

数据库系统概论新技术篇

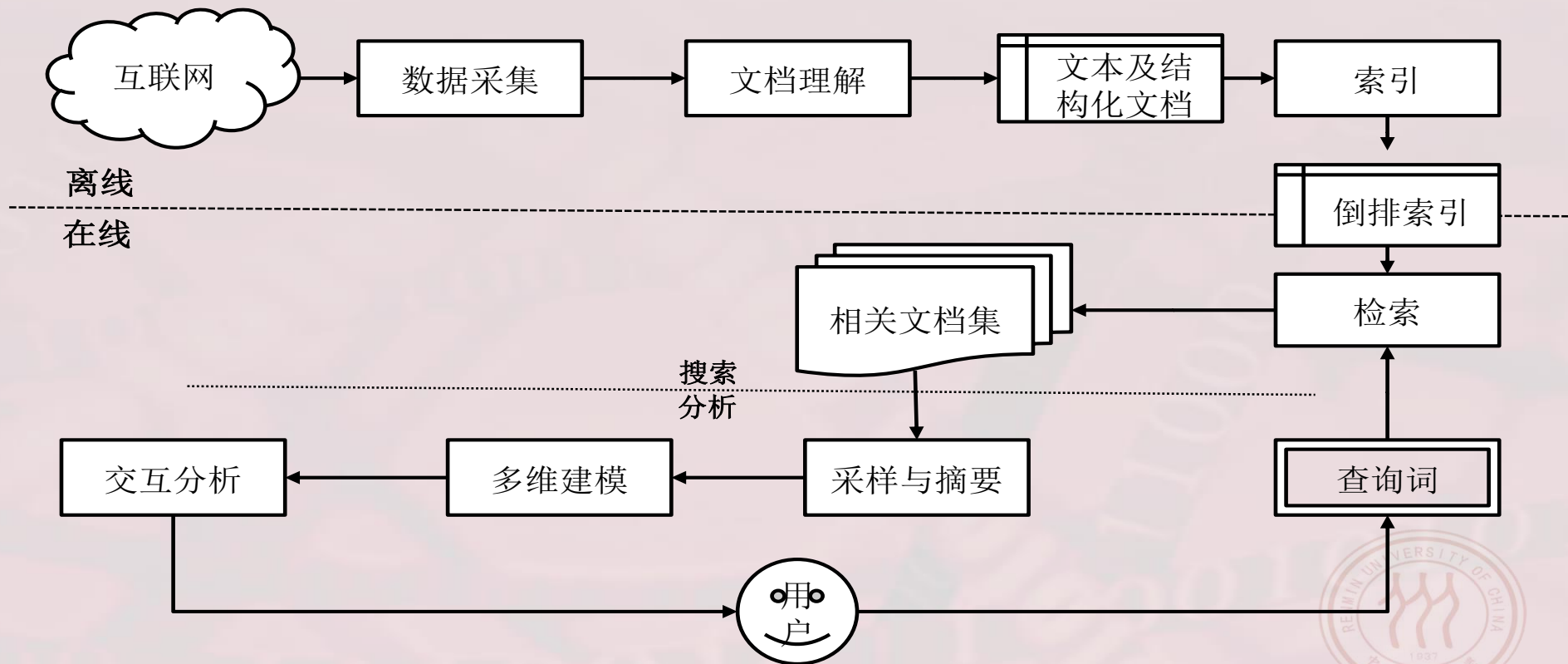
文本大数据分析及应用案例

窦志成

中国人民大学信息学院

2017年7月

文本大数据分析引擎 - 系统构架



文本大数据分析及应用案例

❖ 课程内容

- 交互式文本大数据分析系统：时事探针
- 自然语言处理与文本挖掘基础算法
- 文本搜索、文本分析系统构建



问题

❖ 给定一个文本集合，如何像“时事探针”那样，能对文本内容进行实时检索和分析？

- 全文检索：查询“中国人民大学明德图灵”，文档中只要包含“中国人民大学”，“明德”和“图灵”即可，不需要这三个词连续出现

- 分析：能够对重要维度（如人、地点等）包含的内容进行统计排序

❖ 分析

- 检索是在非结构化的文本数据上进行，统计分析是在结构化字段上进行

- 前面：我们已经从每个文档中抽取出了结构化的信息，如人名、地名、机构名、情感类别等

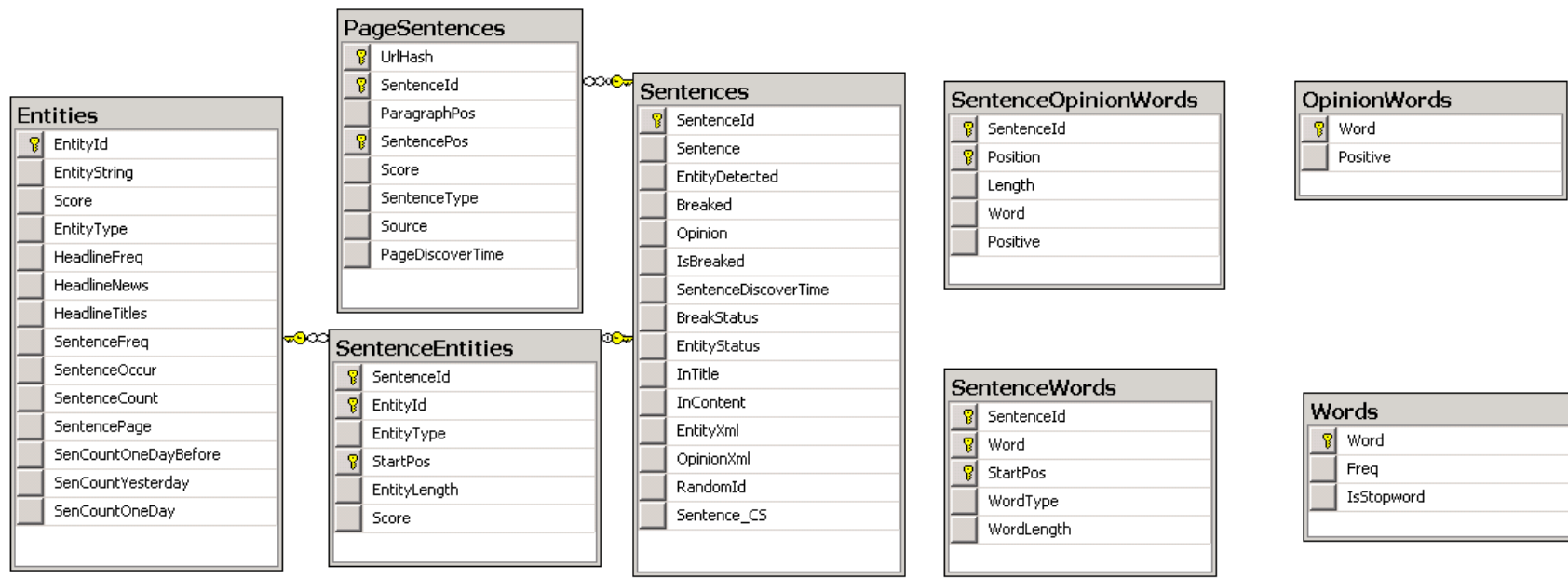
- 接下来：如何基于这些结构化数据，并结合非机构化的文本数据，完成上述目标

❖ 方案：基于关系型数据库，基于分布式大数据系统，基于全文搜索系统



基于关系型数据库

- ❖ 建立类似于下面的数据库表
- ❖ 通过join和group by等实现实体等信息的汇总



基于关系型数据库

❖ 示例：获取一段时间内指定类型（人、地点、机构等）的热门实体

```
SELECT TOP(@Count)
    Entities.EntityId,
    MAX(Entities.EntityType) AS EntityType,
    MAX(EntityString) AS EntityString,
    COUNT(DISTINCT Sentences.SentenceId) as Freq,
    COUNT(DISTINCT Sentences.SentenceId) / MAX(LOG(1+isnull(Entities.SentenceCount,1))) AS TFIDF,
FROM SentenceEntities, Entities, Sentences
WHERE Sentences.SentenceId = SentenceEntities.SentenceId
    AND Entities.EntityId = SentenceEntities.EntityId
    AND Sentences.SentenceDiscoverTime >= @DateBegin
    AND Sentences.SentenceDiscoverTime <= @DateEnd
GROUP BY Entities.EntityId
HAVING COUNT(Distinct Sentences.SentenceId) > 20
ORDER BY TFIDF desc, Freq desc
```



