Consultas federadas sobre dados abertos conectados

Federated queries on connected open data

DOI:10.34117/bjdv7n1-451

Recebimento dos originais: 15/12/2020
Aceitação para publicação: 15/01/2021

Gabriel Lucas Pimenta
Formação acadêmica mais alta: Graduação em Ciência da Computação
Instituição: UNICENTRO
Endereço: Alameda Élio Antonio Dalla Vecchia, 838, Vila Carli, Guarapuava PR
E-mail: p1ment4_1337@hotmail.com

Gisane Aparecida Michelon
Doutora em Informática
Instituição: PUC PR
Endereço: Alameda Élio Antonio Dalla Vecchia, 838, Vila Carli, Guarapuava - PR
E-mail: gisane@unicentro.br

Lúcelia de Souza
Doutora em Informática
Instituição: UFPR
Endereço: Alameda Élio Antonio Dalla Vecchia, 838, Vila Carli, Guarapuava - PR
E-mail: luceliasz@yahoo.com.br

Josiane Michalak Hauagge Dall'Agnol
Doutora em Engenharia Elétrica e Informática Industrial
Instituição: UTFPR
Endereço: Alameda Élio Antonio Dalla Vecchia, 838, Vila Carli, Guarapuava - PR
E-mail: josianehauagge@gmail.com

Sandro Rautenberg
Doutor em Engenharia do conhecimento
Instituição: UFSC
Endereço: Alameda Élio Antonio Dalla Vecchia, 838, Vila Carli, Guarapuava - PR
E-mail: srautenberg@unicentro.br

RESUMO

Dados abertos conectados podem ser livremente acessados, modificados e compartilhados. Para melhorar a tomada de decisão, a otimização de processos e a descoberta de novas tendências, esses dados podem ser utilizados com a aplicação de novos métodos de processamento. O processamento de modo distribuído de consultas em várias bases é chamado de consultas federadas. Como ferramentas para o processamento, o gerenciamento e a pesquisa das consultas podem ser utilizadas a SPARQL (*Protocol and RDF Query Language*), o OpenLink Virtuoso e o Hadoop. A linguagem SPARQL utiliza arquivos em RDF (*Resource Description Framework*) para representar recursos na web para o processamento de consultas federadas. O OpenLink Virtuoso é um Sistema Gerenciador de Banco de Dados com várias funcionalidades sendo conhecido como servidor universal. O Hadoop utiliza o MapReduce para processar um conjunto de dados

e para fazer o armazenamento distribuído. Para a realização das consultas foram escolhidas como base de dados os índices de classificação de periódicos, tais como, o Qualis, da CAPES e a DBpedia, a qual extrai conteúdos estruturados contidos na Wikipedia. Na DBpedia a ideia é buscar periódicos que possuem o Fator de Impacto (FI) baseado em outros índices como o JCR (*Journal Citation Reports*) para a comparação com o índice Qualis. Essas duas bases, o Qualis e a DBpedia podem ser integradas mediante o ISSN (*International Standart Serial Number*) com o objetivo de gerar novas informações que auxiliam na tomada de decisões. Neste trabalho foi relacionado os índices registrados em ambas as bases de dados, integrando-as através do ISSN, comparando a classificação de periódicos Qualis com o Fator de Impacto da DBpedia. Isso resultou na possibilidade de uso de mais de um índice de classificação de periódicos para pontuação das publicações dos pesquisadores, pois nem sempre um periódico está classificado em todos os índices.

Palavras-chave: Dados abertos conectados, processamento distribuído e SPARQL.

