

文章编号:1673-0062(2014)04-0056-05

基于 RDF/XML 的微博知识表达与语义检索系统

罗凌云,史 淼,阳小华,刘志明

(南华大学 计算机科学与技术学院,湖南 衡阳 421001)

摘 要:语义检索技术有别于传统的基本文本的检索方式,它通过为信息添加语义而使其能够被计算机理解,从而实现检索的智能化. 本文通过对新浪微博数据进行分析,设计合适的资源描述框架(Resource Description Framework, RDF)结构,将其转换成富含语义关系的 RDF 格式,利用 Virtuoso 实现微博 RDF 数据的存储,并在此基础上设计实现基于微博数据的语义检索系统.

关键词:微博;RDF;知识表达;语义检索

中图分类号:TP31 **文献标识码:**B

Semantic Search Engine for Weibo Data Based on RDF/XML Oriented Knowledge Representation

LUO ling-yun, SHI Miao, YANG Xiao-hua, LIU Zhi-ming

(School of Computer Science and Technology, University of South China,
Hengyang, Hunan 421001, China)

Abstract: Different from traditional search methods based on text, semantic-based search provides a methodology to accomplish intelligent search by annotating information with semantics so that it can be “understood” by computers. In this paper, after analysis on Sina weibo data, we design a suitable RDF (Resource Description Framework) Schema and translate the data into RDF datasets. We leverage the use of Virtuoso to store the RDF graphs, based on which a semantic search engine for weibo data is established.

key words: Weibo; RDF; knowledge representation; semantic search

收稿日期:2014-07-02

基金项目:湖南省教育厅优秀青年基金资助项目(14B153);衡阳市应用基础研究基金资助项目(2014KJ15);南华大学校级启动基金资助项目(2013XQD07);南华大学创新团队建设基金资助项目(NHCXTD16)

作者简介:罗凌云(1981-),女,湖南衡阳人,南华大学计算机科学与技术学院讲师,博士后. 主要研究方向:语义 Web 技术及其应用,医学信息学,理论计算机科学.

0 引言

语义 Web 技术为 Web 上日益增长的大数据提供了有效的智能化处理手段,有别于传统的基于文本的信息处理方式,它通过为信息添加语义而使其能够被计算机所“理解”,从而实现智能信息检索. 目前,Google 等各大搜索引擎推出的“知识图谱”,IBM 开发的智能机器人 Watson 等,均为该技术的应用. 语义 Web 通常使用资源描述框架 (Resource Description Framework, RDF) 来描述网络数据. 本文以新浪微博为素材,对微博数据进行分析,根据 RDF 三元组规则,为其中所需有用信息数据添加语义,并以 XML 为语法框架,编写程序将其转换为 RDF/XML 文件;此外,通过 Virtuoso 实现微博 RDF 数据的存储,并利用 SPARQL 查询语言,设计实现基于微博数据的语义检索系统.

1 语义 Web 和 RDF

1.1 语义 Web

在计算机科学特别是网络技术高速发展的今天,如何将人类可以理解的网络中的信息和内容,变成计算机可理解和处理的信息,从而实现网络的智能化,显得日益重要. 基于此,“WWW 之父”Tim Berners-Lee 于 1998 年提出了语义 Web 的概念,目的是扩展和延伸现有的网络,通过对文档中数据的可扩展标识,描述事物间的明显关系,且包含语义信息,以利于机器的自动处理. 语义 Web 的研究和发展兴起了网络的新革命,正在快速地改变未来的网络^[1].

1.2 语义 Web 体系结构

语义 Web 的七层模型结构如图 1 所示,前三层如下:

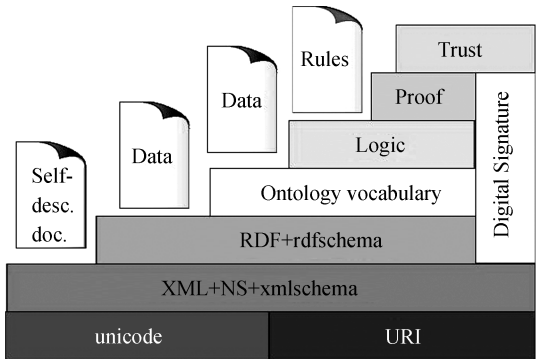


图 1 语义 Web 体系结构
Fig.1 Semantic Web layer cake

1)最底层为统一编码 Unicode 和统一资源标识符 URI.

2)第二层主要为标签语言 XML 和命名空间 (NameSpace). 为了避免冲突,每个标签都是在特定的命名空间中指定的. XML 解决了底层文档交换的格式问题.

3)RDF^[2] 和 RDFS (RDF Schema). RDF 解决如何无二义性的描述资源对象的问题,使得描述的资源元数据信息成为机器可以理解的信息. RDFS 是对 RDF 的抽象,提出了类的概念,定义了类和性质,并表达了各种类、性质之间的关系. RDF/RDFS 解决了语义模型和通用语义的问题.

