**IntegraWeb: uma proposta de arquitetura baseada em
mapeamentos semânticos e técnicas de mineração de dados**

Felipe Lombardi Pierin

DISSERTAÇÃO APRESENTADA
AO
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
DA
UNIVERSIDADE DE SÃO PAULO
PARA
OBTENÇÃO DO TÍTULO
DE
MESTRE EM CIÊNCIAS

Programa: Pós-Graduação em Ciência da Computação

Orientador: Prof. Dr. Jaime Simão Sichman

São Paulo, Janeiro de 2018

IntegraWeb: uma proposta de arquitetura baseada em mapeamentos semânticos e técnicas de mineração de dados

Esta versão da dissertação contém as correções e alterações sugeridas pela Comissão Julgadora durante a defesa da versão original do trabalho, realizada em 05/12/2017. Uma cópia da versão original está disponível no Instituto de Matemática e Estatística da Universidade de São Paulo.

Comissão Julgadora:

- Prof. Dr. Jaime Simão Sichman - EP-USP
- Profa. Dra. Renata Wassermann - IME-USP
- Profa. Dra. Ana Cristina Bicharra Garcia - UNIRIO

Agradecimentos

Agradeço a minha família pelo apoio ao longo de todos esses anos ao longo dessa grande jornada. Nesse período tive ainda mais amor, compreensão e todo o apoio possível durante essa fase de minha vida.

Agradeço também aos meus amigos por toda a paciência que tiveram comigo durante os momentos mais críticos desta etapa, pelo companheirismo sempre presente, pela confiança, pela alegria e pela força transmitida. Todos esses gestos me ajudaram a alcançar resultados melhores e também construíram alicerces para não desistir deste projeto.

Agradeço ao meu amigo e colega de trabalho Daniel Deimann que dedicou parte do seu tempo para sugerir e repassar o conhecimento necessário para a construção da camada de apresentação que implementa a proposta do projeto aqui apresentado.

Agradeço aos meus professores do IME-USP e da EP-USP que mostraram pontos de vista diferentes do meu e me auxiliaram a tomar decisões mais assertivas em relação a este trabalho por meio das matérias que assisti, pela sugestão de artigos e também com a contribuição com ideias.

Agradeço ao meu orientador Jaime Sichman que acreditou no meu potencial e ofereceu uma oportunidade única para participar de um universo de conhecimento. Por sua dedicação com este projeto, pelas sugestões e por estar sempre presente.

Resumo

Pierin, F. L. **IntegraWeb: uma proposta de arquitetura baseada em mapeamentos semânticos e técnicas de mineração de dados**. 2017. 75 f. Dissertação (Mestrado) - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2017.

Atualmente uma grande quantidade de conteúdo é produzida e publicada todos os dias na Internet. São documentos publicados por diferentes pessoas, por diversas organizações e em inúmeros formatos sem qualquer tipo de padronização. Por esse motivo, a informação relevante sobre um mesmo domínio de interesse acaba espalhada pela Web nos diversos portais, o que dificulta uma visão ampla, centralizada e objetiva sobre esta informação. Nesse contexto, a integração dos dados espalhados na rede torna-se um problema de pesquisa relevante, para permitir a realização de consultas mais inteligentes, de modo a obter resultados mais ricos de significado e mais próximos do interesse do usuário. No entanto, tal integração não é trivial, sendo por muitas vezes custosa devido à dependência do desenvolvimento de sistemas e mão de obra especializados, visto que são poucos os modelos reaproveitáveis e facilmente integráveis entre si. Assim, a existência de um modelo padronizado para a integração dos dados e para o acesso à informação produzida por essas diferentes entidades reduz o esforço na construção de sistemas específicos. Neste trabalho é proposta uma arquitetura baseada em ontologias para a integração de dados publicados na Internet. O seu uso é ilustrado através de casos de uso reais para a integração da informação na Internet, evidenciando como o uso de ontologias pode trazer resultados mais relevantes.

Palavras-chave: Web semântica, integração de informação, ontologias

Abstract

Pierin, F. L. **IntegraWeb: An architectural proposal based on semantic mappings and data mining techniques**. 2017. 75 f. Dissertação (Mestrado) - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2017.

A lot of content is produced and published every day on the Internet. Those documents are published by different people, organizations and in many formats without any type of established standards. For this reason, relevant information about a domain of interest is spread through the Web in various portals, which hinders a broad, centralized and objective view of this information. In this context, the integration of the data scattered in the network becomes a relevant research problem, to enable smarter queries, in order to obtain richer results of meaning and closer to the user's interest. However, such integration is not trivial, and is often costly because of the reliance on the development of specialized systems by professionals, since there are few reusable and easily integrable models. Thus, the existence of a standardized model for data integration and access to the information produced by these different entities reduces the effort in the construction of specific systems. In this work we propose an architecture based on ontologies for the integration of data published on the Internet. Its use is illustrated through experimental cases for the integration of information on the Internet, showing how the use of ontologies can bring more relevant results.

Keywords: semantic Web, data-integration, ontologies.

Preâmbulo

Após a sessão de defesa da dissertação, os membros da banca sugeriram alterar parcialmente o título do trabalho. O novo título, mais adequado à ênfase do trabalho seria "IntegraWeb: uma proposta de arquitetura baseada em mapeamentos semânticos". Tal alteração, entretanto, não é possível na legislação vigente da USP.

Conteúdo

Lista de Abreviaturas	xi
Lista de Figuras	xiii
1 Introdução	1
1.1 Motivação	1
1.2 Panorama Tecnológico	3
1.3 Objetivos	5
1.3.1 Objetivo geral	5
1.3.2 Objetivos específicos	5
1.4 Metodologia	5
1.5 Organização do Trabalho	6
2 Conceitos Básicos	7
2.1 Web Semântica	7
2.1.1 Ontologias	8
2.1.2 Exemplos de ontologias	9
2.1.3 Linguagens da Web Semântica	11
2.2 Mineração de dados	16
2.3 Integração de dados	19
2.4 Trabalhos relacionados	20
2.5 Síntese	21
3 Arquitetura proposta	23
3.1 Escolha do domínio de dados	24
3.2 Definição de uma ontologia	24
3.3 Preparação e recuperação de dados	26
3.4 Persistência e centralização de dados	27
3.5 Resolução de duplicações e conflitos	29
3.6 Regras para inferências sobre dados	31
3.7 Camada de apresentação	31
4 Experimentos e Resultados	33
4.1 Implementação	33
4.2 Experimento 1: Integração de dados	36
4.3 Experimento 2: Resolução de conflitos	38

4.4	Experimento 3: Consultas combinadas entre diferentes conceitos	41
5	Conclusões e Trabalhos Futuros	49
5.1	Conclusões	49
5.2	Possíveis Extensões	50
A	Consulta SPARQL para integração de dados	53
B	Consulta com várias restrições	55
	Bibliografia	59

Lista de Abreviaturas

AJAX	Javascript e XML Assíncrono (<i>Asynchronous Javascript and XML</i>)
API	Interface de Programação de Aplicativos (<i>Application Programming Interface</i>)
BD	Banco de dados
HTML	Linguagem de Marcação de Hipertexto (<i>HyperText Markup Language</i>)
OBDA	Acesso a dados baseado em ontologia (<i>Ontology-based data access</i>)
RDF	Framework de descrição de recursos (<i>Resource Description Framework</i>)
REST	Transferência de Estado Representacional (<i>Representational State Transfer</i>)
SOAP	Protocolo Simples de Acesso a Objetos (<i>Simple Object Access Protocol</i>)
SPARQL	Protocolo SPARQL para consulta RDF (<i>SPARQL Protocol and RDF Query Language</i>)
URI	Identificador Padrão de Recursos (<i>Uniform Resource Identifier</i>)
URL	Localizador Padrão de Recursos (<i>Uniform Resource Locator</i>)
XPATH	Linguagem de consulta para XML (<i>XML Path Language</i>)
WSDL	Linguagem de Descrição de Serviços Web (<i>Web Services Description Language</i>)

Lista de Figuras

2.1	Exemplo de taxonomia	8
2.2	Exemplo de inferência de regras	9
2.3	Exemplo da definição de uma organização pelo projeto Schema.org	10
2.4	Integração de dados no DBpedia (<i>Auer et al., 2007</i>)	11
2.5	Integração de fontes de dados abertos em agosto de 2014 (<i>Cyganiak e Jentzsch, 2014</i>)	12
2.6	Exemplo de código HTML	13
2.7	Exemplo de um documento XML	14
2.8	Estrutura básica de uma tripla RDF	14
2.9	Exemplo de grafo RDF	15
2.10	Exemplo de um documento RDF em XML	15
2.11	Exemplo de consulta SPARQL	16
2.12	Exemplo de documento XML	17
2.13	Categorias da Mineração Web	18
2.14	Processo de Web Mining (<i>Zhang, 2011</i>)	18
3.1	Proposta de arquitetura de sistema de integração de dados da Web	23
3.2	Fontes de dados sobre eventos e restaurantes	24
3.3	Processo de extração e persistência de dados	25
3.4	Hierarquia de conceitos	25
3.5	Extração de latitude e longitude em diferentes portais	26
3.6	Repetição de padrão no portal Guia da Semana	27
3.7	Processo de extração de um endereço no portal Guia da Semana	28
3.8	Centralização e ponto de acesso a dados	28
3.9	Conflito de informações em diferentes portais	29
3.10	Processo de resolução de conflitos	30
3.11	Regra que define o conceito perto (near)	31
3.12	Exemplo de apresentação de resultado de consulta	32
4.1	Processo de extração e conversão em triplas RDF	34
4.2	Estações de metrô em São Paulo	35
4.3	Camada de apresentação da proposta	36
4.4	Interface de apresentação do resultado da consulta	36
4.5	Concentração de dados do portal Eventos USP	37
4.6	Concentração de dados do portal Guia da Semana	37
4.7	Concentração de dados do portal Guia da Folha	38

4.8	Concentração de eventos próximos ao Marco Zero de São Paulo	38
4.9	Consulta SPARQL para recuperar eventos próximos ao Marco Zero de São Paulo . .	39
4.10	Resultado do armazenamento de dados provenientes de diferentes portais	39
4.11	Evento "A Era do Rock" em diferentes portais	40
4.12	Geolocalização de evento incorreta no portal Guia da Semana	40
4.13	Geolocalização de evento correta no portal Guia da Folha	41
4.14	Consulta sobre o evento "A Era do Rock"	42
4.15	Informação inconsistente antes da resolução de conflitos	43
4.16	Informação consistente após resolução de conflitos	43
4.17	Dados obtidos após extração nos diferentes portais	44
4.18	Restaurantes próximo a um evento	44
4.19	Restaurantes de comida alemã próximo a um evento	45
4.20	Bares próximos a estações de metrô	45
4.21	Bares próximos à estação de metrô Sé	45
4.22	Consulta relacionando estabelecimentos distintos	46
4.23	Busca por evento, bar e restaurante próximos	46
4.24	Busca integrando diferentes conceitos sobre dados de diferentes portais	47

Capítulo 1

Introdução

Nesse capítulo são apresentados a motivação deste trabalho, o panorama tecnológico associado, a metodologia utilizada em seu desenvolvimento e a organização do documento.

1.1 Motivação

O progresso das pesquisas científicas, somado ao constante avanço do poder computacional, tem impulsionado a humanidade a uma nova era digital na qual a capacidade de armazenar, correlacionar e produzir informação se torna fundamental para garantir vantagem competitiva para as organizações públicas e privadas. Uma grande quantidade de dados é diariamente recolhida e publicada na Internet por meio de incontáveis entidades distribuídas ao longo do planeta formando a Web atual, também conhecida por Web 2.0. Não por outro motivo, nos últimos dez anos tornaram-se cada vez mais populares os portais online para ensino a distância (e-learning), comércio eletrônico (e-commerce) e a transparência de dados públicos (e-government) (Klischewski, 2006; Mustapaşa *et al.*, 2010) que se valem da rede da Internet para disponibilizar dados em documentos nos mais diferentes formatos como, por exemplo, HTML, CSV ou XML.

A maioria dos dados que formam a Internet como a conhecemos é composta por documentos (semi-) estruturados cujo conteúdo é gerado dinamicamente através de consultas em banco de dados. De maneira geral, uma parcela representativa desses dados não possui uma estrutura bem definida de maneira que só pode ser compreendida por humanos; entretanto, dada a grande quantidade desses dados, esses só podem ser processados por computadores (Stumme *et al.*, 2006). Além disso, as informações não são centralizadas e ficam compreendidas em diferentes silos de informação, tais como Wikipédia, Facebook, Flickr, Google Maps, entre outros. Essas aplicações geram e reutilizam a informação de maneira colaborativa por meio de mecanismos que permitem a atualização de dados em tempo real usando tecnologias como o AJAX. Em contrapartida, toda essa informação acaba sendo confinada nessas fontes de conteúdo, dificultando uma maior integração de diferentes de conteúdos e gerando situações como duplicação da informação, informação incompleta ou excessivamente distribuída.

Um exemplo de duplicação é o perfil de um indivíduo, que pode ser encontrado em redes sociais como o Facebook, o Twitter e até mesmo o LinkedIn. Nesse caso, organizações distintas armazenam um mesmo conteúdo, o que implica diretamente não somente em maior custo para as empresas que não compartilham e centralizam essa informação mas, também, para o usuário que acaba sendo levado a preencher o próprio perfil repetidas vezes em todos os portais que demandem seu cadastro. Assim, as informações publicadas na Internet carecem de mecanismos que as inter-relacionem automaticamente. Em decorrência deste fato, muita informação na Internet é incompleta; tal situação poderia ser mitigada pela união dos dados armazenados em diferentes fontes. A título de exemplo, considere-se os portais de divulgação de eventos no Brasil: é comum encontrar cenários em que um determinado portal possui a informação da existência e local de uma palestra mas não informa o horário; um segundo portal indica o local da execução da mesma palestra, o horário e o palestrante. Deste modo, não é possível identificar as palestras de um determinado indivíduo senão pela busca

em ambos os sites e composição da informação. Embora trabalhosa, tal pesquisa ainda é possível se limitada a uma palestra específica de um indivíduo em especial; no entanto, ao estender a todas as palestras, contidas em diferentes sites da Internet, a mesma pesquisa passa a ser inviável, uma vez que são muitos os portais e informações que precisam ser avaliados. Ao combinar as informações de todos os portais acerca desse mesmo domínio, torna-se possível entender e pesquisar melhor as informações sobre o assunto e automatizar processos como, por exemplo, montar a grade de apresentações de um determinado palestrante sem o ônus de pesquisar em diferentes portais.

A rapidez na obtenção de informação faz com que as pessoas não percam tempo. Como nos dias de hoje a informação está espalhada na Internet, para tomar uma decisão mais assertiva muitas vezes o usuário é obrigado a juntá-las, interpretá-las e correlacioná-las de maneira a obter uma visão mais ampla sobre um determinado domínio de interesse. Atualmente um sujeito que gosta de eventos culturais e tenha o interesse em decidir entre ir a uma palestra dentro de uma faculdade ou a um evento artístico que acontece em um parque precisa navegar por diferentes portais para entender a localização, o horário e poder escolher entre uma, outra, ou ambas. Nesse caso, ao menos um portal de uma faculdade e um portal de eventos artísticos acaba sendo visitado na Internet uma vez que na realidade do Brasil e na de outros países do mundo, o portal que concentra dados sobre cerimônias dentro de uma organização é muitas vezes mantido na própria organização. Nesse sentido, uma nova abordagem para captura e pesquisa da informação distribuída na Internet é necessária.

Considere a escolha de um passeio cultural para um final de semana. Existem vários portais na Internet que divulgam dados sobre restaurantes, palestras, peças, etc., em uma dada região para um determinado dia e horário. O *Catraca Livre*¹, por exemplo, possui uma grande quantidade de informações deste tipo sobre São Paulo, Rio de Janeiro e outras seis capitais do Brasil, mas não contempla os eventos que acontecem em cidades ou países fora desse escopo. Já o site *Visite São Paulo*², contém outras informações sobre a capital e completa as informações do *Catraca Livre*. A própria USP possui sua própria agenda de eventos, o *USP Eventos*³. No entanto, embora existam portais que agregam informações sobre eventos, não há uma visão ampla e centralizada sobre tudo o que ocorre numa área próxima ao usuário, o que o obriga a escolher um entre todos os portais ou então a visitar cada um deles para entender e escolher a opção mais adequada a ele naquele momento. Ao centralizar essa informação sob um mesmo domínio da informação, torna-se possível realizar buscas mais relevantes sobre um determinado tema.

A integração entre fontes com domínios distintos é outro ponto relevante de atenção. Levando em consideração o exemplo dos eventos distribuídos por diferentes portais na Internet, podemos tornar essa consulta ainda mais rica unindo a essa base de conhecimento os dados sobre outros domínios como, por exemplo, a informação sobre transporte público. Em grandes metrópoles como São Paulo é cada vez mais frequente a adoção de transportes públicos como ônibus, metrô ou táxi para se deslocar pela cidade. No entanto, para pessoas que dependem exclusivamente desses meios de transporte, muitas vezes a escolha de um passeio, restaurante ou estabelecimento em geral pode depender da proximidade, por exemplo, de uma estação de metrô. Indo além, podemos querer saber os restaurantes abertos localizados perto de uma determinada palestra que desejamos assistir. Atualmente os portais de divulgação de bares e restaurantes não dispõem de inteligência para definir o significado de "perto" ou "longe" e por isso não são capazes de trazer esses dados com precisão. Desse modo, ao combinar informações como estações de metrô com os diferentes eventos na cidade e adicionar significado a essa informação de maneira a permitir estabelecer questões como proximidade entre diferentes pontos pode tornar a pesquisa de um indivíduo ainda mais rica e relevante.

¹<https://catracalivre.com.br/sp/agenda/#data=13/12/2016>.

²<http://www.visitesaopaulo.com/>.

³<http://www.eventos.usp.br/>.

1.2 Panorama Tecnológico

Em geral, a representação dos dados na Internet é feita através de páginas geradas dinamicamente através de consultas pré-definidas em bases de dados e usualmente representadas por uma linguagem de marcação como o HTML. São documentos raramente acessíveis fora da API proprietária (Mika, 2015), o que contribui para dificultar a busca de dados de maneira mais integrada. Desse modo, a interoperabilidade da informação proveniente de fontes distintas na Internet acaba limitada à camada de apresentação na qual a informação é representada em documentos pouco estruturados e sem uma semântica comum, o que potencializa a complexidade de unificação desses repositórios de conteúdo. Essa característica é facilmente observável ao buscar informações sobre eventos em portais brasileiros, uma vez que essa informação está espalhada na Internet, i.e, informações sobre eventos acadêmicos estão geralmente restritas aos próprios portais de universidades que os estão sediando, informações sobre shows estão espalhadas por diferentes portais privados, informações sobre restaurantes estão contidas em outros silos de informação distintos assim como acontece quando buscamos por bares e assim por diante.

A busca da informação na Internet pode ser melhorada a partir da integração e correlacionamento das informações publicadas na Internet. Nesse sentido, a Web Semântica busca transcender as barreiras atuais da integração e se vale de iniciativas como o Schema.org⁴, projeto mantido pelas grandes empresas de indexação e busca de dados na Internet como Google, Microsoft, Yandex e Pinterest que passaram a definir terminologias de referência para atribuir significado à informação. Assim o conteúdo produzido na Internet pode agora ser marcado de maneira a permitir que o computador passe a compreender conceitos mais abstratos como Teatro, Cinema, Restaurante, entre outros. Apesar disso, uma terminologia equivalente não é suficiente para alcançar a integração dos dados, uma vez que é necessária a existência de mecanismos capazes de recuperar as informações contidas nos diferentes portais de dados para então convertê-las para uma terminologia homogênea pré-definida. Nesse contexto, a recuperação de dados juntamente com a respectiva conversão para uma terminologia homogênea é uma solução capaz de incentivar a interconexão da informação.

Os diferentes formatos de representação da informação na Internet são outra barreira para a unificação da informação na Web. Isso porque diferentes linguagens de marcação como o HTML, o XML ou o CSV podem ser utilizadas para publicar uma informação. Tal dificuldade pode ser superada por meio da exploração semântica com o uso de ontologias (Gali *et al.*, 2004) que acabam promovendo um entendimento único e compartilhado dos dados que existem dentro de um determinado domínio de integração. Não por outro motivo, um dos objetivos mais comuns para o desenvolvimento de ontologias é o de compartilhar o entendimento comum da estrutura da informação (Noy e McGuinness, 2001). Assim, a Web semântica aproxima a possibilidade de realização de repositórios com dados organizados que podem ser utilizados para a pesquisa de informação inteligente para humanos e agentes computacionais (Gali *et al.*, 2004) na Internet.

As ontologias são um passo importante para a integração da informação. Elas consistem de diversos conceitos a respeito de um domínio que são expressos em uma linguagem que possui semântica bem definida, de fácil compreensão e leitura (Gali *et al.*, 2004). Também são conhecidas como a especificação explícita de uma conceitualização (Gruber, 1993), possuindo uma expressividade semântica que reside na inferência sobre o seu próprio conteúdo e permitindo a construção com menor esforço de representação redundante. Já o conteúdo de ontologias publicadas na Internet pode ser obtido a partir de técnicas recuperação de dados como a mineração de dados para a Web. No entanto, ao extrair a informação de diversos portais há muita repetição que precisa ser tratada.

Informações sobre uma mesma peça de teatro, uma sessão de cinema ou um evento cultural qualquer podem muitas vezes serem encontradas dentro de diferentes portais na Internet. Desse modo, a recuperação desses dados que estão espalhados na rede deve levar em consideração essa condição. Um sistema capaz de centralizar esses dados dentro da ótica de ontologias deve ser capaz de identificar e tratar esses casos podendo até mesmo ir além permitindo inclusive a mesclagem dos conteúdos encontrados propiciando informações mais completas.