ABSTRACT

Open connected data can be freely accessed, modified and shared. To improve decision making, process optimization and the discovery of new trends, this data can be used by applying new processing methods. Distributed mode processing of queries on various bases is called federated queries. SPARQL (Protocol and RDF Query Language), OpenLink Virtuoso and Hadoop can be used as tools for processing, managing and searching queries. The SPARQL language uses files in RDF (Resource Description Framework) to represent web resources for processing federated queries. OpenLink Virtuoso is a multi-functional database management system known as a universal server. Hadoop uses MapReduce to process a data set and to do distributed storage. To perform the queries, the database was chosen from the classification indexes of journals, such as Qualis, from CAPES and DBpedia, which extracts structured contents contained in Wikipedia. In DBpedia the idea is to search for journals that have the Impact Factor (FI) based on other indexes such as JCR (Journal Citation Reports) for comparison with the Qualis index. These two bases, Qualis and DBpedia can be integrated through the ISSN (International Standard Serial Number) in order to generate new information that helps in decision making. In this work, the indexes registered in both databases were listed, integrating them through the ISSN, comparing the Qualis journals classification with the DBpedia Impact Factor. This resulted in the possibility of using more than one journal classification index to score the researchers' publications, since not always a journal is classified in all the indexes.

Keywords: Open data connected, distributed processing and SPARQL.

1 INTRODUÇÃO

Dados abertos conectados são dados que podem ser livremente acessados, modificados e compartilhados, para qualquer propósito, sendo apenas necessário creditar a autoria dos dados e distribuí-los usando a mesma licença [Rautenberg et al, 2018]. Porém, nem todo dado conectado é aberto. or exemplo, uma entidade privada pode conectar seus dados com outras, sem torná-los abertos. Portanto, dados abertos conectados são os ideais para publicar dados na web, sendo a combinação das

características dos dados abertos com os dados conectados [Rautenberg et al, 2018]. Um esquema de classificação que identifica a quantidade, a maturidade e a qualidade dos dados abertos na web foi criado por Tim Berners-Lee [2011] sendo classificado em Cinco Estrelas, e de de forma progressiva.

Dados abertos são caracterizados por dois aspectos, de acordo com [Open Knowledge, 2020]:

- Abertura legal: a licença dos dados precisa permitir o seu livre acesso, reúso e redistribuição.
- Abertura técnica: os dados devem estar em formatos legíveis por máquinas, sendo possível acessá-los por uma faixa de seleção. Não deve haver cobrança pelo uso dos dados. Se existir algum custo, deve ser apenas referente à sua reprodução, mas os dados devem estar disponíveis em um formato de especificação aberto.

Assim, esses dados podem ser utilizados com a aplicação de novos métodos de processamento para melhorar a tomada de decisão, otimização de processos e descoberta de novas tendências [Hurwitz et al, 2013]. O processamento de consultas de modo distribuído em várias bases é chamado de consultas federadas. Nesse contexto, este trabalho teve com um dos objetivos estudar ferramentas para o processamento distribuído de dados abertos conectados.

Motivados pelo sucesso da iniciativa *Linking Open Data* e pelo grande crescimento da quantidade de fontes de dados disponíveis na web, novas abordagens de processamento de consultas estão surgindo. Enquanto o processamento de consultas no contexto do modelo RDF (*Resource Description Framework*) era tradicionalmente realizado usando armazenamento centralizado, ultimamente pode-se observar uma mudança de paradigma [Schwarte et al, 2011] para a adoção de abordagens federadas em decorrência da estrutura descentralizada da web. Na prática, inúmeros cenários existem em que mais de uma fonte de dados pode contribuir com informações, tornando o processamento de consultas mais complexo. O caminho natural segue a busca por soluções eficientes para o processamento de consultas federadas sobre fontes de dados distribuídas na web.

As consultas podem ser executadas em várias bases, para que assim calcule um conjunto de resultados por meio de dados que são disponibilizados pelas bases, essas conhecidas também como *endpoints* de SPARQL. Buil-Aranda *et. al* (2011) afirmam que o processamento de consultas SPARQL distribuídas sobre *Linked Data* é uma tarefa complexa.

Há vários tipos de consultas federadas, nesta pesquisa o foco será em SPARQL

federado. Nesta, uma consulta de entrada é enviada a um analisador, o qual verifica completamente a consulta para analisar se é ou não SPARQL, até mesmo reescrevendo-a, se necessário. Então, a consulta é encaminhada para uma abordagem de seleção da fonte de padrão triplo que são consultados por meio de um federador, e em seguida mesclado e retornado.