1.3 RDF

RDF 定义了一种通用的资源描述框架,用三元组 <资源,属性,属性值> 灵活地描述 Web 上的资源. 其中,所有资源均通过唯一的 URI (Uniform Resource Identifiers) 来进行标识. 每个三元组叫做一个 RDF Statement, RDF 模型中所有被描述的资源以及用来描述资源的属性值都可以看成是“节点”(Node), 因此, 整个数据模型构成一张 RDF 图,一个 RDF 数据库通常由一张或多张 RDF 图构成^[3-5].

2 微博数据的 RDF 表达

2.1 微博数据的获取

新浪开放平台是一种基于新浪微博客系统的开发平台,主要用来实现信息的订阅、资源的分享和交流,广大开发者或网站只要登录平台网站并在其中创建应用,便可通过新浪开放平台提供的各种开放接口 (Open API) 对微博系统进行读写,获取自己想要的数据和实现某些应用功能. 通过新浪开发平台获取微博数据的具体流程如图 2 所示.

按照图 2 所示流程,合理利用软件开发包就能成功地获取微博数据. 每条微博数据均为普通文本格式,其内容包括两大部分:一部分是关于用户的信息,如用户 ID,名字,粉丝数,好友数等等;另一部分是关于该微博本身的信息,如微博 ID,创建时间,微博内容,评论数和转发数等.

为降低实验设计复杂性,本项目根据实际需要,适当舍去了一些次要的信息量,而只提取了其中关键的信息量,主要包含以下内容:用户 id,微博昵称 (screenName), 个人描述 (description), 粉丝数 (followersCount), 关注数 (friendsCount), 微博数 (statusCount), 收藏数 (favouritesCount), 是否

为认证用户 (verified), 认证理由 (verifiedReason), 微博 id, 微博内容 (text), 发布时间 (create-

dAt), 微博内容来源 (resource), 微博转发数 (repostsCount), 微博评论数 (commentsCount).

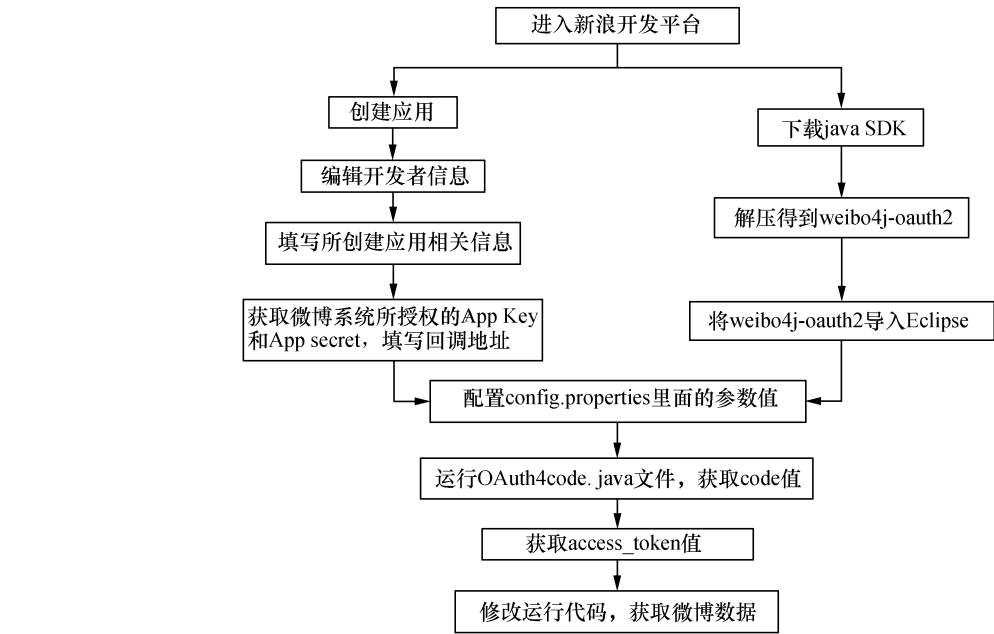


图 2 微博数据获取流程图

Fig. 2 Flow chart of Weibo data achieve process

2.2 RDFS 的设计

为了将普通文本格式的微博数据转换成 RDF 数据,并清晰完整地呈现出各事物之间的关联,首先需要设计合理有效的 RDF Schema. 自定义两个新的命名空间 (NS) “http://weibo. com/ user#” 和 “xmlns:txt = http://weibo. com/txt#”, 用来区分用户信息和微博信息. 其中,命名空间 user 下包含 10 个属性,而命名空间 txt 下包含 7 个属性. 最终设计的 RDFS 如图 3 所示.

