
بسم الله الرحمن الرحيم

استخراج اطلاعات و دانش از حجم عظیم داده‌های RDF با ارسال پرس‌وجوهای SPARQL

شکوفه قالیبافان
تابستان ۱۳۹۴

فهرست مطالب

- مقدمه
- استخراج اطلاعات و مفاهیم آن
- معماری کلی یک سیستم استخراج اطلاعات
- استخراج اطلاعات و داده‌های عظیم
- حجم عظیم داده‌های RDF
  - ذخیره
  - پردازش
- معرفی روش‌ها و ابزار

مقدمه

- RDF
  - یک استاندارد نمایش داده از وب معنایی
- نگاشت و نمایش حقایق در قالب معنایی RDF
  - در نتیجه دست‌یافتن به اطلاعات مورد نیاز با ارسال پرس‌وجوهای SPARQL
- افزایش داده‌های RDF
  - مساله‌ای کلیدی برای توسعه وب معنایی
- ارائه روش مقیاس‌پذیر از مهمترین چالش‌های موجود در سازماندهی این داده‌ها

استخراج اطلاعات

- اطلاعات با ارزش در اسناد متنی
- میزان قابل توجه متون
- مشکل بودن یافتن اطلاعات مربوط و لازم از منابع مختلف
- ذخیره به شکل ساخت‌یافته‌تر در پایگاه دانش
  - مدیریت ساده‌تر
- تبدیل متن به اطلاعات قابل استفاده توسط ماشین

برخی کاربردهای مهم سیستم‌های استخراج اطلاعات

- ایجاد یک پایگاه دانش با دقت بالا
- ابهام‌زدایی
- جستجوی معنایی بر روی حجم انبوه مستندات
- استفاده در سیستم‌های پرسش و پاسخ
  - پاسخگویی به سوالات زبان طبیعی با توجه به موجودیت‌ها و روابط آن‌ها در سوالاتی مثل کی، کجا، چه موقع و ...
- سیستم‌های ترجمه ماشینی، سازماندهی، نگهداری و رشد خودکار پایگاه دانش