Neste trabalho, para a realização das consultas foram escolhidas como base de dados os índices de classificação de periódicos, tais como, o Qualis, da CAPES e a DBpedia, a qual extrai conteúdos estruturados contidos na Wikipedia. Na DBpedia a ideia é buscar periódicos que possuem o Fator de Impacto (FI) baseado em outros índices como o JCR (*Journal Citation Reports*) para a comparação com o índice Qualis. Essas duas bases, o Qualis e a DBpedia podem ser integradas mediante o ISSN (*International Standart Serial Number*) com o objetivo de gerar novas informações que auxiliam na tomada de decisões.

2 MATERIAIS E MÉTODOS

Em todas as fases do projeto, esta pesquisa foi realizada em conjunto com o grupo de pesquisa LASeD (Laboratório de Aplicações Semânticas e Distribuídas) da Unicentro. Foram identificadas algumas ferramentas para estudo: Hadoop, OpenLink Virtuoso e SPARQL. A seguir uma descrição de cada uma delas.

O Hadoop é uma ferramenta *open-source*, utilizada para processar conjuntos de dados e também para armazenamento distribuído, ambos usando MapReduce. O MapReduce é um modelo de programação, onde há um conjunto de bibliotecas que são utilizadas para processar os dados em paralelo. Este modelo diz respeito a duas tarefas distintas, sendo *Mapper* e *Reducer*, onde é feito o mapeamento e a validação dos dados que se encontram no HDFS (*Hadoop Distributed File System)* e tem como entrada o resultado da fase de mapeamento [Apache MapReduce, 2020]. Estes dados podem ser processados por meio de *clusters* de computadores.

O OpenLink Virtuoso é um Sistema Gerenciador de Banco de Dados (SGBD) multiparadigma que apresenta várias funcionalidades como [OpenLink Software, 2020]:

- Gerenciamento de tabelas de dados relacionais.
- Gerenciamento de conteúdo.
- Serviços web e de documentos.
- Implantação de dados abertos conectados no nível Cinco Estrelas.
- Gerenciamento de propriedades de grafos RDF.

- Servidor de aplicações web.

Como há várias funcionalidades, o OpenLink Virtuoso é conhecido como um servidor universal, sendo distribuído em duas versões: uma comercial e uma gratuita *open source*, que apresenta as mesmas funcionalidades da versão comercial, exceto a replicação de dados e o motor de base de dados virtual [OpenLink Software, 2020].

A linguagem SPARQL fornecer protocolos e conjuntos de especificações, que manipulam e consultam arquivos em RDF (*Resource Description Framework*), por meio de correspondência. Ela segue um padrão de triplas que permitem ao usuário fixar uma ou mais características, gerando como resultado um conjunto de elementos que cumpre o padrão especificado [DuCharme, 2013].

Com o surgimento da versão 1.1 da SPARQL, tornou-se possível formular uma consulta federada que procura informações em diversas fontes de dados. Desta forma, a federação de consultas baseia-se no processamento distribuído de consultas de múltiplas fontes de dados autônomas. Para isso, é necessário o conhecimento prévio da localização, o vocabulário heterogêneo e como combinar os dados das várias fontes. Além disso, uma determinada consulta federada precisa ser decomposta em subconsultas, onde cada subconsulta é enviada a um serviço de consulta específico (*endpoint*). A seguir, os resultados obtidos são integrados e a resposta final é enviada ao usuário [Macedo et al, 2012].

Os grafos RDF servem para representar recursos na web, onde as consultas são processadas por aplicações e não apenas visualizadas, modelando declarações em nós e arcos, ou em triplas, sendo sujeito, predicado e objeto. Para isso, é usado XML - RDF/XML e também um vocabulário, chamado de RDF *Schema*, que é composto de classes, propriedades e restrições. Portanto, o RDF *Schema* especifica os relacionamentos entre classes. Há mecanismos em RDF que possibilitam a descrição de grupos de recursos, estes chamados de *containers*, os quais são recursos que possuem objetos, chamados de membros, sendo recursos ou literais. Recursos descritos como um *container* recebem uma propriedade cujo valor determina seu tipo. Podem ser divididos em três diferentes tipos [Lima et al, 2005]:

- *Bag*, representando um grupo de recursos ou de valores não ordenados.
- *Sequence* ou Seq, representando um grupo de recursos ou de valores ordenados.
- *Alternative*, representando um grupo de recursos ou valores que apresentam valores possíveis para uma propriedade.

