporte à decisão, entre outras possibilidades.

Este trabalho propõe uma abordagem que combina tecnologias da Web Semântica (ferramentas de anotação semântica e *Linked Open Data - LOD*) (NGOMO et al., 2014) e tecnologias de *Business Intelligence (BI)* (uso do modelo dimensional e *OLAP*) (GOLFARELLI; MAIO; RIZZI, 1998; KIMBALL; ROSS, 2011) para analisar as incidências de menções de interesse em Clipes Textuais (*Textual Clips - TCs*). A principal contribuição deste trabalho consiste em um processo de Extração, Transformação e Carrega (ETL) que anota semanticamente *TCs* - como postagens de mídias sociais - e constrói *data warehouses* para analisar os dados anotados. Após anotar os *TCs*, as incidências de recursos de *LOD* (instâncias e classes) oriundas das anotações de *TCs* são contadas. Em seguida, hierarquias de *subsumption* de classes e hierarquias de composição de instâncias já existentes em coleções de *LOD* são adaptadas para servir como dimensões de análise de acordo com as incidências de suas classes e instâncias nas anotações de *TC* e usando pontes para uma ontologia específica de domínio para filtrar as classes e instâncias de interesse para análise usadas nas anotações do *TCs*. Este trabalho parte de um pequeno número de pontes fornecidas por especialistas de domínio entre as principais classes de *LOD* encontradas nas anotações de *TCs* e a ontologia de domínio para validar tais pontes, derivar novas pontes, e então usar as pontes resultantes para filtrar os recursos de interesse.

O processo proposto é capaz de responder às seguintes questões acerca de conceitos mencionados em Clipes Textuais (*TCs*):

1. "Quais conceitos (classes ou instâncias) são mais mencionados em um conjunto de *TCs* (e.g. *tweets*)?"
2. "Quais classes são candidatas para construir dimensões em um cubo de dados de um domínio específico?"
3. "Considerando uma determinada dimensão, quais subclasses desta dimensão são mais mencionadas nos *tweets*?"

Este trabalho tem as seguintes delimitações de escopo:

- a proposta se concentra em anotações semânticas de cliques textuais, com classes e instâncias extraídas de repositórios de *Linked Open Data* (e.g. DBpedia) e catálogos e vocabulários específicos de domínio (e.g. eBay, GoodRelations Ontology - GRO (HEPP, 2008));
- análise de sentimento, criação de perfis de usuários a partir das postagens, recomendação e *data mining* estão fora do escopo deste trabalho. Estes temas podem ser abordados em trabalhos futuros.

## 1.2 PERGUNTA DE PESQUISA

Este trabalho busca responder à seguinte questão de pesquisa: “É possível criar um processo de *ETL* semanticamente expandido que usa anotações semânticas de Cliques Textuais (*TCs*) para criar um modelo dimensional que traga mais expressividade às análises *OLAP* desses *TCs* de modo a melhor subsidiar a tomada de decisão?”.

## 1.3 OBJETIVOS

O objetivo geral deste trabalho é desenvolver e avaliar um processo de *ETL* para anotação semântica e análise de menções em *TCs* que sejam relevantes para domínios de aplicação específicos, com um estudo de caso em negócios.

Este trabalho tem os objetivos específicos, listados a seguir:

1. Melhor compreender o estado-da-arte em análise de informações de Cliques Textuais (*Textual Clips - TCs*) utilizando tecnologias de *Web* semântica e *BI* através de uma revisão bibliográfica sistemática, com seleção e comparação de abordagens e ferramentas.
2. Investigar possibilidades de expansão do processo de *ETL* para anotar *TCs* semanticamente e montar um modelo dimensional para analisar o conteúdo anotado.
3. Pesquisar métodos e algoritmos para gerar, verificar a consistência e completar pontes entre classes de *LOD* e classes de uma ontologia de domínio, para então usar essas pontes para adaptar hierarquias existentes de recursos de *LOD*, de modo a servir como dimensões de análises *OLAP*.

4. Verificar em experimentos a viabilidade da proposta em um estudo de caso no domínio de negócios, comparando a abordagem apoiada pelo processo proposto e as funcionalidades das atuais ferramentas de análise de mídias sociais.

## 1.4 MATERIAIS E MÉTODOS

A metodologia empregada neste trabalho compreende a sequência de passos descritos a seguir:

1. Revisão sistemática bibliográfica em áreas de *Social CRM* e *e-business*, uso de tecnologias web semânticas para construção de *DW* e análise de informações em mídias sociais em *DW*.
2. Concepção do processo *ETL* de referência que suporta a abordagem proposta;
3. Realização de experimentos para anotar semanticamente tweets com classes de *LOD*. Conjuntos de dados (*datasets*) de tweets foram obtidos a partir de trabalhos anteriores no LISA/UFSC ((SORATO et al., 2016), (SACENTI et al., 2015), (FILETO et al., 2015)).
4. Desenvolvimento e implementação de algoritmos para verificar e completar pontes entre os recursos e classes de *LOD* de uma ontologia de domínio de alto nível, filtrando menções de interesse para um domínio específico (estudo de caso na área de negócios).
5. Desenvolvimento e implementação de algoritmos para adequar as dimensões a um Cubo de Dados Semânticos (*Semantic Data Cube - SDC*) com anotações e pontes obtidas nas etapas anteriores;
6. Experimentos para validar os algoritmos e toda a proposta.
7. Redação de artigo científico relacionado ao trabalho proposto e sua publicação em evento com o Qualis-CC CAPES, com estrato superior ou equivalente a B3.
8. Escrita e defesa da dissertação.

## 1.5 ESTRUTURA DO TRABALHO

O restante deste trabalho está organizado em seis capítulos. O Capítulo 2 descreve os fundamentos, tecnologias, estruturas básicas e



definições necessárias para entender a proposta. O Capítulo 3 apresenta novas definições para resolver o problema abordado, e em seguida descreve o processo proposto para a realização do enriquecimento semântico em cliques textuais e a análise dos dados enriquecidos em modelos multidimensionais. O Capítulo 4 relata experimentos aplicados em um estudo de caso com uma instância do processo proposto, na área de negócios (*business*) e comércio eletrônico. O Capítulo 5 compara a proposta deste trabalho e os resultados com trabalhos correlatos. Finalmente, o Capítulo 6 resume as contribuições e apresenta perspectivas sobre trabalhos futuros.



## 2 FUNDAMENTOS

Este capítulo discute os fundamentos da Web Semântica e do modelo dimensional, conceitos basilares para compreender a proposta deste trabalho. Primeiro descreve, na Seção 2.1, ontologias e Dados Abertos Ligados (*Linked Open Data - LOD*) como fontes primárias de recursos (descrições de conceitos e instâncias com semântica bem definida) para anotações semânticas.

Este enriquecimento semântico é detalhado na Seção 2.2, definindo e exemplificando as anotações semânticas e os recursos envolvidos. Tais recursos podem ser armazenados em grafos de conhecimento (*Knowledge Graphs - KGs*) e usados como valores de anotação por métodos e ferramentas de anotação semântica, apresentados na Seção 2.3. A organização dos recursos conforme alguns de seus relacionamentos semânticos básicos (*e.g.*, *subClassOf*, *typeOf*, *is-a*) é crucial para a filtragem e análise de informações em modelos dimensionais, conforme demonstrado na Seção 2.4.

### 2.1 ONTOLOGIAS, DADOS ABERTOS LIGADOS (*LOD*) E GRAFOS DE CONHECIMENTO (*KGS*)

Anotações semânticas associam semântica bem-definida com porções de dados de destino, como textos e multimídia, vinculando-os a recursos (*e.g.*, classes, instâncias, *synsets*) descritos em um Grafo de Conhecimento (*KG*). Tais recursos podem ser descritos em ontologias, coleções de Dados Ligados (*LOD*) (*e.g.*, DBpedia<sup>1</sup>, Wikidata<sup>2</sup>), em *lexicons* (*e.g.* WordNet<sup>3</sup>), ou na combinação de coleções de *LOD* e *lexicons* (*e.g.* Babelnet<sup>4</sup>).

Nas últimas décadas, progressos em áreas como mineração de texto e processamento de linguagem natural (*Natural Language Processing - NLP*) permitiram a anotação semântica automática de textos usando tarefas como Reconhecimento de Entidades Nomeadas (*Named Entity Recognition - NER*), Desambiguação de Entidades Nomeadas (*Named Entity Desambiguation - NED*) e *Entity Linking (EL)* (MORO; RAGANATO; NAVIGLI, 2014). Várias técnicas e ferramentas estão agora

---

<sup>1</sup><http://wiki.dbpedia.org>

<sup>2</sup><http://www.wikidata.org>

<sup>3</sup><https://wordnet.princeton.edu/>

<sup>4</sup><https://babelnet.org/>

disponíveis para localizar menções no texto e ligá-las aos recursos de Grafos de Conhecimento (*KGs*) de propósito geral (SPECK; NGOMO, 2014; MENDES et al., 2011; MORO; RAGANATO; NAVIGLI, 2014).

A Figura 2 ilustra as anotações semânticas obtidas após a utilização da ferramenta DBpedia-Spotlight em um tweet, apresentado no canto superior esquerdo da figura. As duas anotações geradas para menções distintas identificadas neste tweet (indicadas pelos respectivos links rotulados com *target*) apontam para os recursos apresentados à direita da Figura 2 (através dos respectivos links rotulados com *body*). Outro detalhe apresentado nesta figura é a relação entre a instância [http://dbpedia.org/page/Dell\\_XPS](http://dbpedia.org/page/Dell_XPS) e a classe <http://dbpedia.org/page/Laptop>, através de um *link* rotulado com o nome *rdfs:type*.

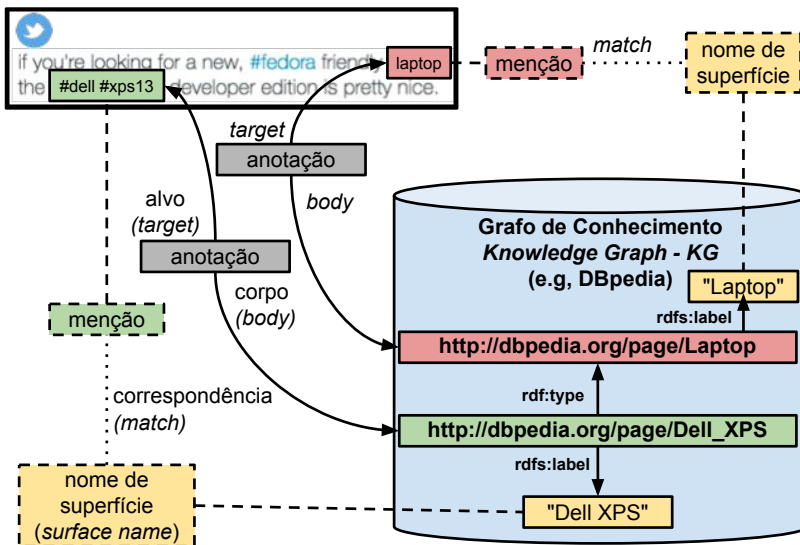


Figura 2 – Exemplo de anotação semântica em um Clipe Textual

No escopo deste trabalho, a DBpedia é a coleção de *LOD* que organiza e representa as anotações direcionadas nas postagens de mídias sociais. Adicionalmente, podemos usar ontologias de domínio de alto nível para descrever conceitos e relações entre conceitos para um dado domínio de conhecimento, como o DeCS <sup>5</sup>(descritores em Ciências da Saúde), que representa as hierarquias e relações entre termos e tam-

<sup>5</sup><http://decs.bvs.br/>

bém conceitos relacionados ao campo da medicina, ou a *GoodRelations Ontology (GRO)* <sup>6</sup>, destinada à representação de conceitos e instâncias de negócio (*business*).

Para complementar a proposta, é necessário determinar filtros para as anotações geradas pela ferramenta de *NER/NED* adotada, selecionando resultados de classes e instâncias de conceitos relacionados à área de negócios. Esse processo de filtragem é o núcleo da proposta desta dissertação, determinado através do conceito de *pontes* entre os recursos de *LOD* e classes de ontologias de domínio de alto nível.

## 2.2 ANOTAÇÕES SEMÂNTICAS

Enriquecimento semântico é o processo que consiste em adicionar metadados aos dados, a fim de facilitar sua descrição ou caracterização. Este processo complementa os dados com metadados descritivos sobre ele. Este enriquecimento pode ser feito através de anotações semânticas. Anotação é qualquer informação anexada a um objeto ou parte de um objeto. Uma anotação liga dois fragmentos de dados, alvo (*target*) e corpo (*body*). O alvo identifica quem é anotado (o fragmento do documento que receberá a anotação) e o corpo aponta para metadados (e.g., recursos, que podem ser conceitos ou instâncias descritas em uma base de conhecimento).

