NExT Search Center
A NUS-Tsinghua Joint Center on Extreme Search

CHUA Tat-Seng 蔡达成
KITHCT Chair Professor
School of Computing,
National University of Singapore
Rich UGC Contents

- We are constantly being exposed to huge amount of UGCs (user-generated contents) from a variety of social networks:
  - which collectively offer insights on interests/concerns of users and pulses of a city

- But just how big is the UGC?

- In every 60 seconds:
  - two million videos are viewed on YouTube
  - 700,000 messages are delivered by way of Facebook
  - 175,000 tweets are fired off into the ether
  - 2,000 Foursquare check-ins tell the world where we are!

(http://venturebeat.com/2012/02/25/60-seconds-social-media/)
The NExT Project

• Collectively, the UGC’s provide a rich source of contents that signal the pulses of a city

• Uniqueness of NExT Center:
  o Gather and analyze LIVE UGC’s related to a city
  o Infer relationships between topics and among people in a city/organization
  o Leverage on our multilingual, multimedia and multi-cultural research
  o Key research focuses: Live; Big Data; Multi-faceted

• Gathering data on Beijing and Singapore
Types of UGC’s we Gathered

**Type 1:** Images/Videos & Check-ins
- Images/Videos
- Check-in Venues

**Type 2:** Contents: User Comments, cQA, Tweets
- User Comments/cQA
- Twitters
- Social News

**Type 3:** Local Apps
- Mobile Apps

**Type 4:** Structured Data
- Str. Data

**Structured Contents**

**People, Domain, Social, Loc & Mobile**

**Users**

From Unstructured Info to Structured Knowledge
CONTENTS

• The Live Observatory
• First Order Analytics
• Large-Scale Image Search
• Bridging User Intention Gap
• Higher Order Analytics
• Summary
Live Social Observatory
Window to outside world (137.132.145.151)

To explore live data and events as they unfold
NExT Live UGC Data Crawling
The Sites that we crawled
Live Crawlers
- Crawl multiple source UGC data every minute and second

Mango DB
- Insert downloaded text data into MangoDB

Solr Index
- Build the index for text search by Solr in real time

Image Index
- Build hash-based index for content-based image search

Multiple APIs
- Provide multiple APIs to access these data
CONTENTS

• The Live Observatory
• First Order Analytics
• Large-Scale Image Search
• Bridging User Intention Gap
• Higher Order Analytics
• Summary
Live Social Observatory

Live Crawler  Location Sense  People Sense  Topics Sense  ORG Sense

ORG Sense takes care of the WWW of your organization, in a word, what people are saying about your organization, who these people are, and where they are.

Try it!
## Live Data Statistics (up to 8 October)