Em conjunto com o OpenLink Virtuoso, a linguagem SPARQL será utilizada para

gerar grafos e consultas, e então processar todos esses conjuntos de dados por meio das consultas federadas.

Após a fase de estudo das ferramentas foram definidas as bases de dados para realizar as consultas federadas: os índices de classificação de periódicos, tais como, o Qualis e a DBpedia. A seguir são descritas essas duas bases de dados:

O Qualis é um sistema brasileiro de avaliação de periódicos oferecida pela CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) a toda comunidade. Refere-se ao sistema de classificação de periódicos nacionais e internacionais, representando a produção intelectual dos programas de pós-graduação brasileira de todas e quaisquer áreas do conhecimento. Esses periódicos são avaliados quanto ao âmbito da circulação (local, nacional ou internacional) e à qualidade (A, B, C), por área de avaliação [Wikipedia, 2020, Qualis, 2020].

A DBpedia é um projeto que tem por objetivo extrair conteúdo estruturado das informações contidas na Wikipedia, que em seguida é disponibilizada na Web. Essa base de dados cria uma rede de ligações entre os dados, permitindo então, ao usuário, realizar consultas sobre o conteúdo da Wikipedia de forma similar a consultas de banco de dados. Neste caso específico, as consultas serão realizadas por meio da linguagem SPARQL, que é responsável pelo processamento de toda esta quantia de dados. Na DBpedia estão disponíveis outros índices de classificação de periódicos através do Fator de Impacto (FI). Ele mede o número médio de citações de artigos científicos publicados em um determinado periódico. É empregado frequentemente para avaliar a importância de um dado periódico em sua área, sendo que aqueles com um maior FI são considerados mais importantes do que aqueles com um menor FI [DBpedia, 2018].

A integração das bases é feita através da variável ISSN (*International Standard Serial Number*), que ambas as bases de dados possuem, comparando os periódicos do DBpedia que tem o Fator de Impacto ao sistema de classificação de periódicos Qualis. As bases de dados estão disponíveis no *endpoint* http://lod.unicentro.br/QualisBrasil/, que disponibiliza o índice Qualis e o *endpoint* da DBpedia http://dbpedia.org, que disponibiliza o Fator de Impacto.

3 RESULTADOS E DISCUSSÃO

A linguagem SPARQL foi utilizada para realizar as consultas nas bases de dados definidas na fase anterior. Para isso, foi estudado o funcionamento desta linguagem e como realizar as consultas, desde as mais básicas até as federadas [World Wide Web

Consortion, 2018a, World Wide Web Consortion, 2018b].

A seguir serão apresentadas algumas das consultas que foram executadas:

Consulta 1. Descrição: esta consulta pesquisa o nome e o ISSN dos periódicos nos dois *endpoints*, o Fator de Impacto na DBpedia e o índice Qualis.

Texto da Consulta 1:

```
PREFIX qualis: <http://lod.unicentro.br/QualisBrasil/>
PREFIX dc: <http://purl.org/dc/elements/1.1/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX bibo: <http://purl.org/ontology/bibo/>
PREFIX dbo:<http://dbpedia.org/ontology/>

SELECT DISTINCT ?issn ?name ?qualis ?issnDBPedia ?nameDBPedia ?impactFactor
COUNT(*) AS ?quantidade  WHERE {
?EvaluationQualis qualis:hasJournal ?JournalQualis .
?EvaluationQualis qualis:hasScore ?ScoreQualis .
?JournalQualis bibo:issn ?issn .
?JournalQualis foaf:name ?name .
?ScoreQualis rdf:value ?qualis.
SERVICE <http://dbpedia.org/sparql/> {
?journal dbo:issn ?issnDBPedia .
?journal foaf:name ?nameDBPedia .
?journal dbo:impactFactor ?impactFactor .
}
FILTER(?issnDBPedia = ?issn)
 }
GROUP BY ?issn ?name ?qualis ?issnDBPedia ?nameDBPedia ?impactFactor
limit 10000
```

O resultado da Consulta 1 é apresentado na Figura 1.