Textos em linguagem natural podem mencionar nomes de superfície de entidades, i.e., referências léxicas para tais entidades. Uma mesma entidade (e.g., a cidade de “Florianópolis”) pode ser mencionada em um texto através de vários nomes de superfície alternativos (e.g., “Florianópolis”, “Floripa”, “Fpolis”, “capital de Santa Catarina”). Uma menção a uma entidade em um texto consiste em um fragmento do texto correspondente a um nome de superfície. Tal menção pode ser semanticamente anotada mesmo que haja ambiguidade, i.e., várias semânticas possíveis para a menção (e.g., “Santa Catarina” pode ser referir ao estado brasileiro, à ilha com tal nome do litoral de tal estado, a uma santa ou até à marca de algum produto ou nome de alguma empresa. Em uma anotação semântica, o alvo (*target*) é uma menção no texto e o corpo (*body*) é uma URI que aponta para uma descrição da menção anotada em um Grafo de Conhecimento (*KG*). Cabe ao interpretador/anotador (humano ou computacional) selecionar para cada menção o melhor alvo candidato, i.e., o recurso do *KG* que tenha um nome de superfície associação à menção e que melhor descreva a

---

<sup>6</sup><http://www.heppnetz.de/projects/goodrelations/>

semântica da menção no contexto em que ela foi detectada.

O lado esquerdo da Figura 2 mostra um exemplo em que um *Textual Clip (TC)* (e.g. um tweet) é anotado por uma ferramenta de *NER/NED (NERD)*. Nesta figura, “laptop” é uma menção de uma classe e “#dell/xps13” é uma menção de uma instância no Grafo de Conhecimento (*KG*) da DBpedia. Ambos mencionam nomes de superfície (*surface names*) na DBpedia, associados a propriedades de recursos da DBpedia (entidades nomeadas ou conceitos). A anotação semântica consiste em uma relação com um alvo (*target*) que aponta para a menção e um *body* que aponta para a *URI* do recurso. Outros relacionamentos na Figura 2 são representados por conexões entre recursos da DBpedia. Este exemplo mostra algumas das formas de enriquecimento semântico através de anotações semânticas.

As entidades nomeadas encontradas nas anotações apontam para descritores no *KG*. A Figura 3 mostra dois recursos (uma instância e uma classe) ligados pelo relacionamento *rdf:type*, e as respectivas propriedades.

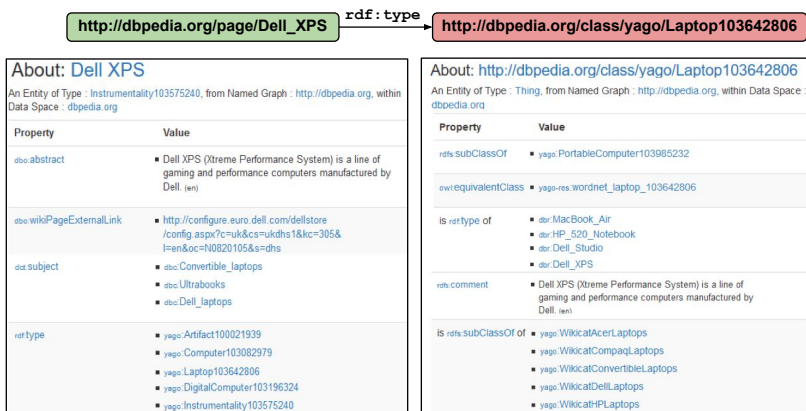


Figura 3 – Exemplo de Entidades Nomeadas anotadas em um Clipe Textual

As anotações semânticas podem ser geradas manualmente, semi-automaticamente ou automaticamente. O processo de anotar entidades nomeadas (e.g. instituição, produto, serviço) no texto pode ser automatizado com ferramentas de *NER/NED (NERD)* (MENDES et al., 2011) (SPECK; NGOMO, 2014). Em nosso trabalho, selecionamos ferramentas automatizadas de *NERD* com ênfase em código-fonte aberto, gratuito

e maior adoção na literatura, conforme detalhado na Seção 2.3 a seguir.

### 2.3 FERRAMENTAS PARA ANOTAÇÃO SEMÂNTICA AUTOMÁTICA

Este trabalho investiga a incidência de menções relacionadas a negócios através de um processo proposto, mais detalhado no Capítulo 3. Para realizar algumas tarefas neste processo proposto, é necessário escolher um conjunto de ferramentas que funcionem de maneira coordenada. Foram escolhidas ferramentas que fazem as anotações através de *Entity Linking (EL)*, via tarefas de Reconhecimento de Entidade Nomeada (*NER*) e de Desambiguação de Entidades Nomeadas (*NED*), também agrupadas como tarefas *NERD* (Reconhecimento de Entidades Nomeadas e Desambiguação).

O objetivo é usar ferramentas de *NERD* para obter anotações semânticas automatizadas para *Textual Clips (TCs)* a partir de uma amostra de *tweets* de um *dataset*. A primeira escolha de ferramentas para anotação semântica compreende duas opções: DBpedia-Spotlight (MENDES et al., 2011) e *Federated Knowledge Extraction Framework - FOX* (SPECK; NGOMO, 2014). O DBpedia-Spotlight é uma ferramenta para anotação de menções na DBpedia a partir de textos em linguagem natural. Esta ferramenta anota instâncias ou conceitos da Coleção *LOD* DBpedia e fornece funcionalidades para *NER/NED*, resolução de nomes e outras tarefas de extração de informações. O DBpedia-Spotlight permite aos usuários configurar as anotações para suas necessidades específicas por meio da ontologia DBpedia e medidas de qualidade como destaque, pertinência do tópico, ambiguidade contextual e confiança na desambiguação.

A Figura 4 mostra um exemplo de anotação gerada pelo DBpedia-Spotlight, onde o *TC* é fornecido a partir do mesmo tweet dos exemplos anteriores. A ferramenta recebe como entrada o mesmo texto do tweet anteriormente apresentado na Figura 2, com a configuração dos parâmetros *confidence* (0.5) e *language* (inglês). Como resultado, a ferramenta identificou oito menções no texto (palavras sublinhadas). Assim, identificamos manualmente algumas referências que não estão relacionadas a área de *business*, como “fedora” (chapéu), e achamos que a menção correta (neste contexto) deve ser para o Sistema Operacional Fedora. Ao filtrar (ainda manualmente) conceitos e instâncias relacionadas aos negócios, são obtidas menções à empresa Dell e à classe laptop, destacadas na Figura 4. Também notamos que o DBpedia-Spotlight, neste

exemplo, não identificou a instância “Dell XPS”, como mostrado nos exemplos anteriores. Portanto, conclui-se que estudos são necessários para parametrizar esta ferramenta, bem como uma comparação com outras ferramentas *NERD* disponíveis, como a *FOX*.

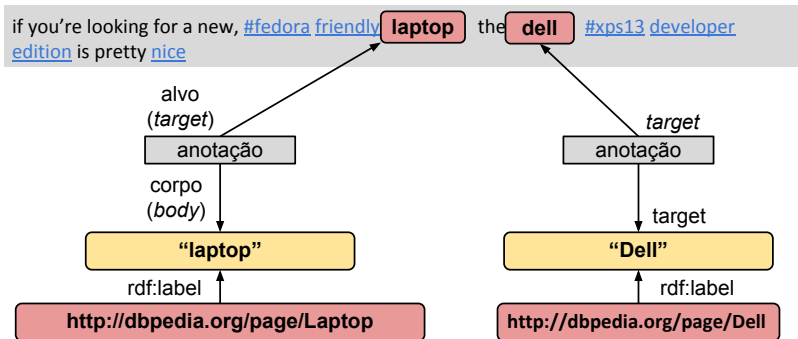


Figura 4 – Exemplo de anotação da ferramenta DBpedia-Spotlight

A ferramenta DBpedia-Spotlight permite definir alguns parâmetros configuráveis, tais como:

1. o nível de confiança das anotações (*confidence*);
2. o idioma do texto de entrada (*e.g.*, idioma de *textual clips - language*);
3. uma opção para trazer os *n*-candidatos para a anotação (não apenas o melhor candidato);
4. uma opção para selecionar os tipos de anotação (*type*). Esses tipos são as taxonomias e o Grafo de Conhecimento *KG* que são utilizados para determinar os possíveis corpos (*bodies*) das menções.

Este trabalho tem um cuidado especial com os parâmetros *type* na ferramenta DBpedia-Spotlight, *language* e *confidence*. Estes parâmetros permitem ajustar a ferramenta para obter melhores anotações direcionadas para os recursos específicos do domínio (entidades ou classes) considerando que o conjunto de *TCs* está em sua maior parte em um único idioma.

O trabalho também planejou usar o *Federated Knowledge Extraction Framework (FOX)*, uma ferramenta *NER* que integra várias



ferramentas para obter entidades nomeadas. Todavia, esta ferramenta funciona apenas com entidades nomeadas. A Figura 5 mostra as entidades encontradas por esta ferramenta usando o mesmo Clipe Textual de entrada usado anteriormente no DBpedia-Spotlight.

The screenshot shows the FOX interface with the following configuration: Lang: en, Input Format: text/html, Extraction Type: ner. The input text is "Dell XPS 13 13.3" QHD+ IPS Touchscreen Notebook Core i5 8GB Ram 256GB SSD 2.3GHz Save \$500". The entities "Dell XPS", "Core i5 8GB", "Ram", and "Save \$500" are highlighted. Below the screenshot, a diagram illustrates the annotation process: a box labeled "alvo (target)" points to "Dell XPS" in the text, which is also labeled as "anotação". This "anotação" points to the "corpo (body)" of the entity, which is "Dell XPS". The "corpo (body)" is linked via "rdf:label" to the URI "http://dbpedia.org/page/Dell\_XPS".

Figura 5 – Exemplo de Clipe Textual anotado com o FOX

O FOX combina resultados (através do *ensemble*) de outras ferramentas do estado-da-arte em *NED/NED (NERD)*: Reconhecedor de Entidades Nomeadas Stanford <sup>7</sup>, Identificador de Entidades Nomeadas Illinois <sup>8</sup>, Extrator de Informações de Baseline de Ottawa (Balie) <sup>9</sup>, Localizador de Nomes Apache OpenNLP (OpenNLP) <sup>10</sup>. No entanto, o FOX tem um fator impeditivo para este trabalho: a ferramenta não está pronta para o Português, limitando o trabalho com o posterior estudo de caso com *textual clips* coletados no Brasil, já que a maio-

<sup>7</sup><https://nlp.stanford.edu/software/CRF-NER.shtml>

<sup>8</sup>[https://cogcomp.org/page/software\\_view/NETagger](https://cogcomp.org/page/software_view/NETagger)

<sup>9</sup><http://balie.sourceforge.net/>

<sup>10</sup><https://opennlp.apache.org/>

ria dos textos publicados em nosso país estão nesse idioma. Por esta razão e também por razões de facilidade e flexibilidade, a ferramenta DBpedia-Spotlight foi escolhida para a aplicação dos experimentos deste trabalho.

## 2.4 MODELO DIMENSIONAL

Anotações semânticas sobre um conjunto de *TCs* resultam em um conjunto de descritores de recursos em *KGs*. Estes recursos têm relações com outros recursos no *KG*, constituindo assim uma Rede Semântica (*Semantic Network - SN*) (SCHIEL, 1989). Esta *SN* é um grafo de conceitos unidos por relações hierárquicas. As hierarquias geradas devem ser organizadas para melhorar a qualidade da análise de BI através de Data Warehouses. Neste trabalho, adotamos a nomenclatura Hierarquia de Recursos Atingidos (*Hierarchy of Hit Resources (HHR)*) para definir uma *SN* mais específica, detalhada no Capítulo 3.

*Data Warehouses (DWs)* (KIMBALL; ROSS, 2011) são bancos de dados multidimensionais com o objetivo de analisar e prever tendências em um domínio. O modelo multidimensional de *DWs* organiza dados em fatos e dimensões de análise, a fim de produzir um cubo de dados, que tenha medidas de interesse. Os fatos podem ser organizados de forma hierárquica, como o Modelo Dimensional de Fatos (*Domain Fact Model - DFM*), apresentado por Golfarelli (GOLFARELLI; MAIO; RIZZI, 1998). Este trabalho é baseado no *DFM* definido por Sacenti *et al.* (SACENTI *et al.*, 2015).

Técnicas de *Business Intelligence (BI)*, como o modelo dimensional (GOLFARELLI; MAIO; RIZZI, 1998) e o *OLAP*, são úteis para averiguar informações extraídas da mídia social, para uma variedade de domínios e aplicativos, incluindo *business* e *social CRM* (WITTWER *et al.*, 2016). A combinação de BI com tecnologias semânticas da Web pode alavancar a análise de menções relevantes em *TCs*.

A Figura 6 ilustra um esquema dimensional para analisar menções de interesse para negócios em tweets, na notação de Golfarelli (GOLFARELLI; MAIO; RIZZI, 1998), com ordenação parcial inversa dos níveis de dimensão (de maior para menor granularidade). Este esquema apresenta uma tabela fato **TCMeasures** com as medidas **#TCs**, **#users** e **#mentions**. O modelo apresenta as dimensões espaço-temporais comuns à *Data Warehouses*, **Time** e **Place**. Por sua vez, as dimensões **Product or Service** e **Business Entity** são derivadas de hierarquias de classes e instâncias DBpedia mencionadas no *TCs*. Por exemplo, a

dimensão `Product or Service` inclui em seus níveis as classes DBpedia `Place`, `information appliance`, `battery charger` e `laptop`.

O processo proposto deriva automaticamente dimensões de domínio específico considerando os recursos de *LOD* existentes usados para anotar *TCs* e construir um *data warehouse*. Com tal esquema dimensional, é possível analisar menções de interesse para um domínio particular nos *TCs*, conforme descrito no Capítulo 3. Resultados experimentais detalhados são apresentados no Capítulo 4.

