Image Retrieval in the Unstructured Data Management System AUDR

Junwu Luo, Bo Lang, Chao Tian, Danchen Zhang
Dept. of Computer Science and Engineering, Beihang University
Outline

• Review of related work

• The scalable architecture of image management
  ✓ Tetrahedral data model of image
  ✓ An advanced unstructured data repository-AUDR
  ✓ Image management system in AUDR

• A composite image retrieval algorithm
  ✓ Visual retrieval
  ✓ Textual retrieval
  ✓ Fusion techniques

• Distributed storage and paralleled retrieval of images
  ✓ Distributed image storage engine
  ✓ Paralleled image retrieval engine

• Results and discussions
Related work

- **Motivation**
  - Exponential increase in digital image database sizes
  - Exponential increase in computing power and storage capacity
  - Increased use of image in entertainment, education, medicine, commercial fields.

- **Text-Based Image Retrieval**
  - Proposed in 1970s, such as Google, Yahoo! etc.
  - Manual annotation of images
  - Use text-based retrieval methods

- **Content-Based Image Retrieval**
  - Proposed in 1990s, such as QBIC, VisualSEEK, Photobook etc.
  - Extract visual features: color, shape, textual etc.

- **Image data management system**
  - Relational database: scalability and efficiency problem
  - Unstructured data management system
Tetrahedral data model of image

• Definition
Tetrahedron = (V, BA, SF, LF, RD, CONJS)
- V: identifier of tetrahedron
- BA: basic attributes facet
- SF: semantic feature facet
- LF: low-level feature facet
- RD: raw data
- CONJS: association between two facets

• Property
✓ integrity and data independence
✓ inner-correlation, extensibility, easy to implement
An advanced unstructured data repository

- Hierarchical structure
  - data processing layer
  - data operation layer
  - data control layer
  - data interface layer
  - data service layer

- Advantage
  - excellent expansion
  - high-efficient
Image management system in AUDR

- **Distributed image storage server**
  - extract features in parallel
  - provide access interface

- **Paralleled image retrieval sub-engine**
  - master-slave architecture
  - index and memory cache

- **Advantage**
  - massive data storage
  - real-time retrieval
A composite image retrieval algorithm

- **Visual Retrieval**
  - ✓ Simple Color Histogram
  - ✓ Tamura Texture Feature
  - ✓ Fuzzy Color and Texture Histogram
  - ✓ SIFT local feature

- **Textual Retrieval**
  - The Vector Space Model
    - ✓ Apache Lucene Tool
    - ✓ BM25 scoring approach
  - The Topic Model
    - ✓ Latent Dirichlet Allocation
    - ✓ Bayes chain: topic~word and document~topic

- **Fusion Techniques**
  - combSUM: visual retrieval, textual retrieval
  - combMNZ: mixed retrieval

- Local index:
  - Locality-sensitive hashing index
  - Inverted index
Distributed image storage engine

- **Storage strategy**
  - ✓ HDFS: raw data, including original image and thumbnails
  - ✓ HBase: basic attributes, semantic and visual features

- **MapReduce Processing:**
  - ✓ Namenode: split data into segments
  - ✓ Datanode: start one map task to process each segment.
Paralleled image retrieval engine

- **Master**
  - ✓ maintain information of all the slaves
  - ✓ schedule the retrieval task
  - ✓ merge retrieval results

- **Slave**
  - ✓ undertake computing task on local data
  - ✓ send local results to Master
Experimental results

- **Dataset**
  - ✓ ImageCLEF 2011 medical dataset: 231,000 images associated with metadata
  - ✓ ImageNet: 111,489 images and add annotations manually

- **Implementation**
  - ✓ storage cluster: 9 nodes using PC, 1 master and 8 slaves (3GHz, 4G Memory)
  - ✓ computing cluster: 3 nodes, 1 master (2.13GHz, 16G memory), 2 slaves (2.93GHz, 4G memory)

- **Evaluation of visual retrieval on ImageCLEF dataset**

<table>
<thead>
<tr>
<th>Features</th>
<th>MAP</th>
<th>P10</th>
<th>P20</th>
<th>Bpref</th>
</tr>
</thead>
<tbody>
<tr>
<td>SCH</td>
<td>0.0047</td>
<td>0.0367</td>
<td>0.0267</td>
<td>0.0497</td>
</tr>
<tr>
<td>Tamura</td>
<td>0.0039</td>
<td>0.0297</td>
<td>0.0263</td>
<td>0.0454</td>
</tr>
<tr>
<td>FCTH</td>
<td>0.0089</td>
<td>0.0733</td>
<td>0.0517</td>
<td>0.0567</td>
</tr>
<tr>
<td>SIFT</td>
<td>0.0032</td>
<td>0.0265</td>
<td>0.0234</td>
<td>0.0412</td>
</tr>
<tr>
<td>FUSE</td>
<td>0.0112</td>
<td>0.0700</td>
<td>0.0617</td>
<td>0.0555</td>
</tr>
</tbody>
</table>

✓ the global features are better than the local features
✓ fusing results obtained by single feature can improve performance
Experimental results

Evaluation of textual retrieval on ImageCLEF dataset

<table>
<thead>
<tr>
<th>Method</th>
<th>MAP</th>
<th>P10</th>
<th>P20</th>
<th>Bpref</th>
</tr>
</thead>
<tbody>
<tr>
<td>lucene</td>
<td>0.1758</td>
<td>0.3133</td>
<td>0.27</td>
<td>0.2187</td>
</tr>
<tr>
<td>lucene_and_tm</td>
<td>0.1917</td>
<td>0.34</td>
<td>0.305</td>
<td>0.2237</td>
</tr>
<tr>
<td>BM25_and_tm</td>
<td>0.0878</td>
<td>0.19</td>
<td>0.165</td>
<td>0.125</td>
</tr>
</tbody>
</table>

- Topic model could discover the abstract topics.
- The results produced by lucene and topic model are more semantic related.

Evaluation of mixed retrieval on ImageCLEF dataset

<table>
<thead>
<tr>
<th>Method</th>
<th>MAP</th>
<th>P10</th>
<th>P20</th>
<th>Bpref</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear fusion</td>
<td>0.1556</td>
<td>0.2764</td>
<td>0.2350</td>
<td>0.1925</td>
</tr>
<tr>
<td>Pseudo-relevance</td>
<td>0.2341</td>
<td>0.3643</td>
<td>0.3327</td>
<td>0.2405</td>
</tr>
</tbody>
</table>

- Linear fusion: reduce the performance for the semantic gap.
- Pesudo-relevance: take the caption of top 20 in visual runs as query expansion to boost the text runs, it enhances the result obviously.
Experimental results

- System Performance on ImageNet dataset

✓ Storage performance: the advantage of MapReduce processing mode is obvious.
✓ Retrieval performance: parallel way and index can accelerate efficiency greatly.
Conclusions

- Propose a scalable architecture for image management based on Tetrahedral Data Model in an advanced unstructured data repository-AUDR
  - distributed storage engine
  - parallel computing engine
- Propose a new image retrieval algorithm incorporating rich visual features, two text models and the specific fusion techniques
- Feature work
  - query expansion using specific terminologies
  - machine learning methods to support more intelligent management
Thank You

Questions?
luojunwu1988@163.com