Figura 1: Resultado da Consulta 1.

issn	name	qualis	issnDBPedia	nameDBPedia	impactFactor
1460-7425	Journal of Artificial Societies and Social Simulation	B3	1460-7425	Journal of Artificial Societies and Social Simulation	1.733
1053-8135	Neurorehabilitation	A1	1053-8135	NeuroRehabilitation	1.736
1536-2442	Journal of Insect Science	B2	1536-2442	Journal of Insect Science	0.921
1673-5374	Neural Regeneration Research.	B5	1673-5374	Neural Regeneration Research	0.18
0218-2718	International Journal of Modern Physics D	B3	0218-2718	International Journal of Modern Physics A	1.343
1573-7373	Journal of Neuro-Oncology	A2	1573-7373	Journal of Neuro-Oncology	3.07
0022-2372	Journal of Mammalogy	A1	0022-2372	Journal of Mammalogy	1.558
0138-9130	Scientometrics	B2	0138-9130	Scientometrics	2.274
2164-554X	Human Vaccines & Immunotherapeutics	B1	2164-554X	Human Vaccines & Immunotherapeutics	3.643
1520-541X	Evolution & Development	B1	1520-541X	Evolution & Development	3.179
1475-4916	Homeopathy	B2	1475-4916	Homeopathy	0.758
0214-8358	Scientia Marina	A2	0214-8358	Scientia Marina	1.144
0218-2718	International Journal of Modern Physics D	A2	0218-2718	International Journal of Modern Physics E	0.455

Obs.: os resultados mostrados nas consultas são parciais, pois, limitou-se o número de linhas devido ao seu tamanho.

Consulta 2. Descrição: esta consulta pesquisa o nome, ISSN e o índice dos periódicos que estão no índice Qualis, mas não estão na DBpedia.

Texto da Consulta 2:

```
PREFIX qualis: <http://lod.unicentro.br/QualisBrasil/>
PREFIX dc: <http://purl.org/dc/elements/1.1/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX bibo: <http://purl.org/ontology/bibo/>
PREFIX dbo:<http://dbpedia.org/ontology/>
SELECT DISTINCT ?issnQualis ?nameQualis ?notaQualis WHERE {

?EvaluationQualis qualis:hasJournal ?JournalQualis .
?EvaluationQualis qualis:hasScore ?ScoreQualis .
?JournalQualis bibo:issn ?issnQualis .
?JournalQualis foaf:name ?nameQualis .
?ScoreQualis rdf:value ?notaQualis .

SERVICE <http://dbpedia.org/sparql/> {
?journal dbo:issn ?issnDBPedia .
?journal foaf:name ?nameDBPedia .
?journal dbo:impactFactor ?impactFactorDBPedia .

}
FILTER(?issnDBPedia = ?issnQualis)
}
limit 1000
```

O resultado da Consulta 2 é apresentado na Figura 2.

Figura 2: Resultado da Consulta 2.

