1
Introdução

1.1
Linked Data

Nos últimos anos, a Web está evoluindo da Web de documentos para a Web de Dados. Este processo de evolução teve inicio quando os princípios da *Linked Data* foram formulados com a visão de criar um espaço de informação global e conectado [1].

A meta de integrar dados semanticamente similares através do estabelecimento de *links* entre recursos relacionados é especialmente perseguida pela *Linking Open Data Iniciative* (LODI). A LODI promove e suporta a interconexão de dados do tipo *Linked Open Data* (LOD) a partir de várias fontes de dados e domínios. Seu principal objetivo é abrir silos de dados, de forma a publicar seu conteúdo [2]. Para conseguir este objetivo, a LODI usa um mecanismo de intercâmbio de dados estruturados via a criação de *links* do tipo *Resource Description Framework* (RDF). Tais *links* são criados entre entidades relacionadas usando identificadores únicos de recursos (URI - *Uniform Resource Identifier*) dereferenciáveis [3].

Seguindo a LODI, as organizações e comunidades de diferentes domínios começaram a publicar dados de maneira intensa, resultando em um crescimento excessivo da nuvem de LOD [4]. A maior parte dos dados publicados como LOD é encontrada na forma de catálogos Web. Estes dados estão disponíveis em formatos como CSV, JON, PDF, XML, RSS, XLS, XLSX, os quais representam dados brutos do usuário. Temos também dados em RDF que optam pelos formatos N3, Turtle e RDF/XML. A maioria dos dados governamentais são publicados como dados brutos, de tal forma que o processo de análise e extração de informação seja muito difícil e custoso. Para enfrentar este desafio é necessário o uso de abordagens de suporte semântico [4].

Os líderes dos Estados Unidos (*United States*) e o Reino Unido (*United Kingdom*) em parceria com o governo e acadêmicos, aproveitando muitas das tecnologias Web baseadas nos princípios de *Linked Data*, começaram um processo de publicação massivo [5, 6, 7]. Transformando muitos conjuntos

de dados publicados como dados brutos em formato RDF, para eles serem publicados dentro de `data.gov` [4, 7, 8]. Tudo isto, seguindo a abordagem de *Linked Open Government Data* (LOGD) [5, 6, 7]

Dentro do processo de publicação definido por Villazón em [8], recomenda-se conectar os dados entre diferentes fontes através de recursos similares que descrevem um domínio em comum. Têm-se diferentes abordagens para enfrentar este desafio:

Consultar iterativamente a fonte de dados: Esta abordagem utiliza a linguagem SPARQL [9] para identificar possíveis recursos similares. Esta abordagem é muito custosa pela quantidade de dados disponíveis nestas fontes de dados e pela complexidade das consultas executadas para recuperar os dados [2, 10]. Deve-se considerar também nesta abordagem as restrições que os responsáveis das fontes de dados aplicam sobre elas[2].

Motores de busca semânticos na Web: Esta abordagem é utilizada para encontrar fontes de dados através das palavras-chave coincidentes com os tópicos de interesse indexados pelos motores de busca. Nesta abordagem tem-se que tratar a ambiguidade e a quantidade considerável de dados recuperados pelo motor de busca usado [11, 12].

***Frameworks* de descobrimentos de *links* baseados em abordagens de *ontology matching*:** Estes *frameworks* são limitados à comparações entre dois conjuntos de dados e que dependem diretamente do tamanho do conjuntos de dados e da complexidade das métricas usadas para medir a similaridade dos dados [13, 14].

1.2 Motivação

A motivação principal deste trabalho é desenvolver uma abordagem de referência, baseada em técnicas de processamento distribuído e indexação dados, para recomendar fontes de dados RDF, deste modo melhorar o processo de publicação, interligação e exploração de dados na LOD.

A abordagem apresentada mostra como dados publicados na LOD, ou dados que pretendem ser publicados, podem ser acessados usando diferentes tecnologias disponíveis na Web. Esta abordagem permite o processamento de complexas e diversas quantidades de dados e a geração de metadados e estatísticas para o entendimento dos dados. Levando em conta o esforço computacional que uma solução deste tipo demanda, em nosso trabalho iremos aproveitar as vantagens oferecidas pelo paradigma MapReduce.

Neste trabalho propomos um processo de recomendação de fontes RDF é para orientar a busca de fontes relevantes RDF [15]. Para descrever o processo desenvolvemos um estudo de caso que utiliza uma fonte de dados publicada na Web com conteúdo acadêmico. Mostraremos um cenário prático de como usuário pode interagir com as diversas funcionalidades oferecidas pela nossa ferramenta desenvolvida, sem precisar ter conhecimento específico dos dados publicados ou explorados, conseguindo uma interligação dos tópicos de interesse.

A abordagem proposta utiliza o paradigma de computação na nuvem. Este estilo de computação fornece uma abstração útil que ajuda a focar na tarefa em questão (processamento paralelo e distribuído, armazenamento ou indexação de dados). A abordagem usada torna possível solicitar recursos computacionais de um provedor de serviços, sem necessidade de se preocupar com os detalhes da oferta computacional.

Existem poucas pesquisas que tentam resolver o problema de recomendação de fontes RDF [15, 16] e nenhum desses trabalhos considera a computação de dados intensivos na nuvem do inglês *Data Intensive Computing Cloud* [17]. Também não encontramos uma ferramenta para resolver esta necessidade, em especial no ecossistema governamental.

1.3 Objetivos

O objetivo principal deste trabalho consiste na criação de uma abordagem para recomendar fontes de dados RDF como parte do processo de publicação, interligação e exploração de dados na LOD. Em detalhe, os objetivos são:

Integrar um conjunto de componentes disponíveis na Web: Para a construção da nossa abordagem proposta neste documento, precisa-se integrar um conjunto de componentes disponíveis na Web sobre um fluxo de operações, onde eles consigam interagir de forma independente. Assim, foi desenvolvido uma ferramenta baseada em uma arquitetura que permite a comunicação dos diferentes componentes.

Gerar estatísticas de fontes de dados RDF: Gerar estatísticas do vocabulário usado pela fonte de dados em RDF em questão. Conseguindo assim um melhor entendimento da fonte, ajudando o usuário a escolher os recursos representativos que serão interligados.

Definir estratégias para a extração de informação: É necessário o processamento e análise textual das entidades de interesse de modo a extrair as estruturas representativas que ajudarão na busca.

Definir um processo de recomendação de fontes RDF: Definir um processo que guie o usuário na busca de fontes RDF através de uma interface amigável. Assim conseguiremos integrar o responsável pela publicação e o consumidor dos dados em um mesmo fluxo de operações.

Desenvolver uma ferramenta baseada em computação na nuvem: Tanto um processo de busca de dados, quanto um processo de análise das entidades de interesse demandam alto custo computacional. Para lidar com grandes quantidades de dados é indispensável o uso de processamento paralelo e distribuído.

1.4 Organização da dissertação

Este trabalho está organizado da seguinte forma. No capítulo 2 apresentamos os conceitos básicos usados neste trabalho. No capítulo 3 descrevemos os diferentes trabalhos relacionados com a conexão (do inglês *interlinking*) de dados. No capítulo 4 apresentamos a abordagem para o processo de recomendação de fontes RDF e um estudo de caso. No capítulo 5 apresentamos o resultado de cada uma das etapas do processo de recomendação. No capítulo 6 apresentamos a implementação de uma ferramenta que dá suporte ao processo de recomendações de fontes RDF. Finalmente, no capítulo 7 listamos as conclusões do trabalho e indicamos as possibilidades para trabalhos futuros.