How to generate large-scale data from small-scale real-world data sets?

Gang Lu

Institute of Computing Technology, Chinese Academy of Sciences

BigDataBench Tutorial
MICRO 2014 Cambridge, UK
Motivation

- Benchmarking big data systems
  - The first thing is to obtain BIG data
- Obtaining REAL big data sets?
  - Large companies possess a lot of data
    - Confidentiality issue (User privacy)
  - Transferring big data sets is rather expensive

Is it possible to use large scale synthetic data?
Goals

- Generating synthetic data to satisfy 4V properties of big data
  - Volume
  - Velocity
  - Variety
  - Veracity

Big Data Generator Suite (BDGS)
Architecture of BDGS

Various Real data sets → Modeling → Parameters

Parallel generators → Big synthetic Data

Format conversion tools → Various workloads

Veracity
Variety

Velocity
Volume
**Veracity and Variety**

- From real world data, we can get:

<table>
<thead>
<tr>
<th>Veracity</th>
<th>Variety</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data set</td>
<td>Data type</td>
</tr>
<tr>
<td>Wikipedia Entries</td>
<td>un-structured</td>
</tr>
<tr>
<td>Amazon Movie Reviews</td>
<td>semi-structured</td>
</tr>
<tr>
<td>Google Web Graph</td>
<td>un-structured</td>
</tr>
<tr>
<td>Facebook Social Graph</td>
<td>un-structured</td>
</tr>
<tr>
<td>E-commerce Transaction</td>
<td>structured</td>
</tr>
<tr>
<td>Profsearch Person Resume</td>
<td>semi-structured</td>
</tr>
</tbody>
</table>
## Original size of real data sets

<table>
<thead>
<tr>
<th>data sets</th>
<th>data size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wikipedia Entries</td>
<td>4,300,000 English articles</td>
</tr>
<tr>
<td>Amazon Movie Reviews</td>
<td>7,911,684 reviews</td>
</tr>
<tr>
<td>Google Web Graph</td>
<td>875713 nodes, 5105039 edges</td>
</tr>
<tr>
<td>Facebook Social Network</td>
<td>4039 nodes, 88234 edges</td>
</tr>
<tr>
<td>E-commerce Transaction</td>
<td>table1: 4 columns, 38658 rows. table2: 6 columns, 242735 rows</td>
</tr>
<tr>
<td>Person Resumes Data</td>
<td>278956 resumes</td>
</tr>
</tbody>
</table>

Use BDGS to scale up these data sets
What does BDGS provide?

- Text generator
- Graph generator
- Table generator
Text generator

- Use LDA (Latent Dirichlet Allocation) (David M Blei, et al.) to generate text corpus.
  - **Topic model**
    - To model the information of semantic level
  - Widely used in machine learning and natural language processing
How to generate a new document

1. Select a topic randomly following a multinomial distribution.
2. Select words randomly following a multinomial distribution under the selected topic.

Example words:
- machine
- evaluate
- big
- data
- architecture
- benchmarking
- memory
- system
- mining
- learning
- CPU
- machine learning
- memory system
- data benchmarking

Example topics:
- topic1
- topic2
- topic3

Example output:
- New document generated
- CPU used

Progress:
- Select topic
- Select words
- New document

BigDataBench Micro 2014 ICT Institute of Computing Technology, Chinese Academy of Sciences
Latent dirichlet allocation

Latent dirichlet allocation

Dirichlet distribution

Multinomial distribution

Hyper parameters

three-level hierarchical Bayesian model

\[
p(\theta, z, w | \alpha, \beta) = p(\theta | \alpha) \prod_{n=1}^{N} p(z_n | \theta) p(w_n | z_n, \beta)
\]

We can use expectation-maximization algorithm to determine $\alpha$ and $\beta$. 
How to use it to generate texts?

- Go into the directory of `BigDataGeneratorSuite`

```
gen_text_data.sh
  <model name>
  <number of files>
  <number of lines>
  <number of words>
  <output dir >
```

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>model name</td>
<td>the name of model used to generate new data (lda_wiki1w or amazonMR)</td>
</tr>
<tr>
<td>No. of files</td>
<td>the number of files to be generated</td>
</tr>
<tr>
<td>No. of lines</td>
<td>number of lines in each file</td>
</tr>
<tr>
<td>No. of words</td>
<td>number of words in each line</td>
</tr>
<tr>
<td>Output dir</td>
<td>output director</td>
</tr>
</tbody>
</table>

- An example
  - sh gen_text_data.sh `lda_wiki1w 10 100 1000 gen_data/`

Note: Installation of the GSL-GNU Scientific Library is needed.
Graph generator

- Use the Stochastic Kronecker Graph model (Jure Leskovec, et al.) to generate graph
  - Used also by graph 500
  - Different from Graph 500, our graph is application specific, the stochastic kronecker initiator is obtained from real representative data set of
Deterministic Kronecker Graph

1: has edge
0: no edge

(a) Graph $K_1$

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>1</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

(b) Intermediate stage

(c) Graph $K_2 = K_1 \otimes K_1$

<table>
<thead>
<tr>
<th></th>
<th>$K_1$</th>
<th>$K_1$</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>$K_1$</td>
<td>$K_1$</td>
<td>$K_1$</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>$K_1$</td>
<td>$K_1$</td>
<td></td>
</tr>
</tbody>
</table>

(d) Adjacency matrix

(e) Adjacency matrix

Stochastic Kronecker Graph

The probability with which the cell generate a edge

Application-specific

Specific real data → Estimate parameters → Scale Big graph → input Specific application

Google Web Graph: KronFit → Stochastic Kronecker → Synthetical Web Graph → PageRank

Facebook Social Graph: KronFit → Stochastic Kronecker → Synthetical Social Graph → Connected components
How to use it to generate graphs?

Go into the directory of `BigDataGeneratorSuite`

```
gen_kronecker_graph
<output file>
<matrix>
<iteration>
<random seed>
```

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>output file</td>
<td>output file name (default: 'graph.txt')</td>
</tr>
<tr>
<td>matrix</td>
<td>Matrix (in Matlab notation) (default: [0.9, 0.5; 0.5, 0.1])</td>
</tr>
<tr>
<td>iteration</td>
<td>Iteration of using kronecker product (default: 5)</td>
</tr>
<tr>
<td>random seed</td>
<td>Time seed of random algorithm (default: 0)</td>
</tr>
</tbody>
</table>

An example

```
sh gen_kronecker_graph -o:../data-outfile/amazon gen.txt
-m:"0.7196 0.6313; 0.4833 0.3601" -i:23
```
Table generator

- Related structured table
  - Parallel Data Generation Framework (Tilmann Rabl, et al.)
    - PDGF is also used by BigBench and TPC-DS
    - using XML configuration files for data description and distribution

- Semi-structured resumes
  - choose a mix of fields, each field follows bernoulli distribution
How to use it to generate tables?

- Go into the directory of BigDataGeneratorSuite

```
pdgf.jar  
-l schema.xml  
-l generation.xml  
-sf 2000
```

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Explannation</th>
</tr>
</thead>
<tbody>
<tr>
<td>schema.xml</td>
<td>the schema configuration: the structure of the data and the generation rules</td>
</tr>
<tr>
<td>generation.xml</td>
<td>the generation configuration defines the output and post-processing of generated data</td>
</tr>
<tr>
<td>sf</td>
<td>A multiple increase in the reference data base 100 000</td>
</tr>
</tbody>
</table>

- An example

```
```
Any Questions