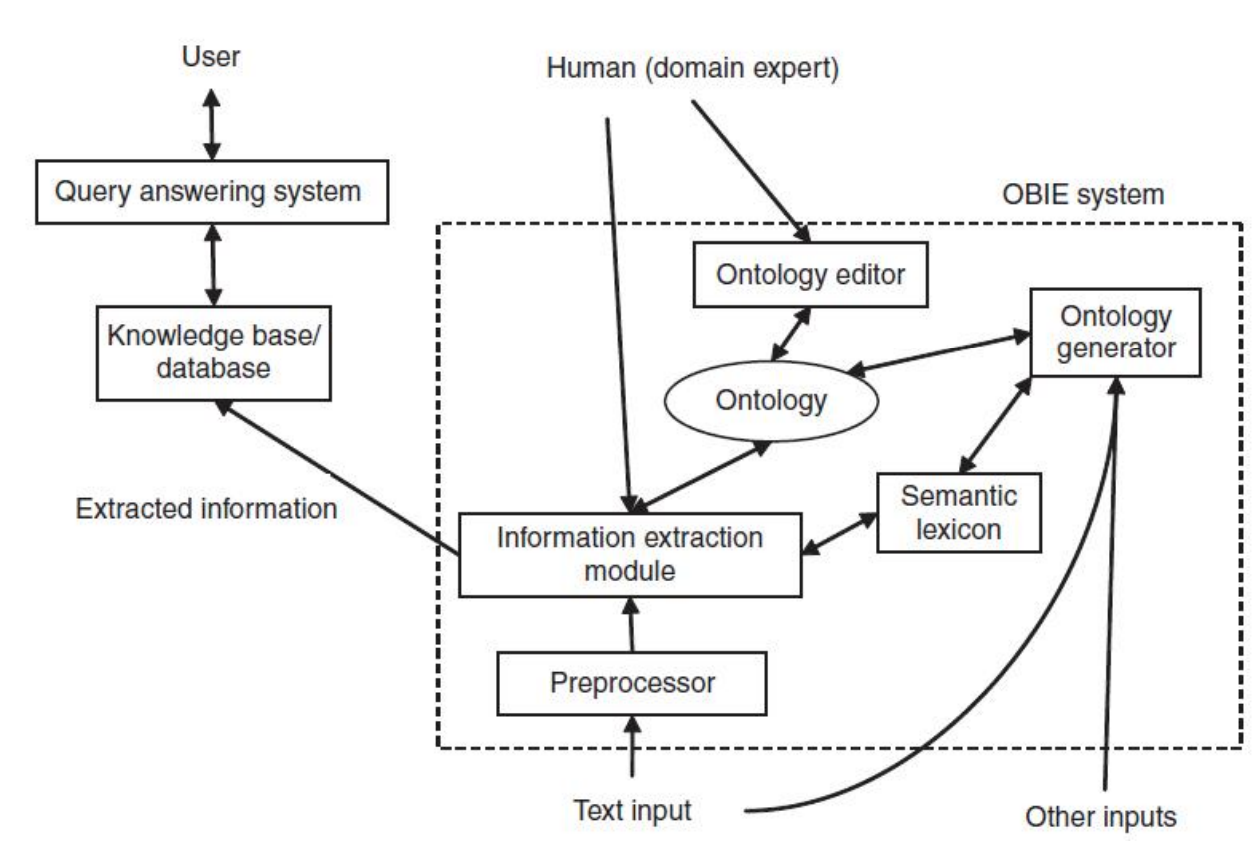
چالش استخراج اطلاعات

- روبه‌رو بودن با مقدار بسیار زیادی اطلاعات متنی
- استخراج اطلاعات از میلیون‌ها و حتی میلیاردها سند
- سنگین و گران بودن استخراج موجودیت‌ها و روابط

روش‌های استخراج اطلاعات

- مبتنی بر قاعده
- مبتنی بر الگو
- مبتنی بر استدلال
- مبتنی بر پردازش زبان طبیعی
- مبتنی بر آنتولوژی

معماری کلی یک سیستم استخراج اطلاعات مبتنی بر آنتولوژی (سیستم OBIE)



دانش ساخت‌یافته و استفاده از آن

- مورد توجه قرار گرفتن تبدیل متون به دانش قابل فهم برای ماشین
  - با توجه به حجم عظیم دانش و اطلاعات بشر و رشد روزافزون مستندات در زمینه‌های مختلف
- ایجاد پایگاه دانش ساخت‌یافته از متون
  - با استفاده از سیستم استخراج اطلاعات
- هدف سیستم استخراج اطلاعات
  - استخراج حقایق از متون غیرساخت‌یافته
  - نمایش آن‌ها در قالب ساخت‌یافته مثل سه‌گانه‌های RDF
- به دست آوردن اطلاعات مورد نیاز با ساخت و ارسال پرس و جوهای SPARQL

استخراج اطلاعات و داده‌های عظیم

- مدیریت مقدار زیاد داده
  - بدون تاثیرپذیری قابل توجه کارایی
- ذخیره تعداد کلان سه‌گانه‌های RDF و توانایی Query زدن روی آن‌ها
- عدم توانایی لازم چارچوب‌های موجود
  - اجرا بر روی یک ماشین واحد مثل Jena
- استفاده از Hadoop

حجم عظیم داده‌های RDF

- رشد سریع مجموعه داده‌های RDF
  - هم در تعداد مخازن RDF و هم سایز مجموعه داده‌های RDF
- بسیاری از مجموعه داده‌های شناخته شده RDF شامل بیلیون‌ها سه‌گانه
  - سه گانه RDF شامل Subject، Predicate و Object
- استفاده از داده‌های RDF در بسیاری از کاربردها
  - شامل تجارت، مهندسی، علوم مانند هوش کسب و کار، شبکه‌های اجتماعی، زیست‌شناسی و ...
- یکی از چالش‌های اساسی برای مدیریت حجم عظیم داده‌های RDF
  - چگونگی اجرای پرس وجوهای RDF به صورت کارا