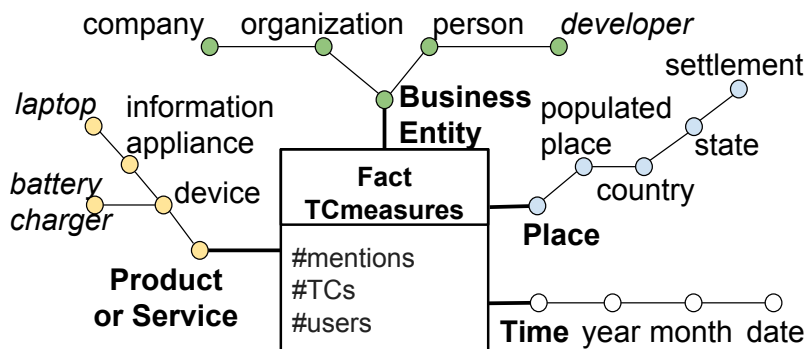


Figura 6 – Exemplo de modelo dimensional

## 2.5 CONSIDERAÇÕES FINAIS

Este capítulo apresentou os fundamentos que amparam o objetivo geral deste trabalho. Para tal, foram apresentados conceitos de Web Semântica, incluindo: (i) organização de representação do conhecimento (ontologias, Dados Abertos Ligados (*Linked Open Data - LOD*) e Grafos de Conhecimento (*Knowledge Graphs - KGs*); (ii) anotações semânticas automáticas (através de ferramentas *open source*) e, por fim, (iii) o modelo dimensional para organizar os conceitos das bases de conhecimento selecionados por anotações sobre *textual clips*, viabilizando assim análises *OLAP* sobre uma entrada de dados originalmente não-estruturada. Por conseguinte, o Capítulo 3 a seguir apresenta a proposta deste trabalho, detalhando cada etapa do processo de *ETL* semântico proposto.



### 3 PROCESSO DE BI SEMÂNTICO PROPOSTO

Este capítulo descreve a principal contribuição deste trabalho, um processo de de Extração, Transformação e Carga (*ETL*) que permite detectar e analisar menções de interesse para um domínio particular em *TCs*. Este processo percorre todas as anotações geradas automaticamente sobre um conjunto de *TCs*, constrói dimensões de análise baseadas em hierarquias de *LOD* adaptadas a um domínio. O resultado é um Cubo de Dados Semântico que possibilita análise das informações contidas no *TCs* usando *OLAP* com hierarquias de instâncias e classes de *LOD* mencionadas direta ou indiretamente nos dados anotados. A Figura 7 ilustra o fluxo de tarefas do processo proposto agrupadas em 3 estágios/etapas: 1:Enriquecimento Semântico e Contagem, 2:Construção do Cubo Semântico (*SDC*) e 3:Análise da Informação. As principais contribuições deste trabalho estão destacadas em negrito (tarefas 1d, 2b e 2c).

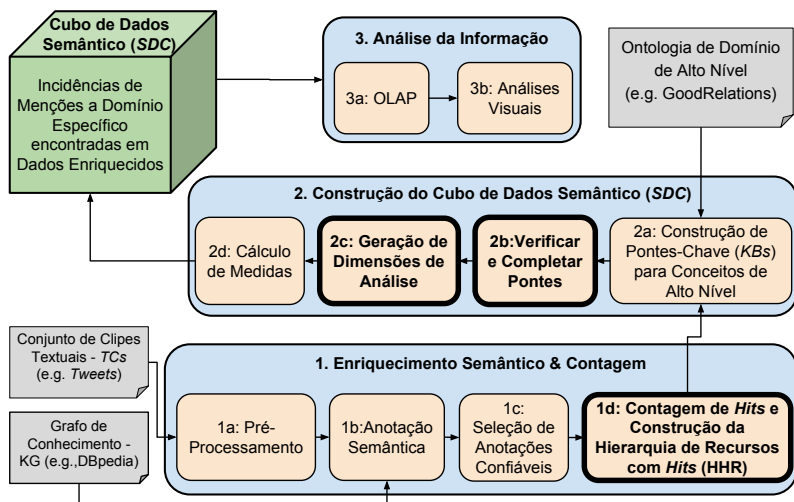


Figura 7 – O processo de *ETL* semanticamente estendido proposto.

O processo inicia pelo estágio 1:Enriquecimento Semântico e Contagem, que compreende uma sequência de quatro tarefas, 1a a 1d. A entrada do processo é um conjunto (*dataset*) de Clipes Textuais (*TCs*), como por exemplo: um arquivo de texto contendo um conjunto de twe-

ets, uma planilha no formato .CSV contendo registros de prontuários médicos, entre outros. No estudo de caso apresentado no Capítulo 4, foi utilizado um *dataset* de tweets para testar a viabilidade do processo.

A tarefa **1a:Pré-processamento** elimina alguns componentes indesejados (*e.g.* emoticons, acrônimos, abreviações) dos *TCs*, ou os substitui por palavras normalizadas a fim de diminuir ruídos. Em seguida, a tarefa **1b:Anotação Semântica** usa ferramentas de *NERD* existentes (*e.g.* DBpedia-Spotlight, Babelify) para vincular menções em textos *TC* aos recursos de um Grafo de Conhecimento - *KG* (*e.g.* DBpedia, Babelnet). A tarefa **1c:Seleção de Anotações Confiáveis** normaliza os valores das anotações geradas, considerando apenas um dentre os recursos conectados pela propriedade *owl:sameAs* e, em seguida, seleciona apenas as anotações mais confiáveis de acordo com critérios da própria ferramenta (*e.g.*, recursos obtidos com um grau de confiança retornado pela anotador *NERD* acima de um determinado limiar). Em seguida, a Tarefa **1d:Contagem de hits & Construção da Hierarquia de Recursos com Hits HHR** constrói a *Hierarchy of Hit Resources - HHR*, isto é, a hierarquia de recursos com *hits* oriundos das anotações selecionadas.

O conceito de *hits* e duas de suas classes (diretos e indiretos) vêm do trabalho de Sacenti *et al.* (SACENTI *et al.*, 2015). *Hits* diretos são aqueles dos valores (*bodies*) das anotações das menções no texto diretamente com classes e instâncias da coleção de *LOD* (*KG*) utilizada pela ferramenta de *NERD*. Os *hits* indiretos, por outro lado, são atingidos através de caminhos indiretos das menções a valores de anotações até outras classes do *KG* via relações semânticas entre classes na ontologia subjacente ao *KG*. O trabalho de Sacenti não faz distinção dos *hits* indiretos atingidos via relações semânticas distintas, *i.e.*, define *hits* indiretos independentemente das relações semânticas percorridas para chegar a eles. Nesta dissertação, os tipos de *hits* indiretos foram refinados para diferenciar aqueles que são atingidos somente via relação *rdf:type* daqueles cujo caminho da menção ao *hit* também passa por pelo menos uma relação *rdfs:subclassOf*. O conceito de *hits*, as descrições dos seus tipos e de sua contabilização são detalhados na Seção 3.1.

A próxima etapa do processo proposto, **2:Construção do Cubo Semântico (SDC)**, começa com a Tarefa **2a:Construção de Pontes Chave (KBs) para Conceitos de Alto Nível**, que depende da interação de especialistas de domínio para fornecer Pontes-Chave (*Key Bridges*) entre as principais classes do HHR (as classes *LOD* com mais *hits* pelas anotações semânticas selecionadas no *dataset* de *TCs*) e

conceitos de uma ontologia de alto nível. A Tarefa 2b: **Verificar e Completar Pontes** é realizada através de algoritmo proposto neste trabalho, que verifica a consistência de *Key Bridges* e gera novas pontes de acordo com as pontes consistentes com a *HHR*. Em seguida, as Tarefas 2c: **Geração de Dimensões de Análise** e 2d: **Cálculo de Medidas** usam as pontes aprimoradas para derivar dimensões de análise do *HHR*, bem como o Cubo de Dados Semânticos (SDC). A seção 3.3 fornece mais detalhes sobre esse estágio, que contempla as principais contribuições desta dissertação.

Finalmente, durante a Etapa 3: **Análise da Informação** ocorrem a Tarefa 3a: **OLAP**, que suporta análise dimensional executando consultas no *SDC* em linguagens como SQL, MDX e SPARQL. Em seguida, Interfaces Gráficas de Usuário (*Graphical User Interfaces - GUI*) da Tarefa 3b: **Análises Visuais** fornecem meios adequados para que os especialistas do domínio obtenham informações do *SDC* para análise estratégica e tomada de decisão informada (e.g., consultas gráficas e navegação em idiomas, gráficos, mapas).

### 3.1 DEFINIÇÕES BÁSICAS

A estrutura de um Clipe Textual (*Textual Clip - TC*) é descrita na Definição 1. Para exemplificar *TCs*, pode-se considerar postagens de mídias sociais ou anotações de prontuários médicos.

**Definição 1** *Um Clipe Textual (**Textual Clip - TC**) é uma tupla:*

$$tc = \langle idTC, idUser, time, text, OM \rangle$$

onde:

*idTC* é o identificador exclusivo de *tc*;

*idUser* é o identificador exclusivo do usuário que produziu o *tc*;

*time* é o timestamp em que *tc* foi criado;

*text* é o conteúdo do texto

*OM* =  $\langle m_1, \dots, m_n \rangle$  ( $n \geq 1$ ) é um conjunto de atributos de metadados associados a *tc*. Cada atributo de metadados  $m_j$  ( $1 \leq j \leq n$ ) é um par  $\langle property, value \rangle$ , em que **property**  $\in P$  é uma propriedade de um conjunto de propriedades *P* e **value** é o valor da propriedade *property*.

Seja  $KG(V, E)$  um grafo de conhecimento (*Knowledge Graph - KG*), onde cada vértice  $r \in V$  representa um recurso do  $KG$ , e cada borda  $\langle r, r', \rho \rangle \in E \subseteq V \times V \times R$  representa um *link* em  $KG$  de  $r$  para  $r'$ , indicando que eles estão conectados pela relação  $\rho \in R$ . Uma anotação semântica usada neste trabalho vincula uma menção identificada no texto de um  $TC$  a um recurso  $r \in V$  do grafo de conhecimento  $KG$ , conforme estabelecido pela Definição 2. Assim, o recurso  $r$  pode ser usado para descrever semanticamente a menção.

**Definição 2** *Dado um Grafo de Conhecimento  $KG(V, E)$  e um Clipe Textual  $tc$ , uma **anotação semântica (SA)** de  $tc$  é uma tupla:*

$$sa = \langle IdTC, m, r \rangle$$

onde:

$IdTC$  é o identificador de  $tc$ ;

$m$  é a menção de  $tc$  que é o alvo (*target*) de  $sa$ ;

$r \in V$  é o recurso do  $KG$  que é o valor de  $sa$ .

As menções de interesse em  $TCs$  podem ser contabilizadas de acordo com os valores de suas anotações semânticas, bem como as classes e/ou superclasses desses valores. Definimos um *hit direto* em um recurso de  $r$  do  $KG$ , uma anotação semântica  $TC$  cujo valor é  $r$ , conforme descrito na Definição 3.

**Definição 3** *Dado um Grafo de Conhecimento  $KG(V, E)$  e um clipe textual  $tc$ , um **hit direto** de  $tc$  em um recurso  $r \in V$  é um par*

$$\langle sa, r \rangle$$

onde:

$sa$  é uma anotação semântica de uma menção no texto de  $tc$ ;

$r$  é o valor de  $sa$ .

Uma ferramenta de *spotting* como o DBpedia-Spotlight gera anotações semânticas cujos valores podem ser instâncias ou classes, pois seu processo de anotação é amparado em correspondências das menções exibidas no texto com nomes de superfície de instâncias ou classes (MENDES et al., 2011). Assim, os *hits* diretos dessas anotações podem mencionar instâncias ou classes da DBpedia.



Além dos *hits* diretos (associações diretas de menções em *TCs* com recursos do *KG*, por meio de anotações semânticas), dois tipos de *hits* indiretos também são relevantes para a análise das menções semanticamente anotadas em *TCs*: *hits* indiretos via *Type* (ou seja, a relação `rdfs:type`) e indiretos via *Subclass* (ou seja, a relação `rdfs:subClassOf`). Um *hit* indireto via *Type* é apresentado na Definição 4.

**Definição 4** *Dado um Grafo de Conhecimento  $KG(V, E)$ , um clipe textual  $tc$  e um hit direto  $\langle sa, r \rangle$  de  $tc$  no recurso  $r \in V$ , um **hit indireto via type** de  $tc$  em uma classe  $r' \in V$  é uma tripla*

$$\langle sa, r, r' \rangle$$

onde:

*$sa$  é uma anotação semântica de uma menção no texto de  $tc$ ;*

*$r \in V$  é o valor de  $sa$ ;*

*$\langle r, r', rdfs:type \rangle \in E$ , indicando que  $r$  é uma instância da classe  $r'$  em  $KG$ .*

Finalmente, um *hit* indireto através de um vínculo de classe para superclasse ocorre quando uma classe  $v''$  do Grafo de Conhecimento *KG* é uma superclasse de uma classe  $v'$  com um *hit* direto ou um *hit* indireto via *type*, como detalhado na Definição 5.