基于关系型数据库

❖ 示例：获取一段时间内和某个话题相关的指定类型（人、地点、机构等）的热门实体

```
SELECT TOP(@Count)
    Entities.EntityId,
    MAX(Entities.EntityType) AS EntityType,
    MAX(EntityString) AS EntityString,
    COUNT(DISTINCT Sentences.SentenceId) as Freq,
    COUNT(DISTINCT Sentences.SentenceId) / MAX(LOG(1+isnull(Entities.SentenceCount,1))) AS TFIDF,
FROM SentenceEntities, Entities, Sentences
WHERE Sentences.SentenceId = SentenceEntities.SentenceId
    AND Entities.EntityId = SentenceEntities.EntityId
    AND Contains(Sentence, @Keyword)
    AND Sentences.SentenceDiscoverTime >= @DateBegin
    AND Sentences.SentenceDiscoverTime <= @DateEnd
GROUP BY Entities.EntityId
HAVING COUNT(Distinct Sentences.SentenceId) > 20
ORDER BY TFIDF desc, Freq desc
```

→ 全文索引



基于关系型数据库

❖ 优点

- 使用SQL进行操作，熟悉数据库的开发人员可以快速上手

❖ 缺点

- 当数据规模增大时，数据库性能明显下降
- 传统的关系型数据库不方便分布式扩展以及备份
- 关系型数据库的全文索引功能有限，不支持复杂字段类型，不支持复杂查询语法和复杂排序算法

❖ 实践

- 当文档数或者句子数达到几千万时，SQL Server的性能明显达不到在线应用的需求，通常需要几秒甚至几十秒才能返回结果



基于分布式大数据系统

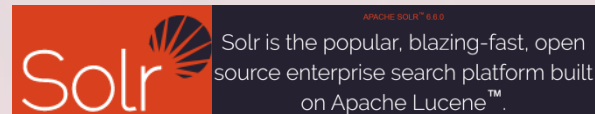
- ❖ 使用NOSQL数据库进行数据处理和计算，通过NOSQL数据库对应的方法进行聚合运算
 - 数据库或文件系统：MongoDB, HBASE, HDFS等
 - 聚合语言：聚合pipeline, map reduce, SQL, MDX等
 - 大数据OLAP系统：HIVE, Kylin, Mondrian等
- ❖ 典型方案：
 - HDFS + HIVE + Kylin + Mondrian + (Saiku)
- ❖ 优点和缺点
 - 基于大数据系统，分布式扩展比较容易，成熟的社区支持
 - 文本格式灵活，经常有嵌套格式的对象数据。HBASE和HDFS等对嵌套格式的对象存储支持的不是非常完善
 - 不支持全文索引或者全文索引能力差，查询能力受限



基于搜索系统

❖ 全文搜索系统（基于Lucene）

- Solr: <http://lucene.apache.org/solr/>



- ElasticSearch: <https://www.elastic.co/>



❖ 已经不仅仅是一个全文搜索系统

- 除文本外，还支持各种数据类型
- 支持自定义复杂数据类型，以及自定义的存储、排序和分析函数
- 强大的统计计算功能，支持数据库上的Join和Group



基于搜索系统

❖ Solr简介

- Faceting 和Pivot Faceting: 类似于SQL中的Group
- Parallel SQL: 开始支持类似SQL的查询方式
 - 索引集合被抽象为关系表
 - 全面支持WHERE语句
 - 分组聚合操作并行化, 使用MapReduce机制, 并自动使用Facet API提升性能
- 分布式Join
- SolrCloud: 良好的数据分布策略和分片机制



基于全文搜索系统



Dashboard

Logging

Core Admin

Java Properties

Thread Dump

newsdoc

Overview

Analysis

Dataimport

Documents

Files

Ping

Plugins / Stats

Query

Replication

Schema Browser

Statistics

Last Modified: 42 minutes ago
Num Docs: 17028042
Max Doc: 19414881
Heap Memory Usage: 335036968
Deleted Docs: 2386839
Version: 1579552
Segment Count: 69

Optimized:
Current:

optimize now

Replication (Slave)

	Version	Gen	Size
Master (Searching)	1499314232289	800531	197.95 GB
Master (Replicable)	1499314232289	800531	-
Slave (Searching)	1499311749918	800499	197.96 GB

Admin Extra

Request-Handler (qt)

/select

— common

q

.

fq

sort

start, rows

0

10

fl

df

Raw Query Parameters

key1=val1&key2=val2

wt

json

☒ indent

☐ debugQuery

☐ dismax

☐ edismax

☐ hl


☐ facet

☐ spatial

☐ spellcheck

Execute Query

搜索



Dashboard

Logging

Core Admin

Java Properties

Thread Dump

newsdoc

Overview

Analysis

Dataimport

Documents

Files

Ping

Plugins / Stats

Query

Replication

Schema Browser

Request-Handler (qt)