<table>
<thead>
<tr>
<th>Data Source</th>
<th>Data Type</th>
<th>Start Date</th>
<th>Records</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amazon</td>
<td>Images</td>
<td>2012-05-01</td>
<td>1,577,839</td>
</tr>
<tr>
<td></td>
<td>Products</td>
<td>2012-05-01</td>
<td>2,911,538</td>
</tr>
<tr>
<td></td>
<td>Reviews</td>
<td>2012-05-01</td>
<td>3,615,097</td>
</tr>
<tr>
<td></td>
<td>Images</td>
<td>2012-05-01</td>
<td>2,236,682</td>
</tr>
<tr>
<td></td>
<td>Reviews</td>
<td>2012-05-01</td>
<td>5,067,873</td>
</tr>
<tr>
<td></td>
<td>Shops</td>
<td>2012-05-01</td>
<td>142,320</td>
</tr>
<tr>
<td></td>
<td>Images</td>
<td>2012-05-01</td>
<td>234,816</td>
</tr>
<tr>
<td></td>
<td>Reviews</td>
<td>2012-05-01</td>
<td>727,250</td>
</tr>
<tr>
<td></td>
<td>Shops</td>
<td>2012-05-01</td>
<td>106,322</td>
</tr>
<tr>
<td></td>
<td>Images</td>
<td>2012-05-21</td>
<td>66,257,149</td>
</tr>
<tr>
<td></td>
<td>Reviews</td>
<td>2012-05-21</td>
<td>109,745,000</td>
</tr>
<tr>
<td></td>
<td>Users</td>
<td>2012-05-05</td>
<td>2,060,357</td>
</tr>
<tr>
<td></td>
<td>Checkins</td>
<td>2012-04-28</td>
<td>5,437,393</td>
</tr>
<tr>
<td></td>
<td>Venues</td>
<td>2012-04-28</td>
<td>2,010,944</td>
</tr>
<tr>
<td>Foursquare</td>
<td>Images</td>
<td>2012-08-15</td>
<td>615,848</td>
</tr>
<tr>
<td></td>
<td>Images</td>
<td>2012-05-05</td>
<td>292,490</td>
</tr>
<tr>
<td></td>
<td>Reviews</td>
<td>2012-05-05</td>
<td>73,078</td>
</tr>
<tr>
<td>Flickr</td>
<td>Images</td>
<td>2012-05-01</td>
<td>109,745,000</td>
</tr>
<tr>
<td></td>
<td>Products</td>
<td>2012-05-01</td>
<td>2,911,538</td>
</tr>
<tr>
<td></td>
<td>Reviews</td>
<td>2012-05-01</td>
<td>3,615,097</td>
</tr>
<tr>
<td></td>
<td>Images</td>
<td>2012-05-01</td>
<td>2,236,682</td>
</tr>
<tr>
<td></td>
<td>Reviews</td>
<td>2012-05-01</td>
<td>5,067,873</td>
</tr>
<tr>
<td></td>
<td>Shops</td>
<td>2012-05-01</td>
<td>142,320</td>
</tr>
<tr>
<td></td>
<td>Images</td>
<td>2012-05-01</td>
<td>234,816</td>
</tr>
<tr>
<td></td>
<td>Reviews</td>
<td>2012-05-01</td>
<td>727,250</td>
</tr>
<tr>
<td></td>
<td>Shops</td>
<td>2012-05-01</td>
<td>106,322</td>
</tr>
<tr>
<td></td>
<td>Images</td>
<td>2012-05-21</td>
<td>66,257,149</td>
</tr>
<tr>
<td></td>
<td>Reviews</td>
<td>2012-05-21</td>
<td>109,745,000</td>
</tr>
<tr>
<td></td>
<td>Users</td>
<td>2012-05-05</td>
<td>2,060,357</td>
</tr>
<tr>
<td></td>
<td>Checkins</td>
<td>2012-04-28</td>
<td>5,437,393</td>
</tr>
<tr>
<td></td>
<td>Venues</td>
<td>2012-04-28</td>
<td>2,010,944</td>
</tr>
<tr>
<td>Instagram</td>
<td>Images</td>
<td>2012-08-15</td>
<td>615,848</td>
</tr>
<tr>
<td></td>
<td>Images</td>
<td>2012-05-05</td>
<td>292,490</td>
</tr>
<tr>
<td></td>
<td>Reviews</td>
<td>2012-05-05</td>
<td>73,078</td>
</tr>
<tr>
<td>Panoramio</td>
<td>Images</td>
<td>2012-05-01</td>
<td>109,745,000</td>
</tr>
<tr>
<td></td>
<td>Products</td>
<td>2012-05-01</td>
<td>2,911,538</td>
</tr>
<tr>
<td></td>
<td>Reviews</td>
<td>2012-05-01</td>
<td>3,615,097</td>
</tr>
<tr>
<td></td>
<td>Images</td>
<td>2012-05-01</td>
<td>2,236,682</td>
</tr>
<tr>
<td></td>
<td>Reviews</td>
<td>2012-05-01</td>
<td>5,067,873</td>
</tr>
<tr>
<td></td>
<td>Shops</td>
<td>2012-05-01</td>
<td>142,320</td>
</tr>
<tr>
<td></td>
<td>Images</td>
<td>2012-05-01</td>
<td>234,816</td>
</tr>
<tr>
<td></td>
<td>Reviews</td>
<td>2012-05-01</td>
<td>727,250</td>
</tr>
<tr>
<td></td>
<td>Shops</td>
<td>2012-05-01</td>
<td>106,322</td>
</tr>
<tr>
<td></td>
<td>Images</td>
<td>2012-05-21</td>
<td>66,257,149</td>
</tr>
<tr>
<td></td>
<td>Reviews</td>
<td>2012-05-21</td>
<td>109,745,000</td>
</tr>
<tr>
<td></td>
<td>Users</td>
<td>2012-05-05</td>
<td>2,060,357</td>
</tr>
<tr>
<td></td>
<td>Checkins</td>
<td>2012-04-28</td>
<td>5,437,393</td>
</tr>
<tr>
<td></td>
<td>Venues</td>
<td>2012-04-28</td>
<td>2,010,944</td>
</tr>
<tr>
<td>Tencent Weibo</td>
<td>Famous People</td>
<td>2012-09-25</td>
<td>395</td>
</tr>
<tr>
<td></td>
<td>Images</td>
<td>2012-06-01</td>
<td>78,957,657</td>
</tr>
<tr>
<td></td>
<td>Hot Topics</td>
<td>2012-08-03</td>
<td>564</td>
</tr>
<tr>
<td></td>
<td>Tweets</td>
<td>2012-06-01</td>
<td>448,240,067</td>
</tr>
<tr>
<td></td>
<td>Users</td>
<td>2012-04-17</td>
<td>22,128,045</td>
</tr>
<tr>
<td></td>
<td>Users.relation</td>
<td>2012-04-17</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Images</td>
<td>2012-05-07</td>
<td>10,555,689</td>
</tr>
<tr>
<td></td>
<td>Hot Topics</td>
<td>2012-06-01</td>
<td>22,719</td>
</tr>
<tr>
<td></td>
<td>Tweets</td>
<td>2012-04-14</td>
<td>790,625,909</td>
</tr>
<tr>
<td></td>
<td>Users</td>
<td>2012-05-07</td>
<td>68,612,688</td>
</tr>
<tr>
<td></td>
<td>Famous People</td>
<td>2012-09-01</td>
<td>662</td>
</tr>
<tr>
<td></td>
<td>Images</td>
<td>2012-05-07</td>
<td>90,819,164</td>
</tr>
<tr>
<td></td>
<td>Hot Topics</td>
<td>2012-06-01</td>
<td>1,003</td>
</tr>
<tr>
<td></td>
<td>Tweets</td>
<td>2012-05-07</td>
<td>227,659,668</td>
</tr>
<tr>
<td></td>
<td>Users</td>
<td>2012-05-07</td>
<td>44,036,023</td>
</tr>
<tr>
<td></td>
<td>Reviews</td>
<td>2012-08-15</td>
<td>38,061,979</td>
</tr>
<tr>
<td></td>
<td>Videos</td>
<td>2012-08-15</td>
<td>52,259</td>
</tr>
<tr>
<td></td>
<td>Youtube</td>
<td>2012-08-15</td>
<td>92,457</td>
</tr>
</tbody>
</table>
## Non-Live Data Statistics (Up to 9 June)