**Definição 5** *Dado um grafo de conhecimento  $KG(V, E)$  e um Clipe Textual (textual clip)  $tc$ , há um **hit indireto via subclasse** de um  $tc$  em um recurso  $r'' \in V$  se e somente se  $\exists \langle r', r'', rdfs:subClassOf \rangle \in E$ , indicando que  $r' \in V$  é uma subclasse de  $r''$ , e há um hit direto de  $tc$  em  $r'$  ou um hit indireto via *type* de  $tc$  em  $r'$ .*

Esses *hits* são contados e usados para extrair hierarquias de recursos do *KG* que possuem *hits*. Estes recursos são usados para anotar *Textual Clips (TCs)* em nossa proposta. Os *hits* diretos induzem acessos indiretos de classes e superclasses. Consideramos apenas o impacto indireto por tipo de uma instância para sua classe mais profunda na hierarquia de *subsumption* da DBpedia. Então, acumulamos o número de *hits* indiretos via subclasse no sentido ascendente, percorrendo do nível da hierarquia de subclasses da DBpedia até a sua raiz `owl:Thing`.

A Figura 8 apresenta um extrato da hierarquia de recursos *LOD* da DBpedia que contém *hits* diretos e/ou indiretos a partir de anotações semânticas produzidas pelo DBpedia-Spotlight para dois tweets

mostrados na parte inferior da figura. Este extrato constitui uma Hierarquia de Recursos com *Hits* (*Hierarchy of Hit Resources - HHR*), cujas folhas possuem *hits* diretos e cujos nós superiores são possuem *hits* partindo das menções de tweets anotados semanticamente. Os números entre parênteses abaixo do rótulo de cada nó representam o número de ocorrências no respectivo recurso DBpedia. O primeiro é o número de *hits* diretos, o segundo, o número de *hits* indiretos por tipo (*type*) e o terceiro, o número de *hits* indiretos via subclasse (*subclassOf*). As instâncias **Brazil** e **Nice** são diretamente atingidas, cada uma com um *hit*. Portanto, suas respectivas classes na hierarquia, **Country** e **Settlement**, têm apenas um *hit* indireto via *type*. Suas respectivas superclasses têm apenas um *hit* indireto via subclasse, e assim por diante.

Embora este exemplo mostre uma *HHR* construída seguindo apenas as relações *type* e *subclassOf*, a partir dos recursos mencionados diretamente em direção ao recurso mais geral, qualquer relação de ordenação parcial pode ser considerada para construir a *HHR* em nossa abordagem.

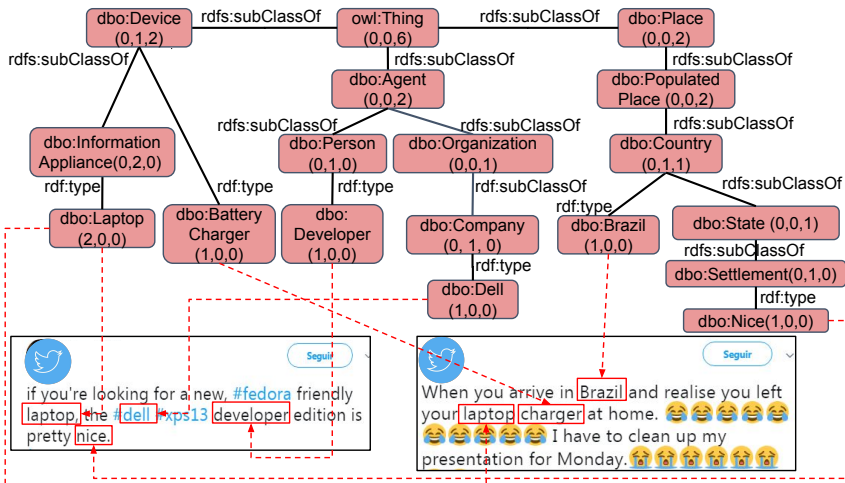


Figura 8 – Classes da DBpedia com *hits* de menções semanticamente anotadas em *tweets*, organizadas em uma Hierarquia de Recursos com *Hits* (HHR)

## 3.2 ENRIQUECIMENTO SEMÂNTICO E CONTAGEM

As subseções a seguir detalham as quatro tarefas que compõem a primeira etapa do processo proposto: **Enriquecimento Semântico e Contagem**. Esta primeira etapa tem como objetivo construir uma hierarquia com os recursos *LOD* (e.g. classes e instâncias da DBpedia) encontrados nas anotações que possuem fragmentos de cliques textuais como alvos. As tarefas do processo envolvem ajustes no *dataset* de *TCs* adotado como entrada do processo (subseção 3.2.1), a escolha de uma ferramenta de anotação automática (subseção 3.2.2), a seleção e filtragem das anotações geradas por tal ferramenta de NERD (subseção 3.2.3), finalizando com a geração da hierarquia de recursos mencionados proposta (subseção 3.2.4).

### 3.2.1 Pré-processamento

Nessa tarefa, os *textual clips (TCs)* recebem tratamentos para eliminar palavras-chave, *emoticons* e outros caracteres indesejados, buscando uma melhor eficiência das ferramentas NERD empregadas. Essa tarefa é fundamental para preservar a precisão dos resultados.

Omitir esta etapa pode gerar resultados inconsistentes, pois podem ocorrer tipos de erros distintos. Dados reais geram frequentemente os seguintes tipos de erros: dados incompletos (e.g., abreviaturas), dados que representam ruídos (e.g., erros ortográficos e gramaticais, gírias, etc.) e inconsistências de dados. As técnicas aplicadas para a correção de problemas nos dados de *TCs* são:

- técnicas de limpeza de dados: remoção de pontuação, caracteres especiais, *hashtags* (caso sejam desnecessárias para a ferramenta de anotação ou técnica de análise adotada);
- redução da quantidade de atributos relevantes ou discrepantes para os experimentos;
- integração de dados, utilizando dados de diferentes fontes para maior confiabilidade;
- seleção de textos em um idioma definido, com o auxílio de um dicionário.

Este trabalho utiliza informações de *Textual Clips (TCs)* (e.g., tweets), com dados não estruturados que possuem palavras de natureza informal, potencialmente contendo vários caracteres especiais, tais

como “|” (*pipe*), *emoticons*, *hashtags* (fora do escopo de anotação semântica deste trabalho) e também *URLs* que causaram problemas às ferramentas de anotação selecionadas. Portanto, é necessário usar a técnica de limpeza de dados. A técnica de redução também foi realizada nesta etapa, filtrando os dados retirados da base de dados LISA em um determinado período e direcionados para a área *e-Business*, apresentada no Capítulo 4.

### 3.2.2 Anotação Semântica

Esta tarefa consiste em selecionar e configurar uma das ferramentas NERD apresentadas na Seção 2.3 e em seguida executar as anotações dos *TCs*. Ao final da execução da ferramenta selecionada, a tarefa obtém um conjunto de anotações realizadas sobre um *dataset* de *Textual Clips*. Conforme já apresentado na seção 2.3, as ferramentas de NERD estudadas por este trabalho foram o DBpedia-Spotlight e o FOX.

A ferramenta utilizada nos experimentos e também em todos os exemplos deste trabalho foi o DBpedia-Spotlight. Por conseguinte, nesta tarefa é necessário configurar os parâmetros da ferramenta (*e.g.* grau de confiança, idioma), chamar o método de anotação e tratar os resultados. Tais resultados podem ser obtidos em formatos diversos (SPARQL, MDX, JSON). Neste trabalho foi desenvolvido um protótipo que trata os resultados obtidos no formato JSON.

### 3.2.3 Seleção de anotações confiáveis

Esta tarefa consiste em selecionar as anotações confiáveis da tarefa (1b). A existência desta tarefa é justificada pela ampla variedade de repositórios de *LOD* que cada ferramenta de anotação pode utilizar e também como cada repositório é organizado. Nesta tarefa são removidas relações de igualdade (*same-as*) entre conceitos com *namespaces* distintos e *URIs* duplicadas.

Um exemplo muito recorrente no uso da ferramenta DBpedia-Spotlight consiste na geração de múltiplos resultados para um mesmo nome de superfície anotado em um *textual clip*. Foram identificados vários corpos (*bodies*) distintos para vários idiomas dado um mesmo alvo *target* anotado, identificando assim o mesmo conceito *LOD* em vários idiomas. A solução encontrada foi padronizar o *namespace* esperado

para os resultados do anotador, por exemplo *pt.dbpedia.org/*. Entretanto, estas opções são parametrizadas no código-fonte desenvolvido, permitindo assim ajustes de acordo com as características da ferramenta NERD selecionada.

### 3.2.4 Contagem de *Hits* e Construção da Hierarquia de Recursos com *Hits* (*HHR*)

A última tarefa da primeira etapa do processo proposto faz uma extensão da proposta de (SACENTI et al., 2015). Tal trabalho foi adaptado por meio de um algoritmo que repete os recursos  $r$  encontrados nas anotações, através dos relacionamentos *type* ou *subClassOf*. O resultado do algoritmo é uma Hierarquia de Recursos com *Hits* associados (*Hierarchy of Hit Resources - HHR*), onde a raiz geralmente é *owl:Thing*.

Os resultados obtidos ao final desta etapa alimentam a etapa 2. **Construção do Cubo de Dados Semântico - (SDC)**, que consiste em verificar quais anotações são aderentes às classes de uma ontologia de domínio de alto nível selecionada (*e.g.* GoodRelations Ontology - GRO). Essa ontologia é o modelo de anotação conceitual e será adaptada para personalizar hierarquias de classes e instâncias de acordo com a incidência de anotações, procurando classes de equivalência entre a ontologia de domínio e bases de conhecimento, descrevendo as entidades e conceitos encontrados nos tweets pela anotação semântica. Tais classes de equivalência fazem parte das pontes entre conceitos GRO para conceitos de base de conhecimento como DBpedia, bases de conhecimento de empresas, etc. Na etapa 2 as dimensões são adaptadas com classes da ontologia de domínio, como detalhado na próxima Seção 3.3.

## 3.3 CONSTRUÇÃO DO CUBO DE DADOS SEMÂNTICO (*SDC*)

As tarefas 2a e 2b desta etapa visam estabelecer pontes entre os recursos da Hierarquia de Recursos com *Hits* (*Hierarchy of Hit Resources - HHR*) extraídos da *LOD* (aqueles atingidos por anotações semânticas selecionadas de anotações em *TCs*) e conceitos de um ontologia de domínio de nível considerado de interesse para definir dimensões por especialistas de domínio. Tais pontes são entradas das tarefas 2c e 2d, a fim de delimitar as dimensões da análise na *HHR* e calcular as medidas de fato para construir o Cubo de Dados Semântico *SDC*. A determina-

ção das pontes é um problema de correspondência de ontologias, que não possui uma solução trivial (SHVAIKO; EUZENAT, 2013).

### 3.3.1 Construção de Pontes-Chave (*KBs*) para Conceitos de Alto Nível

Assim, para iniciar a tarefa de construção de pontes (*bridging*), partimos de um alinhamento parcial de conceitos específicos de domínio com recursos de alto nível de um *HHR* grande. Esse alinhamento parcial depende do suporte humano: Pontes-Chave (*Key Bridges - KBs*) definidas manualmente por especialistas de domínio na Tarefa 2a. Ao longo deste trabalho, nos referimos a estas *Key Bridges* como *KB*.

Cada ponte indica uma correspondência entre:

- uma classe da Hierarquia de Recursos com *Hits* (*Hierarchy of Hit Resources - HHR*). Um exemplo de recurso é uma classe *LOD* da DBpedia encontrada em muitos *TCs* anotados semanticamente;
- um conceito de uma ontologia de domínio de alto nível. Por exemplo: uma classe da GoodRelations considerado de interesse para definir as dimensões específicas do domínio.

Os esforços empregados pelos especialistas de domínio para criar as pontes-chave (*KBs*) são otimizados e minimizados pelo próprio processo de *ETL* proposto, pois ele foi construído com duas premissas:

1. os recursos mais mencionados nas anotações de *Textual Clips* (*TCs*) são apresentados na Hierarquia de Recursos com *Hits* (*HHR*). Desta forma, recursos (classes ou instâncias) pouco mencionados não são passíveis de análise manual dos especialistas, que mantêm o enfoque apenas nos recursos mais recorrentes nas anotações semânticas dos *TCs*;
2. normalmente há um número reduzido de conceitos de uma ontologia de domínio específico referentes a dimensões de análise consideradas relevantes (e gerenciáveis para fins de consulta). Este fato colabora tanto para uma quantidade reduzida de classes candidatas para formar pontes com recursos de *LOD* anotados quanto na geração posterior de dimensões de análise voltadas ao domínio escolhido.

Após a criação das *KBs* para um conjunto reduzido de recursos analisados, é necessário expandir as pontes e também verificar a

consistência, explorando as relações semânticas dos recursos da coleção *LOD* selecionada nas anotações. Para tal, foi proposto o algoritmo *CheckComplete*, conforme detalhado na próxima subseção 3.3.2.