⁴<http://schema.org>

Além das ontologias, que definem conceitos equivalentes para a integração da informação sobre um domínio, se faz necessário ainda identificar e extrair os dados relevantes sobre os diferentes portais que publicam a informação para só então anota-los semanticamente. Por sua vez, tal objetivo pode ser alcançado a partir do uso de técnicas de mineração de dados. Em particular, a mineração de dados voltada para a Web semântica (Semantic Web Mining) combina o desafio de fazer com que dados sejam compreensíveis para máquinas com o de extrair conhecimento útil escondido nesses documentos de forma automática por meio de técnicas de mineração em dados (Stumme *et al.*, 2006). É uma técnica que pretende extrair do conteúdo e da estrutura de recursos na Web a informação relevante de modo (semi-)automático (Stumme *et al.*, 2006) por meio do processo não-trivial de identificar padrões úteis, válidos e previamente desconhecidos (Fayyad *et al.*, 1996). Esse modo de recuperação de informação é válido tanto em cenários bem estruturados como tabelas em banco de dados, em textos semi-estruturados, ou até naqueles textos sem estrutura alguma, de maneira que podem ser aplicados para ajudar a criar a Web Semântica.

A Web semântica, também conhecida por Web 3.0, é uma das mais importantes fontes de informação (Berners-Lee *et al.*, 2002). É uma extensão da Web Atual e não uma Web totalmente nova (Yong-gui e Zhen, 2010) e envolve um campo de pesquisa que faz com que os dados possam ser entendidos por máquinas (Stumme *et al.*, 2006) além de prover meios para a criação de uma infraestrutura genérica capaz de realizar a integração e incentivar o reuso de dados (Bojars *et al.*, 2008). Essa fonte, que é similar à já conhecida Web de hipertextos, baseia-se em documentos provenientes da Internet, mas difere no que diz respeito à aplicação da semântica. Diferente da Web tradicional, no qual os links são ancoras de relacionamento em documentos HTML, na Web de dados (Web of Data) é possível adicionar semântica aos sites a fim de realizar a ligação entre eles (Bojars *et al.*, 2008). Cada relacionamento entre entidades arbitrárias também pode ser descrito de maneira tal que uma pessoa ou máquina possa explorar a Web de dados. Desse modo, todas as informações globais poderiam ser relacionadas e recuperadas a partir do mapeamento semântico com o uso de ontologias.

A Web Semântica, em conjunto com a mineração de dados aplicada à Web, tem o potencial não só de centralizar dados esparsos na Internet de modo a promover uma melhor tomada de decisão sobre um assunto compartilhado, mas também pode ser utilizada para integrar informações contidas na rede. Desse modo, tomando como exemplo a centralização dos eventos culturais de uma região contidos nos diversos portais da Internet, é possível não só ter uma visão ampla, única e centralizada sobre informações contidas nos diversos portais de dados mas também uma informação mais completa se unida com outros portais anotados semanticamente. Isso é, se pensarmos em pesquisar de maneira centralizada todos as peças de teatro ou musicais em cartaz produzidos por um mesmo diretor, podemos juntar a informação com dados de portais como o DBPedia que possui informações semânticas sobre os autores, produtores, entre outros diversos dados que podem tornar a busca mais próxima do interesse do usuário. Em outras palavras, torna-se possível também tornar a recuperação de informações mais rica e relevante.

A proposição de uma arquitetura capaz de alcançar a integração do conteúdo de diferentes portais na Internet e de proporcionar consultas mais próximas do interesse do usuário a partir do uso de ontologias, é portanto, tema muito relevante. Ao utilizar as técnicas de mineração de dados para identificar e recuperar a informação relevante dos portais da Internet para então anota-las semanticamente com o uso de ontologias, torna-se possível alcançar uma condição em que os dados podem ser centralizados, correlacionados, enriquecidos e publicados para novas consultas agora com semântica agregada. Com isso é possível responder perguntas que envolvem buscas complexas que dependem da informação que está inicialmente distribuída por entre diferentes portais como, por exemplo, quais restaurantes de comida italiana estão mais próximos a uma exposição que ocorre em São Paulo, quais eventos acontecem próximo ao metrô Butantã, qual o tipo de evento que mais ocorre em determinada região, quais estações de metrô possuem maior quantidade de opções de entretenimento, entre outras. Esse é objetivo deste trabalho.

1.3 Objetivos

1.3.1 Objetivo geral

O objetivo geral desse trabalho é a proposição de uma arquitetura para a integração de informação proveniente de fontes de dados heterogêneas na Internet. Como estudo de caso, a arquitetura é utilizada para integrar informações sobre eventos como palestras, exposições e peças de teatro, bem como restaurantes, bares e meios de transporte na cidade de São Paulo.

1.3.2 Objetivos específicos

Pra obter o objetivo geral, pretende-se:

- Reaproveitar modelos de ontologias existentes;
- Relacionar dados semelhantes, resolver duplicidades e inconsistências;
- Centralizar informações sobre um dado domínio a partir do uso da Web-semântica;
- Tornar possível o acesso a informação capturada por meio de consultas semânticas;
- Correlacionar a informação de diferentes domínios.

Como estudo de caso, a ontologia Schema.org. será utilizada e aplicada para caracterizar eventos diversos como palestras, peças de teatro, shows, exposições, restaurantes, bares e meios de transporte na cidade de São Paulo. A partir de seu uso, dados provenientes de fontes diversas da Internet serão integrados e centralizados em uma única fonte de informação capaz de permitir a realização de consultas semânticas.

1.4 Metodologia

Via de regra, todo o conhecimento científico produzido é cerceado por metodologias que auxiliam a primorizar a qualidade da obra apresentada. Este trabalho não é diferente de modo que segue uma sistematização bem definida para propor uma solução para o problema de pesquisa abordado. Nesse contexto, dentre os diferentes métodos de pesquisa existentes, optou-se por um processo rigoroso que, entre outras características, fosse capaz de guiar a apresentação da contribuição, a avaliação dos resultados e que permitisse a elaboração de artefatos, como sistemas ou modelos, para resolver problemas observáveis. Este trabalho se baseia na metodologia de pesquisa Design Science Research (DSR).

A DSR é uma forma de pesquisa em sistemas de informação que é cada vez mais aceita como uma abordagem legítima para pesquisas nessa área (Gregor e Hevner, 2013). Trata-se de um paradigma voltado a resolução de problemas e que busca criar inovações que definem ideias, práticas, capacidades técnicas ou produtos na qual análise, projeto, implementação, gerenciamento e uso de sistemas de informação pode ser efetivamente e eficientemente aplicados (Hevner *et al.*, 2004). E que pode ser particionada em seis atividades principais: a identificação e a motivação do problema, definição dos objetivos da solução, o projeto e desenvolvimento da solução, a demonstração, avaliação e a comunicação (Peffers *et al.*, 2007).

A primeira fase da aplicação da DSR é a identificação e motivação do problema define o que será estudado. É o ponto de partida para a elaboração de uma proposta solução. O problema a ser solucionado neste trabalho é a integração dos dados sobre eventos culturais e restaurantes que estão espalhados na Internet. A definição dos objetivos para a solução do problema encontrado anteriormente é outra característica da DSR. Para isso foi definido para este trabalho que uma possível solução deveria ser capaz de centralizar informações acerca do domínio de dados escolhido por meio de uma ótica com semântica agregada com o uso de ontologia e que fosse capaz de identificar e resolver duplicidades e inconsistências de dados.

A elaboração de uma proposta seguida do desenvolvimento é a fase que inclui as atividades que envolvem a criação de um artefato incluindo aí as funcionalidades desejadas e a arquitetura proposta. Neste trabalho a definição das características desejadas e a proposição da arquitetura teve início com tarefas de caráter teórico e bibliográfico na qual foram estudados os conceitos, tecnologias e trabalhos envolvidos com Web Semântica, com a mineração de dados e a integração da informação. Em seguida foi elaborado um modelo de arquitetura capaz de realizar a integração de dados a partir do uso de ontologias.

A fase de demonstração apresenta o uso do artefato para resolver o problema exposto e pode ser feito com o uso de experimentações, simulações, estudos de caso ou outra atividade apropriada (Pefferes *et al.*, 2007). Neste sentido a proposta de arquitetura é instanciada para eventos, restaurantes e para o metrô de São Paulo a partir de dados provenientes de portais amplamente conhecidos na Internet: o Guia da Folha, o Guia da Semana e o portal de eventos da USP. A informação obtida é então correlacionada e enriquecida por meio de inferências. A partir de então consultas mais elaboradas são realizadas sobre a informação extraindo respostas para perguntas como “Quais são os eventos que ocorrem próximo ao metrô Brigadeiro” e “Quais são os eventos que ocorrem próximos a pizzarias” conforme explicado na seção 1.1.

Finalmente, a última etapa do trabalho envolve os processos de análise e comunicação. A fase de análise leva em consideração a observação e medidas que verificam o quanto a proposta suporta a solução para o problema definido. Em outras palavras, o artefato é avaliado em termos de critérios como validade, na qual o artefato produz o que deveria produzir, utilidade, na qual se avalia se os resultados atingidos possuem valor, ou até em questões como qualidade ou eficácia (Gregor e Hevner, 2013). Desse modo, os resultados demonstrados foram avaliados levando em consideração a capacidade de unificar a informação, tratar duplicidades e o enriquecimento da informação. Além disso, é discutido possíveis pontos de melhoria e sugestões de continuidade para a pesquisa aqui exposta. Já a fase de comunicação diz respeito a própria aplicação da metodologia para a estruturação do trabalho (Pefferes *et al.*, 2007).

1.5 Organização do Trabalho

Este trabalho foi estruturado de maneira a proporcionar ao leitor um melhor entendimento sobre o problema abordado, os fundamentos utilizados, a solução proposta e os resultados obtidos. Nesse sentido, no Capítulo 2 abordam-se os principais conceitos sobre a próxima aposta relacionada à integração de informação para a Web atual. Nele são caracterizadas as áreas de pesquisa da Web Semântica e da mineração de dados na Web, e discutido como essas duas áreas podem convergir para auxiliar a resolver alguns dos problemas da integração da informação exposta em meio ao caos da rede de dados. O Capítulo 3 contém a proposta de solução desse desafio para a integração da informação de fontes de dados na Web. Por sua vez, o Capítulo 4 detalha a implementação da proposta e mostra os experimentos realizados, bem como uma avaliação dos resultados obtidos. Finalmente, o Capítulo 5 abrange uma discussão acerca dos resultados gerais do trabalho e apresenta propostas para continuação da pesquisa no futuro. Completam o documento dois apêndices A e B, onde são apresentadas as consultas semânticas (utilizando a linguagem SPARQL) efetuadas nos experimentos apresentados no Capítulo 4.

Capítulo 2

Conceitos Básicos

Nesse capítulo serão abordados os conceitos básicos utilizados neste trabalho visando a integração dos dados expostos na Internet. Em um primeiro momento, é realizado um estudo sobre o surgimento da Web Semântica e como as ontologias são utilizadas para agregar significado à informação visando à correlação dos dados. Abordam-se então as tecnologias e padronizações que auxiliam na propagação dessa proposta. Em um segundo momento, explora-se o contexto da Mineração de Dados, suas principais características, e como esta pode ser aplicada para auxiliar a evolução da Web Semântica. Por fim, verificam-se os diferentes modos de integração de dados e realiza-se uma síntese de tais técnicas, ilustrando como todos os conceitos abordados no capítulo podem ser aplicados de maneira a permitir a proposição de uma arquitetura voltada para a centralização de dados na Web capaz de permitir consultas mais ricas dentro de um determinado domínio de informação.

2.1 Web Semântica

Atualmente a maior parcela dos dados disponibilizados na Internet é voltada para a leitura de humanos (Berners-Lee *et al.*, 2002) e seu conteúdo é geralmente constituído por significados que não são compreendidos por máquinas. Além disso, nessa Web de dados os documentos estão interconectados por meio de âncoras de navegação, estruturas lógicas que apontam endereços da Web conhecidos por URLs (Uniform Resource Locator) que em conjunto formam uma grande infra-estrutura global. No entanto, dada a grande quantidade de informações exposta na rede se faz necessária a existência de um processo automatizado capaz de processar todo esse conteúdo disponível. Por esse motivo, tornou-se importante estabelecer maneiras que possibilitem às máquinas não só processar todo o dado disponível, mas também interpretar a informação assim como ela é compreendida pelos humanos. A Web Semântica é uma área de estudos que busca superar essa limitação de maneira a fazer não só com que a informação possa ser entendida pelas máquinas, mas também compartilhada e reutilizada na Web.

Para Berners-Lee *et al.* (2002) a Web Semântica, ou simplesmente Web 3.0, é uma das mais importantes fontes de informação na Internet. É considerada uma extensão da Web Atual (Web 2.0) e não uma Web totalmente nova (Yong-gui e Zhen, 2010) uma vez que reutiliza os meios estruturais pré-existentes da informação, agregando-lhes uma camada de informação. Envolve um campo de pesquisa que tem por objetivo fazer com que os dados possam ser entendidos pelas máquinas (Stumme *et al.*, 2006) e busca promover meios para permitir uma infra-estrutura genérica capaz de realizar a integração de dados e incentivar o reuso de informação (Bojars *et al.*, 2008). No entanto, a proposta central da Web Semântica não está em tornar a infra-estrutura da Internet atual mais inteligente, mas sim em utilizar formas de atribuir significado ao conteúdo dos documentos na Web a ponto de que a integração da informação entre eles não fique limitada apenas ao nível da camada de apresentação com os *hyperlinks* (Allemang e Hendler, 2011).

A Web Semântica habilita as máquinas a compreenderem e manipularem de maneira mais eficiente os dados expostos na Internet. Isso decorre porque ao trazer suporte ao significado do

conteúdo nos documentos, os computadores passam a trabalhar a informação de maneira mais próxima do que aquela que seria praticada por uma pessoa. A fonte de informação é similar àquela já utilizada na Web de hipertextos e baseia-se em documentos provenientes da Internet, mas diverge no que diz respeito à aplicação da semântica. Diferente da abordagem tradicional, no qual as ligações entre documentos são baseadas em âncoras de navegação que realizam o relacionamento em documentos HTML, na Web de dados é possível adicionar semântica aos sites a fim de construir uma ligação entre eles (Bojars *et al.*, 2008). Desse modo, cada relacionamento entre entidades arbitrárias também pode ser descrito de maneira tal que uma pessoa ou máquina possa explorar a Web de dados; assim, ao ter acesso a uma pequena porção de informação, seria possível encontrar outros dados relacionados, formando uma teia de informações interligada, denominada Linked-Data. Nesta teia, todas as informações globais estão relacionadas e podem ser recuperadas a partir do mapeamento semântico com o uso de ontologias, de modo a tornar possível interligar dados diversos, como por exemplo bancos de dados mantidos por organizações geograficamente distribuídas (Bizer *et al.*, 2009).

2.1.1 Ontologias

Ontologia é uma palavra que deriva do grego “onto” que significa ser e “logia” que representa o discurso escrito ou falado. É uma expressão cujas raízes estão situadas no campo da Filosofia a partir do estudo de teorias da natureza da existência (Berners-Lee *et al.*, 2002). Valendo-se desse contexto, os pesquisadores da Web de dados e da inteligência artificial passaram a adotar o termo para definir uma teia de conceitos, e relações entre eles, de maneira que possam ser processados pelas máquinas. Desse modo, podem ser criadas taxonomias, como apresentado na Figura 2.1, classificações, agrupamentos ou outras formas de representar domínios de conhecimento, geralmente visando o compartilhamento e reuso de dados.

Figura 2.1: *Exemplo de taxonomia*

As ontologias são o coração da Web Semântica. São elas que formalizam o conhecimento sobre um domínio específico de maneira a tornar possível a interpretação do significado de dados contidos em diferentes documento pelos computadores (Bienvenu *et al.*, 2013) permitindo dessa maneira um processamento automatizado e mais eficiente da informação. É frequentemente definida como a especificação explícita de uma conceitualização (Gruber, 1993) consistindo de diversos conceitos, expressos em uma linguagem que possui semântica bem definida, de fácil compreensão e leitura (Gali *et al.*, 2004; Kabir *et al.*, 2014) e que dizem respeito a um contexto delimitado. Em outras palavras, trata-se da descrição formal explícita de conceitos que podem ser descritos como classes, formadas por atributos, e que descrevem características e restrições (Jain *et al.*, 2011).

O problema da unificação de diferentes formas de dados pode ser superado por meio da exploração semântica com o uso de ontologias. Por serem caracterizadas por diversos conceitos a respeito de um domínio, acabam promovendo um entendimento único e compartilhado dos dados que existem

dentro deste domínio. Desse modo, um dos objetivos mais comuns para o desenvolvimento de ontologias é o de compartilhar o entendimento comum da estrutura da informação (Noy e McGuinness, 2001). Assim, a Web semântica aproxima a possibilidade de realização de repositórios com dados organizados que podem ser utilizados para a pesquisa de informação inteligente para humanos e agentes computacionais (Gali *et al.*, 2004) na Internet. Regras de inferências nas ontologias adicionam ainda mais poder. Uma ontologia pode expressar uma regra para que um programa capaz de interpretá-la possa deduzir coisas como “Se a Universidade de São Paulo está situada na cidade de São Paulo” e “São Paulo é uma cidade do Brasil”, então “A Universidade de São Paulo está localizada no Brasil” assim como representado na Figura 2.2. Assim, as ontologias podem melhorar a funcionalidade da Web, aumentando a acurácia dos indexadores de informação e permitindo buscas mais ricas e sem ambigüidade acerca de um conceito.

Figura 2.2: Exemplo de inferência de regras

2.1.2 Exemplos de ontologias

O crescente aumento no interesse da aplicação de ontologias para atribuição de significado aos dados incentivou a criação de projetos de padronização de conceitos. Essa normalização de significado busca alcançar um entendimento único e reaproveitável de conceitos a fim de alcançar a integração dos dados expostos na Internet. Valendo-se desse contexto é que diversos esforços passaram a aparecer com maior força não só na comunidade científica mas inclusive na iniciativa privada.

Alguns dos projetos mais importantes nessa área são o **Schema.org** patrocinado pelos grandes indexadores de informação como Google e Bing, o vocabulário FOAF (Brickley e Miller, 2010) (acrônimo Friend of a Friend) que busca descrever pessoas e relações e atividades entre elas, o DBPedia¹ que busca elevar a informação da Wikipédia para a Web Semântica e o projeto Linking Open Data² que pretende contribuir com o inter-relacionamento da informação na Internet a partir da aplicação de Web Semântica.

Schema.org

Interessados pela possibilidade de entregar conteúdo mais relevante e personalizado para seus usuários, os maiores portais de pesquisa de dados correntes, Google, Bing, Yahoo e o Yandex³

¹<http://wiki.dbpedia.org/>

²<http://linkeddata.org/>

³<http://www.yandex.com/>

criaram a iniciativa Schema.org⁴. A idéia por trás desse esforço conjunto se traduz em melhorar a acurácia e apresentação de resultados da pesquisa nesses portais por meio da criação de marcações padronizadas que podem ser adicionadas pelos desenvolvedores das páginas HTML fazendo com que a máquina possa interpretar melhor os dados ali contidos (Mika, 2015; Tort e Olivé, 2014). Desse modo, o portal Schema.org se estabeleceu com um conjunto padronizado de ontologias, constituído pela documentação dos conceitos e exemplos de uso das classes que representam os tipos de dados mais populares contidos na Web e construídas a partir de esforços já pré-existentes para a criação de terminologias, que estão sendo cada vez mais adotadas em larga escala (Mika, 2015). Entre os exemplos de definições contidas nesse repositório estão, por exemplo, o perfil de uma pessoa bem como as propriedades que o definem, informações sobre produtos, entidades de negócio, eventos, organizações, restaurantes, entre muitos outros.

A Figura 2.3 ilustra o que seria uma organização do ponto de vista semântico e exemplifica como essas ontologias podem ser utilizadas em documentos na Internet. Nesse exemplo, podemos notar a descrição de uma organização (Organization⁵), a quantidade de vezes em que ela foi utilizada em diferentes portais da Web, as propriedades que ela possui, i.e, endereço ("*address*"), email, entre outros; além disto, descreve-se também o tipo de cada uma dessas propriedades, juntamente com uma breve descrição de cada uma delas.

Property	Expected Type	Description
Properties from Organization		
<u>address</u>	<u>PostalAddress</u> or <u>Text</u>	Physical address of the item.
<u>aggregateRating</u>	<u>AggregateRating</u>	The overall rating, based on a collection of reviews or ratings, of the item.
<u>alumni</u>	<u>Person</u>	Alumni of an organization. Inverse property: <u>alumniOf</u> .
<u>areaServed</u>	<u>AdministrativeArea</u> or <u>GeoShape</u> or <u>Place</u> or <u>Text</u>	The geographic area where a service or offered item is provided. Supersedes <u>serviceArea</u> .
<u>award</u>	<u>Text</u>	An award won by or for this item. Supersedes <u>awards</u> .
<u>brand</u>	<u>Brand</u> or <u>Organization</u>	The brand(s) associated with a product or service, or the brand(s) maintained by an organization or business person.
<u>contactPoint</u>	<u>ContactPoint</u>	A contact point for a person or organization. Supersedes <u>contactPoints</u> .
<u>department</u>	<u>Organization</u>	A relationship between an organization and a department of that organization, also described as an organization (allowing different urls, logos, opening hours). For example: a store with a pharmacy, or a bakery with a cafe.
<u>dissolutionDate</u>	<u>Date</u>	The date that this organization was dissolved.
<u>duns</u>	<u>Text</u>	The Dun & Bradstreet DUNS number for identifying an organization or business person.
<u>email</u>	<u>Text</u>	Email address.