ذخیره سازی داده‌های عظیم RDF

- عدم مقیاس‌پذیری یکی از مهمترین مشکلات
  - در مواجهه با مخازن داده RDF ماشین واحد
- نامناسب بودن چارچوب‌های موجود برای گراف‌های عظیم RDF
- تلاش برای ذخیره بهینه داده‌های RDF
- معرفی ابزارهای ذخیره‌سازی

پردازش داده‌های عظیم RDF

- جستجو و استخراج اطلاعات از گراف RDF با استفاده از SPARQL
- پرس‌وجوی توزیع‌شده داده RDF
  - جستجوی داده RDF بر روی سیستم توزیع‌شده
- سرعت و کارایی دو جنبه مهم برای ارزیابی پرس‌وجوی RDF
- فرآیند متداول پرس‌وجو
  - ورود داده
  - تولید جدول شاخص
  - پردازش پرس وجو
  - ادغام نتایج

Map Reduce

- یک چارچوب نرم‌افزاری برای پردازش حجم زیاد داده به صورت موازی
- تقسیم کردن مجموعه داده ورودی به تکه‌های مستقل توسط MapReduce job
  - پردازش به صورت موازی با map taskها
  - خروجی مرتب شده map به عنوان ورودی reduce taskها
- یکسان بودن نودهای ذخیره‌سازی و نودهای محاسباتی
- عمل کردن بر روی جفت (کلید، مقدار)

- (input) <k1, v1> -> **map** -> <k2, v2> -> **combine** -> <k2, v2> -> **reduce** -> <k3, v3> (output)

ابزار موجود

- پردازش داده RDF
  - SPIDER
  - RAPID
  - PigSPARQL
  - Sparqlify
- دیگر ابزار
  - HiveQL
    - تبدیل SPARQL به HiveQL و اجرای کوئری Hive
  - Jena-Hbase
  - rdfstore-js
  - CliqueSquare
  - H2RDF

ذخیره‌سازی داده RDF

- استفاده از Hbase توسط SPIDER
- SHARD (مبتنی بر Hadoop)
- HadoopRDF
- CumulusRDF (یک مخزن کلید، مقدار مبتنی بر Apache Cassandra)
- CouchDB and MongoDB (برای ذخیره‌سازی و پرس وجوی داده RDF)
- PRESTO-RDF
- Sempala

SPIDER

- برای ذخیره‌سازی و پردازش داده RDF
  - شامل لود و پرس‌وجوی گراف RDF
- پیاده‌سازی با استفاده از MapReduce
  - جستجوی بخش‌های تطبیق‌یافته زیرگراف توسط بخش Map
  - و سپس انتقال نتیجه تطابق به بخش Reduce و در نهایت ارائه نتیجه مطلوب