/select

common

q

中国人民大学

fq

sort

Time desc

start, rows

010

fi

Url, Title, Entity_Person, Entity_Organization

df

Raw Query Parameters

key1=val1&key2=val2

wt

json

☒ indent

☐ debugQuery

☐ dismax

☐ edismax

☐ hl

☐ facet

☐ spatial

☐ spellcheck

Execute Query

http://183.174.228.10:9993/solr/newsdoc/select?q=%E4%B8%AD%E5%9B%BD%E4%BA%BC

```
{
  "responseHeader": {
    "status": 0,
    "QTime": 0,
    "params": {
      "q": "中国人民大学",
      "indent": "true",
      "fl": "Url, Title, Entity_Person, Entity_Organization",
      "sort": "Time desc",
      "wt": "json",
      "_": "1499314530299"
    }
  },
  "response": {
    "numFound": 82778,
    "start": 0,
    "docs": [
      {
        "Url": "http://news.stnn.cc/hk_taiwan/2017/0706/446780.shtml",
        "Title": "干强：美对台售武 给蔡英文双重打击",
        "Entity_Person": [
          "蔡英文",
          "布什",
          "奥巴马",
          "特朗普",
          "赖清德",
          "叶望辉"
        ],
        "Entity_Organization": [
          "中评社",
          "国际关系学院公共管理系",
          "中国人民大学公共管理学院",
          "美国国务院"
        ]
      },
      {
        "Url": "http://news.ifeng.com/a/20170706/51383926_0.shtml",
        "Title": "北京市委书记蔡奇出席清华大学党代会",
        "Entity_Organization": [
          "北京市委",
          "清华大学",
          "清华新闻网",
          "中国共产党清华大学第十四次党员代表大会",
          "清华学堂",
          "中共北京市委",
          "中共教育部党组",
          "市委教育工委",
          "市委",
          "北京市委教育工委",
          "中共清华大学第十四次党员代表大会",
          "教育部",
          "北京大学",
          "中国人民大学",
          "北京师范大学",
          "中国农业大学",
          "北京协和医学院",
          "北京工业大学",
          "首都师范大学"
        ]
      }
    ]
  }
}
```

分析聚合



Dashboard

Logging

Core Admin

Java Properties

Thread Dump

newsdoc

Overview

Analysis

Dataimport

Documents

Files

Ping

Plugins / Stats

Query

Replication

Schema Browser

Request-Handler (qt)

/select

common

q

中国人民大学

fq

sort

Time desc

start, rows

0

10

fl

Url, Title, Entity_Person, Entity_Orgnization

df

Raw Query Parameters

key1=val1&key2=val2

wt

json

☒ indent

☐ debugQuery

☐ dismax

☐ edismax

☐ hl

☒ facet

facet.query

facet.field

Entity_Person

facet.prefix

☐ spatial

☐ spellcheck

Execute Query

http://183.174.228.10:9993/solr/newsdoc/select?q=%E4%B8%AD%

```
{
  "responseHeader": {
    "status": 0,
    "QTime": 109,
    "params": {
      "q": "中国人民大学",
      "facet.field": "Entity_Person",
      "indent": "true",
      "fl": "Url, Title, Entity_Person, Entity_Orgnization",
      "sort": "Time desc",
      "rows": "0",
      "wt": "json",
      "facet": "true",
      " _": "1499314646977"
    }
  },
  "response": {
    "numFound": 82778,
    "start": 0,
    "docs": []
  },
  "facet_counts": {
    "facet_queries": {},
    "facet_fields": {
      "Entity_Person": [
        "习近平",
        6689,
        "李克强",
        2476,
        "马克思",
        2298,
        "刘俊海",
        2273,
        "赵锡军",
        1993,
        "奥巴马",
        1571,
        "刘元春",
        1545,
        "王义棉",
        1426,
        "金灿荣",
        1208,
        "董希淼",
        1036,
        "郑风田",
        997,
        "韩大元",
        941,
        "汪洋",
        932,
        "邱宝昌",
        931,
        "沃尔玛"
      ]
    }
  }
}
```

An Introduction to Database system

交互式分析

❖ 切块

Request-Handler (qt)