Figura 2.3: Exemplo da definição de uma organização pelo projeto Schema.org

DBPedia

A enciclopédia digital Wikipédia⁶ é hoje um dos portais mais importantes do planeta, contando com informações nas mais diversas linguagens. Nele é disponibilizado o esforço de uma produção colaborativa de pessoas que produzem o conteúdo e que o revisam para garantir boa qualidade dos dados ali apresentados. No entanto, a informação ali contida está limitada à busca por palavras-chave,

⁴<http://schema.org/>

⁵<http://1.schemaorgae.appspot.com/Organization>

⁶<https://www.wikipedia.org/>

gerando uma série de problemas que vão desde dados contraditórios, inconsistência de convenções taxonômicas, erros ou até spam (Auer *et al.*, 2007). Levando isso em consideração, o DBpedia foi criado com o objetivo de colocar uma camada semântica sobre o conteúdo ali exposto e, com isso, solucionar parte destes problemas a partir de consultas mais sofisticadas.

O projeto DBpedia é um esforço de extração dos dados da Wikipédia e conversão para documentos de formato RDF. De acordo com Auer *et al.* (2007), é um trabalho que hoje conta com muitos milhões de triplas RDF e que se interliga com outras bases de informação, conforme representado na Figura 2.4, podendo ser expandido para um conjunto de bilhões de triplas. Todos os dados extraídos são expostos para consultas na Internet a partir de uma infra-estrutura refinada que conta com bancos relacionais para armazenamento da informação tais como o MySQL e também com bancos de dados que lidam com informação semânticamente anotada assim como a solução Virtuoso⁷ permite. Esse esforço faz parte da comunidade que mantém o projeto Linking Open Data⁸ (LOD) que tem o interesse de criar uma grande quantidade de conjunto de dados e ontologias interligados, contribuindo com a Web Semântica; em agosto de 2014 já assimilava quase 600 conjuntos, alcançando em 2017 a marca de 1139 conjuntos além do DBPedia conforme ilustra a Figura 2.5.

Figura 2.4: Integração de dados no DBpedia (Auer *et al.*, 2007)

2.1.3 Linguagens da Web Semântica

De acordo com Allemang e Hendler (2011), as tecnologias necessárias para habilitar o novo paradigma da Web já estão disponíveis e envolvem a utilização das linguagens de marcação XML (*Extensible Markup Language*) e HTML (*HyperText Markup Language*), a linguagem RDF (*Resource Description Framework*) para a descrição de ontologias e uma linguagem para consultas em RDF, o SPARQL (*SPARQL query language for RDF*). Em conjunto, essas tecnologias tornam possível a integração de dados provenientes de diferentes bases de dados, documentos, motores de inferência ou qualquer outras fontes de informação que possam expressar conhecimento (Wang *et al.*, 2009).

⁷<https://virtuoso.openlinksw.com/>

⁸<http://linkeddata.org/>

Figura 2.5: Integração de fontes de dados abertos em agosto de 2014 (Cyganiak e Jentzsch, 2014)

HyperText Markup Language (HTML)

Uma parcela considerável da informação publicada mundialmente na Internet esta disponibilizada através de documentos no formato HTML. Isso ocorre porque essa linguagem foi projetada para estruturar a informação para que um navegador possa interpretar a estrutura dos dados ali definida, e exibi-la de maneira visual para um usuário por meio de botões, textos e outros elementos, que permitem a um usuário entender e manipular a página com maior facilidade. É baseada em um padrão de marcação de dados estabelecido pela norma ISO8879 que define o SGML (*Standard Generalized Markup Language*) e tem por responsabilidade representar os diversos elementos de uma página, conforme ilustra a Figura 2.6.

```
1 <!DOCTYPE html>
2 <html>
3 <head>
4   <title>Titulo da pagina</title>
5 </head>
6 <body>
7
8   <h1>Cabecalho</h1>
9   <p>Paragrafo</p>
10
11   <!-- comentario -->
12
13   <br />
14
15   <a href="http://www.usp.br">ancora de navegacao (hyperlink)</a>
16
17 </body>
18 </html>
```

Figura 2.6: Exemplo de código HTML

eXtended Markup Language (XML)

Em virtude da publicação de informação em larga escala nos meios eletrônicos, tornou-se necessário a existência de um mecanismo capaz de facilitar a troca de uma grande variedade de dados (W3C, 2008). Nesse contexto, o XML surge como solução para este problema, diferentemente do HTML que foi projetado para a marcação de estrutura de exibição em navegadores. Exemplos da versatilidade da aplicação do XML vão desde páginas na Internet, livros, interfaces de programação ou a representação bancos de dados relacionais e seu conteúdo permitindo inclusive o processamento de dados que não estavam previamente armazenados como XML (Wang *et al.*, 2009). Além disso, a manipulação de um documento XML é feita de maneira trivial pela grande maioria das linguagens de programação (Heath e Motta, 2008),

Um documento XML consiste em unidades de armazenamento chamadas de entidades e que formam, em conjunto, uma árvore de nós ordenados. Essas entidades, formadas por caracteres, delimitam os dados através de marcações formando assim o modelo de armazenamento da informação e a estrutura lógica do documento. Além disso, a ordem entre as entidades é relevante, sendo que sua alteração pode inviabilizar a interpretação do documento por ferramentas de software. Dessa maneira, permite-se a criação de rótulos como <cep> ou <nome> , como ilustrado pela Figura 2.7, que podem ser interpretados a partir da implementação de programas.

Um documento XML conciso respeita certas condições simples, como é possível verificar na Figura 2.7. A primeira condição é que toda marcação tem seu início e fim. Isso significa que um rótulo do tipo “<nome>” obrigatoriamente tem seu contraponto “</nome>” a menos que ele seja auto-contido como o exemplo da “geolocalizacao” na linha 6. Além disso o documento está contido em uma única marcação e todas elas estão aninhadas. A linha 7, por exemplo, mostra uma tag “<

```

1 <?xml version="1.0" encoding="UTF-8"?>
2 < Pessoa>
3   < nome>Felipe</ nome>
4   < sobrenome>Pierin</ sobrenome>
5   < email>fpierin@ime.usp.br</ email>
6   < geolocalizacao latitude="-23.558745" longitude="-46.731859" />
7   < endereco>
8     < cep>05508-090</ cep>
9     < cidade>Sao Paulo</ cidade>
10    < rua>Rua do Matao, 1010</ rua>
11  </ endereco>
12 </ Pessoa>

```

Figura 2.7: Exemplo de um documento XML

endereco>” que contém outras tags filhas. No entanto, como essas estruturas não possuem semântica sobre o significado, faz-se necessário o entendimento da estrutura por parte do projetista do sistema (Allemang e Hendler, 2011).

Resource Description Framework (RDF)

Na Web Semântica, quando queremos nos referir a algo ou a algum conceitos utilizamos recursos (resources). Desse modo, todas as entidades são definidas como recursos, bem como os seus relacionamentos (Allemang e Hendler, 2011). Nesse sentido, o RDF é uma recomendação da W3C para padronizar o uso de descrições de metadados de recursos baseados na Web (Deursen *et al.*, 2008) para representar, entre outras coisas, informações pessoais, redes sociais, metadados de artefatos digitais e promover meios de integrar fontes diferentes de informação visando facilitar a troca, fusão e compartilhamento de dados por aplicações diversas. As iniciativas DBPedia, Linked Open Data e o projeto Schema.org auxiliam a transformação dos dados para RDF, permitindo assim a interligação entre as diversas fontes existentes (Heath e Motta, 2008).

O RDF é uma extensão do mecanismo de *hyperlinks* e faz a conexão entre recursos da Web atual usando identificadores chamados URI (*Uniform Resource Identifier*) que definem o relacionamento entre as entidades que desejamos descrever. Desse modo, conforme exemplificado pela Figura 2.8, a estrutura básica de um documento RDF possui três elementos chaves: o sujeito, um predicado e o objeto de referência. O sujeito ao qual estabeleceremos atributos como, por exemplo, uma pessoa, um automóvel ou uma faculdade, uma entidade objeto que define algum valor tal qual um nome, uma cor ou uma matéria, e um predicado que estabelece a relação entre esse sujeito e a entidade objeto como, por exemplo, “conhece”, “é amigo de”, “é filho de”. No exemplo anterior, é possível observar que o sujeito é “Felipe Pierin”, representado pelo URI “http://www.example.org/persons/Felipe_Pierin”, o predicado é “mora em” (http://www.example.org/properties/mora_em) e o objeto “São Paulo” (http://www.example.org/places/Sao_Paulo).

Figura 2.8: Estrutura básica de uma tripla RDF

Ao expandir esse simples modelo para inúmeras conexões, alcançamos uma rede de recursos de informações, inter-relacionados por propriedades que estabelecem relações entre triplas. O resultado disso é um grafo RDF conforme exemplificado pela Figura 2.9. Essa condição é fundamental

porque permite que as informações sejam associadas com facilidade e favorece a interpretação por máquinas, uma vez que descreve uma parcela considerável dos dados que são processados por elas. (Allemang e Hendler, 2011). Além disso, como a publicação e conversão de documentos em RDF adiciona semântica aos recursos, ela também promove a integração da informação na Internet proveniente de fontes de dados diversas, uma vez que unifica o entendimento a respeito de um conteúdo, permite um melhor entendimento das máquinas a respeito de um determinado domínio e ainda torna possível a busca de conteúdo relevante por meio de linguagem padrão de consulta, o SPARQL (Heath e Motta, 2008).

Figura 2.9: Exemplo de grafo RDF

Por fim, outra propriedade relevante de um documento RDF é que ele pode ser representado em uma estrutura XML, como ilustrado pela Figura 2.10. Nesse exemplo há a definição de um objeto do tipo “Person” e que possui um nome definido na marcação “name” e um email definido por “mbox”. Essa forma de apresentação é denominada RDF/XML e têm sido amplamente utilizada para retratar informações semanticamente anotadas.

```

1 <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
2   xmlns:foaf="http://xmlns.com/foaf/0.1/">
3   <foaf:Person>
4     <foaf:name>Felipe Pierin</foaf:name>
5     <foaf:mbox rdf:resource="fpierin@ime.usp.br"/>
6   </foaf:Person>
7 </rdf:RDF>

```

Figura 2.10: Exemplo de um documento RDF em XML

SPARQL Protocol and RDF Query Language (SPARQL)

Em virtude do documento RDF poder ser entendido como um grafo direcionado para a representação de informação, iniciativas como o SPARQL nasceram para aproveitar essa característica. O protocolo de busca em RDF é, de maneira simplificada, uma linguagem de consulta de correspondência em grafos. Desse modo, levando em consideração uma fonte de dados em RDF, a consulta seria um padrão combinado contra essa fonte e os valores obtidos a partir dessa correspondência são processados para dar a resposta (Pérez *et al.*, 2006). Além disso, uma consulta SPARQL não precisa ser restrita a uma única fonte de informação (*endpoint* SPARQL), isso é, vários grafos RDF podem ser combinados dentro de uma única pesquisa, sejam eles documentos RDF nativos ou gerados em tempo de execução via outros sistemas intermediários (W3C, 2013). Por esse motivo, essa

linguagem é fundamental no que diz respeito ao Linked Data pois é ela que permite a extração de conhecimento dos diversos endpoints SPARQLs distribuídos (Singh e Singh, 2010).

Conforme observado por Pérez *et al.* (2006), uma consulta SPARQL apresenta três partes fundamentais: a correspondência de padrões, os modificadores de resultado e o resultado final. No que diz respeito a correspondência de padrões, a linguagem não só apresenta características de correspondência de padrões em gráficos tais quais partes opcionais, união, aninhamento e filtragem como também permite restringir a qual fonte de dados será aplicada. Já os modificadores de resultado permitem modificar os valores encontrados uma vez que a saída padrão já foi computada. Por fim, o resultado final pode ser tanto um valor específico, como sim ou não, ou até mesmo um sub-grafo das fontes iniciais utilizadas na consulta. Um exemplo de um código SPARQL é ilustrado na Figura 2.11. Na primeira linha o prefixo (PREFIX) “foaf” indica que será utilizado para representar a fonte de dados “http://xmlns.com/foaf/0.1/”. Na segunda linha há uma restrição de informação com o comando “SELECT” que possui um filtro “WHERE” para recuperar todas as triplas conectadas pela ligação “foaf:name”.

```
1 PREFIX foaf: <http://xmlns.com/foaf/0.1/>
2 SELECT ?name
3 WHERE {
4     ?person foaf:name ?name .
5 }
```

Figura 2.11: Exemplo de consulta SPARQL

2.2 Mineração de dados

A mineração de dados é uma área de pesquisa que almeja a extração do conhecimento útil existente em uma grande quantidade de dados. Seu foco está concentrado em buscar padrões e regras previamente desconhecidos usando, por exemplo, métodos estatísticos, modelos matemáticos e algoritmos de aprendizado de máquina (Kabir *et al.*, 2014; Quboa e Sarae, 2013). Pode ser dividida em três sub-áreas: mineração de dados (*data mining*), mineração de textos (*text mining*) e mineração na Web (*Web mining*). Além disso, pode manipular dados estruturados ou não estruturados (Singh e Singh, 2010). Enquanto a mineração de dados lida com dados organizados e estruturados que estão geralmente armazenados em uma base de dados, a mineração de textos, por sua vez, lida com dados não-estruturados. Já a mineração na Web se situa entre os dois extremos, combinando técnicas de ambas para aplicar no vasto repositório de documentos da Internet (Quboa e Sarae, 2013; Singh e Singh, 2010).

De acordo com o trabalho de Sleiman *et al.* (2011), a recuperação da informação útil pode ser facilitada por meio de extratores classificados em dois grupos distintos: os baseados em heurística ou em regras. Além disso, regras de extração de informação podem ser supervisionadas, quando há necessidade de treinamento de um algoritmo, ou não-supervisionadas quando essa necessidade inexistente. As regras comuns de extração de informações e reconhecimento de padrões variam de expressões regulares a gramáticas livres de contexto, regras de primeira ordem, tradutores ou modelos baseados em linguagens específicas como XQuery ou XPath.

O reconhecimento de padrões pode ser feito de modo manual ou a partir de técnicas supervisionadas de aprendizado de máquina que variam desde a extração de palavras chaves usando o método tf.idf, conhecido da área de Recuperação de Informação (Information Retrieval), a técnicas que avaliam as estruturas sintáticas do HTML ou a linguagem natural em consideração, até técnicas que extraem a referência para uma estrutura modelo alvo, como uma ontologia (Stumme *et al.*, 2006). Sistemas como o RAPIER (Califf e Mooney, 2004) e o BWI (Freitag e Kushmerick, 2000), por exemplo, fazem essa tarefa a partir de comparação de padrões em torno da informação relevante com a estrutura geral, calculando a frequência relativa das palavras. Já o modelo de May (2002) se baseia na estrutura sintática dos documentos analisados usando XPath.

A linguagem XPath foi construída de forma a permitir a recuperação e a manipulação simples de textos, números ou valores booleanos em documentos XML (Clark e DeRose, 1999); entretanto, também pode ser aplicada a outros documentos com estrutura parecida, como o HTML. Constituída de uma sintaxe compacta, lembra a notação de caminhos em URLs uma vez que é utilizada tendo em foco a localização de um conteúdo na estrutura hierárquica do arquivo alvo. O XPath modela um documento XML como uma árvore de nós, incluindo nós de elemento, nós de atributo e nós de texto. Em resumo, define uma maneira de calcular um valor de sequência de caracteres para cada tipo de nó. Tomando como exemplo a Figura 2.12, utilizando a expressão XPath “/universidades/universidade/sigla”, obteremos como resultado a sequência de valores “USP”, “Unicamp”, “UFRJ” e “UFSC”.

```
1 <?xml version="1.0" encoding="UTF-8"?>
2 <universidades>
3   <universidade>
4     <nome lang="pt">Universidade de São Paulo</title>
5     <sigla>USP</sigla>
6     <alunos>94875</price>
7   </universidade>
8   <universidade>
9     <nome lang="pt">Universidade Estadual de Campinas</title>
10    <sigla>Unicamp</sigla>
11    <alunos>40850</price>
12  </universidade>
13  <universidade>
14    <nome lang="pt">Universidade Federal do Rio de Janeiro</title>
15    <sigla>UFRJ</sigla>
16    <alunos>55887</price>
17  </universidade>
18  <universidade>
19    <nome lang="pt">Universidade Federal do Santa Catarina</title>
20    <sigla>UFSC</sigla>
21    <alunos>46225</price>
22  </universidade>
23 </universidades>
```

Figura 2.12: *Exemplo de documento XML*

Embora existam diferentes técnicas para a mineração de dados em documentos, todas elas técnicas podem ser encapsuladas de maneira tal que uma requisição para um documento, por exemplo escrito em RDF, possa ser reescrita através de um mediador entre cliente e servidor que irá realizar a tradução da requisição da página, extrair a informação relevante e enviá-la ao cliente (Berendt *et al.*, 2004; Kushmerick *et al.*, 1997). Isso permite que elas possam ser reaproveitadas dentro de arquiteturas com contextos específicos que não só a mineração de dados. Neste trabalho, as informações extraídas de portais na Internet são reescritas como triplas RDF.

Mineração na Web

A mineração na Web é um processo de extração de informação ou conhecimento usando diferentes técnicas de mineração, como regras de descoberta de associação, agrupamento e classificação aplicadas a fontes de dados expostas na Internet (Quboa e Saraee, 2013; Zhang, 2011). Conforme Quboa e Saraee (2013) e Stumme *et al.* (2006) a área pode ser dividida em três categorias (Figura 2.2): a mineração de uso na Web, a mineração de estrutura na Web e a mineração de conteúdo na Web. A primeira busca entender o comportamento de um determinado usuário na Internet valendo-se dos registros deixados por ele em arquivos de logs como o do navegador Web. Com isso, essa categoria de pesquisa visa criar ou melhorar serviços, tornando-os mais personalizados. A segunda é uma técnica que tenta reconhecer informações de topologia entre fontes de informações na Internet. Já a mineração de conteúdo atem-se à mineração de informação em textos e imagens a

fim de auxiliar uma pessoa a encontrar resultados mais precisos conforme o seu interesse.

Figura 2.13: *Categorias da Mineração Web*

A mineração de conteúdo é um processo que aproveita a natureza da maioria dos documentos semi-estruturados contidos na Web a partir da exploração de anotações HTML e de suas marcações XML (Stumme *et al.*, 2006). Isso ocorre porque essas fontes de dados carregam não só a informação de como uma determinada página na Web deve ser apresentada para um usuário no navegador mas também a estrutura lógica de como esses dados são apresentados. Dessa maneira, ao identificar os padrões de como um determinado texto ocorre em um recurso Web, é possível elaborar processos automatizados para extrair o conteúdo relevante e então mapeá-los para uma nova estrutura que permita um melhor gerenciamento do domínio da informação. Em geral, as técnicas de extração de informação visando à construção da Web semântica dependem muito do reconhecimento da informação em documentos que geralmente é feita a partir da descoberta de padrões (Berendt *et al.*, 2004).

De acordo com Zhang (2011), a mineração de dados na Web é alcançada por meio de quatro passos conforme ilustra a Figura 2.14: (i) a descoberta de recursos que compreende a escolha e obtenção de dados de um recurso na Web seja ele um documento eletrônico, um e-mail ou registros de navegação de um usuário; (ii) a escolha da informação e o pré-processamento que envolve remover os dados irrelevantes à informação escolhida; (iii) a busca automatizada do padrão por meio de ferramentas de mineração como, por exemplo, o XPath e (iv) a análise do padrão encontrado, que envolve a validação por meios automatizados ou a partir da avaliação humana.

Figura 2.14: *Processo de Web Mining (Zhang, 2011)*

Mineração de dados para a Web Semântica

Embora a mineração Web forneça meios para extrair conhecimento relevante na Web, a maioria dos dados espalhados na rede é desestruturada, a ponto de tornar complicada a agregação de conteúdos sob uma estrutura comum (Kabir *et al.*, 2014). Nesse contexto, a Web Semântica pode ser combinada com essas técnicas de mineração na Web formando uma área de pesquisa que engloba as áreas de Mineração de dados e Web Semântica para suprir essa deficiência.

A mineração de dados voltada para a Web semântica (*Semantic Web Mining*) combina o desafio de fazer com que dados sejam compreensíveis para máquinas com o de extrair conhecimento útil escondido nesses documentos de forma automática por meio de técnicas de mineração em dados (Stumme *et al.*, 2006). Esse modo de recuperação de informação é válido tanto em cenários bem estruturados, como tabelas em banco de dados, em textos semi-estruturados, ou até naqueles textos sem estrutura alguma, de maneira que podem ser aplicados para ajudar a criar a Web Semântica.

2.3 Integração de dados

Vivemos em meio a um cenário em que a informação é constantemente exposta não só pelos meios tradicionais da Web, a partir de páginas dinâmicas HTML, mas também facilitada nos diversos aplicativos por meio de interfaces para a recuperação de informação em software chamadas API (*Application Programming Interface*), cujo propósito é o de permitir ao usuário final a criação de sistemas mais ricos que integram o conteúdo do aplicativo com outros de seu interesse. Exemplos de empresas que trabalham dessa forma são os mais diversos e compreendem grandes empresas de tecnologia como Facebook, LinkedIn, Amazon, Youtube, Dropbox, Instagram, Flickr, Twitter, entre outros. Por outro lado, toda essa informação é uma considerável parcela do conteúdo da Internet está confinado a esses silos de informação e isso dificulta uma visão homogênea e extração de dados sobre um determinado domínio de interesse (Civili *et al.*, 2013; Heath e Motta, 2008).