### 3.3.2 Verificar e Completar Pontes

O Algoritmo *CheckComplete* realiza a Tarefa 2b, verificando a consistência das pontes-chave (*KBs*), reportando pontes inconsistentes (*IBs*) e derivando novas pontes a partir das *KBs* consistentes (*NBs*). Este algoritmo recebe as seguintes entradas (*Input*):

- um recurso  $r \in V_{HHR}$ ;
- uma classe  $c$  indicando a classe de ontologia de domínio associada a  $r$  e seus filhos (*null* na primeira chamada);
- o conjunto de *Key Bridges*  $KB$ .

Denotamos o conjunto de nodos da *HHR* como  $V_{HHR}$  e o conjunto de nodos da ontologia de domínio como  $V_{DO}$ ,  $r \in V_{HHR}$  e  $c \in V_{DO}$ .

O *CheckComplete* executa recursivamente uma Busca em Profundidade (*Depth-First Search - DFS*) na Hierarquia de Recursos com *Hits - HHR*, partindo de sua raiz (o valor do parâmetro  $r$  na chamada inicial do *CheckComplete*). Para cada nó  $r$  de cada nível da *HHR*, isto é, cada chamada recursiva do *CheckComplete* (linhas 11 e 13), o *CheckComplete* verifica se existe uma ponte entre  $r$  e um conceito de ontologia de domínio  $c$  (verificações nas linhas 2, 3, 6, 7 e 9). Conforme o tipo de ponte encontrado, o algoritmo adiciona o  $r$  a um dos três possíveis conjuntos de saída (*Output*):

- um conjunto de Novas Pontes (*New Bridges*)  $NB \subseteq KB$  (linhas 4 e 5);
- um conjunto de Pontes Inconsistentes (*Inconsistent Bridges*)  $IB$  (linha 8);
- um conjunto de Recursos com *Hits* e sem Correspondências na ontologia de alto nível (*Unmatched Hit Resources*)  $UHR \subseteq V_{HHR}$  (linha 10).

Esses conjuntos são apresentados aos especialistas de domínio, permitindo por conseguinte correções das *KBs* de maneira iterativa.

Portanto, o algoritmo *CheckComplete* pode contribuir para a consistência e integridade das dimensões de análise Cubo de Dados Semântico (*Semantic Data Cube - SDC*).

---

**Algorithm 1:** *CheckComplete*( $r, c, KB, NB, IB, UHR$ )

---

```

Input:  $r \in V_{HHR}, c \in V_{DO}, KB$ ;
Output:  $NB, IB \subseteq KB, UHR \subseteq V_{HHR}$ ;
1  $dc \leftarrow getDomainClass(r, KB)$ ;
2 if  $c = null$  then
3   if  $dc \neq null$  then
4      $NB.addBridge(r, dc)$ ;
5      $c \leftarrow dc$ ;
6 else
7   if  $dc \neq null \wedge dc \neq c$  then
8      $IB.addBridge(r, dc)$ ;
9   else
10     $UHR.add(r)$ ;
11  $children \leftarrow getChildren(r)$ ;
12 for each  $child \in children$  do
13    $CheckComplete(child, c, KB, NB, IB, UHR)$ ;

```

---

Em termos de complexidade computacional na execução, o *CheckComplete* tem o custo assintótico de uma *DFS* ( $\mathcal{O}(|V| + |E|)$ ), onde o conjunto de vértices é composto pelos recursos da *HHR*, e as arestas representam relações semânticas entre os recursos. Supondo que a hierarquia é uma árvore (um gráfico acíclico direcionado - *DAG*), o número de arestas é o número de vértices - 1. Assim, podemos simplificar a complexidade em  $\mathcal{O}(|V|)$ , onde  $|V|$  representa o número de recursos na hierarquia.

A Figura 9 a seguir apresenta do lado esquerdo os recursos da *HHR* (da DBpedia) com mais hits, ordenados de forma decrescente. No lado direito são ilustradas as principais classes da ontologia de alto nível *GoodRelations Ontology*. As arestas contínuas representam as pontes-chave (*Key Bridges - KB*) criadas pelos especialistas de domínio. A aresta tracejada representa uma ponte nova (*New Bridge - NB*) derivada pelo algoritmo *CheckComplete*. O algoritmo utiliza a ponte-chave (*KB*) entre `dbo:Place` e `gr:Location` e a relação semântica *rdf:subClassOf* entre `dbo:Place` e `dbo:PopulatedPlace` para derivar esta nova ponte.

Finalmente, a aresta pontilhada apresenta uma ponte inconsistente (*Inconsistent Bridge - IB*), pois a superclasse de `dbo:Broadcaster` (`dbo:Organization`) possui uma *KB* com `dbo:BusinessEntity`. Considerando que as pontes são determinadas no sentido da especialização



(partindo do recurso mais geral e mais mencionado em direção ao mais específico), o algoritmo detectou que a ponte entre `dbo:Broadcaster` e `dbo:Brand` é inconsistente (*IB*).

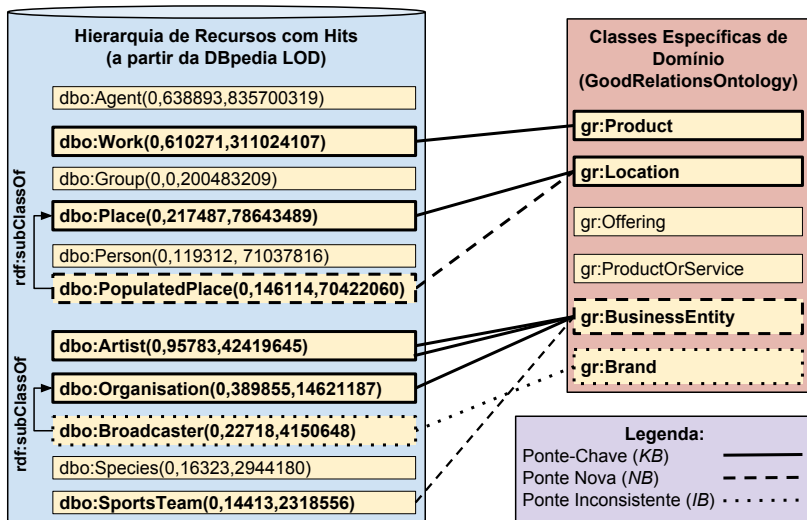


Figura 9 – Exemplos de pontes entre classes de *LOD* (à esquerda) e específicas de domínio (à direita):

- (1) pontes-chave (*KB*), criadas por especialistas de domínio e representadas por linhas contínuas;
- (2) pontes novas (*NB*), inferidas pelo algoritmo *CheckComplete* e representadas por linhas tracejadas;
- (3) única ponte inconsistente (*IB*) neste exemplo, detectada pelo algoritmo *CheckComplete* e representada por uma linha pontilhada.

### 3.3.3 Geração de Dimensões de Análise

A Tarefa 2c deriva as dimensões de análise, ao delimitar na *HHR* subhierarquias cujos nodos são todos ligados a um mesmo conceito de alto nível da ontologia de domínio, e montá-los na dimensão correspondente ao conceito de domínio respectivo. A Figura 10 mostra sub-hierarquias da *HHR* apresentadas na Figura 8 coligadas com os conceitos de domínio `gr:ProductOrService`, `gr:BusinessEntity` e `gr:Location`. Cada dimensão gerada pode ser removida de acordo com o número de ocorrências de cada recurso, como descrito em (SACENTI et al., 2015), para diminuir o número de células no *SDC*.

### 3.3.4 Cálculo de Medidas

A Tarefa 2d consolida na tabela-fato as medidas como a quantidade de cliques textuais ( $\#TCs$ ), quantidade de usuários ( $\#users$ ) e de menções ( $\#mentions$ ) para as combinações de nodos das dimensões derivadas, como ilustrado na Figura 6. As dimensões *Time* e *Place* de tal *SDC* podem se referir à hora e origem da criação do *TC*, respectivamente, anotando cada *TC* como um todo nessas dimensões, com base em seus metadados além de menções (FILETO et al., 2015). Naturalmente, a análise da informação no *SDC* gerado envolve algumas generalizações do modelo *OLAP* tradicional (FILETO et al., 2014).

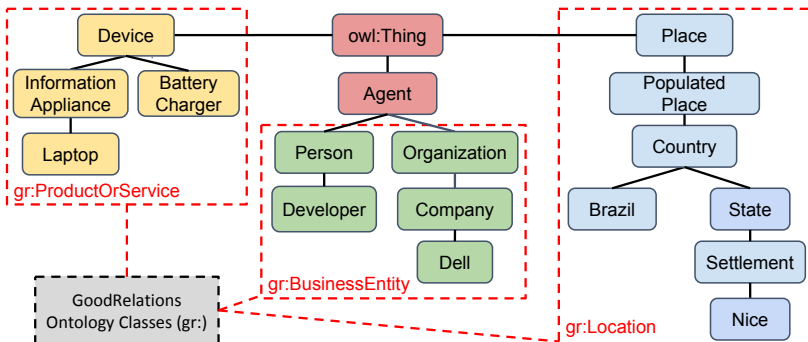


Figura 10 – Hierarquias de dimensões delimitadas em uma *HHR*

O Capítulo 4 apresenta os resultados experimentais em um estudo de caso em *business*, com exemplos de análise suportadas nas

dimensões construídas pela proposta, além de exemplos de consultas SQL afim de responder às perguntas iniciais do trabalho.

### 3.4 ANÁLISE DA INFORMAÇÃO

O estágio final do processo contém duas tarefas para executar a análise de informações sobre o Cubo de Dados Semânticos gerado: (3a) a análise OLAP, combinada com as consultas SPARQL e (3a) a geração de gráficos para os especialistas do domínio.

O modelo dimensional deve ser construído conforme o modelo estrela proposto por Kimball (KIMBALL; ROSS, 2011), contendo uma tabela-fato com as medidas úteis para o domínio escolhido (e.g., quantidade de cliques textuais, quantidade de usuários, *hits* de cada menção, etc.) e dimensões para análise. As dimensões comuns a todos os modelos dimensionais são as espaço-temporais (e.g. *Place*, *Time*). As demais dimensões são construídas através das hierarquias de classes da ontologia de alto nível que foram associadas à *HHR* durante a etapa 2 do processo proposto.

A Figura 11 a seguir apresenta o esquema relacional da base de dados do protótipo construído para os experimentos. A principal tabela deste esquema é *KGNODE*, que representa um recurso (um nodo) em uma base de conhecimento (e.g. *DBpedia LOD*). Esta tabela tem função similar à uma tabela-fato no modelo estrela proposto, pois contém atributos para representar:

1. valores agregados das menções oriundas das anotações semânticas (*hits* diretos e indiretos);
2. o tipo de recurso mencionado (classe ou instância);
3. o próprio recurso, por meio de *URI* e *label*.

A tabela *TWEET* representa uma especialização de um Clipe Textual (*TC*) usada nos experimentos, contendo o texto do clipe, a data de criação do tweet, o identificador do usuário e se foi ou não fruto de um *retweet*. A tabela associativa *KGNODE\_TWEET* representa as anotações encontradas pela ferramenta de *NERD*, pois contém atributos para armazenar:

1. as chaves do alvo (*target*) e do corpo (*body*) da anotação semântica, *IDTWEET* e *IDNODE*, respectivamente;

- o nome de superfície encontrado na anotação e a posição inicial deste nome no *TC* (*offset*).

Estas tabelas são essenciais para as tarefas 1a, 1b e 1c do processo proposto. Por sua vez, a tarefa 1d exige adicionalmente as tabelas *KGNODE\_TYPE*, *KGNODE\_SUPERCLASS*, *SUPERCLASSES\_PATH* e *HIERARCHY*. Com estas tabelas é possível construir a Hierarquia de Recursos com *Hits* (*HHR*), contando os *hits* à medida em que os resultados das anotações são processados.

Finalmente, a tabela *BRIDGE* armazena todas as pontes utilizadas na etapa 2 do processo. Isso inclui tanto as pontes-chave (*KB*) criadas por especialistas de domínio, quanto as pontes novas *NB* e as pontes inconsistentes *IB* identificadas pelo algoritmo *CheckComplete*.