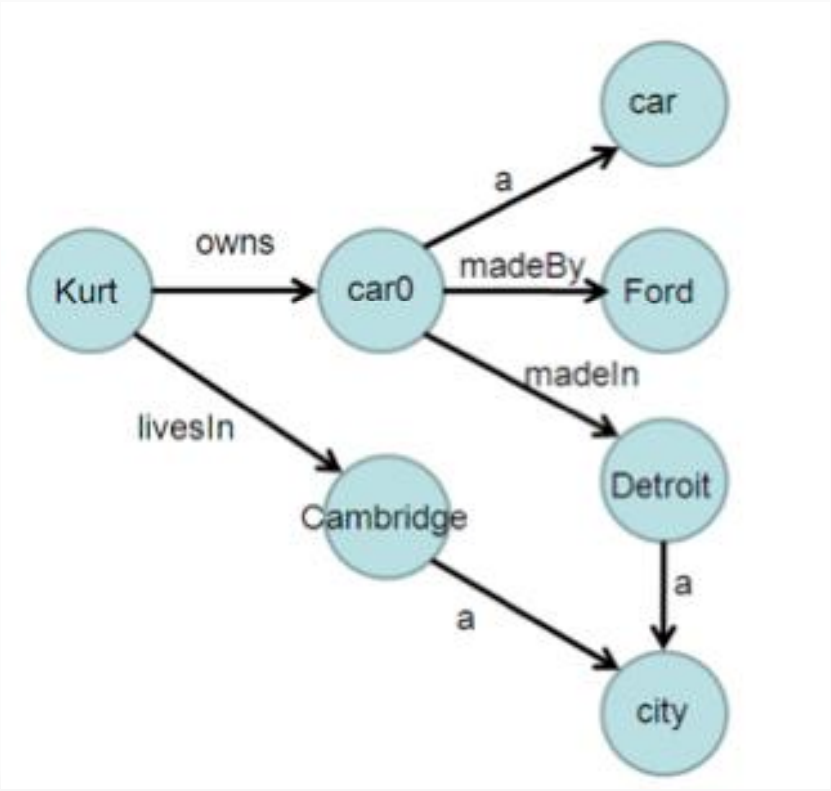
CumulusRDF

- یک مخزن RDF بر روی معماری مبتنی بر ابر
- فراهم کردن امکانی برای مدیریت داده RDF
- استفاده از Apache Cassandra
- لود سه‌گانه‌ها و ذخیره در مخزن CulumusRDF
- اجرای پرس‌وجوی SPARQL

SHARD

- تکنولوژی مبتنی بر ابر
  - ذخیره‌سازی سه‌گانه‌ها
  - امکان پردازش داده مقیاس‌پذیر با Hadoop و MapReduce
- در نظر گرفتن داده به صورت استاندارد RDF و اجرای پرس‌وجوها با SPARQL