A necessidade em gerenciar vastas quantidades de informações provenientes de fontes de dados heterogêneas têm incentivado a pesquisa em maneiras mais inteligentes de se realizar a integração de dados sobre um mesmo domínio (Vettor *et al.*, 2014; Wang *et al.*, 2009). A integração e busca em fontes de dados heterogêneas tem sido um tópico muito explorado uma vez que é cada vez mais complicado o acesso a informações relevantes nesse contexto dada a velocidade de exposição de novos conteúdos e a grande quantidade de diferentes tipos de formatos. São dados em grande quantidade, desorganizados e desconectados. Desse modo, o objetivo da integração de dados é prover ao usuário uma maneira de acesso uniforme a múltiplas fontes de dados heterogêneas (Wang *et al.*, 2009). Entretanto, ainda são poucas as soluções genéricas para converter e conectar fontes de dados de origens diversas em um conjunto de dados único e coerente mas todas elas podem ser categorizadas em um modelo de integração no qual todos os dados são recuperados e agrupados dentro de um mesmo banco de dados ou em um modelo que se vale de um processo bem definido para a integração de informação mas que não depende de um banco de dados centralizado.

A integração de dados pode ser alcançada por meio de uma dentre duas estratégias distintas conhecidas como Global As View (GAV) ou Local As View (LAV) (Abdellaoui e Nader, 2015; Putra e Khalil, 2017; Wang *et al.*, 2017). A estratégia GAV é tradicionalmente utilizada para aplicações em que há consultas federadas nas quais uma única consulta dispara pesquisas em múltiplas fontes de dados e unifica a informação recuperada por meio de múltiplas camadas de abstrações. Já o método LAV realiza a materialização desses dados em um banco de dados único. Contudo, cada uma dessas técnicas possuem vantagens e desvantagens. O método GAV, por exemplo, pode não ser aplicável em situações em que existam fontes de dados faltando para a visão completa de um determinado domínio como, por exemplo, um portal estiver fora do ar e passar por modificações nos dados dele. Por outro lado não depende de materialização dos dados em um banco centralizado o que pode ser um estímulo do ponto de vista financeiro (Putra e Khalil, 2017). A estratégia LAV, por sua vez, trabalha melhor em fontes de dados incompletas ou que não possam ser acessadas em um determinado momento, como no contexto da Internet (Putra e Khalil, 2017), ou em situações nas quais a realização de uma consulta Federada seja lenta demais para trazer um resultado dentro de um tempo aceitável. A fim de evitar problemas com fontes de dados incompletas, portais inacessíveis ou lentidão para consultas federadas, neste trabalho adotou-se uma abordagem LAV.

A combinação de dados não é um processo trivial e exige soluções que vão desde formas para lidar com as divergências entre documentos, duplicações e ruídos. É um processo que exige adaptação das fontes de informação a nível sintático, estrutural e semântico (Vettor *et al.*, 2014). A necessidade de atuar no nível sintático é justificada porque o formato dos dados pode ser diferente já que podemos, por exemplo, querer combinar a informação contida em um banco de dados com outra contida em documentos XML. Mesmo que os documentos sejam do mesmo tipo, por exemplo páginas HTML, a estrutura pode não ser igual, e portanto deve-se atuar no nível estrutural. Por fim, é necessário atuar no nível semântico porque a representação do conhecimento nos documentos que terão a informação combinada pode ser diferente. Além disso, ao recuperar e combinar informações provenientes de múltiplos recursos é possível que exista repetição de informação, ruídos, imprecisão ou outros tipos de inconsistência que podem causar perda de significado ou mal-entendidos, que são prejudiciais ao resultado final da integração (Vettor *et al.*, 2014).

2.4 Trabalhos relacionados

A fim de auxiliar a resolução de possíveis problemas decorrentes da integração de dados, máquinas podem ser empregadas para desempenhar um papel relevante. Em outras palavras, há a necessidade de oferecer algoritmos e descrições semânticas precisas dos dados trabalhados para que o computador possa interpreta-las corretamente e entregar melhores resultados (Vettor *et al.*, 2014). Por esse motivo, vários linhas de estudos convergem e vão desde o uso de ontologias para mapear um domínio comum visando a solução do problema da integração de dados heterogêneos (Ahmed, 2008), a integração baseada em Sistemas Multi-Agentes (Sui *et al.*, 2009) ou o acesso a informação baseado em ontologias (Civili *et al.*, 2013; Kharlamov *et al.*, 2013; Lembo *et al.*, 2014).

Levando em consideração que a maior parte dos documentos existentes na Web está definida valendo-se de formatos semi-estruturados, e.g. XML, é de se esperar que a integração de dados seja feita por meio de anotações semânticas (May, 2002). Por essa razão, iniciativas como o SIOC (Bojars *et al.*, 2008) buscam uma proposta valendo-se do apontamento ontológico em Resource Description Framework (RDF) (RDF Working Group, 2014) mas visando à conexão de informação proveniente de comunidades sociais tais como Flickr, Youtube, Facebook e Wikipédia por meio do mapeamento das APIs disponibilizadas por estes sites. Já trabalhos como o de Cardoso e Sheth (2006) buscam criar uma ontologia para o a indústria do turismo.

Utilizar o conteúdo já existente na Web atual e convertê-lo para a Web semântica é outra linha de pesquisa. Ela visa à integração de informação, a reutilização de dados e a evolução da Web atual para a Web semântica. O Deep Annotation, por exemplo, é um framework para prover anotação para um vasto conjunto de dados a partir de uma ferramenta que é simples e intuitiva (Handschuh *et al.*, 2003). Trabalhos como o de Heß, Andreas e Dieric (2008) buscam este objetivo por meio de um mapeamento ontológico da informação a partir de um mecanismo semi-automático de marcação, no qual os usuários escolhem “tags” dentre uma lista de sugestões. Outros trabalhos como o de May (2002) fazem o uso da linguagem de consulta XPath⁹ para selecionar os nós de um documento XML, encontrar estruturas relevantes do documento para então mapeá-las ontologicamente. Já trabalhos como o Bottari (Balduini *et al.*, 2012) aliam a interpretação da diversidade de conteúdo produzido das publicações de pessoas no Twitter, seguida de mapeamento semântico desses dados em uma ontologia padronizada. A partir disso aplica o resultado obtido a fim de proporcionar uma realidade aumentada capaz de sugerir pontos de interesse como restaurantes.

A orquestração de serviços disponibilizados na Internet é outra abordagem de estudo amplamente explorada por pesquisadores. Trabalhos como o SBWS¹⁰ e o ASSAM (Heße Kushmerick, 2003; Heße *et al.*, 2004) buscam realizar mapeamento semântico sobre uma descrição de serviços WSDL¹¹ que funcionam sobre o protocolo SOAP¹². Com isso, um usuário pode anotar de forma semântica um Webservice a partir de sugestões de classes ontológicas para marcar cada elemento da descrição do serviço exposto. O projeto Satine¹³ é outro exemplo. Ele busca promover a criação e desenvolvimento de serviços Web enriquecidos semanticamente mas visando a integração de informação no domínio de viagens de maneira a padronizar e simplificar o processo de busca de passagens aéreas e hospedagem no domínio de viagens.

Outros estudos buscam mapear serviços REST¹⁴ em RDF de maneira a permitir a busca de conteúdo através de consultas SPARQL (Battle e Benson, 2008). Isso porque ao publicá-los em RDF, os dados ficam acessíveis via uma linguagem de pesquisa padrão (W3C, 2013) e isto torna mais fácil a integração de diferentes fontes, uma vez que os dados passam a ser entendidos por máquinas (Heath e Motta, 2008). Já trabalhos como o SOARI (Nardin *et al.*, 2011) buscam tornar possível a interoperabilidade entre sistemas heterogêneos e distribuídos por meio de uma ontologia comum a partir da qual podem trocar informação com um mesmo entendimento de significado.

⁹<https://www.w3.org/TR/xpath-31/>

¹⁰<http://asio.bbn.com/sbws.html>

¹¹<https://www.w3.org/TR/wsdl>

¹²<https://www.w3.org/TR/soap/>

¹³Dogac2004

¹⁴<https://www.w3.org/2001/sw/wiki/REST>

A exposição de dados ocultos é outro grande problema relacionado à integração da informação na Internet. Ao passo em que é cada vez maior o número de aplicações fazendo o uso de bancos de dados relacionais, o mapeamento semântico pode ser considerado como um instrumento para a relação e extração de informação relevante a partir de diferentes fontes de dados, tornando mais clara a importância em revelar dados relacionais como RDF ou Linked-Data (Hellmann *et al.*, 2009). Ferramentas já conhecidas como o VirtuosoRDF¹⁵, D2RQ¹⁶, entre outras, estão aptas para uso com a capacidade de gerar representações RDF que derivam diretamente de acordos implícitos e explícitos dos bancos de dados (BD) relacionais (Hellmann *et al.*, 2009), permitindo assim o acesso à informação baseado em ontologias.

O paradigma para a integração de dados provenientes de fontes distintas é o OBDA (*Ontology-based data access*), que baseia o acesso à informação por meio de ontologias sobre um mesmo domínio de conhecimento de interesse (Bienvenu *et al.*, 2013; Civili *et al.*, 2013; Lembo *et al.*, 2014). Essa capacidade se dá a partir de uma arquitetura de três níveis constituída pela ontologia, fontes de dados e mapeamento entre ambas (Civili *et al.*, 2013; Lembo *et al.*, 2014). Dessa forma é possível extrair informações de uma fonte de dados a partir de consultas que usam ontologias (Bienvenu *et al.*, 2013). Nesse contexto, um dos primeiros representantes desse modelo é o D2RQ (Calvanese *et al.*, 2016) que realiza o mapeamento do banco de dados em estruturas semanticamente anotadas que permitem o acesso à informação por meio dessas ontologias.

O OBDA pode ser usado para enriquecer o conhecimento em uma fonte de dados incompleta a ponto de trazer um conjunto mais completo de respostas para uma consulta aproveitando as capacidades de inferências por meio do raciocínio lógico dos bancos de dados semânticos (Bienvenu *et al.*, 2013; Calvanese *et al.*, 2016). Suponha uma fonte de dados que descreva um paciente com um Linfoma e outro paciente com Melanoma. Uma pesquisa sobre essa base de dados apenas de posse apenas dessas informações impede que seja pesquisado de forma satisfatória todos os pacientes que possuem algum tipo de câncer. Por outro lado, podemos enriquecer esses dados a partir do uso de ontologias que classificam ambos como tipos de câncer. Agora é possível extrair um resultado com um conjunto mais completo de informações úteis para um hospital ou um médico.

Sistemas como o Ontop (Calvanese *et al.*, 2016; Rodríguez-Muro *et al.*, 2013) e o o MastroStudio (Civili *et al.*, 2013) são exemplos de implementações ODBA e que expõem bancos de dados relacionais como grafos RDF virtuais a partir de mapeamentos entre ontologias e o banco de dados. O grafo RDF pode ser então consultado usando SPARQL, já que o sistema traduz a consulta para a linguagem SQL de maneira transparente ao usuário, de maneira tal que ele não precisa ter conhecimento sobre os bancos de dados nem sobre o relacionamento entre eles.

2.5 Síntese

A proposta de integração de informação dos portais na Web almejada neste trabalho utiliza tecnologias apresentadas neste capítulo e envolve técnicas de recuperação de dados e mapeamento semântico para a integração dos dados espalhados na Web. Ao passo em que as fontes de conteúdo são escolhidas, a recuperação da informação nesses portais acontece por meio de rotinas de extração de dados que, por sua vez, são desenvolvidas manualmente e baseadas na análise dos padrões de repetição dos assuntos relacionados ao domínio escolhido. Desse modo, informações como o nome, endereço ou outros dados relacionados a um determinado evento ou restaurante são recuperadas de maneira automatizada de acordo com regras pré-estabelecidas por meio de expressões capazes de interpretar a estrutura sintática do documento HTML dos diferentes portais e recuperar conteúdo relevante.

O processo de análise da estrutura dos documentos HTML expostos pelos diferentes portais escolhidos em busca de padrões de repetição de informação é realizado manualmente. A interpretação das fontes de dados escolhidas, bem como a transferência da inteligência para a expressão regular, são de responsabilidade do indivíduo que aplica o modelo proposto neste trabalho. Isso também se

¹⁵<http://virtuoso.openlinksw.com/>

¹⁶<http://d2rq.org/>

aplica no que diz respeito a transferência da informação obtida através da recuperação dos dados aplicada em uma ontologia de maneira a obter dados anotados semanticamente.

Uma vez que os dados são obtidos, uma camada de informação semântica é adicionada utilizando as ontologias descritas no portal Schema.org. Nesse momento, a informação está descrita como triplas RDF que são armazenadas em um banco de dados semântico. As triplas, ao serem persistidas no banco de dados, se integram com os dados que foram obtidos de diferentes portais da Web e são enriquecidas a partir do processo realizado pelo motor de inferência. Nesse momento, o acesso à informação por meio de ontologias pode ser realizado utilizando a linguagem de consulta SPARQL. A informação de diferentes fontes de dados pode então ser recuperada sobre um ponto de vista amplo e centralizado.

No capítulo a seguir, a proposta da arquitetura para integração de dados de portais na Web que utiliza tais tecnologias é detalhada.

Capítulo 3

Arquitetura proposta

Como visto no capítulo 2, a proposição de uma arquitetura visando à integração de dados na Internet não é um processo trivial e demanda o conhecimento em diferentes áreas de pesquisa tais como a Web Semântica e a Integração de Dados.

Um dos principais pontos para a integração da informação na Internet é a definição de uma estrutura que represente com clareza um determinado domínio de conhecimento, como explicado por May (2002) e Ahmed (2008). Além disso é preciso estabelecer mecanismos capazes de interpretar e identificar os dados relevantes sobre os documentos de dados expostos na Web e convertê-los sobre a semântica comum estabelecida anteriormente. É preciso também considerar o acesso à informação pois, uma vez que os dados estão centralizados, eles podem (e serão) consultados por diferentes usuários interessados na informação ali contida. Nesse contexto, uma arquitetura capaz de promover a interoperabilidade de informações na Internet deve conter ao menos uma camada de recuperação de dados, uma camada de persistência e centralização da informação e uma camada de apresentação para o acesso à informação, conforme ilustrado pela Figura 3.1.

Figura 3.1: Proposta de arquitetura de sistema de integração de dados da Web

Cada uma destas camadas, bem como os diversos processo envolvidos, são descritos a seguir.

3.1 Escolha do domínio de dados

Do ponto de vista pragmático, uma primeira decisão a ser tomada num projeto do tipo proposto neste trabalho é delimitar o escopo de dados em um domínio previamente estabelecido. Além disto, devem-se determinar as fontes de dados que serão utilizadas.

Neste trabalho os domínios de informação escolhidos são eventos e restaurantes. São vários os portais que divulgam exposições, peças de teatros, palestras, aulas, entre outros. No entanto, uma informação sobre um determinado evento está distribuída nessas diferentes fontes de dados, o que dificulta uma visão centralizada sobre os eventos que acontecem em determinada data, horário ou região. Nesse contexto, a existência de um sistema capaz de responder esses questionamentos de maneira centralizada pode ser importante para alcançar uma visão mais ampla e auxiliar a tomada de decisão do usuário final. Neste trabalho, espera-se responder a um usuário, de maneira centralizada, os eventos de uma determinada região, os eventos que acontecem em um determinado dia e os eventos de um determinado tipo como, por exemplo, os diferentes eventos que acontecem perto e uma determinada referência de geolocalização.

Quanto às fontes de dados utilizadas, neste trabalho adotam-se os portais Guia Da Semana¹, Guia da Folha² e o portal de Eventos da USP³ conforme apresentado na Figura 3.2. Os dois primeiros portais foram escolhidos em razão de serem amplamente conhecidos e utilizados, além de possuírem uma ampla base de dados sobre eventos e restaurantes. Já o portal da USP foi escolhido por manter uma grande quantidade de eventos acadêmicos dentro do contexto da Universidade de São Paulo.

Figura 3.2: Fontes de dados sobre eventos e restaurantes

3.2 Definição de uma ontologia

A fim de integrar fontes provenientes de diferentes portais de eventos dos quais não temos domínio sobre o conteúdo exposto, é necessário adotar uma estrutura comum previamente definida, i.e. uma ontologia, bem como construir um modelo de conversão de dados de cada uma destas fontes para esta estrutura comum. Neste trabalho, o processo de extração, transformação e persistência dos dados é ilustrado na Figura 3.3. Em primeiro lugar, os dados são recuperados dos portais de eventos escolhidos a partir de técnicas de recuperação de informação usando expressões regulares.

¹<https://www.guiadasemana.com.br/>

²<http://guia.folha.uol.com.br/>

³<http://www.eventos.usp.br/>

O conteúdo selecionado é então mesclado com uma ontologia, gerando assim uma informação semanticamente anotada. Por fim, a informação é então armazenada em um repositório de dados para acesso baseado em ontologias. Como resultado deste processo, obtém-se um repositório de dados semanticamente anotado que pode ser consultado de maneira centralizada.

Conforme já relatado na seção 2.1.2 do Capítulo 2, o portal Schema.org (Mika, 2015) apresenta definições de conceitos padronizadas e amplamente revisados, inclusive por esforços de iniciativa privada como no caso de Google e Microsoft.

Figura 3.3: *Processo de extração e persistência de dados*

Neste trabalho serão utilizados os conceitos “Estabelecimento de alimentos” (FoodEstablishment), “Evento” (Event) e suas respectivas derivações, como apresentado na hierarquia de conceitos da Figura 3.4. Nesta figura, todos os conceitos são derivados do conceito abstrato “Thing”. Definem-se conceitos tais como exposições (“ExhibitionEvent”), sessões de cinema (“ScreeningEvent”), e peças de teatro (“TheaterEvent”), entre outros. Herda-se também as declarações de propriedades para cada classificação, i.e, o título, latitude, longitude, descrição, data e horário do evento, além de especificações mais restritas aos níveis mais inferiores da hierarquia; por exemplo, o tipo de comida, que só será válido para “Restaurant”, para classificar se o restaurante oferece comida italiana, mexicana ou japonesa.

Figura 3.4: *Hierarquia de conceitos*

3.3 Preparação e recuperação de dados

Uma vez definido o domínio de dados, as fontes de informação e as ontologias, deve-se então identificar e extrair a informação relevante nos portais escolhidos. Na proposta deste trabalho, a camada de recuperação de dados contém toda a inteligência para a extração de informação relevante nos portais selecionados. Em outras palavras, é o módulo responsável pela mediação da requisição para uma fonte de dados na Internet e a consequente transformação daquele conteúdo escolhido para um documento RDF válido conforme representado pelas setas 1 e 4 na Figura 3.1. Para tal, é necessário criar um conjunto de regras baseadas em expressões regulares específicas a cada fonte de dados escolhida, transformando assim em tempo real as informações relevantes em dados semanticamente anotados a partir de triplas RDF coerentes que podem ser entendidas por máquinas.

O reconhecimento do conteúdo significativo nos diferentes portais não é uma tarefa trivial. Para cada fonte de dados, é necessário analisar e identificar os padrões de repetição de dados apresentados dentro de cada domínio de interesse. Embora cada portal da Web utilize em comum o protocolo HTML para expor a informação que é dinamicamente produzida, a estrutura desses documentos é geralmente muito diferente. Tal característica é esperada, uma vez que são diferentes profissionais construindo modelos de apresentação específicos distanciando-se assim de padronizações. Desse modo, o mapeamento entre onde estará um determinado conteúdo não será o mesmo para dois portais distintos; por outro lado, uma vez identificado um padrão para uma fonte de dados, este pode ser reaproveitado para todas as outras páginas do mesmo portal.

Figura 3.5: Extração de latitude e longitude em diferentes portais

Os exemplos contidos nas Figuras 3.5 e 3.6 ilustram com maior clareza essa condição. Neles é possível observar que a apresentação do conteúdo nessas páginas é diferente e isso ocorre não só por causa dos dados ali publicados, mas também porque a estrutura do documento que é interpretado pelo navegador difere em cada portal. Tomando como exemplo a Figura 3.5 vemos que a captura dos dados de latitude e longitude é diferente para cada portal. Apesar disso, esse padrão se repete para todo conteúdo específico daquele portal, como é possível notar através da Figura 3.6. Nesse exemplo a captura do título de um evento sempre está delimitada por marcações específicas no documento HTML (`<h1>`). Sendo assim, embora seja uma tarefa simples para os humanos identificar e interpretar a informação de coordenadas geográficas da latitude e longitude dos exemplos propostos, isto não ocorre para um processamento automático.