issnQualis	nameQualis	notaQualis
"0001-5113"	"Acta Adriatica"	"B5"
"0001-5385"	"Acta Cardiologica"	"B2"
"0001-5504"	"Acta Cientifica Venezolana"	"B4"
"0001-5504"	"Acta Cientifica Venezolana"	"B5"
"0001-5512"	"Acta Clinica Belgica"	"B2"
"0001-5547"	"Acta Cytologica"	"B1"
"0001-5547"	"Acta Cytologica"	"B2"
"0001-5547"	"Acta Cytologica"	"B4"
"0001-5709"	"Acta Geologica Polonica"	"B1"
"0001-5938"	"Acta Leprologica"	"B3"
"0001-5962"	"Acta Mathematica"	"A1"
"0001-6160"	"Acta Metallurgica"	"B5"
"0001-6365"	"Acta Odontologica Venezolana"	"B2"
"0001-6365"	"Acta Odontologica Venezolana"	"B3"
"0001-6489"	"Acta Oto-Laryngologica"	"B1"
"0001-6489"	"Acta Oto-Laryngologica"	"B2"
"0001-656X"	"Acta Paediatrica Scandinavica"	"B3"
"0001-656X"	"Acta Paediatrica Scandinavica"	"B5"
"0001-6640"	"Acta Pediatrica Espanola"	"B3"
"0001-6772"	"Acta Physiologica Scandinavica"	"B1"
"0001-6896"	"Acta Psiquiátrica y Psicológica de América Latina"	"B1"
"0001-6969"	"Acta Scientiarum Mathematicarum"	"B4"
"0001-7051"	"ACTA THERIOLOGICA"	"B1"
"0001-7051"	"ACTA THERIOLOGICA"	"B2"
"0001-7051"	"ACTA THERIOLOGICA"	"B4"
"0001-706X"	"Acta Tropica"	"A1"
"0001-706X"	"Acta Tropica"	"A2"
"0001-706X"	"Acta Tropica"	"B1"

Poderia ser utilizada para classificar periódicos que estão no Qualis, mas não se encontram classificados no índice da DBpedia.

Na sequência foram realizadas três consultas. A diferença entre elas é que, no primeiro caso, os valores exibidos são somente da própria base de dados (*endpoint local* – lod.unicentro), enquanto a segunda é uma consulta federada que, através da função SERVICE obtém os dados da DBpedia.

A última consulta utiliza funções que realizam operações lógicas sobre os dados, exibindo as publicações contidas em ambas as bases de dados, sem repetições. Para tal, é preciso utilizar a função FILTER para selecionar os registros com ISSN igual nas duas bases de dados.

O objetivo é comparar as consultas na questão da velocidade e quantidade de informações exibidas. Quando necessário foi limitado o número de linhas exibidas na função LIMIT de forma a aproximar-se do valor limite de exibição de linhas para ambos os casos. Os resultados são exibidos a seguir.

Consulta 3. Descrição: esta consulta mostra o ISSN e o índice extraídos do *endpoint* local. A função LIMIT não foi usada. Isso é devido ao fato de que a consulta não sendo federada, não precisa de tal, ela tem a capacidade de retornar todos os valores principalmente pelo motivo de não ter o atraso da rede e assim ser mais rápida a sua execução. A Figura 3 mostra o resultado da Consulta 3.

Texto da Consulta 3:

```
PREFIX sjr: <http://lod.unicentro.br/SJR/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX bibo: <http://purl.org/ontology/bibo/>

SELECT DISTINCT ?issn ?indice WHERE {

?sjr sjr:hasJournal ?journalSJR .
?sjr sjr:hasScore ?scoreSJR .
?journalSJR bibo:issn ?issn .
?scoreSJR rdf:value ? indice .
 }
order by desc(?issn)
```

O resultado da Consulta 3 é apresentado na Figura 3.

Figura 3: Resultado da Consulta 3.

issn	indice
"8821-1127"	"0.1080"
"8756-9728"	"0.6530"
"8756-9728"	"0.9790"
"8756-9728"	"0.8240"
"8756-9728"	"0.2180"
"8756-9728"	"0.5920"
"8756-9728"	"1.4730"
"8756-971X"	"0.6870"
"8756-971X"	"0.4640"
"8756-971X"	"0.5240"
"8756-971X"	"0.5250"
"8756-971X"	"0.8470"
"8756-971X"	"0.6260"
"8756-971X"	"0.5810"
"8756-971X"	"0.7040"
"8756-971X"	"0.7510"
"8756-971X"	"0.6440"

Consulta 4. Descrição: nesta consulta a função SERVICE busca os dados da DBpedia. Neste caso, também não foi necessário limitar os dados, todos os registrados puderam ser exibidos. A Figura 4 mostra o resultado da Consulta 4.