<table>
<thead>
<tr>
<th>Data Source</th>
<th>Data Type</th>
<th>Data Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flickr (Beijing)</td>
<td>Images</td>
<td>2,622,134</td>
</tr>
<tr>
<td>Flickr (Singapore)</td>
<td>Images</td>
<td>2,621,677</td>
</tr>
<tr>
<td>Baidu Zhidao</td>
<td>QA Pairs</td>
<td>33,636</td>
</tr>
<tr>
<td>Yahoo Answer</td>
<td>QA Pairs</td>
<td>3,158,912</td>
</tr>
<tr>
<td>Wiki Answer</td>
<td>QA Pairs</td>
<td>49,018,100</td>
</tr>
<tr>
<td>27 Singapore Food Forums</td>
<td>Food Info pages</td>
<td>&gt; 600GB</td>
</tr>
<tr>
<td>Buzzillion</td>
<td>Product</td>
<td>5,601</td>
</tr>
<tr>
<td>CNet</td>
<td>Product</td>
<td>597</td>
</tr>
<tr>
<td>Gamarena</td>
<td>Product</td>
<td>51,055</td>
</tr>
<tr>
<td>Newegg</td>
<td>Product</td>
<td>159</td>
</tr>
<tr>
<td>Reevoo</td>
<td>Product</td>
<td>11,082</td>
</tr>
<tr>
<td>Viewpoint</td>
<td>Product</td>
<td>1,032</td>
</tr>
</tbody>
</table>
Where are the check-in data?
(Distribution in Singapore)
Where are the UGC data?
(Distribution in 北京)

52,949 records are found in All source and this search catch 1,000 of them (0.563 seconds).
Generation of First Order Analytics
(Map of UGCs on “Mother’s Day”)
Generation of First Order Analytics
(Cloud Maps of “Instagram” around 10th April 2012)