Neste trabalho, foram elaborados mecanismos com capacidade de interpretar e transformar o conteúdo de cada um dos portais escolhidos em informações com semântica agregada. O primeiro passo para a sua construção é a análise da estrutura do documento HTML produzido dinamicamente

1) Identificação da repetição de uma propriedade na estrutura do documento HTML

```

▼ <div itemprop="address" itemscope itemtype="http://schema.org/PostalAddress">
  ▼ <address>
    <span itemprop="streetAddress">Praça do Capivari, s/n </span>
    <span itemprop="addressLocality"></span>
    <span itemprop="postalCode">09111-100</span>
    <meta itemprop="addressRegion" content>
    <meta itemprop="addressCountry" content="Brasil">
  </address>
</div>

```

2) Definição de expressão regular capaz de extrair o conteúdo

```
<span itemprop="streetAddress">(.*?)</span>
```

3) Resultado da aplicação da expressão regular sobre o conteúdo

Praça do Capivari, s/n

Figura 3.7: Processo de extração de um endereço no portal Guia da Semana

Figura 3.8: Centralização e ponto de acesso a dados

nada em diferentes servidores distribuídos na Web, um mecanismo de consultas e inferência sobre toda essa informação deve levar em consideração essa característica. Entre os diferentes problemas que a pesquisa da informação na rede enfrenta está a velocidade de transferência de dados; esta é mais afetada quanto mais distante estão dois pontos alvos da recuperação da informação. Há também o caráter técnico que envolve a própria estrutura da Internet que não garante que um determinado servidor de informações estará acessível em um dado momento, seja por problemas da própria rede ou outros quaisquer. Além disso, ainda há o processo de inferência que demanda a recuperação de todo o conteúdo das diferentes bases de dados de todas as organizações escolhidas. Esta opção se justifica pois uma simples busca feita por um usuário em um modelo descentralizado pode não ser atendida em tempo hábil, uma vez que toda a informação contida nos diferentes portais escolhidos precisará ser obtida, armazenada e inferida a cada consulta. Neste trabalho, foi utilizado portanto um modelo centralizado de dados.

O modelo centralizado proposto neste trabalho se baseia em um banco de dados único construído a partir das informações obtidas diretamente das fontes de informações pré-estabelecidas.

Ele é o resultado de um processo constante, realizado diariamente, da aplicação dos mecanismos de interpretação e recuperação discutidos anteriormente. Desse modo as informações sobre os eventos e restaurantes são adicionadas ou atualizadas de forma contínua dentro do banco de dados semântico o que garante que alterações no conteúdo exposto pelos portais escolhidos sejam refletidas no banco de dados centralizado.

O acesso aos dados persistidos é outro fator importante e está representado pelas setas 6 e 7 da proposta de arquitetura ilustrada na Figura 3.1. Para alcançar essa capacidade, a arquitetura proposta neste trabalho disponibiliza serviços para consultas SPARQL que são feitas diretamente sobre a base de informações elaborada e possibilitam consultas baseadas em ontologias. Além disso, também há serviços que executam a inferência sobre os dados contidos na base de dados. Desse modo, a inferência sobre os dados existentes pode ser realizada a qualquer momento.

3.5 Resolução de duplicações e conflitos

A resolução de duplicações e conflitos é um fator importante no que diz respeito à centralização de dados. Tomando como base o contexto de publicadores de eventos na Internet, embora o conteúdo existente em cada um desses portais seja geralmente complementar, isso é, cada site focado em uma região ou tipo de informação específica, pode ocorrer que a mesma informação esteja duplicada em diferentes portais ao mesmo tempo, conforme ilustra o exemplo da Figura 3.9. Na figura, a informação sobre a peça de teatro “Lés Misérables” aparece ao mesmo tempo em portais distintos.

Figura 3.9: *Conflito de informações em diferentes portais*

Há diferentes formas para a identificação de dados similares. No entanto, a escolha de uma forma em detrimento de outra pode ser pautada pelo tipo de dado que está sendo tratado. No caso dos eventos e restaurantes tratados neste trabalho, o título de uma peça de teatro, o nome de um filme no cinema, o nome de um restaurante e até mesmo o título de uma exposição é sempre fixo, mesmo que exposto em diferentes portais. Exemplo disso é a peça “Les Misérables”, que é discriminada da mesma maneira nos diferentes portais em que esta publicada. Essa característica dos eventos foi explorada neste trabalho para identificar que se trata do mesmo evento.

Além de duplicações, pode ocorrer conflito de informação, quando duas fontes distintas fornecem informações diferentes sobre um mesmo evento. No domínio tratado neste trabalho, um exemplo de tal conflito poderia ser o fato de dois portais fornecerem informações distintas a respeito do horário de início do mesmo evento.

Além das responsabilidades de centralizar a informação, realizar inferências e expor um ponto de acesso aos dados, a camada de persistência ainda dispõe de inteligência para a resolução de conflitos. Essa característica é definida a partir de um módulo especializado capaz de decidir, a partir de comparações simples, se um determinado dado deve prevalecer sobre outro semelhante, evitando assim a inconsistência da informação. Nele, o processo de decisão emerge a partir de uma sequência de regras bem definidas.

A resolução de conflitos neste trabalho se dá através de um processo de avaliação baseado em regras pré-definidas, criadas a partir do domínio dos dados escolhido, conforme ilustrado na Figura 3.10. O primeiro passo da resolução de conflitos é a comparação das informações, através de valores da tripla RDF no qual uma consulta SPARQL é realizada diretamente na base de dados, utilizando como parâmetros de filtragem os dados do sujeito e do predicado da tripla RDF. Quando a consulta não retorna um resultado, a informação pode ser incluída na base de dados, já que se trata de uma nova informação; entretanto, quando dados são retornados, inicia-se uma segunda fase de validações.

Figura 3.10: *Processo de resolução de conflitos*

A segunda fase de validação envolve a proveniência da informação. Isso significa que alguns dados são escolhidos em detrimento de outros, baseado em uma ordem de confiabilidade das fontes de informação. Desse modo, a decisão entre duas informações contidas em triplas que possuem objetos diferentes sendo uma extraída de um site A e de outro site B dependem da ordem de precedência pré-estabelecida. Supondo que se estabeleça que a informação do portal B é mais relevante que a do portal A pelo fato deste último ter conteúdo mais preciso, possui menor desvio gramatical, maior abrangência territorial ou qualquer outro, então a informação de A será sempre substituída pela informação de B. Por fim, se a informação veio do próprio site e portanto não há ordem de preferência da informação, então os dados são substituídos e tratados como uma atualização da informação.

A fim de ilustrar o processo de resolução de conflitos, vamos tomar como base a peça “A Era do Rock” que existe tanto no portal Guia da Semana quanto no portal Guia da Folha. Neste trabalho, as informações provenientes do portal Guia da Folha foram definidas como prioritárias em relação às informações do Guia da Semana em virtude da abrangência de informação e da quantidade de revisores desse portal, que é maior que a do segundo. O primeiro passo é recuperar as informações do portal com prioridade menor, nesse caso o Guia da Semana. Como as triplas RDF geradas a partir da análise do conteúdo desse evento ainda não estão registradas, toda a informação descoberta é prontamente adicionada no banco de dados. A partir desse momento, a informação passa a ser extraída do portal Guia da Folha. Todas as propriedades que não forem encontradas no banco de dados são prontamente adicionadas na base de dados. Esse processo determina a mesclagem da informação dos diferentes portais sobre o mesmo evento. Por outro lado, as propriedades que já existem tais como título, data ou endereço, são prontamente atualizadas já que o portal Guia da Folha tem prioridade em relação ao Guia da Semana.

Como a informação na Internet pode sofrer alterações, há um processo de atualização que realiza novamente a recuperação dos dados dentro de um período pré-definido. Levando em consideração o exemplo anterior, quando a informação é novamente recuperada do portal Guia da Semana, caso uma nova propriedade que ainda não exista no banco de dados seja detectada, ela será prontamente adicionada. Se a propriedade já existe, então a informação extraída do portal será descartada uma

vez que sua prioridade de persistência é menor. Por outro lado, quando a informação é recuperada do portal Guia da Folha e caso a propriedade não exista, ela é prontamente adicionada; quando ela já existe, ela é atualizada. O ciclo de recuperação de informação é então mantido por tempo pré-definido.

A geolocalização é um caso especial do processo de substituição de dados. Diferentemente das outras propriedades do conceito, a geolocalização não pode simplesmente substituir um valor pré-existente de um Restaurante, por exemplo. Um exemplo típico são franquias, onde a informação de latitude e longitude de um estabelecimento específico não pode simplesmente sobrescrever os dados pré-existentes. Nesses casos, a informação sobre a latitude e longitude só é atualizada caso se encontre num raio de 5m da posição original. Caso contrário, uma nova informação é adicionada na base de dados.

3.6 Regras para inferências sobre dados

A proposta deste trabalho busca realizar a integração de dados disponibilizados na Internet a partir da centralização do conteúdo exposto em diferentes portais da Internet, adicionando uma camada semântica que permite realizar a busca de dados mais ricos no contexto de eventos culturais que acontecem na cidade de São Paulo. Nesse cenário, saber se um determinado lugar, como por exemplo um teatro ou uma exposição, é próximo a outro como um restaurante ou um pub pode ser interessante para que o usuário final possa programar sua agenda de acordo com a proximidade entre esses dois pontos. Uma das formas de se alcançar esse conhecimento é verificar se dois pontos são suficientemente próximos. Essa forma é custosa, pois exige um grande esforço manual por parte do indivíduo interessado nessa consulta. Outra forma mais eficiente é a definição de um conceito que explique o que é estar “próximo” e depois, por inferência sobre os dados, alcançar esse objetivo a partir de uma consulta baseada nesse significado. O conceito “perto” foi definido neste trabalho para atingir tal finalidade.

A regra “perto de um lugar” possibilita a busca de pontos de geolocalização que estão próximos. Essa condição é calculada a partir dos resultados de uma função que delimita a área na qual os estabelecimentos devem estar para serem considerados próximos e é ilustrada pela Figura 3.11. No entanto, para que funcione corretamente, é preciso definir qual é a distância que deve ser considerada para poder inferir que dois locais estão localizados perto um do outro. Neste trabalho, a distância de quinhentos metros foi escolhida por tratar-se de um espaço curto que pode ser percorrido em poucos minutos. Assim, caso dois pontos “a” e “b” que possuem respectivamente as latitudes e longitudes (x1,y1) e (x2,y2) estiverem distantes em um raio de até quinhentos metros, então uma nova tripla “?a iweb:near ?b” é adicionada na base de dados sugerindo que “a” está perto de “b”. A segunda regra da figura implementa o aspecto transitivo desta relação.

```

1 @prefix schema: <http://schema.org/> .
2 @prefix iWeb: <http://integraWeb.ddns.net/> .
3
4 [rule1: (?a schema:latitude ?y1), (?a schema:longitude ?x1), (?b schema:latitude
      ?y2), (?b schema:longitude ?x2), near(?x1, ?y1, ?x2, ?y2) -> (?a iweb:near ?
      b)]
5 [rule2: (?a iweb:near ?b)-> (?b iweb:near ?a)]

```

Figura 3.11: Regra que define o conceito perto (near)

3.7 Camada de apresentação

A última camada da arquitetura proposta neste trabalho é a camada de apresentação. É nesta camada que se espera a interação entre o usuário e a fonte dos dados integrados a partir de uma página Web. Assim, a partir de consultas SPARQL, o usuário consegue obter os dados de interesse

de forma centralizada, sem a necessidade de consultar cada uma dos outros portais. Essa condição está representada pelas setas 5 e 8 na Figura 3.1, que representam um usuário interagindo com um sistema que implementa a arquitetura apresentada neste trabalho.

A camada de apresentação simplifica a interação do usuário leigo com o sistema, ao gerar consultas utilizando a linguagem SPARQL de maneira transparente sem a necessidade de conhecimento prévio na linguagem. Essa condição se dá através de campos de pesquisa pré-definidos como, por exemplo, o tipo do evento ou do estabelecimento de interesse do usuário. No entanto, o sistema não se limita a essas condições e possibilita ao usuário mais avançado realizar uma consulta SPARQL diretamente no sistema, permitindo assim pesquisas mais elaboradas sobre o banco de dados.

O resultado da consulta é desenhado em cima de uma mapa, que contém informações sobre ruas e estabelecimentos na região dos eventos e estabelecimentos encontrados. Um exemplo de tal saída é apresentado na Figura 3.12.

Figura 3.12: Exemplo de apresentação de resultado de consulta

Tal interface facilita a interação do usuário, pois atualmente são muitos os sistemas baseados em ferramentas que se valem de mapas, como o aplicativo Google Maps⁴, Uber, TripAdvisor, entre outros. Além disso, tal modo de apresentação auxilia no entendimento da informação do evento e sua distância em relação a outros pontos de referência identificados pelo usuário no próprio mapa.

No capítulo a seguir, apresentam-se os experimentos realizados com esta arquitetura.

⁴<http://maps.google.com/>

Capítulo 4

Experimentos e Resultados

Neste capítulo, será descrita a implementação da proposta da arquitetura e apresentada uma série de experimentos realizados para avaliá-la. Como domínio de aplicação, utilizou-se a integração de eventos e estabelecimentos de alimentação na região de São Paulo. O primeiro experimento indica que a integração dos dados provenientes de diferentes portais da Internet é possível e independe do acesso ao banco de dados das organizações das quais o conteúdo foi extraído. O segundo experimento mostra que o conteúdo sobre um mesmo evento pode ser construído pela junção das informações extraídas de diferentes fontes de informação. O terceiro experimento ilustra que é possível para um usuário obter todas as informações sobre restaurantes de um determinado tipo de comida, ou sobre eventos específicos que acontecem próximo a ele. Por fim, o último experimento mostra que é possível realizar consultas mais avançadas a partir da combinação de diferentes interesses como, por exemplo, restaurantes italianos perto de um determinado evento que acontecerá em determinado dia ou bares que estão perto de estações de metrô.

4.1 Implementação

A arquitetura de integração de dados, implementada conforme a proposta apresentada na Figura 3.1, contempla três diferentes camadas: a camada de recuperação de dados, a camada de persistência e acesso aos dados e a camada de apresentação. A implementação realizada é modular, sendo o funcionamento de cada camada independente das outras, dividindo as responsabilidades de cada uma delas em sistemas distintos. Essa característica torna o sistema mais flexível e permite uma alteração menos custosa para a adaptação de novas tecnologias, estratégias ou necessidades específicas, já que permite a troca de componentes por partes bem delimitadas sem a obrigação de reescrever o projeto por completo.

A camada de recuperação de dados é o módulo do sistema que é responsável por extrair os dados dos portais de conteúdo e alimentar a camada de persistência. Para tanto, ela está implementada na linguagem de programação Java e utiliza de forma exaustiva a biblioteca de expressões regulares da própria linguagem. Além disso, é constituída de diversos micro-sistemas, um para cada portal específico, que possuem a inteligência da extração de informação relevante de cada fonte de conteúdo escolhido. Em outras palavras, cada processo de extração de dados é um sistema totalmente novo que agrega apenas o conhecimento de interpretação de dados e conversão da informação nos conceitos de ontologias do Schema.org específicos para aquele portal. É nessa camada que os dados são transformados em triplas RDF utilizando a API RDF¹ da biblioteca Jena. Assim, uma vez que as triplas são formadas, elas são prontamente enviadas para a camada de persistência dos dados. O processo de recuperação de dados, por sua vez, acontece de forma contínua. Dessa forma, uma vez que a recuperação da informação de uma determinada fonte de dados é finalizada, um novo processo de recuperação de dados se inicia naquele mesmo portal criando assim um ciclo de obtenção e atualização de dados.

¹<https://jena.apache.org/documentation/rdf/index.html>

A camada de persistência é outro sistema independente na implementação da proposta sugerida neste trabalho. Ela foi construída com base no componente TDB² do portfólio Jena, um arcabouço mantido pela organização sem fins lucrativos Apache Software Foundation que é capaz de armazenar, consultar e realizar inferências sobre dados anotados semanticamente. Trata-se de uma biblioteca gratuita, robusta, bastante conhecida e amplamente utilizada em trabalhos científicos. Desse modo, todo conteúdo obtido na preparação de dados é convertido no formato de triplas RDF sob um identificador único e armazenado em um sistema que utiliza essa biblioteca seguindo o processo ilustrado na Figura 4.1. Nesse processo, o título do evento é convertido em um código hash e utilizado como sujeito da Tripla RDF; o tipo da informação é então interpretado juntamente com o valor da informação de acordo com as regras de análise da estrutura do documento HTML.

Figura 4.1: Processo de extração e conversão em triplas RDF

A biblioteca Jena permite armazenar e consultar triplas em RDF. No entanto uma propriedade torna esta biblioteca mais especial: a capacidade de realizar inferências sobre os próprios dados. Essa condição permite a realização de consultas que percorrem a estrutura de conceitos relacionada a um dado conceito alvo que oferece um resultado final mais rico, uma vez que inferem-se informações que antes não existiriam sem esse processo. A biblioteca TDB vai além, permitindo também a adição de novas regras que possam tornar o modelo ainda mais completo.

A capacidade de inferência da biblioteca Jena se dá por meio dos motores de inferência. Ela conta com motores pré-definidos que incluem um raciocinador transitivo que implementa apenas as propriedades transitivas e reflexivas da ontologia, um raciocinador RDF que implementa um subconjunto de implicações do RDF, um raciocinador OWL, OWL Mini, OWL Micro Reasoners que implementa um subconjunto de implicações do OWL/Lite e OWL/FULL e o raciocinador genérico baseado em regras definidas pelo usuário. Neste trabalho foram utilizados os raciocinadores RDF e Genérico para derivar informações a partir dos dados recuperados nos portais selecionados.

A camada de apresentação é o último sistema implementado neste trabalho e apresenta uma

²<https://jena.apache.org/documentation/tdb/>

interface de pesquisa e exposição de informação que facilita a interação do usuário com a camada de persistência de dados, através da qual ele irá extrair informações relevantes. Esse objetivo foi alcançado por meio de uma área na qual as consultas são realizadas e os resultados são colocados em um mapa, conforme ilustrado pelo exemplo de uma consulta das estações de metrô em São Paulo ilustrado pela Figura 4.2. Desse modo o sistema alterna entre consulta e apresentação de dados.

A consulta de dados no sistema acontece pela interação do usuário com as opções de consultas disponíveis na área 1 ilustrada na Figura 4.3. Através dela é possível realizar diversos tipos de consultas, tanto as voltadas para usuários que não possuem domínio da linguagem SPARQL e que podem utilizar as pesquisas pré-definidas por tipos de eventos, i.e, eventos em teatros, exposições, restaurantes entre outros quinze diferentes tipos de consultas pré-definidas utilizando a seleção ilustrada na área 2 ou, para usuários mais experientes, inserir a consulta SPARQL diretamente utilizando o campo de texto logo abaixo dessa área. Enquanto no primeiro caso a consulta por tipo de estabelecimento gera uma consulta SPARQL e depois é enviada para a camada de persistência, na segunda opção a consulta elaborada pelo usuário pula essa etapa e vai diretamente para a camada de persistência de dados. O resultado encontrado é então desenhado no mapa. Por fim, toda a consulta construída pode ser recuperada a qualquer momento clicando no botão “Q” (apresentado na área 3, no canto superior direito da tela). Ao clicar nesse botão, a consulta SPARQL que foi gerada no sistema é mostrada conforme apresentado na área 4 da interface.

Figura 4.2: Estações de metrô em São Paulo

O resultado obtido a partir de uma interação de um usuário com a camada de apresentação é representado como um conjunto de pontos no mapa, e também como uma listagem no canto esquerdo da tela conforme ilustrado na Figura 4.4 nas áreas 1 e 2. Ao clicar em um desses pontos, é possível buscar mais detalhes sobre aquele estabelecimento específico. Além disso, cada item possui uma marcação específica indicando a proveniência da informação, isso é, de qual portal entre os escolhidos ela foi obtida conforme ilustra a área 3; ao clicá-los, é possível obter mais informações tais como data e hora, endereço, telefone, preço, link do site do qual a informação foi recuperada, entre outros. Por fim, é possível obter a consulta SPARQL gerada pelo sistema ao clicar no botão “Q” localizado na área 4.

Por fim, a implementação da arquitetura proposta está disponível ao público na Internet no endereço “<http://integraweb.ddns.net>”. Nesse portal, é possível interagir com o sistema a partir da base de dados que foi utilizada para a avaliação dos resultados neste trabalho, utilizando as consultas exemplos ou criando novas. Além disso, todo o código-fonte produzido também está exposto e pode ser acessado, analisado e carregado através do portal GitHub no endereço “<https://github.com/fpierin/mantichub>”. A partir do conteúdo do repositório, é possível propor modificações na estrutura do projeto ou tomá-lo como base para novas implementações.

Figura 4.3: Camada de apresentação da proposta

Figura 4.4: Interface de apresentação do resultado da consulta

4.2 Experimento 1: Integração de dados

Um primeiro experimento foi realizado de modo a ilustrar a integração de dados provenientes de fontes heterogêneas de informação. A Figura 4.5 mostra todos os eventos recuperados do portal Eventos USP. Nessa figura, é possível observar que esses eventos estão concentrados em regiões próximas a um campus da USP, como na região do Butantã, São Carlos ou Ribeirão Preto. Na Figura 4.6 vemos que os eventos recuperados pelo portal Guia Da Semana estão aglomerados na região central da cidade de São Paulo e abrangem em sua maioria peças de teatro, exposições ou shows. Já o Guia da Folha possui uma base de conteúdo mais diversificado e cobre praticamente toda a cidade de São Paulo, com informações sobre restaurantes, peças de teatro, exposições e bares

como é possível observar na Figura 4.7.

Como neste trabalho todo o conteúdo dos portais citados foi recuperado e centralizado em um única base de dados, é possível realizar consultas sem que o usuário tenha de consultar cada um dos portais individualmente, minimizando o seu trabalho manual de busca e provendo resultados mais abrangentes.

Figura 4.5: Concentração de dados do portal Eventos USP

Figura 4.6: Concentração de dados do portal Guia da Semana

A busca de informações sobre eventos que ocorrem em uma determinada região da cidade tende a ser relevante para um indivíduo. Desse modo, para alcançar resultados sobre uma determinada região, é necessário saber os valores de latitude e longitude de um determinado ponto. Tomando como base o marco zero da cidade de São Paulo, situado na região da Praça da Sé, podem-se recuperar informações sobre eventos que ocorre num raio pré-definido com centro nessa região, por exemplo a 1km. Desse modo, utilizando a consulta SPARQL apresentada no anexo A alcançamos o resultado apresentado na Figura 4.8, que mostra restaurantes recuperados do portal Guia da Folha, eventos culturais dos portais Eventos USP e Guia da Semana e também algumas estações de Metrô que foram adicionadas à mesma base de dados.