Texto da Consulta 4:

```
PREFIX dbo:<http://dbpedia.org/ontology/>

SELECT DISTINCT ?issn ?fatorDeImpacto WHERE {

SERVICE <http://dbpedia.org/sparql/> {
?journal dbo:issn ?issn .
?journal dbo:impactFactor ?fatorDeImpacto .
}
}
```

O resultado da Consulta 4 é apresentado na Figura 4.

Figura 4: Resultado da Consulta 4.

issn	fatorDeImpacto
"0028-0836"	38.138
"1476-4687"	38.138
"0741-8329"	2.006
"1873-6823"	2.006
"0007-0963"	4.275
"1365-2133"	4.275
"0008-6223"	6.196
"1054-1500"	2.049
"1089-7682"	2.049
"0160-9009"	0.196
"1536-0334"	0.196
"0378-1119"	2.319
"1879-0038"	2.319
"2073-4425"	3.242
"0300-8126"	2.618

Consulta 5. Descrição: esta consulta mostra os índices por ISSN buscando nas duas bases de dados que tem o mesmo ISSN (base local lod.unicentro.br e DBpedia). A Figura 5 mostra o resultado da Consulta 5.

Texto da Consulta 5:

```
PREFIX qualis: <http://lod.unicentro.br/QualisBrasil/>
PREFIX sjr: <http://lod.unicentro.br/SJR/>
PREFIX dc: <http://purl.org/dc/elements/1.1/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX bibo: <http://purl.org/ontology/bibo/>
PREFIX dbo:<http://dbpedia.org/ontology/>

SELECT DISTINCT ?issn ?indice WHERE {

?sjr sjr:hasJournal ?journalSJR .
?sjr sjr:hasScore ?scoreSJR .
?journalSJR bibo:issn ?issn .
?scoreSJR rdf:value ?indice .

SERVICE <http://dbpedia.org/sparql/> {
?journal dbo:issn ?issnDBPedia .
}

FILTER(?issnDBPedia = ?issn)
}
LIMIT 238000
```

Como mostrado na Consulta 5, foi necessário o uso da função LIMIT (que limita o número de linhas retornadas na consulta). O limite encontrado foi de 238000. O primeiro valor utilizado na consulta foi com o limite de 10000. E, na sequência foram utilizados os valores nesta ordem: 100000 e 1000000, sendo que para este último ocorreu um *timeout* (tempo máximo de resposta da rede, após este tempo é retornado uma mensagem de erro e abortado a consulta). Então, foi realizada uma média entre 100000 e 1000000, e o novo valor utilizado foi 500000. Porém, novamente ocorreu *timeout*. E, assim sucessivamente, foram realizadas tentativas com a média sendo que com o valor de 238000 não ocorreu mais *timeout*.

O resultado da Consulta 5 é apresentado na Figura 5.

Figura 5: Resultado da Consulta 5.

issn	indice
"0145-2258"	"0.1420"
"0146-4833"	"1.6710"
"0151-9107"	"0.2820"
"0163-5999"	"0.3090"
"0306-3747"	"0.1050"
"0323-0465"	"0.2360"
"0559-8680"	"0.1070"
"0753-4973"	"0.4410"
"0001-4826"	"2.5950"
"0001-8392"	"6.4180"
"0001-8686"	"2.0640"
"0001-9720"	"0.3200"
"0002-7014"	"0.5860"
"0002-7316"	"1.3260"
"0002-7642"	"0.2840"
"0002-8312"	"1.6920"
"0002-9114"	"0.2970"
"0002-9254"	"0.6960"

4 CONSIDERAÇÕES FINAIS

Os dados abertos conectados são dados livres para utilização, apenas deve-se dar os devidos créditos, além de fazer uso da mesma licença. Para que se possa usufruir de dados de melhor qualidade, é possível processá-los por meio de consultas em várias bases, chamadas de consultas federadas.