注意到图 3 中有两个空白节点 (anonymous node), 即两个未标注的资源,它们起到了将图中各个部分连通的必要作用,分别表示了“用户的微博”和“微博影响力”这两个概念. 此外,由于同一用户可发布若干条微博,因此实际的微博数据 RDF 图将出现许多这样的空白节点.

2.3 数据转换

确定好微博数据的 RDFS 之后,便可通过编程实现数据格式的转换. 在实现的过程中,为减少对数据的再次读入,可直接在利用新浪开放平台获取微博数据时,通过设置相关参数,将获取到的微博数据按照 RDFS 框架写入 xml 文件中. 其过程如下:

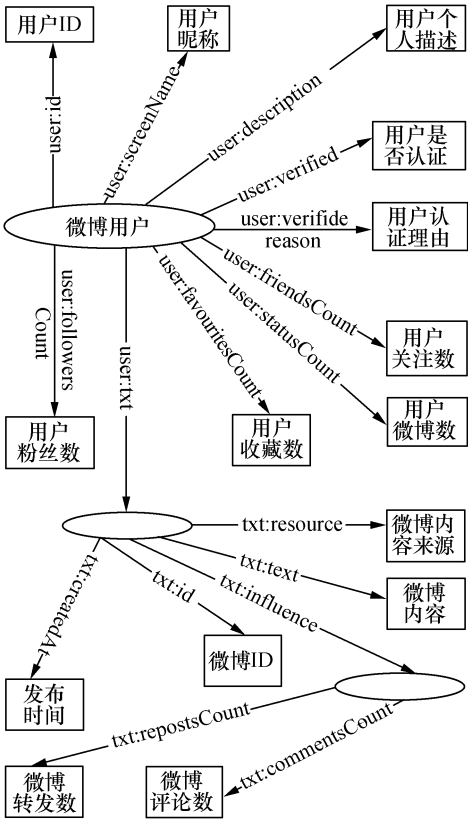


图 3 微博数据的 RDFS 设计

Fig. 3 RDFS for Weibo data

- 1)新建一个 XML 文档
- 2)写入标签“<rdf:RDF>”
- 3)添加三个命名空间“xmlns:rdf”、“xmlns:user”和“xmlns:txt”
- 4)读取第一条微博 ID,将其相关信息写入标签环境“rdf:Description”中
- 5)按照步骤 4 循环处理接下来读取到的每一条微博
- 6)写入结束标签“</rdf:RDF>”

3 基于微博数据的语义检索系统的实现

3.1 微博 RDF 数据的存储

本实验使用 Virtuoso^[6]数据库存储 RDF 数据. Virtuoso 数据库是 OpenLink 公司开发的一个可伸缩的高性能的跨平台对象关系数据库,为用户提供了复杂的 SQL\XML\RDF 数据库管理功能.在本地机器安装好 Virtuoso 数据库后,启动运行,然后在浏览器中输入 `http://localhost:8890/conductor/`,便可获得 Virtuoso 数据库的可视化界面.

为了导入 RDF 文件,用户登录后选择“Linked data”中的“Quad Store Upload”,选择微博 RDF 文件,单击确定后给即将要生成的 RDF 图命名一个 Graph IRI,如: `http://localhost:8890/DAV/testRDF`. 上传后便可完成微博 RDF 数据的导入工作.文件成功导入后,在“Linked data”中的“Graphs”中便会看到新增的 RDF 图 `http://localhost:8890/DAV/testRDF`.

3.2 基于 SPARQL 语言的检索系统

SPARQL (Simple Protocol and RDF Query Language) 是一种面向 RDF 数据模型的查询语言,是 W3C 组织指定的候选推荐标准^[7-10]. 它可以空白结点、无格式和类型文字的形式提取信息,可以提取 RDF 子图,能在查询图中构造新的 RDF 图. SPARQL 作为一种数据访问语言,结合 Virtuoso 提供的 RDF 查询界面,便可实现微博数据的语义检索系统. 例如,选择微博数据的 RDF 图 `http://localhost:8890/DAV/testRDF`,使用如下的查询语句,即可获得本实验所有微博数据的三元组信息:

```
SELECT ?s ?p ?o
WHERE { ?s ?p ?o }
```

要知道杨幂所发微博的评论数和转发数,便可按步骤进行如下的 SPARQL 查询操作:

- 1)使用如下查找语句找到杨幂微博用户的 ID.

```
SELECT ?s
WHERE { ?s ?p “杨幂” }
```

- 2)根据获取到的用户杨幂的微博 ID “`http://weibo.com/u/1195242865`”,可使用如下 SPARQL 查询语句查询其微博的相关信息. 因为用户可发多条微博,所以结果可能有多个.

```
SELECT ?p ?o
WHERE { < http://weibo.com/u/1195242865 > ?p ?o }
```

选取其中一条微博,可得到变量? p 为 user:txt 时对应的第一个空白节点 ID <nodeID://b10010>. 根据 2.2 节的定义,该空白节点表示用户的微博.