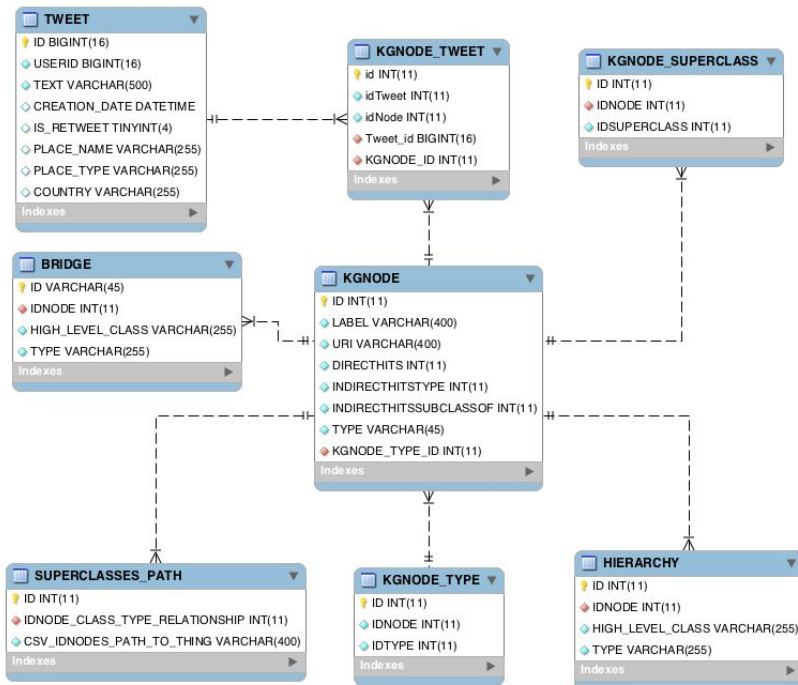


Figura 11 – Diagrama ER da base de dados gerada para análises

A questão *Q1* é respondida com a seguinte consulta SQL:

*Q1: “Quais são as organizações mais mencionadas nos tweets enviados do Brasil, durante um certo período de tempo?”*

```
SELECT NODE.label, SUM(NODE.directHits) AS mencoes_diretas,
       SUM(NODE.indirectHitsSubClassOf + NODE.indirectHitsType)
       AS mencoes_indiretas
FROM   KGNode NODE, KGNode ROOT, Hierarchy H,
       KGNode_Tweet KT, Tweet T, Bridge B
WHERE  H.idNode = NODE.id AND H.idRoot = ROOT.id
       AND B.idNode = ROOT.id AND T.country = 'Brasil'
       AND B.high_level_class = 'gr:BusinessEntity'
       AND T.creation_date BETWEEN '11/30/2015' AND '12/15/2015'
GROUP BY NODE.label
ORDER BY mencoes_diretas, mencoes_indiretas DESC;
```

Consultas análogas podem envolver outras dimensões derivadas automaticamente das hierarquias de recursos existentes nos KGs usados para enriquecer o *TCs*. A seguir temos a consulta que responde à questão *Q2* da introdução:

*Q2: “Quais são os produtos ou serviços mais mencionados nos tweets enviados na cidade de Florianópolis nos últimos 30 dias?”*

```
SELECT NODE.label, SUM(NODE.directHits) AS mencoes_diretas,
       SUM(NODE.indirectHitsSubClassOf + NODE.indirectHitsType)
       AS mencoes_indiretas
FROM   KGNode NODE, KGNode ROOT, Hierarchy H,
       KGNode_Tweet KT, Tweet T, Bridge B
WHERE  H.idNode = NODE.id AND H.idRoot = ROOT.id
       AND B.idNode = ROOT.id
       AND T.country = 'Brasil'
       AND B.high_level_class = 'gr:ProductOrService'
       AND T.place_name = 'Florianópolis'
       AND T.place_type = 'City'
       AND T.creation_date BETWEEN '10/30/2018' AND '11/30/2018'
GROUP BY NODE.label
ORDER BY mencoes_diretas, mencoes_indiretas DESC;
```

Por fim, este trabalho responde à questão *Q3* apresentada na introdução através da consulta SQL a seguir:

*Q3: “Quais são os produtos de informática mais mencionados em tweets no Estado de Santa Catarina no ano de 2018?”*

```

SELECT NODE.label, SUM(NODE.directHits) AS mencoes_diretas,
       SUM(NODE.indirectHitsSubClassOf + NODE.indirectHitsType)
       AS mencoes_indiretas
FROM KGNode NODE, KGNode ROOT, Hierarchy H,
     KGNode_Tweet KT, Tweet T, Bridge B
WHERE H.idNode = NODE.id AND H.idRoot = ROOT.id
     AND B.idNode = ROOT.id
     AND B.high_level_class = 'gr:InformationAppliance'
     AND T.creation_date BETWEEN '01/01/2018' AND '12/31/2018'
     AND T.place_name = 'Santa Catarina'
     AND T.place_type = 'State' AND T.country = 'Brasil'
GROUP BY NODE.label
ORDER BY mencoes_diretas, mencoes_indiretas DESC;

```

A partir das tabelas deste modelo é possível construir o modelo-estrela e popular um *DW* com dados originados de anotações semânticas, permitindo consultas SPARQL, MDX ou análises visuais através de gráficos e *dashboards*.

## 4 EXPERIMENTOS

Este capítulo descreve os materiais e métodos empregados nos experimentos em um estudo de caso em uma instância do processo proposto. A viabilidade desse processo para detectar, filtrar e analisar as menções de interesse de um domínio em Clipes Textuais *TCs* foi investigada em um estudo de caso no domínio de negócios (*business*) apresentado na seção 4.1. A Seção 4.2 descreve o ambiente experimental, *hardware*, *software* e configurações. Em seguida, a Seção 4.3 apresenta e discute os resultados experimentais coletados.

### 4.1 ESTUDO DE CASO: *BUSINESS*

O estudo de caso deste trabalho anota *tweets* usando a ferramenta de *NERD* DBpedia-Spotlight, filtra menções de interesse para o domínio de negócios e constrói dimensões da hierarquia DBpedia de recursos usados como valores de anotações semânticas de interesse de menções, para analisá-los usando um modelo dimensional. Especialistas em domínio de negócios criaram pontes-chave (*Key Bridges - KBs*) entre classes *LOD* e classes de uma ontologia de alto nível, pontes estas voltadas para o domínio de negócios. Neste estudo de caso, as classes *LOD* são originadas da DBpedia e classes de alto nível são da Ontologia GoodRelations (GRO) (HEPP, 2008) <sup>1</sup>. A Figura 12 ilustra algumas das principais classes da ontologia *GoodRelations*.

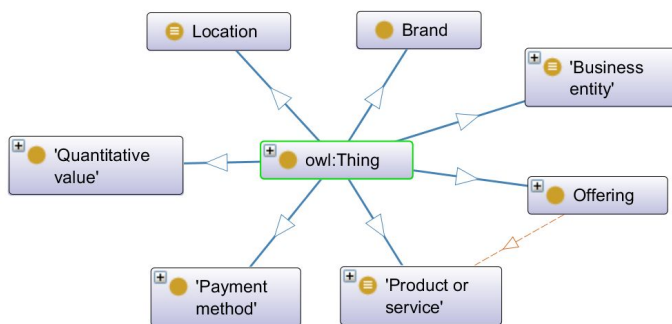


Figura 12 – Principais conceitos da GRO (HEPP, 2008)

<sup>1</sup><http://www.heppnetz.de/projects/goodrelations/>

A GRO é utilizada por empresas como Google, Yahoo!, Best-Buy, Sears e Kmart. As classes da GRO (algumas relacionadas ao *schema.org*) consideradas relevantes por especialistas de domínio em *business* para análise de informação foram:

- **gr:BusinessEntity**: representa um agente, empresa ou indivíduo de negócio;
- **gr:Offering**: representar uma venda, reparar, arrendamento ou oferta;
- **gr:ProductOrService**: identifica produtos ou serviços;
- **gr:Location**: identifica o local de uma loja ou oferta disponível.

A Tabela 1 representa uma amostra com as principais Pontes-chave (*Key Bridges*) criadas pelos especialistas de domínio. Os especialistas iniciaram a análise partindo das 20 classes de *LOD* mais mencionadas nas anotações semânticas do *dataset* de *tweets*, considerando os *hits* acumulados (diretos, indiretos por *type* e indiretos por *subClassOf*):



Tabela 1 – Pontes-chave (*KB*) elaboradas por especialistas de domínio nas 20 classes mais mencionadas da DBpedia

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

## 4.2 CONFIGURAÇÕES DOS EXPERIMENTOS

Os experimentos utilizaram o *dataset* BR-2015, que possui uma amostra aleatória de 100.000 *tweets* enviados do Brasil no período entre 30/11/2015 e 15/12/2015, com conteúdos textuais escritos em português. Este *dataset* foi anotado semanticamente com a ferramenta DBpedia-Spotlight. Em seguida, três pesquisadores do Departamento de Informática Empresarial da Universidade de Leipzig - especialistas em *Social CRM* - construíram por consenso 34 *Key Bridges (KBs)* para os experimentos. Deste total de 34 *KBs* associando classes da GRO aos conceitos mais mencionados na ontologia DBpedia, 10 foram ilustradas na Tabela 1.

A construção das *KBs* foi iniciada com classes de recursos ordenadas de maneira decrescente pelo número total de ocorrências (*hits* diretos ou indiretos nas anotações do conjunto de 100 mil tweets usado nos experimentos). A etapa de construção das *KBs* consumiu menos de duas horas de trabalho de cada especialista no domínio. Em seguida, ocorreu uma reunião de duas horas com os três especialistas do domínio para verificar inconsistências e torná-las compatíveis.

Foi desenvolvida uma aplicação<sup>2</sup> em Java usando a *IDE* Eclipse<sup>3</sup>. A aplicação recebe como entrada um arquivo contendo um documento texto do *dataset* de *TCs* e chama a *API REST* do DBpedia-Spotlight para obter as anotações. Os parâmetros do DBpedia-Spotlight usados em nossos experimentos foram:

- *confidence: 0.5*;
- *language: pt*;
- *selected types: ALL*;

Em seguida, a aplicação constrói a Hierarquia de Recursos de com *Hits (Hierarchy of Hit Resources - HHR)* dessas anotações e chama o algoritmo *CheckComplete*, com a *HHR* e as *KBs* fornecidas pelos especialistas do domínio. Como resultado, o aplicativo gera planilhas para realizar análises estatísticas. Os gráficos das análises selecionadas foram gerados usando a biblioteca Pandas<sup>4</sup>. Os experimentos foram executados em dois computadores (C1 e C2):

<sup>2</sup><https://gitlab.com/vilmarcesarpereira/semantic-bi>

<sup>3</sup>[www.eclipse.org](http://www.eclipse.org)

<sup>4</sup><http://pandas.pydata.org/>

- C1: Core i5 1,6 GHz, 8 Gb RAM, SSD 256 Gb, com o sistema operacional OSX 10.12.3;
- C2: Core i5 1.6GHz, 4Gb RAM, HD 500Gb, com o sistema operacional Ubuntu 17.04.

### 4.3 RESULTADOS EXPERIMENTAIS

Os resultados experimentais são apresentados nas Subseções 4.3.1 a seguir. A explanação foi organizada em consonância às tarefas destacadas em negrito na Figura 7, para destacar as principais realizações deste trabalho, a saber: **1d:Contagem de *Hits* e Construção da Hierarquia de Recursos com *Hits* (HHR)**; **2b:Verificar e Completar Pontes** e **2c:Geração de Dimensões de Análise**.

#### 4.3.1 Contagem de *Hits* e Construção da Hierarquia de Recursos com *Hits* (HHR)

A contagem de *hits* é realizada no sentido ascendente, partindo as folhas (recursos anotados diretamente) em direção à raiz da hierarquia das classes *LOD*, geralmente *owl:Thing*. Assim, *hits* diretos, *hits* indiretos via *type* e *hits* indiretos via *subClassOf* são somados e acumulados para cada nó da *Hierarchy of Hit Resources (HHR)*.

Nos experimentos, observamos uma predominância da classe *dbo:Agent*. Os especialistas do domínio verificaram que não há nenhuma ponte para esta classe DBpedia na ontologia de domínio, já que *dbo:Agent* representa uma classe mais geral do que qualquer classe *GoodRelations Ontology (GRO)*. No entanto, *dbo:Organization* - uma subclasse direta de *dbo:Agent* - tem uma forte aderência à classe da *GRO gr:BusinessEntity*, classe esta que foi escolhida como uma dimensão no Cubo de Dados Semânticos (*SDC*). Portanto, removemos o *dbo:Agent* do ranking apresentado na Figura 13, que contém o *ranking* das 15 classes com mais *hits* nas anotações semânticas geradas pelo DBpedia-Spotlight para o *dataset* BR-2015. A classificação é ordenada em número decrescente de ocorrências acumuladas.

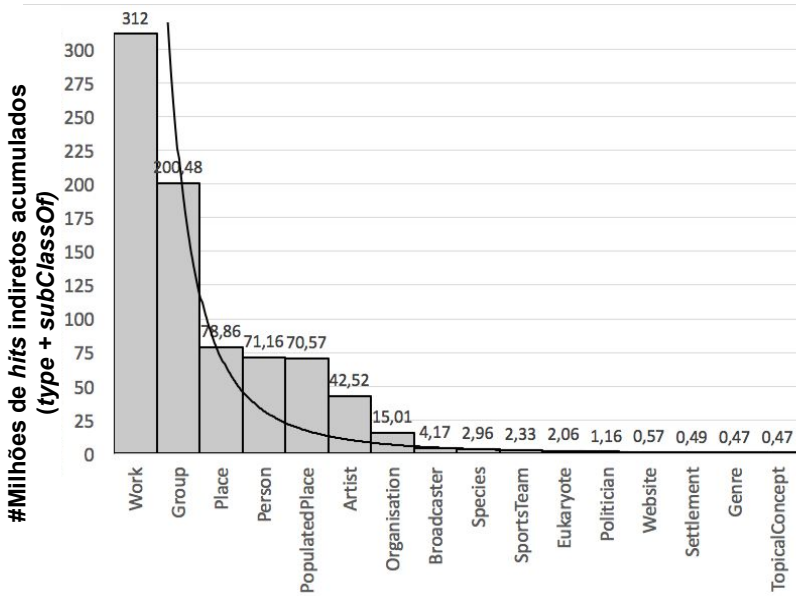


Figura 13 – Classes mais mencionadas nas anotações do *dataset* BR-2015

Outra análise realizada concerne às instâncias afetadas por *hits* diretos. A Tabela 2 a seguir apresenta o *ranking* decrescente das 10 instâncias da DBpedia mais mencionadas nas anotações semânticas dos *TCs* do *dataset* *BR-2015*.