O processamento dos dados ocorreu principalmente por meio das consultas SPARQL federadas, gerenciadas pelo OpenLink Virtuoso, ou seja, de modo distribuído entre o Qualis e a DBpedia. O principal objetivo da realização destas consultas foi de relacionar os índices registrados em ambas as bases de dados, integrando-as através do ISSN, comparando a classificação de periódicos Qualis com o Fator de Impacto da DBpedia. A execução de consultas federadas gera informações relevantes, as quais, só

são possíveis de serem obtidas com as bases integradas. Um exemplo disso é fazer uso de mais de um índice de classificação de periódicos para pontuação das publicações dos pesquisadores, pois nem sempre um periódico está classificado em todos os índices.

AGRADECIMENTOS

Agradecemos à Fundação Araucária que apoiou o desenvolvimento deste projeto.

REFERÊNCIAS

Apache MapReduce. A Programming paradigm that allows for massive scalability of unstructured data across hundreds or thousands of commodity clusters servers in an Apache Hadoop cluster. Disponível em: https://www.ibm.com/analytics/hadoop/mapreduce. Acessado em julho de 2020.

Berners-Lee, T. Linked data-design issues, 2009. Disponível em http://www.w3.org/DesignIssues/LinkedData.html. Acessado em julho de 2020.

Buil-Aranda, C., Arenas, M., and Corcho, O. Semantics and optimization of the sparql 1.1 federation extension. Proceedings of the 8th extended semantic web conference on The semantic web: research and applications - Vol II. Springer-Verlag, pp. 1–15, 2011.

Dbpedia. Disponível em http://wiki.dbpedia.org/. Acesso em janeiro de 2018.
Lima, J. C. e Carvalho, C. L. Resource Description Framework (RDF), 2005. Disponível em: http://www.inf.ufg.br/sites/default/files/uploads/relatorios-tecnicos/RT-INF_003-05.pdf. Acesso em julho de 2020.

DuCharme, B. Learning Sparql. "O'Reilly Media, Inc", 2013.

Hurwitz, Judith; Nugent, Alan; Halper, Fern; Kaufman, Marcia. Big Data For Dummies. 1St edition. For Dummies, 2013.

Lima, J. C. e Carvalho, C. L. Resource Description Framework (RDF), 2005. Disponível em: http://www.inf.ufg.br/sites/default/files/uploads/relatorios-tecnicos/RT-INF_003-05.pdf. Acessado em julho de 2020.

Macedo Sousa Maia, João Carlos Pinheiro, Regis Pires Magalhães, José Maria da Silva Monteiro Filho, Vânia Maria Ponte Vidal. Junções Adaptativas em Consultas Federadas sobre Linked Data Simpósio Brasileiro de Bancos de Dados - SBBD 2012, Short Papers, 2012.
Open Knowledge. Open Definition 2.1. Disponível em: http://opendefinition.org/od/2.1/en/ Acessado em 22 de dezembro de 2017.
OpenLink Software. OpenLink Virtuoso Home Page. Disponível em: https://virtuoso.openlinksw.com. Acesso em julho de 2020.

Rautenberg, S. et al. Guia prático para publicação de dados abertos conectados na web. Editora Appris, 2018.

Schwarte, K.A., J.R. Russell, J.L. Kovar, D.G. Morrical, S.M. Ensley, K.-J. Yoon, N.A. Cornick, and Y.-I. Yoon. Grazing management effects on sediment, phosphorus, and pathogen loading of streams in cool-season grass pastures. Journal of Environment Quality 40:1303-1313, 2011.

Qualis. Qualis Periódicos. Disponível em https://sucupira.capes.gov.br/sucupira/public/consultas/coleta/veiculoPublicacaoQualis/listaConsultaGeralPeriodicos.jsf. Acesso em setembro de 2020.

Wikipedia. Disponível em https://pt.wikipedia.org/wiki/Qualis. Acesso em julho de 2020.

World Wide Web Consortion. W3c - SPARQL Query Language for RDF. Disponível em: https://www.w3.org/TR/rdf-sparql-query/#basicpatterns. Acesso em fevereiro de 2018.

World Wide Web Consortion. W3c - SPARQL 1.1 Overview. Disponível em: https://www.w3.org/TR/2013/REC-sparql11-overview-20130321. Acesso em março de 2018.