Neste exemplo, não ocorre duplicação ou conflito na integração dos dados. Em outras palavras, a informação sobre eventos e estabelecimentos que foi recuperada em um portal é complementar ao outro. A Figura 4.9 apresenta uma simplificação da consulta apresentada no Anexo A. Essa

Figura 4.7: Concentração de dados do portal Guia da Folha

Figura 4.8: Concentração de eventos próximos ao Marco Zero de São Paulo

consulta retorna apenas o identificador do evento, o seu título e a origem dos dados. O resultado da consulta encontra-se representado na Figura 4.10. Nesta figura, destacam-se os dados que são armazenados em cada uma das consultas aos diferentes portais.

Supondo que a recuperação dos dados aconteça primeiro no portal de Eventos da USP, a consulta retorna apenas dois eventos: a “Comemoração da Faculdade de Direito” e a “Exposição no Centro de Preservação Cultural”. Se o sistema então obtiver os dados do portal Guia da Semana, novos eventos são adicionados, entre eles o musical “Lès Miserables”. Finalmente, após o sistema recuperar o conteúdo do Portal Guia da Folha, outros eventos são acrescentados como, por exemplo, o evento “A Experiência Munduruku”. Nesse contexto, a implementação da arquitetura conseguiu integrar os dados dos diferentes portais em uma única base de informações.

4.3 Experimento 2: Resolução de conflitos

A resolução de conflitos é uma tarefa essencial quando estamos lidando com informações provenientes de diferentes fontes de dados. Como conteúdos podem estar registrados de formas distintas nas diferentes fontes de informação, conflitos de informação podem ocorrer, e o contexto de eventos culturais não foge dessa situação. Por exemplo, uma peça que acontece em um determinado teatro pode ter informações sobre preços, localização ou mesmo uma descrição totalmente diferentes

```

1 PREFIX rdf:<http://www.w3.org/1999/02/22-rdf-syntax-ns#>
2 PREFIX rdfs:<http://www.w3.org/2000/01/rdf-schema#>
3 PREFIX schema:<http://schema.org/>
4 SELECT DISTINCT ?s ?title ?url
5 WHERE {
6   ?s schema:title ?titleObj ;
7   schema:latitude ?latitudeObj ;
8   schema:longitude ?longitudeObj ;
9   rdf:type ?typeObj .
10  ?typeObj rdfs:subClassOf ?subClassObj .
11  OPTIONAL { ?s schema:serviceURL ?urlObj }
12  values ?subClassObj { schema:Event schema:FoodEstablishment schema:
13    CivicStructure }
14  BIND (str(?titleObj) as ?title)
15  BIND (str(?latitudeObj) as ?latitude)
16  BIND (str(?longitudeObj) as ?longitude)
17  BIND (str(?urlObj) as ?url)
18  FILTER (?latitude > '-23.539606783940812' && ?latitude < '-23.55759321605919')
19  FILTER (?longitude > '-46.62938980581791' && ?longitude < '-46.6490101941821')
20  FILTER NOT EXISTS {
21    ?s rdf:type ?subtype .
22    ?subtype rdfs:subClassOf* ?typeObj .
23    filter ( ?subtype != ?typeObj )
24  }
25 LIMIT 100

```

Figura 4.9: Consulta SPARQL para recuperar eventos próximos ao Marco Zero de São Paulo

1. Dados armazenados após consulta no portal Eventos USP

Identificador	Fonte de dados	Evento
#70e89905b0dc7b1fe8e52dccf71dab3f	Eventos USP	Biblioteca da Faculdade de Direito da USP comemora 191 anos com e
#c439a85ef2c748159972cf0b3c4a11bd	Eventos USP	Centro de Preservação Cultural promove exposição de 130 anos de D

2. Dados armazenados após consulta no portal Guia da Semana

Identificador	Fonte de dados	Evento
#70e89905b0dc7b1fe8e52dccf71dab3f	Eventos USP	Biblioteca da Faculdade de Direito da USP comemora 191 anos com e
#c439a85ef2c748159972cf0b3c4a11bd	Eventos USP	Centro de Preservação Cultural promove exposição de 130 anos de D
#3f41e1d2633cf5835e104c69b3aad2d	Guia da Semana	Les Misérables
#1760d11bfdb8e0eb80b9b4a7c4a7b859	Guia da Semana	Hotel Mariana
#1ca6602f016bd13cad4a1693d4fe303c	Guia da Semana	Limonada
#e924c59e790e613bed182566a4b4b4e9	Guia da Semana	Comédia Stand Up

3. Dados armazenados após consulta no portal Guia da Folha

Identificador	Fonte de dados	Evento
#70e89905b0dc7b1fe8e52dccf71dab3f	Eventos USP	Biblioteca da Faculdade de Direito da USP comemora 191 anos com e
#c439a85ef2c748159972cf0b3c4a11bd	Eventos USP	Centro de Preservação Cultural promove exposição de 130 anos de D
#3f41e1d2633cf5835e104c69b3aad2d	Guia da Semana	Les Misérables
#1760d11bfdb8e0eb80b9b4a7c4a7b859	Guia da Semana	Hotel Mariana
#1ca6602f016bd13cad4a1693d4fe303c	Guia da Semana	Limonada
#e924c59e790e613bed182566a4b4b4e9	Guia da Semana	Comédia Stand Up
#e6a037d4791281cbee2e5a3a4a58a81d	Guia da Folha	O Cinema de Jerzy Skolimowski
#e7a6304acfbda361de67ddda7d8e4a4d	Guia da Folha	Experiência Munduruku
#d3a61f4ec868502976f03292b2e5d5e8	Guia da Folha	Grande Festa Junina Vegana 2017
#fd3b91c252b494eb73848bf777bc2547	Guia da Folha	DaTchau - Rumo à Estação GrandeAvenida
#aa7f322e84d538e0bf03d18d12a5499a	Guia da Folha	Enquanto as Crianças Dormem
#d265b2c3309e10c36e337e2e45074103	Guia da Folha	Lembro Todo Dia de Você
#43d63bb94abeb24b4fe3c897638dcdfc	Guia da Folha	Pobre Super-Homem, Avesso do Herói

Figura 4.10: Resultado do armazenamento de dados provenientes de diferentes portais

em fonte de dados distintas. Isso evidencia ser necessário um mecanismo capaz de resolver essas situações.

Figura 4.11: Evento "A Era do Rock" em diferentes portais

Figura 4.12: Geolocalização de evento incorreta no portal Guia da Semana

No modelo proposto neste trabalho, há uma hierarquia de preferência da informação. Desse modo, uma fonte de dados por ser mais confiável que as outras sempre terá preferência em relação a sua informação. Um exemplo desta situação se refere à peça de teatro “A Era do Rock”, que aparece tanto no portal Guia da Folha quanto no portal Guia da Semana, como mostra a Figura 4.11. Embora a informação sobre um mesmo evento seja exibida em ambos os portais de eventos, a extração de dados de uma determinada fonte pode ser mais consistente do que outra por motivos diversos que vão desde a quantidade de revisores de conteúdo, investimento a quantidade de atualizações. Nesse caso específico, a informação sobre geolocalização do evento está incorreta no Guia da Semana e correta no Guia da Folha. Enquanto na primeira fonte de dados marca um local de exibição fora do país e claramente incorreto, conforme ilustra a Figura 4.12, no segundo a informação marca corretamente o Teatro Porto Seguro, conforme ilustra a Figura 4.13. Quando a informação é recuperada do portal Guia da Folha e adicionada ao sistema, como o portal Guia da Folha possui preferência sobre a informação do site Guia da Semana, a geolocalização incorreta é então substituída e a informação torna-se mais confiável.

A consulta SPARQL, ilustrada na Figura 4.14, permite observar como os dados são armazenados e modificados conforme a informação é recuperada na Internet. A Figura 4.15 é um exemplo de quando os dados do evento são capturados de uma fonte de dados com inconsistência da informação e adicionados à base de conhecimento. A Figura 4.16 representa os dados ajustados após o processo de resolução de conflitos por fonte mais confiável. Outra vantagem da proposta apresentada neste trabalho é a capacidade de combinar as informações provenientes das diferentes fontes escolhidas. No exemplo anterior, além de atualizar a informação sobre latitude e longitude do evento, a informação

Figura 4.13: Geolocalização de evento correta no portal Guia da Folha

também foi complementada com uma descrição mais específica com o título "overview" o que agrega mais detalhes e consequentemente oferece uma informação mais abrangente ao usuário final.

O processo de atualização da informação sobre o evento “A Era do Rock” pode ser observado também do ponto de vista da camada de persistência de dados, conforme ilustrado pela Figura 4.17. Após a extração do conteúdo do portal Guia da Semana, várias triplas RDF foram geradas definindo propriedades como o título, latitude, longitude, descrição, endereço na Internet, entre outros. Nesse momento ainda não havia conteúdo sobre o evento na base de dados, e portanto todas as triplas foram persistidas. Na sequência, a informação foi recuperada do portal Guia da Folha. Nesse momento, todas as propriedades encontradas no portal e que já estavam armazenadas na base de dados foram substituídas pelo valor encontrado, em especial os que estão marcados nas linhas em negrito com fundo amarelo: latitude, longitude, endereço do site e da imagem de referência. Além disso, a descrição que ainda não existia foi adicionada (propriedade “overview”) e mesclada junto as outras propriedades do espetáculo, conforme destacado na linha destacada com fundo verde.

4.4 Experimento 3: Consultas combinadas entre diferentes conceitos

A combinação de interesses é um dos objetivos mais interessantes para um usuário que realiza uma consulta a respeito de eventos. É bastante comum nos depararmos com situações em que desejamos ter uma visão ampla sobre os estabelecimentos de uma determinada região para que possa ser possível nos programar para alcançar diferentes finalidades pessoais. Um exemplo disso é quando queremos encontrar restaurantes próximos a um determinado evento do qual já sabemos com antecedência que iremos participar. A Figura 4.18 mostra um exemplo em que temos a informação de todos os restaurantes próximos ao evento escolhido. Essa informação é um apoio a decisão final do usuário. Pode-se ir além e restringir ainda mais essas opções, alterando a consulta para mostrar apenas os restaurantes que servem comida alemã, como mostra a Figura 4.19.

A partir dos dados centralizados em uma única base com informação semântica, aumenta-se as possibilidades de pesquisa de um usuário. Outro exemplo interessante concerne indivíduos que estão

```

1 PREFIX rdf:<http://www.w3.org/1999/02/22-rdf-syntax-ns#>
2 PREFIX rdfs:<http://www.w3.org/2000/01/rdf-schema#>
3 PREFIX iweb:<http://integraweb.ddns.net/>
4 PREFIX schema:<http://schema.org/>
5 SELECT DISTINCT ?s ?title ?latitude ?longitude ?startDate ?endDate ?startTime ?
    endTime ?cuisine ?description ?priceRange ?telephone ?overview ?streetAddress
    ?price ?type ?url ?image
6 WHERE {
7   ?s schema:title ?titleObj ;
8   schema:latitude ?latitudeObj ;
9   schema:longitude ?longitudeObj ;
10  rdf:type schema:TheaterEvent ;
11  rdf:type ?typeObj .
12  ?typeObj rdfs:subClassOf ?subClassObj .
13  OPTIONAL { ?s schema:cuisine ?cuisineObj }
14  OPTIONAL { ?s schema:description ?descriptionObj }
15  OPTIONAL { ?s schema:endDate ?endDateObj }
16  OPTIONAL { ?s schema:endTime ?endTimeObj }
17  OPTIONAL { ?s schema:overview ?overviewObj }
18  OPTIONAL { ?s schema:price ?priceObj }
19  OPTIONAL { ?s schema:priceRange ?priceRangeObj }
20  OPTIONAL { ?s schema:startDate ?startDateObj }
21  OPTIONAL { ?s schema:startTime ?startTimeObj }
22  OPTIONAL { ?s schema:streetAddress ?streetAddressObj }
23  OPTIONAL { ?s schema:telephone ?telephoneObj }
24  OPTIONAL { ?s schema:serviceURL ?urlObj }
25  OPTIONAL { ?s schema:image ?imageObj }
26  values ?subClassObj { schema:Event schema:FoodEstablishment schema:
    CivicStructure }
27  BIND (str(?titleObj) as ?title)
28  BIND (str(?latitudeObj) as ?latitude)
29  BIND (str(?longitudeObj) as ?longitude)
30  BIND (str(?cuisineObj) as ?cuisine)
31  BIND (str(?descriptionObj) as ?description)
32  BIND (str(?endDateObj) as ?endDate)
33  BIND (str(?overviewObj) as ?overview)
34  BIND (str(?priceObj) as ?price)
35  BIND (str(?priceRangeObj) as ?priceRange)
36  BIND (str(?startDateObj) as ?startDate)
37  BIND (str(?streetAddressObj) as ?streetAddress)
38  BIND (str(?telephoneObj) as ?telephone)
39  BIND (str(?urlObj) as ?url)
40  BIND (str(?imageObj) as ?image)
41  BIND ( strafter( strafter( str(?typeObj), "http://" ), "/" ) as ?type )
42  BIND ( strafter( str(?endTimeObj), "T" ) as ?endTime )
43  BIND ( strafter( str(?startTimeObj), "T" ) as ?startTime )
44  FILTER (?title = 'A Era do Rock')
45  FILTER NOT EXISTS {
46    ?s rdf:type ?subtype .
47    ?subtype rdfs:subClassOf* ?typeObj .
48    filter ( ?subtype != ?typeObj )
49  }
50 }
51 LIMIT 100

```

Figura 4.14: Consulta sobre o evento "A Era do Rock"

Figura 4.15: Informação inconsistente antes da resolução de conflitos

Figura 4.16: Informação consistente após resolução de conflitos

migrando para o transporte público e deixando de lado os seus carros. Para esse tipo de usuário, ter a informação sobre o que ocorre próximo ao metrô pode vir a definir a escolha de um evento em relação a outro. A Figura 4.20 mostra uma pesquisa por bares próximos ao metrô. Seguindo a ideia de restrição de consultas, é possível melhorar ainda mais a pesquisa e consultar, por exemplo, apenas os eventos que estão próximos a um metrô, por exemplo o da praça da Sé, como ilustra a Figura 4.21.

A combinação de dados pode ser ainda mais complexa de maneira, a tornar possível programar uma agenda mais completa para o dia de uma pessoa. Suponha a situação em que um indivíduo que tenha interesse em conversar com um velho amigo em um bar, levar a esposa para a exposição “Fiaminghi - Pensamentos Compostos” e, por fim, fazer uma surpresa para ela levando-a para um restaurante de comida Natural. Para que o usuário aproveite melhor os momentos, é importante que os eventos estejam próximos uns dos outros para que ele possa se deslocar, andando sem se demorar. Nessa situação, a resposta pode ser obtida por uma consulta mais elaborada sobre a base de dados conforme o exemplo de código SPARQL resumido apresentado na Figura 4.22. Tal consulta retorna as informações ilustradas na Figura 4.23 e auxilia o usuário na tomada de decisão. A consulta completa pode ser encontrada no anexo B.

O exemplo anterior demonstra que é possível realizar uma busca que integra a informação de diferentes conceitos: bares, eventos e restaurantes. No entanto, os dados apresentados eram todos do portal Guia da Folha. Ainda que seja custoso, avaliar os dados contidos dentro de um mesmo portal

Triplas RDF armazenadas após extração de dados no portal Guia da Semana

Sujeito	Predicado	Objeto
#38681703e693eb151523b2afb1923648	schema:title	A Era do Rock
#38681703e693eb151523b2afb1923648	schema:latitude	32.504946
#38681703e693eb151523b2afb1923648	schema:longitude	-105.580439
#38681703e693eb151523b2afb1923648	schema:startDate	2017-06-02
#38681703e693eb151523b2afb1923648	schema:endDate	2017-07-30
#38681703e693eb151523b2afb1923648	schema:description	Diretamente da Broadway, comédia musical...
#38681703e693eb151523b2afb1923648	schema:streetAddress	Alameda Barão de Piracicaba, 740
#38681703e693eb151523b2afb1923648	schema:url	https://www.guiadasemana.com.br/sao-paulo/t
#38681703e693eb151523b2afb1923648	schema:type	TheaterEvent
#38681703e693eb151523b2afb1923648	schema:image	https://s3a4z9x7.map2.ssl.hwcdn.net/contentF

Triplas RDF armazenadas após extração de dados no portal Guia da Folha

Sujeito	Predicado	Objeto	
#38681703e693eb151523b2afb1923648	schema:title	A Era do Rock	
#38681703e693eb151523b2afb1923648	schema:latitude	-23.5301554	Modificado
#38681703e693eb151523b2afb1923648	schema:longitude	-46.645438	Modificado
#38681703e693eb151523b2afb1923648	schema:startDate	2017-06-02	
#38681703e693eb151523b2afb1923648	schema:endDate	2017-07-30	
#38681703e693eb151523b2afb1923648	schema:description	Diretamente da Broadway, comédia musical...	
#38681703e693eb151523b2afb1923648	schema:streetAddress	Alameda Barão de Piracicaba, 740	
#38681703e693eb151523b2afb1923648	schema:url	http://gula.folha.uol.com.br/teatro/musical/a	Modificado
#38681703e693eb151523b2afb1923648	schema:type	TheaterEvent	
#38681703e693eb151523b2afb1923648	schema:image	http://f1.uol.com.br/hunting/2017/06/01/1496	Modificado
#38681703e693eb151523b2afb1923648	schema:overview	O musical se passa nos anos 1980 em Los	Adicionado

Figura 4.17: Dados obtidos após extração nos diferentes portais

Figura 4.18: Restaurantes próximo a um evento

é mais fácil do que quando a informação está espalhada por diferentes portais da Internet. Quanto mais informações precisam ser avaliadas e quanto mais portais precisam ser visitados, mais difícil é chegar a um resultado mais assertivo. A tarefa é ainda maior quando se é preciso avaliar a distância entre locais de diferentes eventos, quando cada um deles é retornado por um portal distinto. Apesar disso, quando a arquitetura proposta nesse trabalho é aplicada, essa tarefa torna-se trivial pois toda a informação está concentrada em um só banco. A Figura 4.24 exemplifica uma variação do exemplo anterior. Vamos agora levar em consideração o evento “Metrópole: experiência paulista” que está sendo divulgado apenas pelo portal de eventos da USP, bares e restaurante de comida coreana próximos. Verifica-se que a arquitetura consegue integrar a informação de diferentes portais e associá-la por meio de conceitos como bar, restaurante, tipo de comida, etc. Essa característica simplifica bastante a interação do usuário com as informações espalhadas na Internet.

Figura 4.19: Restaurantes de comida alemã próximo a um evento

Figura 4.20: Bares próximos a estações de metrô

Figura 4.21: Bares próximos à estação de metrô Sé

```

1 ...
2 PREFIX iweb:<http://integraweb.ddns.net/>
3 PREFIX schema:<http://schema.org/>
4 SELECT ?title ?latitude ?longitude ?startDate ?endDate ?startTime ?endTime ?
   cuisine ?description ?priceRange ?telephone ?overview ?streetAddress ?price ?
   type ?url ?image ?servesCuisine WHERE {
5 {
6   ?s schema:title ?titleObj ;
7   schema:latitude ?latitudeObj ;
8   schema:longitude ?longitudeObj ;
9   iweb:near ?nearObj ;
10  rdf:type schema:Restaurant ;
11  schema:servesCuisine ?servesCuisineObj ;
12  rdf:type ?typeObj .
13  ?nearObj schema:title ?nearObjTitle .
14  ...
15  FILTER (str(?nearObjTitle) = 'Fiaminghi – Pensamentos Compostos')
16  FILTER (?servesCuisine = 'Natural')
17  ...
18 }
19 UNION
20 {
21  ...
22  FILTER (?title = 'Fiaminghi – Pensamentos Compostos')
23  FILTER (?type = 'ExhibitionEvent')
24  ...
25 }
26 UNION
27 {
28  ...
29  iweb:near ?nearObj ;
30  rdf:type schema:BarOrPub ;
31  ...
32  FILTER (str(?nearObjTitle) = 'Fiaminghi – Pensamentos Compostos')
33  ...
34 }
35 }

```

Figura 4.22: Consulta relacionando estabelecimentos distintos

Figura 4.23: Busca por evento, bar e restaurante próximos

Figura 4.24: Busca integrando diferentes conceitos sobre dados de diferentes portais

Capítulo 5

Conclusões e Trabalhos Futuros

A Web Semântica tem o potencial de conectar os dados espalhados nas mais diferentes fontes de informação na Web, contidos em diferentes portais e representados em diferentes formatos. A premissa deste trabalho é que ao combinar a tecnologia de Web Semântica com técnicas de mineração e integração de dados na Web, permite-se que o conteúdo relevante dos portais espalhados pela Internet possa ser extraído de maneira automatizada para oferecer um resultado muito mais expressivo ao usuário final.

A aplicação de ontologias, auxilia a elevar a qualidade da informação exposta na Internet. Através de seu uso, é possível atribuir significado à informação presente na Internet, abrindo assim um leque de possibilidades ao propiciar consultas mais ricas, com semântica agregada, proveniente de fontes e formatos de dados diversas, com informações mais próximas dos interesses dos indivíduos, auxiliando-os na tomada de decisões do dia-a-dia.

Este trabalho dá um primeiro passo nesta direção, com uma proposta de arquitetura capaz de permitir a integração da informação provenientes de portais distintos a partir do mapeamento para ontologias bem estabelecidas e a centralização da informação para consulta aos dados com o uso de ontologias.