Tabela 2 – Lista das 10 instâncias com mais *hits* diretos

#	Instância	<i>Hits</i> Diretos
1	dbo:Deus	1056
2	dbo:Twitter	988
3	dbo:One_Direction	840
4	dbo:Brasil	628
5	dbo:Sono	563
6	dbo:Impeachment	516
7	dbo:YouTube	403
8	dbo:Porrete	384
9	dbo:Jogos_Olímpicos_de_Inverno_de_2010	375
10	dbo:Demi_Lovato	324

### 4.3.2 Verificar e Completar Pontes

Os resultados da execução do algoritmo *CheckComplete* estão resumidos nas Figuras 14 e 15. A Figura 14 e a Figura 15 apresentam a distribuição das instâncias e classes da DBpedia LOD, respectivamente, associadas aos conceitos da *GRO* via Pontes-Chave (*Key Bridges - KB*), Novas (*New Bridges - NB*) e Inconsistentes (*Inconsistent Bridges - IB*).

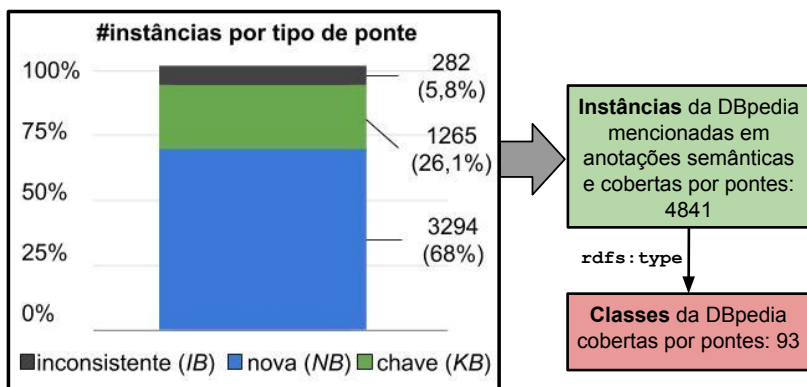


Figura 14 – Resultados do algoritmo *CheckComplete* categorizados por instâncias

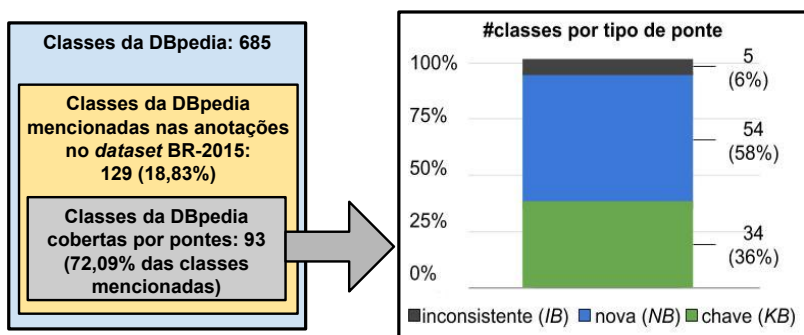


Figura 15 – Resultados do algoritmo *CheckComplete* categorizados por classes

Em ambas as análises (por instâncias ou classes) observa-se uma contribuição significativa do algoritmo *CheckComplete*, através dos seguintes fatos:

- (i) pelo menos 50% de todas as pontes estão em novas pontes (*NB*): 55,81% para classes e 52,94% para instâncias;
- (ii) menos de 5% das pontes verificadas pelo algoritmo *CheckComplete* são inconsistentes (*IB*): 3,1% para classes e 4,85% para instâncias.

### 4.3.3 Geração de Dimensões de Análise

A Hierarquia de Recursos com *Hits* (*HHR*) criada a partir das anotações do *dataset* BR-2015 possui 15 classes de DBpedia associadas diretamente à raiz, *owl:Thing*. A Figura 16 mostra o primeiro nível deste *HHR*, com os recursos com *Key Bridges* associadas destacados com um contorno vermelho.

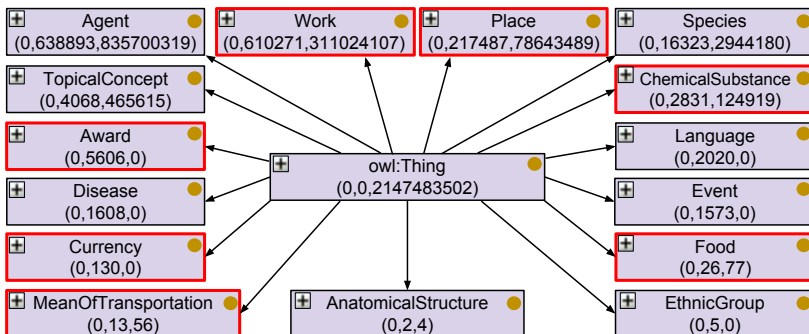


Figura 16 – Subclasses de *owl:Thing* com mais *hits*

A Figura 17 mostra um extrato do *HHR* abaixo da classe *dbo:Organization*. Novamente, os recursos com *KBs* são destacados com um contorno vermelho, com filhos vinculados aos mesmos conceitos de domínio do *GRO*, conforme aplicado pelo algoritmo *CheckComplete*.

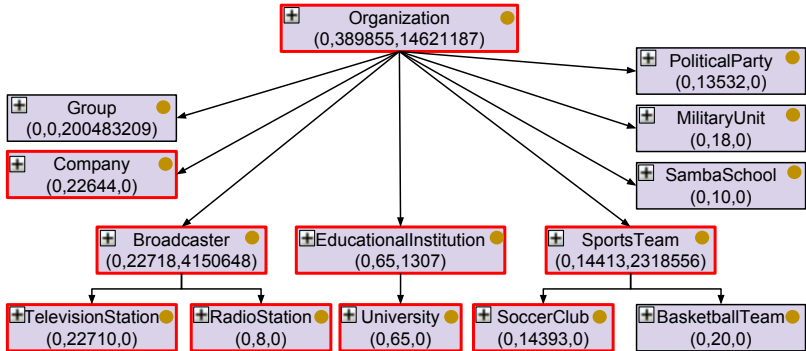


Figura 17 – Hierarquia de subclasses de `dbo:Organization`

A consulta *Q1*, expressa por uma frase em português no Capítulo 1 e em *SQL* na Seção 3.4, pode ser respondida com a seguinte expressão *SQL* no Cubo de Dados Semânticos (*SDC*) conceitual, cujo esquema é mostrado na Figura 6:

```

SELECT B.organization, sum(F.mentions) AS n
FROM TCmeasures AS F
  NATURAL JOIN BusinessEntity AS B
  NATURAL JOIN ProductOrService Pr
  NATURAL JOIN Place AS P NATURAL JOIN Time AS T
WHERE Pr.any = 'ALL' AND P.country = 'Brazil'
  AND T.date BETWEEN '11/30/2015' AND '12/15/2015'
GROUP BY B.organization ORDER BY n DESC;

```

As dimensões *Time* e *Place* se referem à hora e origem da criação do Clipe Textual (*TC*), respectivamente. A Figura 17 ilustra esquematicamente os resultados retornados, embora as medidas na *SDC* não sejam exatamente as mesmas que as quantidades de ocorrências apresentadas nesta figura. A medida *Sum(#mentions)* retornado por esta consulta corresponde à soma dos três números entre parênteses (*hits*) em cada nó da hierarquia na Figura 17.

#### 4.3.4 Discussão

Os resultados experimentais revelaram uma concentração significativa das anotações em algumas instâncias e classes, com uma distribuição de lei de potência típica dos dados de mídias sociais. Algumas

classes de interesse em *business* são bastante proeminentes nas anotações semânticas dos *tweets* enviados do Brasil, incluindo *Broadcaster*, *TelevisionStation*, *Company* e *Schema: Product*. A quantidade média de anotações por *tweet* é de aproximadamente 6 e uma quantidade considerável de menções vinculadas a conceitos e instâncias do domínio de negócios. Este fato indica um potencial para explorar *tweets* anotados em CRM comercial e social. Seguem as demais observações realizadas com os resultados experimentais:

1. Apenas 34 Pontes-Chave (*Key Bridges - KBs*) produzidos pelo homem (consenso de 3 seres humanos) foram suficientes para derivar dimensões úteis;
2. 129 (18,83%) de um total de 685 classes *LOD* da DBpedia receberam *hits* pelas anotações (diretos ou indiretos);
3. Dentre as 129 classes da DBpedias com *hits* associados, 93 (72,09%) têm pontes associadas;
4. 79,41% das *Key Bridges (KBs)* foram usadas pelo algoritmo *Check-Complete* para derivar novas pontes ou detectar inconsistências;
5. Os números de *hits* na classe raiz DBpedia *owl:thing* foram: #direct: 31,219; #indirect via *type*: 3.725,517; #indirect via *sub-ClassOf*: 3.787.780,005;
6. as pontes cobriram 62,38% das instâncias mencionadas nas anotações semânticas dos *tweets*;
7. 37,62% das classes com *hits* não têm correspondentes no GRO (e.g, *dbo:Person* com 22,13% dos resultados). Este fato sugere que as pontes e a filtragem podem se beneficiar da complementação de conceitos de domínio oriundos de outras ontologias de alto nível além da *GRO*.

Finalmente, *Deus* foi a instância mais anotada, seguida de *Twitter*. Tais instâncias indicam um vasto uso de interjeições ou termos religiosos (para o recurso *Deus*) e um grande número de menções à própria ferramenta de microblogging (*Twitter*). Outras instâncias dignas de destaque são *Brasil* e *Impeachment*, pois ilustram o momento político no Brasil via a emissão de *tweets* no período entre 30/11/2015 e 11/12/2015. Esses resultados sugerem que as postagens de mídias sociais podem ser usadas para várias finalidades analíticas, como a identificação de tendências, demandas de produtos ou tópicos emocionais comuns de clientes (e.g. *Deus*).



## 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 OR data warehouse)*

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 descartados 35 artigos. A fim de complementar o conjunto de trabalhos relacionados, foram incluídos artigos indicados pelos especialistas de domínio de *e-commerce* que criaram as pontes nos experimentos. Os artigos indicados (*e.g.*, (ABRAHAMS et al., 2012)) versam sobre *Web* semântica e *data warehousing*. A combinação dos resultados filtrados da busca combinados com as citados pelos especialistas resultou no conjunto de trabalhos correlatos a este descritos e comparados no restante deste capítulo, segundo os seguintes critérios:

- Postagens de mídias sociais (*Social Media Posts - SMP*): este critério está relacionado à origem dos *TCs*. *Datasets* de cliques textuais originados de mídias sociais (*e.g.* *tweets*) são vastos e possuem *APIs* públicas para coleta. Os valores possíveis para os trabalhos correlatos são: usa (✓) ou não usa (-);
- Anotação Semântica (*Semantic Annotation - SA*): esta critério determina se a abordagem usa (✓) ou não (-) técnicas de anotação semântica para enriquecer os *TCs*.
- Método de Anotação (*Annotation Method - AM*): determina a técnica utilizada para anotação (manual: apenas com conhecimento de especialistas; semiautomático: regras e ontologias de especialistas; automático: somente ontologias e base de conhecimento);
- Modelo Dimensional (*Dimensional Model - DM*): a abordagem gera um cubo dimensional (✓), permitindo assim a análise *OLAP*

ou não gera (-);

- Customizável (*Customizable - C*): determina se a abordagem é personalizável 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*) a postagens de fóruns de discussão on-line, procurando características de veículos para alimentar sistemas de recomendação. Uma abordagem semelhante (VILLANUEVA et al., 2016), anota semanticamente os *tweets* para extrair informações para fins de recomendação. Ambos os trabalhos contêm Clipes Textuais (*Textual Clips - TCs*) como entradas para um processo de busca de informações mais detalhadas, entretanto, não usam o modelo dimensional.

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

O trabalho de Francia et al. (FRANCIA; GOLFARELLI; RIZZI, 2014) apresenta um processo iterativo para desenvolver um *BI Social*, cujas entradas são *TCs* semanticamente anotados pela ferramenta Synthema, mas seu processo requer considerável trabalho humano.

O trabalho de Cuzzocrea et al. (CUZZOCREA et al., 2015) considera o uso de operadores *OLAP* em *data warehouses* dimensionais para analisar *tweets*, mas sem enriquecimento semântico. Aplica a Análise Formal de Conceitos (*Formal Concept Analysis - FCA*) nas postagens de mídias sociais, combinada com uma ontologia de domínio para preencher o cubo de dados dimensionais.

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 *HHR* 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 execução do algoritmo *CheckComplete*, que verifica as *KBs* e identifica pontes incorretas (*IBs*) ou novas pontes (*NBs*). Tais pontes orientam a seleção de anotações relevantes e a adaptação de hierarquias de *LOD* usados nas anotações para servir como dimensões de análise de dados. O resultado final é um Cubo de Dados Semântico construído especialmente para o domínio determinado pelos especialistas, no qual dimensões de análise são hierarquias de recursos de *LOD* (e.g., hierarquias de conceitos, hierarquias de composição de objetos) que aparecem como valores nas anotações de interesse para o domínio. Esta etapa de construção de pontes com a *HHR* diferencia este trabalho dos dois trabalhos analisados ((FILETO et al., 2015), (SACENTI et al., 2015)).