(a) Word cloud generated from sampled tweets w.r.t. “instagram” on Apr 09, 2012

(b) Word cloud generated from sampled tweets w.r.t. “instagram” on Apr 10, 2012
Generation of First Order Analytics
(Live Events in Singapore and Beijing)
Generation of First Order Analytics
(Live Location Window to Clark Quay)
• The Live Observatory
• First Order Analytics
• Large-Scale Image Search
• Bridging User Intention Gap
• Higher Order Analytics
• Summary
Given over 235 millions images, we want to develop robust techniques to index and search for images based on both keywords and contents.
Current Image Search Engines

- Google images
- Bing
- SOSO
- Yahoo!
- Baidu
- Gazopa
- Soutugou
- Sogou
- Picitup
- Imagine
- Incogna
- TinEye
- Terra Galleria Photography

Visual + Text + User-log → Image Search Engines

Image Search Engines → Text + User-log

Visual →
Example Search Results
One limitation: The feature must be sparse.

Our Approach
Our Framework

Input Image → Feature Extraction → Hash Code Generation → Hash Code Extension → Index by Lucene → Re-ranking

SIFT Feature Extraction

Sparse coding formulation in the dictionary and codebook:

\[
\min_{U, V} \sum_{m=1}^{M} \| x_m - u_m \|_2^2 + \lambda |u_m |
\]

Hash Code Generation

1. Fit Multidimensional Rectangle

Hash Code Extension

For the given image, the words are generated by the hash code and the code with a Hamming distance of 1 or 2. The number of words for each image is \(1 + C_{32}^1 + C_{32}^2\).

\[
\begin{array}{cccccccccccc}
0 & 1 & 1 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1
\end{array}
\]

011
• The Live Observatory
• First Order Analytics
• Large-Scale Image Search
• Bridging User Intention Gap
• Higher Order Analytics
• Summary
User often cannot express what she wants precisely in a query

“Intention Gap” between user’s search intent and query

Pose difficult in understanding user’s intent by search engine, leading to search results far from the intent

How to narrow down “Intention Gap”? 
Human in the Loop – Feedback

Relevance Feedback
(Rui et al. TCSVT’98)

Visual query suggestion
(Zha et al. MM’09)

Related Sample Feedback
(Yuan et al. TMM’11)

Attribute Feedback
(Zhang et al. MM’12)

Social Sense Media Search
(Cui et al. 2012)
Human in the Loop – Feedback

- **Relevance Feedback**
  - (Rui et al. TCSVT’98)

- **Visual query suggestion**
  - (Zha et al. MM’09)

- **Related Sample Feedback**
  - (Yuan et al. TMM’11)

- **Attribute Feedback**
  - (Zhang et al. MM’12)

- **Social Sense Media Search**
  - (Cui et al. 2012)
Visual Query Suggestion (VQS): a query suggestion scheme tailored to multimedia search

- provide both image and keyword suggestions
- refine the text-based image search using image suggestions as query examples
Visual Query Suggestion (Zha et al. MM’09)
Visual Query Suggestion - Evaluation

- Suggestion quality evaluation
  - Do VQS outperform existing engines with textual query suggestion in terms of eliciting users’ search intent?
- Search Performance Evaluation

- 3 million images, ≈15 million tags from Flickr
- 30 professional users, 10 average users
- 25 experimental queries covering different semantics, such as scene, object, event, etc.

VQS can help user deliver search intent more precisely, leading to better search results.
Human in the Loop – Feedback

Relevance Feedback
(Rui et al. TCSVT’98)

Related Sample Feedback
(Yuan et al. TMM’11)

Attribute Feedback
(Zhang et al. MM’12)

Social Sense Media Search
(Cui et al. 2012)

Visual query suggestion
(Zha et al. MM’09)

Relevance Feedback
(Rui et al. TCSVT’98)
Interactive search performance suffers from lack of relevant sample problem especially for complex queries.

Solution: allows users to feedback on related samples, rather than just +ve and −ve samples.
The related samples are usually visually similar with the relevant ones.