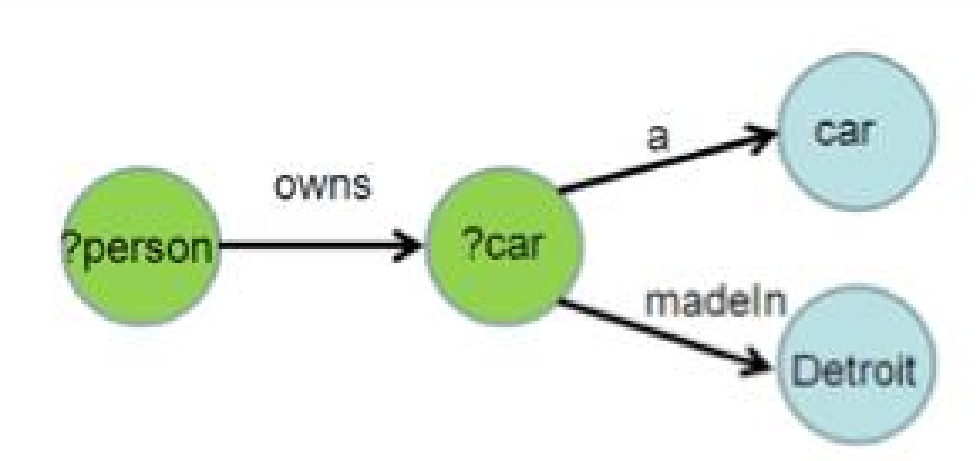
مثالی از گراف کوچکی از داده‌های سه‌گانه



مثالی از پرس‌وجوی SPARQL

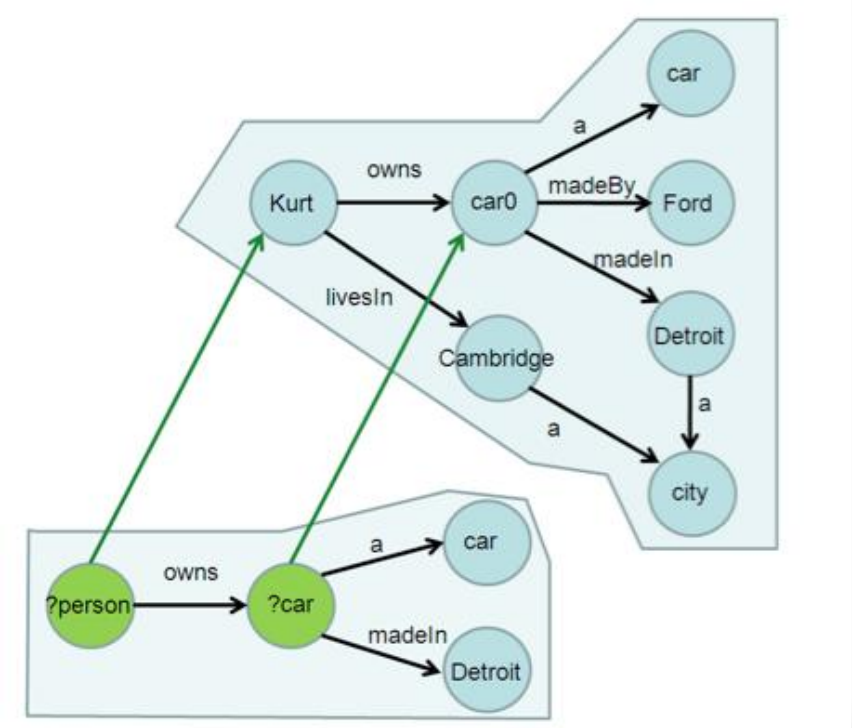
```
SELECT ?person

WHERE {
    ?person :owns ?car .
    ?car :a :car .
    ?car :madeIn :Detroit .
}
```