5.1 Conclusões

Há grandes obstáculos a serem superados no que diz respeito à extração de dados. Como a maioria dos portais não anota semanticamente o seu conteúdo que é gerado dinamicamente, torna-se necessária a aplicação de artifícios para a recuperação de conteúdo a partir da estrutura sintática desses documentos. Propostas como a apresentada neste trabalho são vulneráveis à mudança da maneira pela qual a informação é exposta para os usuários. Em outras palavras, se um determinado portal muda sua forma de apresentação para os seus usuários, o processo de extração de dados deve ser atualizado. Além disso, quanto maior a frequência de modificação da estrutura desses documentos, maior é a quantidade de manutenção na infra-estrutura de recuperação do conteúdo do portal.

A própria natureza da Internet é um obstáculo para a recuperação de dados quando há a necessidade de se percorrer variados portais da Web para a construção de um resultado mais abrangente. Isso porque a navegação link a link, página a página, depende de que o servidor alvo tenha capacidade de expor todo o conteúdo necessário para a consulta de maneira rápida e sem interrupções. No entanto, esse tipo de solução fica a mercê da velocidade do servidor, da disponibilidade da informação, da estrutura de rede ou até mesmo de detalhes como a cota de tráfego de dados de um portal ser excedida. Além disso, é inviável a recuperação de todo o conteúdo de grandes portais a cada consulta, uma vez que essas organizações possuem bases gigantescas de dados. A centralização de conteúdo em uma base de dados semântica diminui os problemas intrínsecos à rede e permite que o conhecimento possa ser construído e atualizado de maneira mais gradual, o que tende a oferecer mais informações a qualquer momento e independe do gargalo ocasionado pela consulta federada em várias fontes de informação de maneira simultânea. Ainda nessa linha, ao centralizar a infor-

mação em uma base de dados única, o processo de inferência sobre dados não precisa ser feito sob demanda a cada consulta realizada por um usuário, mas de maneira assíncrona, o que implica em resultados mais rápidos.

A integração da informação de dados não necessita ser confinada a uma definição de conceito específica e suas derivações. Neste trabalho, demonstrou-se que diferentes domínios da informação podem ser avaliados de maneira conjunta: foram tratados eventos como exposições, peças de teatro e shows, bem como restaurantes, bares ou estações de metrô.

A resolução de conflitos de informação acontece quando um determinado dado possui referências em portais distintos na Internet mas com informações diferentes expostas. Há várias formas de tratar esse problema e a seleção de uma depende, entre outros fatores, do domínio do conteúdo escolhido. No contexto da extração de conhecimento sobre eventos, este trabalho demonstrou que a priorização da fonte dos dados é um mecanismo que pode ser utilizado para decidir qual informação mostrar para o usuário final. Além disso, quando as informações dizem respeito a um mesmo conteúdo, torna-se possível agregá-las de maneira a oferecer uma informação mais completa e consequentemente mais relevante, conforme os experimento visto na seção 4.3.

Neste trabalho, diferentemente de projetos como o VirtuosoRDF e o D2RQ que se valem das definições estruturais de tabelas, colunas entre outras características bem definidas no banco de banco para promover a junção das fontes de informação através do mapeamento semântico para RDF, o foco é o conteúdo publicado na Internet, conteúdo este que muitas vezes não está disponível para o usuário senão por meio de uma camada de apresentação que faz o uso de documentos com padrões bastante distintos. Além disso, o processo de anotação do conteúdo de um portal é feito de forma manual uma vez que o foco da proposta apresentada é mesclar a informação espalhada na Web e não a facilitação de marcação semântica assim como é proposta nos trabalhos de (Heß, Andreas e Dieric, 2008) e no DeepAnnotation. Finalmente, no que diz respeito a extração de informação em documentos estruturados, este trabalho se assemelha ao de (May, 2002) para a integração de documentos distintos mas diferente na forma pela qual a integração é realizada uma vez que na proposta aqui apresentada a informação é recuperada por meio do uso de expressões regulares e então transformada em RDF.

Espera-se que este trabalho possa ser mais um incentivo para o reuso de informação exposta na Internet e para o avanço da Web Semântica. Trata-se de uma amostra de que a nova proposta da Web na qual todos os dados estão interconectados não é uma utopia e pode estar mais próxima do que imaginamos.

5.2 Possíveis Extensões

Almeja-se que a proposta apresentada neste trabalho possa ser tomada como referência para a construção de sistemas mais inteligentes, capazes de unificar sob um ponto de vista mais amplo os mais diferentes domínios da informação para, dessa maneira, oferecer conteúdos mais ricos para os futuros usuários da Internet. São inúmeros os domínios que poderiam ser abordados pela proposta aqui apresentada, entre eles, uma possível aplicação seria o de, por exemplo, facilitar a busca de hotéis e pousadas próximos a aeroportos.

Atualmente a busca de passagens aéreas em diferentes portais não integra a informação da localização dos aeroportos e horário de partida e chegada dos diferentes voos com a informação sobre hotéis e pousadas abertos no dia e no horário da chegada da viagem. Dessa maneira, ao simular uma data de chegada em uma determinada cidade, esses portais indicam outros portais agregadores de conteúdo com as informações sobre hotéis na região. Acontece que muitas vezes esses hotéis estão fechados, sem disponibilidade de vaga, fora de uma faixa de valor de diária aceitável, possivelmente sem serviços adicionais que possam ser de interesse de um indivíduo e até mesmo muito longe do aeroporto do desembarque. Nesse contexto, aplicando a arquitetura proposta neste trabalho e utilizando ontologias que definem o domínio Hotel e Voo como, por exemplo, os

conceitos Flight¹ e Hotel² descritos no portal Schema.org, tal pesquisa poderia ser simplificada pois a partir de uma única consulta tornaria-se possível filtrar todas as opções disponíveis nos diversos portais agregadores de informações dentro do assunto.

A conclusão desse trabalho abre portas para novas pesquisas voltadas para a melhoria da integração da informação na Web. Algumas áreas abordadas na proposta de arquitetura apresentada neste documento ainda podem ser bastante exploradas a fim de torná-la ainda mais eficiente. Entre as sugestões futuras, encontram-se: (i) a pesquisa de mecanismos mais eficientes de reconhecimento e recuperação de informação relevantes, (ii) a auto-deteção e mapeamento em tempo real dos dados contidos em documentos Web para documentos semanticamente anotados, (iii) a construção de motores capazes de converter consultas SPARQL em HTML, assim como hoje algumas consultas SPARQL podem ser traduzidas diretamente para SQL e (iv) mecanismos de extração de informação capazes de reconhecer a mudança da estrutura sintática das páginas HTML e que tenham a habilidade de se auto ajustar a elas, sem a necessidade de intervenção humana, por exemplo através de técnicas de aprendizado de máquina.

¹<http://schema.org/Flight>

²<http://schema.org/Hotel>

Apêndice A

Consulta SPARQL para integração de dados

```
1 PREFIX rdf:<http://www.w3.org/1999/02/22-rdf-syntax-ns#>
2 PREFIX rdfs:<http://www.w3.org/2000/01/rdf-schema#>
3 PREFIX iweb:<http://integraweb.ddns.net/>
4 PREFIX schema:<http://schema.org/>
5 SELECT DISTINCT ?title ?latitude ?longitude ?startDate ?endDate ?startTime
   ?endTime ?cuisine ?description ?priceRange ?telephone ?overview ?
   streetAddress ?price ?type ?url ?image
6 WHERE {
7   ?s schema:title ?titleObj ;
8   schema:latitude ?latitudeObj ;
9   schema:longitude ?longitudeObj ;
10  rdf:type ?typeObj .
11  ?typeObj rdfs:subClassOf ?subClassObj .
12  OPTIONAL { ?s schema:cuisine ?cuisineObj }
13  OPTIONAL { ?s schema:description ?descriptionObj }
14  OPTIONAL { ?s schema:endDate ?endDateObj }
15  OPTIONAL { ?s schema:endTime ?endTimeObj }
16  OPTIONAL { ?s schema:overview ?overviewObj }
17  OPTIONAL { ?s schema:price ?priceObj }
18  OPTIONAL { ?s schema:priceRange ?priceRangeObj }
19  OPTIONAL { ?s schema:startDate ?startDateObj }
20  OPTIONAL { ?s schema:startTime ?startTimeObj }
21  OPTIONAL { ?s schema:streetAddress ?streetAddressObj }
22  OPTIONAL { ?s schema:telephone ?telephoneObj }
23  OPTIONAL { ?s schema:serviceURL ?urlObj }
24  OPTIONAL { ?s schema:image ?imageObj }
25  values ?subClassObj { schema:Event schema:FoodEstablishment schema:
    CivicStructure }
26  BIND (str(?titleObj) as ?title)
27  BIND (str(?latitudeObj) as ?latitude)
28  BIND (str(?longitudeObj) as ?longitude)
29  BIND (str(?cuisineObj) as ?cuisine)
30  BIND (str(?descriptionObj) as ?description)
31  BIND (str(?endDateObj) as ?endDate)
32  BIND (str(?overviewObj) as ?overview)
33  BIND (str(?priceObj) as ?price)
34  BIND (str(?priceRangeObj) as ?priceRange)
35  BIND (str(?startDateObj) as ?startDate)
36  BIND (str(?streetAddressObj) as ?streetAddress)
37  BIND (str(?telephoneObj) as ?telephone)
38  BIND (str(?urlObj) as ?url)
39  BIND (str(?imageObj) as ?image)
```

```
40 BIND ( strafter(strafter( str(?typeObj), "http://" ),"/") as ?type )
41 BIND ( strafter( str(?endTimeObj), "T" ) as ?endTime )
42 BIND ( strafter( str(?startTimeObj), "T" ) as ?startTime )
43 FILTER (?latitude > '-23.539606783940812' && ?latitude < '
    -23.55759321605919')
44 FILTER (?longitude > '-46.62938980581791' && ?longitude < '
    -46.6490101941821')
45 FILTER NOT EXISTS {
46     ?s rdf:type ?subtype .
47     ?subtype rdfs:subClassOf* ?typeObj .
48     filter ( ?subtype != ?typeObj )
49 }
50 }
51 LIMIT 100
```

Apêndice B

Consulta com várias restrições

```
1 PREFIX rdf:<http://www.w3.org/1999/02/22-rdf-syntax-ns#>
2 PREFIX rdfs:<http://www.w3.org/2000/01/rdf-schema#>
3 PREFIX iweb:<http://integraweb.ddns.net/>
4 PREFIX schema:<http://schema.org/>
5 SELECT ?title ?latitude ?longitude ?startDate ?endDate ?startTime ?endTime
   ?cuisine ?description ?priceRange ?telephone ?overview ?streetAddress
   ?price ?type ?url ?image ?servesCuisine WHERE {
6 {
7   ?s schema:title ?titleObj ;
8   schema:latitude ?latitudeObj ;
9   schema:longitude ?longitudeObj ;
10  iweb:near ?nearObj ;
11  rdf:type schema:Restaurant ;
12  schema:servesCuisine ?servesCuisineObj ;
13  rdf:type ?typeObj .
14  ?nearObj schema:title ?nearObjTitle .
15  OPTIONAL { ?s schema:cuisine ?cuisineObj }
16  OPTIONAL { ?s schema:description ?descriptionObj }
17  OPTIONAL { ?s schema:endDate ?endDateObj }
18  OPTIONAL { ?s schema:endTime ?endTimeObj }
19  OPTIONAL { ?s schema:overview ?overviewObj }
20  OPTIONAL { ?s schema:price ?priceObj }
21  OPTIONAL { ?s schema:priceRange ?priceRangeObj }
22  OPTIONAL { ?s schema:startDate ?startDateObj }
23  OPTIONAL { ?s schema:startTime ?startTimeObj }
24  OPTIONAL { ?s schema:streetAddress ?streetAddressObj }
25  OPTIONAL { ?s schema:telephone ?telephoneObj }
26  OPTIONAL { ?s schema:serviceURL ?urlObj }
27  OPTIONAL { ?s schema:image ?imageObj }
28  values ?subClassObj { schema:Event schema:FoodEstablishment schema:
    CivicStructure }
29  BIND (str(?titleObj) as ?title)
30  BIND (str(?latitudeObj) as ?latitude)
31  BIND (str(?longitudeObj) as ?longitude)
32  BIND (str(?servesCuisineObj) as ?servesCuisine)
33  BIND (str(?descriptionObj) as ?description)
34  BIND (str(?endDateObj) as ?endDate)
35  BIND (str(?overviewObj) as ?overview)
36  BIND (str(?priceObj) as ?price)
37  BIND (str(?priceRangeObj) as ?priceRange)
38  BIND (str(?startDateObj) as ?startDate)
39  BIND (str(?streetAddressObj) as ?streetAddress)
40  BIND (str(?telephoneObj) as ?telephone)
41  BIND (str(?urlObj) as ?url)
```

```

42 BIND (str(?imageObj) as ?image)
43 BIND ( strafter(strafter( str(?typeObj), "http://" ),"/") as ?type )
44 BIND ( strafter( str(?endTimeObj), "T" ) as ?endTime )
45 BIND ( strafter( str(?startTimeObj), "T" ) as ?startTime )
46 FILTER (str(?nearObjTitle) = 'Fiaminghi - Pensamentos Compostos')
47 FILTER (?servesCuisine = 'Natural')
48 FILTER (?type = 'Restaurant')
49 FILTER NOT EXISTS {
50     ?s rdf:type ?subtype .
51     ?subtype rdfs:subClassOf* ?typeObj .
52     filter ( ?subtype != ?typeObj )
53 }
54 }
55 UNION
56 {
57     ?s schema:title ?titleObj ;
58     schema:latitude ?latitudeObj ;
59     schema:longitude ?longitudeObj ;
60     rdf:type schema:ExhibitionEvent ;
61     rdf:type ?typeObj .
62     OPTIONAL { ?s schema:servesCuisine ?servesCuisineObj }
63     OPTIONAL { ?s schema:description ?descriptionObj }
64     OPTIONAL { ?s schema:endDate ?endDateObj }
65     OPTIONAL { ?s schema:endTime ?endTimeObj }
66     OPTIONAL { ?s schema:overview ?overviewObj }
67     OPTIONAL { ?s schema:price ?priceObj }
68     OPTIONAL { ?s schema:priceRange ?priceRangeObj }
69     OPTIONAL { ?s schema:startDate ?startDateObj }
70     OPTIONAL { ?s schema:startTime ?startTimeObj }
71     OPTIONAL { ?s schema:streetAddress ?streetAddressObj }
72     OPTIONAL { ?s schema:telephone ?telephoneObj }
73     OPTIONAL { ?s schema:serviceURL ?urlObj }
74     OPTIONAL { ?s schema:image ?imageObj }
75     values ?subClassObj { schema:Event schema:FoodEstablishment schema:
        CivicStructure }
76 BIND (str(?titleObj) as ?title)
77 BIND (str(?latitudeObj) as ?latitude)
78 BIND (str(?longitudeObj) as ?longitude)
79 BIND (str(?servesCuisineObj) as ?servesCuisine)
80 BIND (str(?cuisineObj) as ?cuisine)
81 BIND (str(?descriptionObj) as ?description)
82 BIND (str(?endDateObj) as ?endDate)
83 BIND (str(?overviewObj) as ?overview)
84 BIND (str(?priceObj) as ?price)
85 BIND (str(?priceRangeObj) as ?priceRange)
86 BIND (str(?startDateObj) as ?startDate)
87 BIND (str(?streetAddressObj) as ?streetAddress)
88 BIND (str(?telephoneObj) as ?telephone)
89 BIND (str(?urlObj) as ?url)
90 BIND (str(?imageObj) as ?image)
91 BIND ( strafter(strafter( str(?typeObj), "http://" ),"/") as ?type )
92 BIND ( strafter( str(?endTimeObj), "T" ) as ?endTime )
93 BIND ( strafter( str(?startTimeObj), "T" ) as ?startTime )
94 FILTER (?title = 'Fiaminghi - Pensamentos Compostos')
95 FILTER (?type = 'ExhibitionEvent')
96 FILTER NOT EXISTS {
97     ?s rdf:type ?subtype .
98     ?subtype rdfs:subClassOf* ?typeObj .
99     filter ( ?subtype != ?typeObj )

```

```

100     }
101 }
102 UNION
103 {
104     ?s schema:title ?titleObj ;
105     schema:latitude ?latitudeObj ;
106     schema:longitude ?longitudeObj ;
107     iweb:near ?nearObj ;
108     rdf:type schema:BarOrPub ;
109     rdf:type ?typeObj .
110     ?nearObj schema:title ?nearObjTitle .
111     OPTIONAL { ?s schema:servesCuisine ?servesCuisineObj }
112     OPTIONAL { ?s schema:description ?descriptionObj }
113     OPTIONAL { ?s schema:endDate ?endDateObj }
114     OPTIONAL { ?s schema:endTime ?endTimeObj }
115     OPTIONAL { ?s schema:overview ?overviewObj }
116     OPTIONAL { ?s schema:price ?priceObj }
117     OPTIONAL { ?s schema:priceRange ?priceRangeObj }
118     OPTIONAL { ?s schema:startDate ?startDateObj }
119     OPTIONAL { ?s schema:startTime ?startTimeObj }
120     OPTIONAL { ?s schema:streetAddress ?streetAddressObj }
121     OPTIONAL { ?s schema:telephone ?telephoneObj }
122     OPTIONAL { ?s schema:serviceURL ?urlObj }
123     OPTIONAL { ?s schema:image ?imageObj }
124     values ?subClassObj { schema:Event schema:FoodEstablishment schema:
        CivicStructure }
125     BIND (str(?titleObj) as ?title)
126     BIND (str(?latitudeObj) as ?latitude)
127     BIND (str(?longitudeObj) as ?longitude)
128     BIND (str(?servesCuisineObj) as ?servesCuisine)
129     BIND (str(?cuisineObj) as ?cuisine)
130     BIND (str(?descriptionObj) as ?description)
131     BIND (str(?endDateObj) as ?endDate)
132     BIND (str(?overviewObj) as ?overview)
133     BIND (str(?priceObj) as ?price)
134     BIND (str(?priceRangeObj) as ?priceRange)
135     BIND (str(?startDateObj) as ?startDate)
136     BIND (str(?streetAddressObj) as ?streetAddress)
137     BIND (str(?telephoneObj) as ?telephone)
138     BIND (str(?urlObj) as ?url)
139     BIND (str(?imageObj) as ?image)
140     BIND ( strafter(strafter( str(?typeObj), "http://" ),"/") as ?type )
141     BIND ( strafter( str(?endTimeObj), "T" ) as ?endTime )
142     BIND ( strafter( str(?startTimeObj), "T" ) as ?startTime )
143     FILTER (str(?nearObjTitle) = 'Fiaminghi - Pensamentos Compostos')
144     FILTER (?type = 'BarOrPub')
145     FILTER NOT EXISTS {
146         ?s rdf:type ?subtype .
147         ?subtype rdfs:subClassOf* ?typeObj .
148         filter ( ?subtype != ?typeObj )
149     }
150 }
151 }

```


Bibliografia

- Abdellaoui e Nader(2015)** Sabrina Abdellaoui e Fahima Nader. Semantic Data Warehouse at the heart of Competitive Intelligence Systems: design approach. Em *2015 6Th International Conference on Information Systems and Economic Intelligence (Ssie)*, páginas 141–145. IEEE. ISBN 978-1-4799-8934-8. doi: 10.1109/ISEI.2015.7358736. URL <http://ieeexplore.ieee.org/document/7358736/>. Citado na pág. 19
- Ahmed(2008)** Emdad Ahmed. Resource capability discovery and description management system for bioinformatics data and service integration - An experiment with gene regulatory networks. Em *Proceedings of 11th International Conference on Computer and Information Technology, IC-CIT 2008*, páginas 56–61. IEEE. ISBN 9781424421367. doi: 10.1109/ICCITECHN.2008.4802991. URL <http://ieeexplore.ieee.org/document/4802991/>. Citado na pág. 20, 23
- Allemang e Hendler(2011)** Dean Allemang e Jim Hendler. *Semantic Web for the Working Ontologist*. Elsevier. ISBN 9780123859655. doi: 10.1016/B978-0-12-385965-5.10016-0. URL <http://www.sciencedirect.com/science/article/pii/B9780123859655100160>. Citado na pág. 7, 11, 14, 15
- Auer et al.(2007)** Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak e Zachary Ives. DBpedia: A nucleus for a Web of open data. Em Karl Aberer, , Key-Sun Choi, , Natasha Noy, , Dean Allemang, , Kyung-Il Lee, , Lyndon Nixon, , Jennifer Golbeck, , Peter Mika, , Diana Maynard, , Riichiro Mizoguchi, , Guus Schreiber, e Philippe Cudr{\`e}-Mauroux, editors, *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 4825 LNCS, páginas 722–735. Springer Berlin Heidelberg. ISBN 3540762973. doi: 10.1007/978-3-540-76298-0_52. URL http://link.springer.com/10.1007/978-3-540-76298-0_{_}52. Citado na pág. xiii, 11
- Balduini et al.(2012)** Marco Balduini, Irene Celino, Daniele Dell’Aglío, Emanuele Della Valle, Yi Huang, Tony Lee, Seon Ho Kim e Volker Tresp. BOTTARI: An augmented reality mobile application to deliver personalized and location-based recommendations by continuous analysis of social media streams. *Journal of Web Semantics*, 16:33–41. ISSN 15708268. doi: 10.1016/j.websem.2012.06.004. URL <http://dx.doi.org/10.1016/j.websem.2012.06.004><http://linkinghub.elsevier.com/retrieve/pii/S157082681200073X>. Citado na pág. 20
- Battle e Benson(2008)** Robert Battle e Edward Benson. Bridging the semantic Web and Web 2.0 with Representational State Transfer (REST). *Web Semantics*, 6(1):61–69. ISSN 15708268. doi: 10.1016/j.websem.2007.11.002. Citado na pág. 20
- Berendt et al.(2004)** Bettina Berendt, Andreas Hotho, Dunja Mladenic, Maarten van Someren, Myra Spiliopoulou e Gerd Stumme. A Roadmap for Web Mining: From Web to Semantic Web. Em Bettina Berendt, Andreas Hotho, Dunja Mladenić, Maarten van Someren, Myra Spiliopoulou e Gerd Stumme, editors, *Web Mining: From Web to Semantic Web*, volume 3209 of *Lecture Notes in Computer Science*, páginas 1–22. Springer Berlin Heidelberg, Berlin, Heidelberg. ISBN 3540416714. doi: 10.1007/978-3-540-30123-3_1. URL <http://eprints.pascal-network.org/archive/00000841/{%}5Cnhttp://www.springerlink.com/index/6c87nfe4t1qhc0h6.pdf><http://www.springerlink.com/index/6c87nfe4t1qhc0h6.pdf>