Relevant samples tend to appear in neighborhood of related samples.
Related Sample RF: Experiments

- Formulate as fusion of both visual similarity & temporal ranking
- Leads to substantial improvement in query performance

![Graphs showing performance comparison](a) TV08 Dataset  (b) YouTube Dataset

Figure 5.10: The performance comparison in each iteration between our approach and the state-of-art methods measured by MAP@1000

- Basis for memory-recall search
Human in the Loop – Feedback

- **Relevance Feedback** (Rui et al. TCSVT’98)
- **Visual query suggestion** (Zha et al. MM’09)
- **Related Sample Feedback** (Yuan et al. TMM’11)
- **Attribute Feedback** (Zhang et al. MM’12)
- **Social Sense Media Search** (Cui et al. 2012)
**Attribute Feedback (Zhang et al. MM’12)**

- **Attribute Feedback**: a novel interactive image search scheme:
  - help user express search intent more precisely

- **Limitation of RF**
  - Leaves search engine to infer user’s search intent from her feedbacks, in fact, from the low-level visual features, thus suffering from the “semantic gap” between search intent and low-level visual features.
  - Ineffective in narrowing down the search to user’s target

Traditional interactive search scheme: Relevance Feedback (RF)
Attribute Feedback (Zhang et al. MM’12)

- Allow user to deliver search intent by providing feedbacks on semantic attributes
- Feedbacks on attributes compose a semantic description of user intent
- Allow user to give affinity judgments on images containing the same attribute

- Can shape user’s intent more precisely and quickly
- Can quickly narrow down the search to user’s target with less interaction efforts

Intermediate-level **semantic attributes** act as the bridge connecting user intent and low-level visual features.
Attribute Feedback – Sample Results
Human in the Loop – Feedback

Relevance Feedback
(Rui et al. TCSVT’98)

Visual query suggestion
(Zha et al. MM’09)

Related Sample Feedback
(Yuan et al. TMM’11)

Attribute Feedback
(Zhang et al. MM’12)

Social Sense Media Search
(Cui et al. 2012)

NEXT
Social-Sensed Image Search (Cui et al 2012)

Personalization

- Search with query + intent

- Personalized search
  - Query log + log mining = personalization of general search engines

Lack of personal and social data for intent discovery

- Social media + user profiling = personalization of social media search

Only cover a tiny fraction of the whole web
Social-Sensed Image Search (Cui et al 2012)

A sophisticated offline module and an efficient online module
The social-sensed image search results have more consistent visual style and return more relevant images with favored images.
• The Live Observatory
• First Order Analytics
• Large-Scale Image Search
• Bridging User Intention Gap
• Higher Order Analytics
• Summary
Other Social Image-based Research

• Visual search in vertical domains
  – Fashion search

• Integrate image search and social media
  – Assign semantic tags to images; Social image search
  – Perform MM-based event detection and product/brand tracking
**Event Detection for an Organization**

- Sources of info: Twitter and forums
- Key issues: Crawling and filtering of info
- Approach:
  - Known relevant data → evolving keyword set
  - Known info sources
  - Known user community → key users
  - Find relevant info and identify emerging events
Location Analytics

- Data: location-based UGCs, mobile footprints.
- Mine relations between check-in venues, and local landmarks.
- Identify popular trials (of individuals and their friends).
- Analyze user demographic and social communities.
- A continuous location-based service system towards Live City.
• Mining structure of data from multiple UGC sources
  – Including: Wikipedia, cQA, Forum and Twitters, or equivalent
• OneSearch: Multimedia/multilingual question-answering
  – Turn social QA into dynamic knowledge structures
Differential News and Social TV

- Differential news portal
  - From official newspaper perspectives
  - From social and users’ perspectives
- Social TV
The Live Observatory
First Order Analytics
Large-Scale Image Search
Bridging User Intention Gap
Higher Order Analytics
Summary
Working towards Life Analytics

• We focus on real-world problems:
  o Work on large-scale real-world problems of impact
  o Analyze large scale live UGCs to un-cover events and happenings in city
  o Work towards completeness, reliability and privacy issues of UGC data

• Offer live social media observatory
  o Provide both current and historic data for a wide range of analysis
  o Monitor events and happenings in city to help users lead better life

• Collaboration:
  o A wide range of technologies on live UGC analytics are ready
  o Explore collaboration with industry and government organizations
  o Ready to pilot large-sale applications of national impact
THANKS
Q & A

Visit our Web Observatory Site at: http://137.132.145.151/