گراف نشان‌دهنده پرس‌وجو

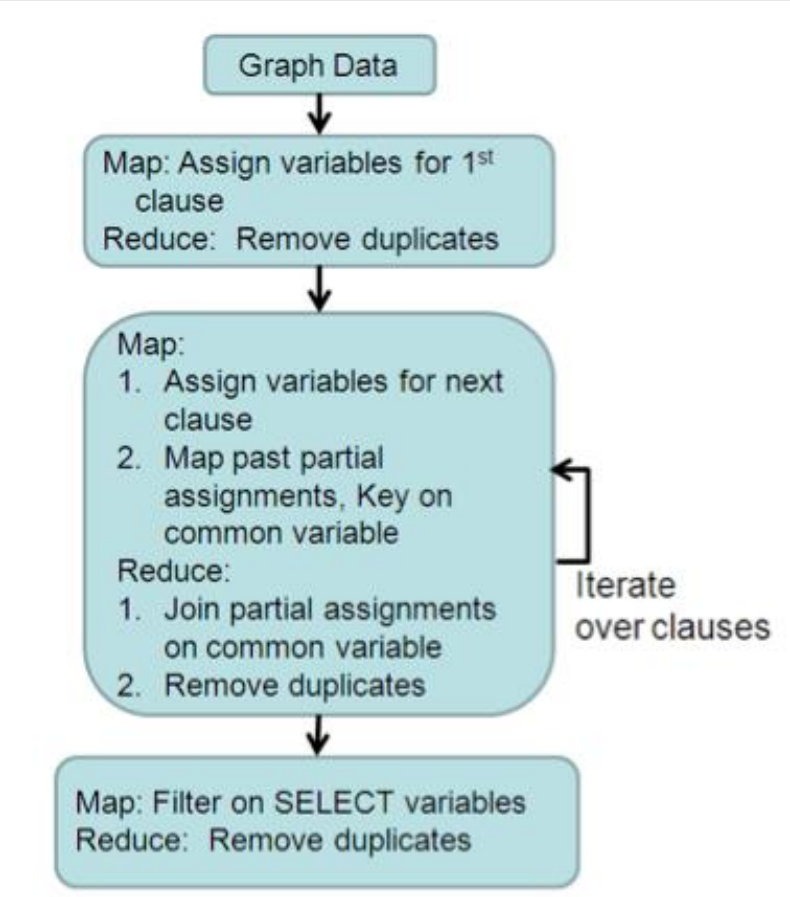


پردازش پرس‌وجو

- شناسایی متغیرهای bind شده عبارات پرس‌وجو به نودهای گراف



الگوریتم MapReduce برای پردازش SPARQL



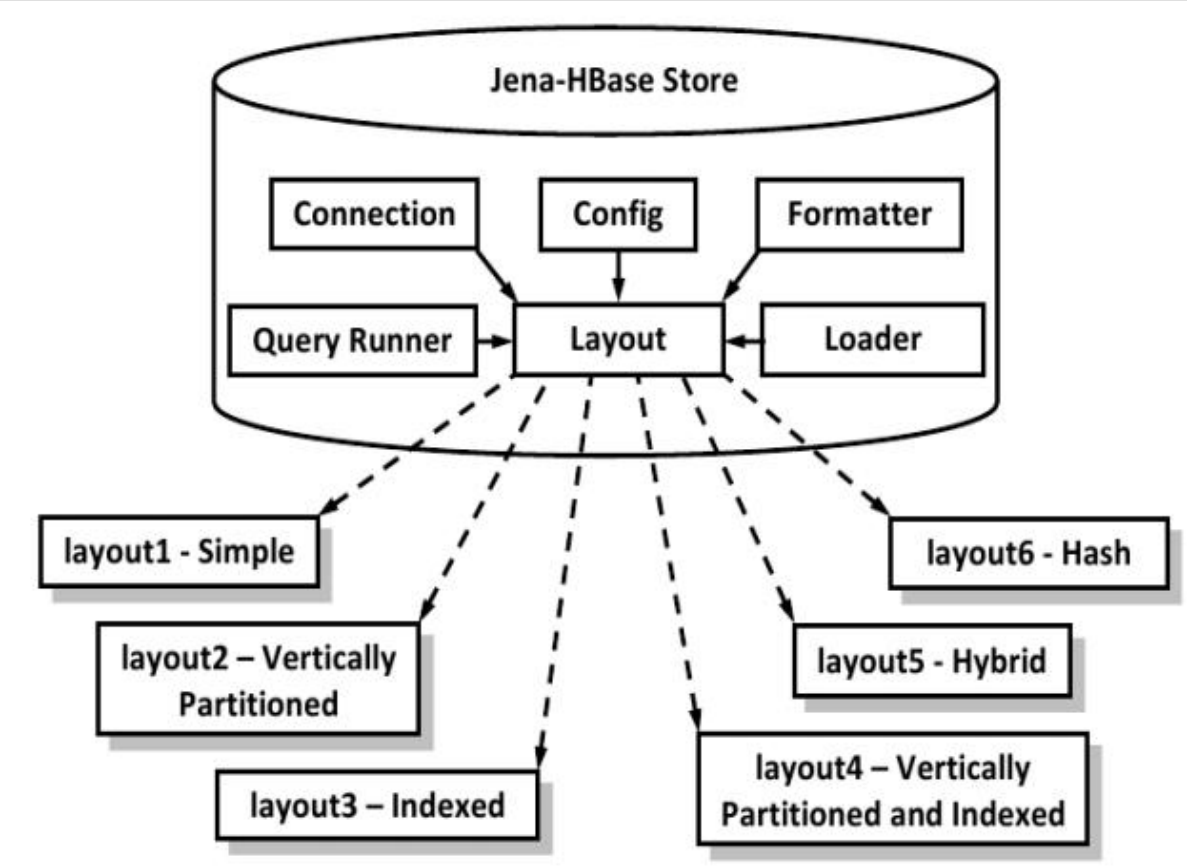
Jena-HBase

- Hbase (برای ذخیره‌سازی سه‌گانه‌ها) که می تواند با چارچوب Jena مورد استفاده قرار گیرد.
- دارا بودن ویژگی‌هایی نظیر
  - امکان دستکاری داده‌های RDF در فرمت‌های مختلف
  - پشتیبانی از زبان پرس وجوی SPARQL
- تنها نیاز به پیاده‌سازی عملیاتی برای افزودن، حذف و بازیابی سه‌گانه‌ها از Hbase

استفاده از Hbase در این ابزار

- استفاده از Hbase به عنوان لایه ذخیره‌سازی به دو دلیل
  - ۱- به طور کلی، Hbase یک مخزن ستون محور است و یک مخزن ستون محور بهتر از مخازن ردیف محور عمل می‌کند.
  - Hadoop شامل HDFS (فایل سیستم توزیع شده برای ذخیره داده) و MapReduce (برای پردازش حجم عظیمی از داده‌های ذخیره شده در HDFS) است.
  - ۲- Hbase از HDFS به عنوان مکانیزم ذخیره‌سازی استفاده می‌کند اما به MapReduce نیازی ندارد.
  - بنابراین Jena-Hbase نیاز به پیاده‌سازی موتور پرس وجو برای اجرای پرس‌وجوها بر روی سه‌گانه‌های RDF ندارد.

Jena-Hbase معماری



پردازش Query در Jena-Hbase

- پردازش عملیات Find
  - برگرداندن همه سه‌گانه‌هایی که با الگوی سه‌گانه داده شده تطابق دارند.
  - متفاوت برای هر طرح ذخیره‌سازی