- 3)接着查看杨幂这条微博的相关信息,可得到表示微博影响力的第二个空白节点 ID <nodeID://b10011>. 利用该 ID 进一步查询,便得知该微博的转发数和评论数分别为 875 和 2018,如图 4 所示.

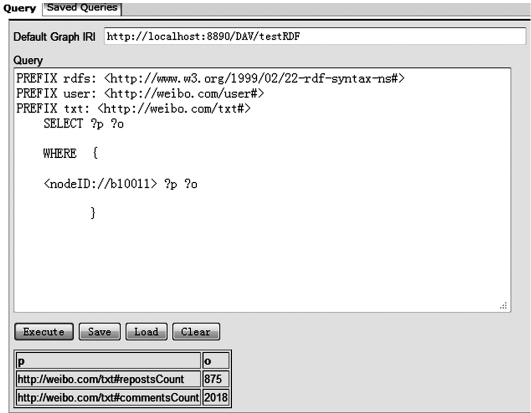


图 4 微博转发和评论数查询结果

Fig. 4 Query results for reposts count and comments count

事实上,也可以将上述步骤合并,在 SPARQL 语言中嵌套,运行一次查询获得所需信息,如图 5 所示.

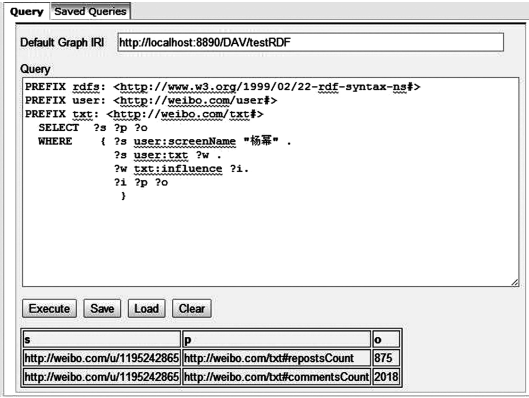


图 5 嵌套的 SPARQL 查询
Fig. 5 Integrated SPARQL query

4 结 论

本文以目前国内流行的社交网络平台,即新浪微博为研究对象,利用语义 Web 技术构建了一个微博数据语义检索原型系统,主要包含以下几个方面:首先介绍了微博数据的获取方法,接着通过对微博数据的分析,构建了合适的 RDF 结构图,为将其转换为包含语义信息的 RDF 数据奠定了基础.紧接着,通过编程将所获取的微博数据转换为 RDF 格式并将其存储在 Virtuoso 数据库中.最后,通过 Virtuoso 提供的查询界面与 SPARQL 查询语言,便可实现对新浪微博数据的语义查询.总之,本系统为微博数据的语义转换与检索提供了借鉴,在网络舆情监控方面也具有重要的应用意义.

下一步将继续从两个方面完善和补充该系统:一方面对 SPARQL 语言进行封装,提供更为

性化的查询界面;另一方面将对微博自身的内容进行自然语言处理和语义标识,以丰富语义查询的内容.

参考文献:

[1] Antoniou G, Harmelen F. A semantic web primer[M]. 2nd ed. London: The MIT press, 2008.

[2] Brian M. Handbook on Ontologies[M]. 1st ed. Berlin: Springer Heidelberg, 2004.

[3] Wu G, Li J, Hu J Z, Hu J Q, et al. System II: a native RDF repository based on the hypergraph representation for RDF data model[J]. Journal of Computer Science & Technology, 2009, 24(4): 652-664.

[4] 朱敏. 基于 HBase 的 RDF 数据存储与查询研究[D]. 南京: 南京大学, 2013 年.

[5] 杜方, 陈跃国, 杜小勇. RDF 数据查询处理技术综述[J]. 软件学报, 2013, 24(6): 1222-1242.

[6] 邹益民, 张智雄, 钱力, 等. 语义仓储 Virtuoso 的技术分析和应用[J]. 图书情报工作, 2012, 56(23): 97-102.

[7] Quilitz B, Leser U. Querying distributed RDF data sources with SPARQL[J]. LNCS, 2008, 5021: 524-538.

[8] Liu C, Wang H, Yu Y, et al. Towards efficient SPARQL query processing on RDF data[J]. 清华大学学报: 自然科学英文版, 2010, 15(6): 613-622.

[9] Gu Y, Liu D. Research on RDF query using SPARQL language[C]//Proceedings of the 2012 International Conference on Convergence Computer Technology. Washington, DC: IEEE Computer Society, 2012: 105-109.

[10] 肖竹军. 基于 SPARQL 的 RDF 数据节点间关系路径检索[J]. 微型机与应用, 2011, 30(9): 50-53.