O trabalho de (TAO et al., 2016) organiza os textos de entrada em um modelo dimensional e propõe consultas de usuário baseadas em classificação, mas não usa anotações semânticas. Finalmente, a abordagem interativa EXODuS proposta em (CHOUDEUR; 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 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 para *ETL* que emprega e filtra as anotações semânticas de Clipes Textuais (*Textual Clips - TCs*) que sejam de interesse para um domínio de aplicação utilizando pontes para uma ontologia de alto nível para tal domínio.

Tabela 3 – Comparação de trabalhos correlatos

Trabalho	SMP	SA	DM	C	AM	ABH	Saída
Abrahams(2012)	-	✓	-	-	Manual	-	Recomendações
Nebot(2012)	-	✓	✓	-	Manual	✓	Cubo dimensional populado
Francia(2014)	✓	✓	✓	-	Semi-Automático	✓	Cubo dimensional populado
Cuzzocrea(2015)	✓	-	✓	✓	-	-	Cubo dimensional populado
Fileto(2015)	✓	✓	-	✓	Automático	-	Dados semânticos de movimento
Sacanti(2015)	✓	✓	✓	✓	Automático	✓	Cubo dimensional não-populado
Villanueva(2016)	✓	✓	-	-	Semi-Automático	-	Recomendações
Tao(2016)	-	-	✓	✓	-	✓	Cubo dimensional populado
Chouder(2019)	✓	-	✓	✓	-	✓	Cubo dimensional populado
<b>Este trabalho</b>	✓	✓	✓	✓	<b>Automático</b>	✓	<b>Cubo dimensional populado</b>

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



## 6 CONCLUSÕES E TRABALHOS FUTUROS

Este trabalho propõe um processo fundamentado nas tecnologias da Web semântica e de modelagem dimensional para detectar e analisar menções a recursos de interesse em Clipes Textuais (*Textual Clips - TCs*) (e.g. postagens em mídias sociais). Pontes entre classes *LOD* e conceitos de alto nível de uma ontologia de domínio desempenham um papel central no processo proposto, especificamente na filtragem de menções de interesse para um domínio específico. Como exemplo de tais pontes, podemos citar associações entre alguma classe da ontologia de alto nível *GoodRelations* (*GRO*) para o domínio de negócios (*business*) e uma classe da ontologia da *DBpedia*, coleção de dados ligados usada para anotar semanticamente os *TCs* em nosso estudo de caso.

Foi observada uma dificuldade em obter automaticamente pontes adequadas em alguns casos, mesmo usando técnicas de alinhamento de ontologia de última geração. Assim, esta proposta baseia-se em um pequeno número de pontes-chaves (*Key Bridges - KBs*) criadas manualmente e emprega um algoritmo especialmente desenvolvido para verificar a consistência dessas pontes-chaves e derivar novas pontes, com base em hierarquias de subsunção (*subsumption*) de ontologias *LOD*, como a ontologia *DBpedia* (*DBO*). Esta abordagem permite o uso de uma variedade de ferramentas atualmente disponíveis para anotar semanticamente textos com recursos (e.g. conceitos e instâncias) de Grafos de Conhecimento (*Knowledge Graphs - KGs*). A abordagem proposta também constrói dimensões de análise a partir de hierarquias existentes nesses *KGs* (e.g. hierarquias de subsunção ou de composição) e usa essas dimensões em um cubo de dados para suportar a análise das incidências de menções de interesse nos *TCs* anotados.

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 *GoodRelations 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.

Quanto à pergunta de pesquisa enunciada no início deste trabalho, os resultados experimentais demonstraram que um conjunto de *TCs* pode ser enriquecido semanticamente de forma automática, iniciando o processo de *ETL* proposto. Este processo passa por uma etapa

de construção, verificação e expansão de pontes entre classes da ontologia subjacente à coleção de *LOD* de uso geral empregados dos valores das anotações semânticas e classes de uma ontologia de alto nível, para então usar tais pontes na seleção de anotações de interesse e na construção de dimensões para análise de informações de interesse anotadas no *TCs* usando num modelo dimensional. As pontes determinam as raízes e a estrutura de cada dimensão, através da associação a classes de alto nível da ontologia de domínio.

As principais contribuições deste trabalho podem ser sumarizadas como se segue.

1. Concepção de um processo geral de *ETL* para anotar semanticamente *TCs*, filtrar anotações de interesse e derivar hierarquias para a análise dos *TCs* em um modelo dimensional.
2. Criação do algoritmo *CheckComplete*, especialmente desenvolvido para verificar a consistência das pontes-chaves, detectar pontes incorretas e derivar outras pontes.
3. Resultados experimentais mostrando a distribuição de classes e instâncias da coleção de *LOD* (*DBpedia Ontology - DBO*) usados como valores de anotações semânticas e classes da ontologia de domínio de alto nível (*GoodRelations Ontology - GRO*) nas anotações de um conjunto de *TCs* (*tweets*).

O processo de *ETL* semântico proposto também viabilizou novos tipos de consultas analíticas, por exemplo, referenciando classes e instâncias de *LOD* mencionadas nos tweets e explorando hierarquias de classificação e composição presentes em coleções de *LOD* e adaptadas para servirem como dimensões de análise de dados em um modelo dimensional. Adicionalmente, os resultados do estudo de caso em *business* sugerem que alguns ramos das hierarquias de classe de *LOD* (e.g. organismos vivos, animais, plantas) são de interesse apenas para linhas específicas de negócios (e.g., certos animais são relevantes para *pet shops* e certas plantas são pertinentes à agricultura). Tais fatos sugerem a necessidade de mais pesquisas e experimentos em domínios e ramos de negócios específicos.

Trabalhos futuros incluem:

1. Encontrar maneiras eficientes e eficazes de determinar pontes para coisas que são relevantes para domínios específicos, em ramos específicas de negócios;



2. Investigar teórica e empiricamente o impacto das dimensões de análise construídas a partir das hierarquias de *LOD* no modelo dimensional e na expressividade do *OLAP*;
3. Avaliar os benefícios das novas capacidades para detectar e analisar menções de interesse em *TCs* habilitadas por esta proposta em uma variedade de domínios de aplicação, como *e-commerce*, e atividades como *CRM* social;
4. Efetuar experimentos mais extensos para avaliar em detalhes a qualidade das informações produzidas pelo processo proposto e as distribuições de anotações de menções relevantes em classes e instâncias de interesse, com possíveis variações de acordo com a localização, tempo, idioma, cultura e demografia, entre outras questões.



## REFERÊNCIAS

- ABRAHAMS, A. S. et al. Vehicle defect discovery from social media. *Decision Support Systems*, Elsevier, v. 54, n. 1, p. 87–97, 2012.
- ATEFEH, F.; KHREICH, W. A survey of techniques for event detection in twitter. *Computational Intelligence*, v. 31, n. 1, p. 132–164, 2015. ISSN 1467-8640. <<http://dx.doi.org/10.1111/coin.12017>>.
- BONTCHEVA, K.; ROUT, D. Making sense of social media streams through semantics: A survey. *Semantic Web*, 2012. ISSN 1570-0844. <<http://dx.doi.org/10.3233/SW-130110>>.
- BROHMAN, M. K. et al. The business intelligence value chain: Data-driven decision support in a data warehouse environment: An exploratory study. In: IEEE. *System Sciences, 2000. Proceedings of the 33rd Annual Hawaii International Conference on*. [S.l.], 2000. p. 10–pp.
- CHARLES-SMITH, L. E. et al. Using social media for actionable disease surveillance and outbreak management. a systematic literature review. *PLoS ONE*, Public Library of Science, v. 10, n. 10, Oct 2015. <<http://www.osti.gov/pages/servlets/purl/1229934>>.
- CHOUDEH, 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>>.
- CUZZOCREA, A. et al. Towards olap analysis of multidimensional tweet streams. In: ACM. *Proceedings of the ACM Eighteenth International Workshop on Data Warehousing and OLAP*. [S.l.], 2015. p. 69–73.
- 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>>.

FRANCIA, M.; GOLFARELLI, M.; RIZZI, S. A methodology for social bi. In: ACM. *Proceedings of the 18th International Database Engineering & Applications Symposium*. [S.l.], 2014. p. 207–216.

GALLINUCCI, E.; GOLFARELLI, M.; RIZZI, S. Meta-stars: Dynamic, schemaless, and semantically-rich topic hierarchies in social bi. In: *EDBT*. [S.l.: s.n.], 2015. p. 529–532.

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.

GUARINO, N. et al. Formal ontology and information systems. In: *Proceedings of FOIS*. [S.l.: s.n.], 1998. v. 98, n. 1998, p. 81–97.

HEPP, M. Goodrelations: An ontology for describing products and services offers on the web. In: GANGEMI, A.; EUZENAT, J. (Ed.). *EKAW*. [S.l.]: Springer, 2008. (Lecture Notes in Computer Science, v. 5268), p. 329–346. ISBN 978-3-540-87695-3.

INMON, W. H. *Building the Data Warehouse*. New York, NY, USA: John Wiley & Sons, Inc., 1992. ISBN 0471569607.

KIMBALL, R.; ROSS, M. *The data warehouse toolkit: the complete guide to dimensional modeling*. [S.l.]: John Wiley & Sons, 2011.

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>>.

LINDA, S.-I. L. Social commerce—e-commerce in social media context. *World Academy of Science Engineering and Technology*, v. 72, p. 39–44, 2010.

MENDES, P. N. et al. Dbpedia-spotlight: shedding light on the web of documents. In: ACM. *Proceedings of the 7th international conference on semantic systems*. [S.l.], 2011. p. 1–8.

MIDDLETON, S. E.; MIDDLETON, L.; MODAFFERI, S. Real-time crisis mapping of natural disasters using social media. *IEEE Intelligent Systems*, IEEE, v. 29, n. 2, p. 9–17, 2014.

- MORO, A.; RAGANATO, A.; NAVIGLI, R. Entity Linking meets Word Sense Disambiguation: a Unified Approach. *Transactions of the Association for Computational Linguistics (ACL)*, v. 2, p. 231–244, 2014.
- 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. N. et al. Introduction to linked data and its lifecycle on the web. In: KOUBARAKIS, M. et al. (Ed.). *10th Intl. Summer School Reasoning on the Web in the Big Data Era*. Springer, 2014. (LNCS, v. 8714), p. 1–99. ISBN 978-3-319-10586-4. <[http://dx.doi.org/10.1007/978-3-319-10587-1\\_1](http://dx.doi.org/10.1007/978-3-319-10587-1_1)>.
- PAUL, M. J. et al. Social media mining for public health monitoring and surveillance. In: *Pacific Symposium on Biocomputing (PSB)*. [S.l.: s.n.], 2016. p. 468–79.
- REINHOLD, O.; ALT, R. Social customer relationship management: State of the art and learnings from current projects. In: *Bled eConference*. [S.l.: s.n.], 2012. p. 26.
- RITTER, A. et al. Named entity recognition in tweets: an experimental study. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the conference on empirical methods in natural language processing*. [S.l.], 2011. p. 1524–1534.
- 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.
- SAKAKI, T.; OKAZAKI, M.; MATSUO, Y. Earthquake shakes twitter users: real-time event detection by social sensors. In: ACM. *Proceedings of the 19th international conference on World wide web*. [S.l.], 2010. p. 851–860.
- SCHIEL, U. Abstractions in semantic networks: axiom schemata for generalization, aggregation and grouping. *ACM SIGART Bulletin*, ACM, n. 107, p. 25–26, 1989.
- SHVAIKO, P.; EUZENAT, J. Ontology matching: state of the art and future challenges. *IEEE Transactions on knowledge and data engineering*, IEEE, v. 25, n. 1, p. 158–176, 2013.

SIGNORINI, A.; SEGRE, A. M.; POLGREEN, P. M. The use of twitter to track levels of disease activity and public concern in the us during the influenza a h1n1 pandemic. *PloS one*, Public Library of Science, v. 6, n. 5, p. e19467, 2011.

SORATO, D. et al. Analysis of methods and tools for relevant words recognition in microblogs. In: BRAZILIAN COMPUTER SOCIETY. *Proceedings of the XII Brazilian Symposium on Information Systems on Brazilian Symposium on Information Systems: Information Systems in the Cloud Computing Era-Volume 1*. [S.l.], 2016. p. 46.

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.

TAO, F. et al. Multi-dimensional, phrase-based summarization in text cubes. *IEEE Data Eng. Bull.*, v. 39, n. 3, p. 74–84, 2016.

VIEIRA, H. S. et al. Towards the effective linking of social media contents to products in e-commerce catalogs. In: ACM. *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. [S.l.], 2016. p. 1049–1058.

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.

WITTWER, M. et al. Social media analytics using business intelligence and social media tools—differences and implications. In: SPRINGER. *International Conference on Business Information Systems*. [S.l.], 2016. p. 252–259.

YIN, J. et al. Using social media to enhance emergency situation awareness. In: *Twenty-Fourth International Joint Conference on Artificial Intelligence*. [S.l.: s.n.], 2015.

ZHAO, W. X. et al. Connecting social media to e-commerce: cold-start product recommendation using microblogging information. *IEEE Transactions on Knowledge and Data Engineering*, IEEE, v. 28, n. 5, p. 1147–1159, 2016.