- پردازش SPARQL Query
  - تبدیل به مجموعه‌ای از عملیات Find
  - شامل متغیرهایی نشان‌دهنده الحاق بین الگوهای سه‌گانه متفاوت از یک Query

پیاده‌سازی بدون ابزار

- ذخیره‌سازی در HBase
  - با استفاده از برنامه MapReduce برای ساخت جدول و لودکردن داده در جدول
  - اجرا با دستورات لازم و استفاده از برنامه‌ها و کتابخانه‌های موجود
- پردازش
  - پیاده‌سازی با استفاده از مدل MapReduce
    - نوشتن توابع Map و Reduce

پردازش پرس‌وجو

- پرس‌وجوی SPARQL شامل مجموعه‌ای از الگوهای سه‌گانه (BGP)
- بررسی BGP
  - شامل پیاده‌سازی دو عملیات
    - MR Selection: به دست آوردن سه‌گانه‌های تطبیق یافته با حداقل یک الگوی سه‌گانه
    - MR Join: ادغام سه‌گانه‌های تطبیق یافته
  - پیاده‌سازی توابع Map و Reduce به صورت جداگانه برای هر عملیات

H2RDF

- پرس‌وجو و شاخص‌گذاری توزیع‌شده داده‌های عظیم RDF
  - با استفاده از NoSQL و MapReduce
- نیاز به Hadoop، Hbase و Zookeeper
- آپلود فایل سه‌گانه‌ها در HDFS و اجرای دستورات لازم برای ذخیره داده RDF
- اجرای پرس‌وجوی SPARQL با دستورات لازم

CliqueSquare

- پلت‌فرم توزیع‌شده مدیریت داده RDF بر روی Hadoop
- شامل دو بخش
  - Data Partitioning: ذخیره داده RDF در فایل سیستم توزیع‌شده
  - Query Processing: پاسخ به پرس‌وجو
- لود مجموعه داده ورودی در HDFS
- لود مجموعه داده ورودی HDFS در CliqueSquare
- اجرای پرس‌وجو