/select

— common

q

中国人民大学 AND Time:[2017-06-01T00:00:00.0000Z TO *]

fq

sort

Time desc

start, rows

0 0

fl

Time, Url, Title, Entity_Person, Entity_Organization

df

Raw Query Parameters

key1=val1&key2=val2

wt

json

☒ indent

☐ debugQuery

☐ dismax

☐ edismax

☐ hl

— ☒ facet

facet.query

facet.field

Entity_Person

facet.prefix

☐ spatial

☐ spellcheck

Execute Query

http://183.174.228.10:9993/solr/newsdoc/select?q=%E4%B8%AD%

```
{
  "responseHeader": {
    "status": 0,
    "QTime": 78,
    "params": {
      "q": "中国人民大学 AND Time:[2017-06-01T00:00:00.0000Z TO *]",
      "facet.field": "Entity_Person",
      "indent": "true",
      "fl": "Time, Url, Title, Entity_Person, Entity_Organization",
      "sort": "Time desc",
      "rows": "0",
      "wt": "json",
      "facet": "true",
      "_": "1499314820328"
    }
  },
  "response": {
    "numFound": 1863,
    "start": 0,
    "docs": []
  },
  "facet_counts": {
    "facet_queries": {},
    "facet_fields": {
      "Entity_Person": [
        "习近平",
        221,
        "刘俊海",
        93,
        "董希淼",
        81,
        "马克思",
        73,
        "刘伟",
        69,
        "刘强东",
        68,
        "靳诺",
        64,
        "赵锡军",
        62,
        "肖中华",
        56,
        "陈卫东",
        51,
        "特朗普",
        47,
        "杨东",

```


Request-Handler (qt)

/select

— common —

q

中国人民大学

fq

sort

start, rows

0 0

fl

df

Raw Query Parameters

key1=val1&key2=val2

wt

json

☒ indent

☐ debugQuery

☐ dismax

☐ edismax

☐ hl

— ☒ facet —

facet.query

facet.field

Entity_Person

facet.prefix

http://183.174.228.10:9993/solr/new

```
{
  "responseHeader": {
    "status": 0,
    "QTime": 313,
    "params": {
      "q": "中国人民大学",
      "facet.field": "Entity_Person",
      "indent": "true",
      "rows": "0",
      "wt": "json",
      "facet": "true",
      "_": "1499336435021"
    }
  },
  "response": {
    "numFound": 82799,
    "start": 0,
    "docs": []
  },
  "facet_counts": {
    "facet_queries": {},
    "facet_fields": {
      "Entity_Person": [
        "习近平",
        6691,
        "李克强",
        2477,
        "马克思",
        2299,
        "刘俊海",
        2274,
        "赵锡军",
        1993,
        "奥巴马",
        1572,
        "刘元春",
        1545,
        "王义槐",
        1426,
        "金灿荣",
        1208,
        "董希淼",

```

Request-Handler (qt)

/select

— common —

q

中国人民大学

fq

Entity_Person:刘伟

sort

start, rows

0 0

fl

df

Raw Query Parameters

key1=val1&key2=val2

wt

json

☒ indent

☐ debugQuery

☐ dismax

☐ edismax

☐ hl

— ☒ facet —

facet.query

facet.field

Entity_Person

facet.prefix

http://183.174.228.10:9993/solr/new

```
{
  "responseHeader": {
    "status": 0,
    "QTime": 62,
    "params": {
      "q": "中国人民大学",
      "facet.field": "Entity_Person",
      "indent": "true",
      "fq": "Entity_Person:刘伟",
      "rows": "0",
      "wt": "json",
      "facet": "true",
      "_": "1499336478617"
    }
  },
  "response": {
    "numFound": 769,
    "start": 0,
    "docs": []
  },
  "facet_counts": {
    "facet_queries": {},
    "facet_fields": {
      "Entity_Person": [
        "刘伟",
        769,
        "靳诺",
        197,
        "习近平",
        178,
        "马克思",
        131,
        "郑水泉",
        118,
        "洪大用",
        99,
        "张建明",
        95,
        "伊志宏",
        93,
        "吴付来",
        89,
        "王利明",
        89,

```

文本大数据分析系统的挑战

❖ 高效实用的文本大数据分析系统

- 高性能（近）实时分析平台（搜索+聚合分析）
- 高效的采样和摘要技术
- 深层次信息抽取和语义理解
- 分析维度的自动发现
- 统计结果排序函数
- 数据质量控制
- 。 。 。

❖ 文本大数据分析需要信息检索、自然语言处理、机器学习、数据库等各个领域相结合



总结

- ❖ 交互式文本大数据分析系统：时事探针
- ❖ 自然语言处理与文本挖掘基础算法
- ❖ 文本搜索、文本分析系统构建

- ❖ 玩转大数据：
 - <http://websensor.playbigdata.com/fss3/>
 - <http://www.bigtextdata.com/>

