Large Scale Log Analysis and Mining

Zheng CHEN
Machine Learning Group, MSRA
Motivation

• Rapid growth of online business
  – Live search, MSN portal, Office Online, Global Foundation Services (GFS), ...

• Log mining (web analytics) is one of ways to make data-driven decisions

• Weak capability on analysis and decision making
  – Outsource to Omniture & WebTrend
  – Analyze by channel, and information is fragmented
Challenges

• Data scale
  – Terabytes in size: Live search (600GB /day), MSN portal (500GB /day)
  – Billions in rows: Live search (500M records /day), MSN portal (1.3 billion records /day)

• Computation complexity
  – Range from $O(n)$ to $O(n^2)$ in most cases

• Streaming data
  – Snapshot vs. incremental

• System extensibility
  – Built-in analysis vs. pluggable analysis
# Case Study – One Typical Analysis

**Data Source:** MSN portal log

<table>
<thead>
<tr>
<th>Time</th>
<th>UserID</th>
<th>Gender</th>
<th>Age</th>
<th>Location</th>
<th>URL</th>
<th>Referer</th>
<th>Browser</th>
<th>OS</th>
<th>Market</th>
</tr>
</thead>
</table>

Each day: 1 billion rows, 500GB

14 hours by single machine if disk read speed is 10MB/s

**Analysis Result:** Traffic stat. by category and market

<table>
<thead>
<tr>
<th>Date</th>
<th>Category</th>
<th>Market</th>
<th>Traffic</th>
</tr>
</thead>
</table>

Each day: 200,000 rows, 2MB

40 seconds by single machine to do an aggregation on 6 month’s data

**UI:** Overall traffic trend
Our Past Investment

- Over two dozens of log mining related research papers from ML Group, published on WWW, SIGIR, AAAI, ICDM, ...

- Research fruits covered
  - Finding potential audience
  - Site performance measurement and recommendation
  - User modeling
    - Interest model
    - Access behavioral model
    - Demographic model
  - Search enhancement
  - Content understanding
    - Translation, summarization, ...
  - Temporal analysis

2008-11-13
Goal: Scale-up Our Research

- Build on top of distributed platform
- Support both statistical analysis and complex mining
- Hourly/daily/monthly incremental updating
- Easily customized to different log, different analysis and different mining algorithms
Our Solution

- A large scale log analysis platform
  - Offline processing pipeline
  - Online query engine

- Build on top of a distributed platform (SILK)

- Easy to be customized to different log data and different analyses tasks
Data Partition Strategy

- Problem to solve
  - Incremental data update
  - Analysis performance
  - Data join for search query log & click log

- Partition method
  - First partition key: day (or can be configured to other time unit)
  - Second partition key
    - User ID: for session analyses, click analyses, user analyses etc.
    - None: for other statistic analyses
Component Stack

Offline Processing Pipeline
- Copy Pipeline
- Importing Pipeline
- Analysis Pipeline
- DTL
- Pipeline Management and Monitoring

Online Query Engine
- MDA Engine
- Cache Management
- Service Monitor

General Purpose Distributed Platform (SILK)
• Live search analysis
  – 3 months’ S-Log and G-Log
• MSN pageview analysis
  – 3 months’ C-Log
• MSW log analysis
• Toolbar log analysis
• Messenger log analysis
• Hotmail log analysis
• ...
Performance

• Offline performance (for daily C-log data)
  - On 8 machines
    • Importing: 6 hours
    • Data analysis (21 analyses): 4 hours
  - On 60 machines (shared with other app.)
    • Importing: 2 hours
    • Data analysis (21 analyses): 1.5 hours

• Online performance
  - Linear with the machine number
Selected Examples

- Smartmatch pipeline
  - Collaborate w/ MSN adCenter

- IME new word pipeline
  - Collaborate w/ Office CTG (China/Japan)

- Web Analytics
  - Collaborate w/ CCTV.COM
  - Global Foundation Services (GFS)
  - AP / Office Online
Make Log Mining Extensible

- Open Log Analysis and Mining Platform (LAMP)
  - Supports large scale and daily updated log volume
  - Pluggable incremental analysis module
  - Customizable and interactive reporting

- Advantage: simplified programming interface
  - Programming efficiently on log mining algorithms
  - Coding at client side, running on clusters
  - Little knowledge on distributed systems required
  - Needn’t care about storage and distribution
Hosted w/ UR team (July 2008)
- 37 interns from 10 research groups attended
- Worked with MSRA.CN site log

11 different topics attended final presentation
- Discovered attacking from hackers!
- Star search engine in China
- Website optimization by sequential mining
- ...

Applying projects from competition to optimize MSRA.CN website
Summary

- Large scale log analysis and mining is critical for online businesses

- Scale up research fruits on log mining, and tested on several products

- Extend effort to LAMP to enable more researchers contribute and/or benefit from log mining research
Q & A