جمع‌بندی

- دست‌یافتن به اطلاعات مورد نیاز
  - با ارسال پرس‌وجوهای SPARQL بر روی حجم عظیم داده‌های RDF
- استفاده از ابزار موجود
  - Jena-Hbase
  - H2RDF
  - CliqueSquare
- پیاده‌سازی با استفاده از مدل MapReduce

مراجع

- Hyunsik Choi, Jihoon Son, YongHyun Cho , SPIDER : “A System for Scalable, Parallel / Distributed Evaluation of large-scale RDF Data” ,2014
- Ke Hou, Jing Zhang and Xing Fang , “Review of Large-Scale RDF Data Processing in MapReduce”,2014
- François Goasdoué, Zoi Kaoudi, Ioana Manolescu, Jorge Quiané-Ruiz, Stamatis Zampetakis, “CliqueSquare: efficient Hadoop-based RDF query processing”, *Journées de Bases de Données Avancées*, Oct 2013
- Kurt Rohloff, Richard E Schantz, “High-performance, massively scalable distributed systems using the MapReduce software framework: the SHARD triple-store”, Programming Support Innovations for Emerging Distributed Applications,ACM NewYourk, 2010
- Nikolaos Papailiou, Ioannis Konstantinou, Dimitrios Tsoumakos, Nectarios Koziris, “H2RDF: adaptive query processing on RDF data in the cloud**”,** Proceedings of the 21st international conference companion on World Wide Web, ACM NewYourk, 2012

مراجع

- G Ladwig, A Harth, “CumulusRDF: linked data management on nested key-value stores” , iswc2011
- DC Wimalasuriya, D Dou, “Ontology-based information extraction: An introduction and a survey of current approaches”, Journal of Information Science, 2010
- Jiewen Huang, Daniel J. Abadi, Kun Ren, “Scalable SPARQL Quer ying of Larg e RDF Graphs”, Seattle, Washington.Proceedings of the VLDB Endowment, Vol. 4, No. 11,2011
- Jaeseok Myung, Jongheum Yeon, Sang-goo Lee, “SPARQL Basic Graph Pattern Processing with Iterative MapReduce”, 2010
- Mohammad Farhan Husain, Pankil Doshi, Latifur Khan, Bhavani Thuraisingham, “Storage and Retrieval of Large RDF Graph Using Hadoop and Map Reduce”, Springer-Verlag Berlin Heidelberg, 2009

مراجع

- Jos´e M. Gim´enez-Garc´ıa A, Javier D. Fern´andez A, Miguel A. Mart´ınez-Prieto B, “MapReduce-based Solutions for Scalable SPARQL Querying” , Open Journal of Semantic Web (OJSW),Volume 1, Issue 1, 2014
- Marcelo Arenas , Jorge Pérez , “Querying Semantic Web Data with SPARQL”, ACM 978-1-4503, 2011
- Alexander Schatzle, Martin Przyjaciel-Zablocki, Antony Neu, Georg Lausen, “Sempala: Interactive SPARQL Query Processing on Hadoop”, Springer International Publishing Switzerland 2014
- Vaibhav Khadilkar, Murat Kantarcioglu, Bhavani Thuraisingham, Paolo Castagna, “Jena-HBase: A Distributed, Scalable and Efficient RDF Triple Store”,
- V.Richard Benjamins,Dieter Fensel, A.Gomez, “”Knowledge Management through Ontologies”, conf. on practical aspects of knowledge management,1998
- Junsong Zhang , Wu Zhao, Gang Xie , Hong Chen, “Ontology- Based Knowledge Management System and Application”, Procedia Engineering 15 (2011)

مراجع

- Martins Zviedris, Aiga Romane, Guntis Barzdins, and Karlis Cerans, “Ontology-Based Information System”, Springer International Publishing Switzerland 2014

با تشکر از توجه شما