- <http://link.springer.com/10.1007/b100615>http://link.springer.com/10.1007/978-3-540-30123-3_{_}1. Citado na pág. 17, 18
- Berners-Lee et al.(2002)** T Berners-Lee, J Hendler e O Lassila. The Semantic Web. Em *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 5926, páginas 402–410. ISBN 978-3-540-89703-3. doi: 10.1007/978-3-540-89704-0. URL <http://www.scientificamerican.com/article.cfm?id=the-semantic-web>. Citado na pág. 4, 7, 8
- Bienvenu et al.(2013)** Meghyn Bienvenu, Balder ten Cate, Carsten Lutz e Frank Wolter. Ontology-based data access. *Proceedings of the 32nd symposium on Principles of database systems - PODS '13*, 39(4):213. ISSN 15574644. doi: 10.1145/2463664.2465223. URL <http://dl.acm.org/citation.cfm?doid=2463664.2465223>. Citado na pág. 8, 21
- Bizer et al.(2009)** Christian Bizer, T Heath e T Berners-Lee. Linked data-the story so far. *International journal on Semantic Web and Information Systems*, 5(3):1–22. ISSN 15526283. doi: 10.4018/jswis.2009081901. URL <http://eprints.soton.ac.uk/271285/>. Citado na pág. 8
- Bojars et al.(2008)** U. Bojars, J. G. Breslin, a. Finn e S. Decker. Using the Semantic Web for linking and reusing data across Web 2.0 communities. *Web Semantics*, 6(1):21–28. ISSN 15708268. doi: 10.1016/j.websem.2007.11.010. Citado na pág. 4, 7, 8, 20
- Brickley e Miller(2010)** Dan Brickley e Libby Miller. FOAF Vocabulary Specification. *Namespace Document*, 3(August):<http://xmlns.com/foaf/spec/>. URL <http://xmlns.com/foaf/spec/>. Citado na pág. 9
- Califf e Mooney(2004)** Mary Elaine Califf e Raymond J Mooney. Bottom-Up Relational Learning of Pattern Matching Rules for Information Extraction. *Journal of Machine Learning Research*, 4:177–210. ISSN 15324435. doi: 10.1162/153244304322972685. URL http://www.crossref.org/jmlr_DOI.html. Citado na pág. 16
- Calvanese et al.(2016)** D Calvanese, B Cogrel e S Komla-Ebri. Ontop: Answering SPARQL queries over relational databases. *Semantic Web*, 0(0). ISSN 22104968. doi: 10.3233/SW-160217. URL <http://content.iospress.com/articles/semantic-web/sw217>. Citado na pág. 21
- Cardoso e Sheth(2006)** Jorge Cardoso e Amit P. Sheth, editors. *Semantic Web Services, Processes and Applications*, volume 3 of *Semantic Web and Beyond*. Springer US, Boston, MA. ISBN 978-0-387-30239-3. doi: 10.1007/978-0-387-34685-4. URL <http://link.springer.com/10.1007/978-0-387-34685-4>. Citado na pág. 20
- Civili et al.(2013)** Cristina Civili, Marco Ruzzi, Valerio Santarelli, Domenico Fabio Savo, Marco Console, Giuseppe De Giacomo, Domenico Lembo, Maurizio Lenzerini, Lorenzo Lepore, Riccardo Mancini, Antonella Poggi e Riccardo Rosati. Mastro Studio: Managing Ontology-based Data Access Applications. *Proceedings of the VLDB Endowment*, 6(12):1314–1317. ISSN 21508097. doi: 10.14778/2536274.2536304. URL <http://dl.acm.org/citation.cfm?doid=2536274.2536304>. Citado na pág. 19, 20, 21
- Clark e DeRose(1999)** James Clark e Steve DeRose. XML path language (XPath) version 1.0, 1999. URL <https://www.w3.org/TR/1999/REC-xpath-19991116/>. Citado na pág. 17
- Cyganiak e Jentzsch(2014)** Richard Cyganiak e Anja Jentzsch. The Linking Open Data cloud diagram, 2014. URL <http://lod-cloud.net/>. Citado na pág. xiii, 12
- Deursen et al.(2008)** Davy Van Deursen, Chris Poppe, Getan Martens, Erik Mannens e Rik Van de Walle. XML to RDF Conversion: A Generic Approach. Em *2008 International Conference on Automated Solutions for Cross Media Content and Multi-Channel Distribution*, pginas 138–144. IEEE. ISBN 978-0-7695-3406-0. doi: 10.1109/AXMEDIS.2008.17. URL <http://ieeexplore.ieee.org/document/4688061/>. Citado na pg. 14

- Fayyad et al.(1996)** Usama Fayyad, G Piatetsky-Shapiro e Padhraic Smyth. From data mining to knowledge discovery in databases. *AI magazine*, páginas 37–54. ISSN 0738-4602. doi: 10.1145/240455.240463. URL <http://www.aaai.org/ojs/index.php/aimagazine/article/viewArticle/1230>. Citado na pág. 4
- Freitag e Kushmerick(2000)** Dayne Freitag e Nicholas Kushmerick. Boosted Wrapper Induction. *Proceedings Of The National Conference On Artificial Intelligence*, 145(15):577–583. doi: 10.2307/3977872. URL <http://www.aaai.org/Papers/AAAI/2000/AAAI00-088.pdf>. Citado na pág. 16
- Gali et al.(2004)** Anuradha Gali, Cindy X. Chen, Kajal T. Claypool e Rosario Uceda-Sosa. From Ontology to Relational Databases. Em *Conceptual Modeling for Advanced Application Domains*, páginas 278–289. Springer, Berlin, Heidelberg. ISBN 978-3-540-23722-8. doi: 10.1007/978-3-540-30466-1_26. URL http://link.springer.com/chapter/10.1007/978-3-540-30466-1_{_}26http://link.springer.com/10.1007/978-3-540-30466-1_{_}26. Citado na pág. 3, 8, 9
- Gregor e Hevner(2013)** Shirley Gregor e Alan R Hevner. POSITIONING AND PRESENTING DESIGN SCIENCE Types of Knowledge in Design Science Research. *MIS Quarterly*, 37(2): 337–355. ISSN 02767783. doi: 10.2753/MIS0742-122240302. Citado na pág. 5, 6
- Gruber(1993)** Thomas R. Gruber. A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5(2):199–220. ISSN 10428143. doi: 10.1.1.101.7493. URL <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.101.7493>. Citado na pág. 3, 8
- Handschuh et al.(2003)** Siegfried Handschuh, Steffen Staab e Raphael Volz. On deep annotation. *Proceedings of the twelfth international conference on World Wide Web - WWW '03*, página 431. doi: 10.1145/775213.775214. URL [http://portal.acm.org/citation.cfm?doid=775152.775214&delimiter="026E30F\\$nh](http://portal.acm.org/citation.cfm?doid=775152.775214&delimiter=)<http://www.isi.edu/info-agents/workshops/ijcai03/papers/Handschuh.pdf>. Citado na pág. 20
- Heath e Motta(2008)** Tom Heath e Enrico Motta. Ease of interaction plus ease of integration: Combining Web2.0 and the Semantic Web in a reviewing site. *Web Semantics*, 6(1):76–83. ISSN 15708268. doi: 10.1016/j.websem.2007.11.009. Citado na pág. 13, 14, 15, 19, 20
- Hellmann et al.(2009)** Sebastian Hellmann, Sebastian Hellmann, Sören Auer, Sören Auer, Sebastian Dietzold, Sebastian Dietzold, Jens Lehmann, Jens Lehmann, David Aumüller e David Aumüller. Triplify – Light-Weight Linked Data Publication from Relational Databases. *Proceedings of the 18th International Conference on World Wide Web, Madrid*, páginas 621–630. doi: 10.1145/1526709.1526793. URL <http://www2009.org/proceedings/pdf/p621.pdf>. Citado na pág. 21
- Heße Kushmerick(2003)** Andreas Heße Nicholas Kushmerick. Learning to Attach Semantic Metadata to Web Services. *The Semantic Web - ISWC 2003*, 2870:258–273. ISSN 03029743. doi: 10.1007/978-3-540-39718-2_17. Citado na pág. 20
- Heß et al.(2004)** Andreas Heß, Eddie Johnston e Nicholas Kushmerick. Assam: A tool for semi-automatically annotating semantic web services. *3rd International Semantic Web Conference (ISWC 2004)*. ISSN 03029743. doi: 10.1007/978-3-540-30475-3_23. URL <http://www.springerlink.com/index/a3g85u9abhjxtbk5.pdf>. Citado na pág. 20
- Heß, Andreas e Dieric(2008)** Christian MaaßHeß, Andreas e Francis Dieric. From Web 2.0 to Semantic Web: A Semi-Automated Approach. *CEUR Workshop Proceedings*, 351:20–34. ISSN 16130073. URL http://www.ibiblio.org/hhalpin/homepage/notes/cisweb2008_submission_6.pdf. Citado na pág. 20, 50
- Hevner et al.(2004)** Hevner, March, Park e Ram. Design Science in Information Systems Research. *MIS Quarterly*, 28(1):75. ISSN 02767783. doi: 10.2307/25148625. URL <http://www.jstor.org/stable/10.2307/25148625>. Citado na pág. 5

- Jain et al.(2011)** Ankita Jain, Ilyas Khan e Bhupendra Verma. Secure and Intelligent Decision making in Semantic web mining. *International Journal of Computer Applications*, 15(7):14–18. ISSN 09758887. doi: 10.5120/1962-2625. URL <http://www.ijcaonline.org/volume15/number7/pxc3872625.pdf>. Citado na pág. 8
- Kabir et al.(2014)** Sumaiya Kabir, Shamim Ripon, Mamunur Rahman e Tanjim Rahman. Knowledge-based Data Mining Using Semantic Web. *IERI Procedia*, 7:113–119. ISSN 22126678. doi: 10.1016/j.ieri.2014.08.018. URL <http://www.sciencedirect.com/science/article/pii/S2212667814000379>. Citado na pág. 8, 16, 18
- Kharlamov et al.(2013)** Evgeny Kharlamov, Ernesto Jiménez-Ruiz, Dmitriy Zheleznyakov, Dimitris Bilidas, Martin Giese, Peter Haase, Ian Horrocks, Herald Killapi, Manolis Koubarakis, Özgür Özçep, Mariano Rodríguez-Muro, Riccardo Rosati, Michael Schmidt, Rudolf Schlatte, Ahmet Soylu e Arild Waaler. Optique: Towards OBDA systems for industry. Em *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 7955 LNCS, páginas 125–140. Springer Berlin Heidelberg. ISBN 9783642412417. doi: 10.1007/978-3-642-41242-4_11. URL http://link.springer.com/10.1007/978-3-642-41242-4_{_}11. Citado na pág. 20
- Klischewski(2006)** Ralf Klischewski. Semantic Web for e-Government. *Electronic Government*, 2739(June):288–295. ISSN 03029743. doi: 10.1007/10929179_52. URL <http://www.springerlink.com/openurl.asp?genre=article&id=69RMTUDPR46Q5B9N>. Citado na pág. 1
- Kushmerick et al.(1997)** Nicholas Kushmerick, Daniel S Weld e R Doorenbos. Wrapper induction for information extraction, 1997. URL citeseer.ist.psu.edu/kushmerick97wrapper.html. Citado na pág. 17
- Lembo et al.(2014)** Domenico Lembo, Jose Mora, Riccardo Rosati, Domenico Fabio Savo e Evgenij Thorstensen. Towards Mapping Analysis in Ontology-Based Data Access. Em Roman Kontchakov e Marie-Laure Mugnier, editors, *Web Reasoning and Rule Systems: 8th International Conference, RR 2014, Athens, Greece, September 15-17, 2014. Proceedings*, páginas 108–123. Springer International Publishing, Cham. ISBN 978-3-319-11113-1. doi: 10.1007/978-3-319-11113-1_8. URL http://link.springer.com/10.1007/978-3-319-11113-1_{_}8. Citado na pág. 20, 21
- May(2002)** Wolfgang May. Linking the Semantic Web with existing sources. *Proceedings - International Workshop on Database and Expert Systems Applications, DEXA*, 2002-January:93–97. ISSN 15294188. doi: 10.1109/DEXA.2002.1045882. Citado na pág. 16, 20, 23, 50
- Mika(2015)** Peter Mika. On Schema.org and Why It Matters for the Web. *Internet Computing, IEEE*, 19(4):52–55. ISSN 1089-7801. doi: 10.1109/MIC.2015.81. Citado na pág. 3, 10, 25
- Mustapaşa et al.(2010)** Oğuz Mustapaşa, Dilek Karahoca, Adem Karahoca, Ahmet Yücel e Hüseyin Uzunboylu. Implementation of Semantic Web Mining on E-Learning. *Procedia - Social and Behavioral Sciences*, 2(2):5820–5823. ISSN 18770428. doi: 10.1016/j.sbspro.2010.03.949. Citado na pág. 1
- Nardin et al.(2011)** Luis G. Nardin, Anarosa a F Brandão e Jaime S. Sichman. Experiments on semantic interoperability of agent reputation models using the SOARI architecture. *Engineering Applications of Artificial Intelligence*, 24(8):1461–1471. ISSN 09521976. doi: 10.1016/j.engappai.2011.05.004. Citado na pág. 20
- Noy e McGuinness(2001)** Natalya F. Noy e Deborah L. McGuinness. Ontology Development 101: A Guide to Creating Your First Ontology. *Stanford Knowledge Systems Laboratory*, página 25. ISSN 09333657. doi: 10.1016/j.artmed.2004.01.014. Citado na pág. 3, 9

- Peffers et al.(2007)** Ken Peffers, Tuure Tuunanen, Marcus A. Rothenberger e Samir Chatterjee. A Design Science Research Methodology for Information Systems Research. *Journal of Management Information Systems*, 24(3):45–77. ISSN 0742-1222. doi: 10.2753/MIS0742-1222240302. URL <http://www.tandfonline.com/doi/full/10.2753/MIS0742-1222240302>. Citado na pág. 5, 6
- Pérez et al.(2006)** Jorge Pérez, Marcelo Arenas e Claudio Gutierrez. Semantics and Complexity of SPARQL. Em *The Semantic Web - ISWC 2006*, volume 4273, páginas 30–43. "Springer Berlin Heidelberg". ISBN 978-3-540-49055-5. doi: 10.1007/11926078_3. URL http://www.springerlink.com/index/10.1007/11926078_3. Citado na pág. 15, 16
- Putra e Khalil(2017)** Syopiansyah Jaya Putra e Ismail Khalil. Context for the intelligent search of information. Em *2017 5th International Conference on Cyber and IT Service Management (CITSM)*, páginas 1–4. IEEE. ISBN 978-1-5386-2739-6. doi: 10.1109/CITSM.2017.8089317. URL <http://ieeexplore.ieee.org/document/8089317/>. Citado na pág. 19
- Quboa e Saraee(2013)** Qudamah K Quboa e Mohamad Saraee. A State-of-the-Art Survey on Semantic Web Mining. *Intelligent Information Management*, 05(January):10–17. ISSN 2160-5912. doi: 10.4236/iim.2013.51002. URL <http://dx.doi.org/10.4236/iim.2013.51002><http://www.scirp.org/journal/iimhttp://www.scirp.org/journal/PaperDownload.aspx?DOI=10.4236/iim.2013.51002>. Citado na pág. 16, 17
- RDF Working Group(2014)** RDF Working Group. RDF - Semantic Web Standards, 2014. URL <http://www.w3.org/RDF/>. Citado na pág. 20
- Rodríguez-Muro et al.(2013)** Mariano Rodríguez-Muro, Roman Kontchakov e Michael Zakharyashev. Ontop at work. Em *PROC. OF OWL: EXPERIENCES AND DIRECTIONS WORKSHOP 2013 (OWLED 2013)*. CEUR-WS. Citado na pág. 21
- Singh e Singh(2010)** Brijendra Singh e Hemant Kumar Singh. Web Data Mining research: A survey. *IEEE International Conference on Computational Intelligence and Computing Research*, páginas 1–10. ISSN 19310145. doi: 10.1109/ICCIC.2010.5705856. URL <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5705856>. Citado na pág. 16
- Sleiman et al.(2011)** Hassan a. Sleiman, Carlos R. Rivero e Rafael Corchuelo. On a proposal to integrate web sources using semantic-web technologies. Em *2011 7th International Conference on Next Generation Web Services Practices*, páginas 326–331. IEEE. ISBN 978-1-4577-1127-5. doi: 10.1109/NWeSP.2011.6088199. URL <http://ieeexplore.ieee.org/document/6088199/>. Citado na pág. 16
- Stumme et al.(2006)** G Stumme, A Hotho e B Berendt. Semantic Web Mining: State of the art and future directions. *Web Semantics: Science, Services and Agents on the World Wide Web*, 4: 124–143. ISSN 15708268. doi: 10.1016/j.websem.2006.02.001. Citado na pág. 1, 4, 7, 16, 17, 18
- Sui et al.(2009)** Xin Sui, Suozhu Wang e Zhaowei Li. Research on the model of Integration with Semantic Web and Agent Personalized Recommendation System. Em *2009 13th International Conference on Computer Supported Cooperative Work in Design*, páginas 233–237. IEEE. ISBN 978-1-4244-3534-0. doi: 10.1109/CSCWD.2009.4968064. URL <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4968064>. Citado na pág. 20
- Tort e Olivé(2014)** Albert Tort e Antoni Olivé. A Computer-Guided Approach to Website Schema.org Design. Em Eric Yu, , Gillian Dobbie, , Matthias Jarke, e Sandeep Purao, editors, *Conceptual Modeling: 33rd International Conference, ER 2014, Atlanta, GA, USA, October 27-29, 2014. Proceedings*, páginas 28–42. Springer International Publishing. ISBN 9783319122052. doi: 10.1007/978-3-319-12206-9_3. URL http://link.springer.com/10.1007/978-3-319-12206-9_3. Citado na pág. 10

- Vettor et al.(2014)** Pierre De Vettor, Michael Mrissa, Djamal Benslimane e Salim Berbar. A Service Oriented Architecture for Linked Data Integration. Em *2014 IEEE 8th International Symposium on Service Oriented System Engineering*, páginas 198–203. IEEE. ISBN 978-1-4799-3616-8. doi: 10.1109/SOSE.2014.30. URL <http://www.scopus.com/inward/record.url?eid=2-s2.0-84903640868&partnerID=tZOtx3y1>. Citado na pág. 19, 20
- W3C(2013)** W3C. SPARQL Query Language for RDF, 2013. URL <http://www.w3.org/TR/rdf-sparql-query/>. Citado na pág. 15, 20
- W3C(2008)** W3C. Extensible markup language (xml) 1.0 (fifth edition), 2008. URL <https://www.w3.org/TR/2008/REC-xml-20081126/>. Citado na pág. 13
- Wang et al.(2017)** Anduo Wang, Jason Croft e Eduard Dragut. Reflections on Data Integration for SDN. Em *Proceedings of the ACM International Workshop on Security in Software Defined Networks & Network Function Virtualization - SDN-NFVSec '17*, páginas 65–68, New York, New York, USA. ACM Press. ISBN 9781450349086. doi: 10.1145/3040992.3041006. URL <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85018293907&doi=10.1145/3040992.3041006&partnerID=40&md5=16dcd67b46c6e6a0b6f69717fd715ec2http://dl.acm.org/citation.cfm?doid=3040992.3041006>. Citado na pág. 19
- Wang et al.(2009)** Jinpeng Wang, Jianjiang Lu, Yafei Zhang, Zhuang Miao e Bo Zhou. Integrating heterogeneous data source using ontology. *Journal of Software*, 4(8):843–850. ISSN 1796217X. doi: 10.4304/jsw.4.8.843-850. Citado na pág. 11, 13, 19
- Yong-gui e Zhen(2010)** Wang Yong-gui e Jia Zhen. Research on Semantic Web Mining. *Computer Design*, 1(Iccda):67–70. doi: 10.1109/ICCDA.2010.5541057. URL http://ieeexplore.ieee.org/xpl/freeabs{__}all.jsp?arnumber=5541057. Citado na pág. 4, 7
- Zhang(2011)** Haiyang Zhang. The Research of Web Mining in E-Commerce. Em *2011 International Conference on Management and Service Science*, volume 3, páginas 1–4. IEEE. ISBN 978-1-4244-6579-8. doi: 10.1109/ICMSS.2011.5999196. URL <http://ieeexplore.ieee.org/document/5999196/>. Citado na pág. xiii, 17, 18