My Journey in Academic Research

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

• The Environment
• My Research Path
• A Reflection
The Development of National University of Singapore (NUS)

- Started as a Teaching University in 1980s
- Started towards Research University in 1990s
- Achieved Research University Status in early 2000
- What did we do that makes the difference:
  - Good remuneration package: generous government supports
  - Good people: 2/3 rule
  - Nurturing environment and Research Support: Allow young ambitious people to prosper
  - Research Metrics: Started conference tiering in 1995
Research Journey of School of Computing, NUS

- Started as a Department of IS/CS (Info Systems & Computer Sc) under the Faculty of Science, 1980
- Became a Faculty (School of Computing) in 1998

Research Targets:
- Research Metrics: Started conference tiering for CS in 1995
  - Emphasize conference publications over journals
- Evolve into a measure in terms of research portfolio over the 3-year cycle
- Future: move towards measuring impact

Official Ranking now:
- NUS: 26th in the World
- School of Computing: Among the top 20
My Academic Career -1

• Mirrored the development of the School of Computing, NUS

• Started as a teacher/administrator in 1983
  o Can be addictive ...

• Began to focus in research in 1990
  o Spent 3 years in a research lab (Institute of Systems Science), 1988-1990
  o Working on Hypermedia and Computer Graphics

• Started Graphic in 1987
  o Organized my first international graphics conference in CGI (Commuter Graphics International) in 1990
My Academic Career -2

• Moved into Multimedia in 1990
  o Chose content-based retrieval as it has a long “future”
  o Started a conference series in 1993 (MMM: Multimedia Modeling)
  o Had my first ACM Multimedia Paper in 1994
  o Organized CIVR and ACM Multimedia in 2005

• Worked on text research in 2002
  o Started working on Question-Answering – hottest topic at that time
  o Participated in TREC-QA for next 4 years – obtaining 2nd position consistently
  o Had first SIGIR Paper in 2003
  o Organized ACM SIGIR in 2008

• Moved towards social media in 2008
  o Started joint research center with Tsinghua: The NExT Center
The Environment
My Research Path
A Reflection
A joint research center between NUS and Tsinghua

Started exploring in 2007, focusing on social media research to analyze huge amount of UGCs (user-generated contents) from a variety of social networks
  
  - Vision of Google: organize the world's information and make it universally accessible and useful
  - Vision of NExT: To index the live social media data and make them universally accessible

Aims and uniqueness of NExT Center:
  
  - Gather and analyze LIVE UGC’s related to a city
  - Infer relationships between topics and among people in a city/organization
  - Leverage on our multilingual, multimedia and multi-cultural research
  - Key research focuses: Live; Big Data; Multi-faceted; Users
AIM of NExT: From Unstructured Info to Structured Knowledge

Types of UGC’s Gathered

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

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

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

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

Structured Contents

People, Domain, Social, Loc & Mobile

Users
Live Observatory
Window to outside world

To explore live data and events as they unfold
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>Dianping</td>
<td>Shops</td>
<td>2012-05-01</td>
<td>142,320</td>
</tr>
<tr>
<td>Fantong</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>Flickr</td>
<td>Images</td>
<td>2012-05-01</td>
<td>66,257,149</td>
</tr>
<tr>
<td></td>
<td>Reviews</td>
<td>2012-05-01</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>Foursquare</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>Panoramio</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>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>2012-06-01</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>Reviews</td>
<td>2012-05-06</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>2012-06-01</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>Users</td>
<td>2012-08-15</td>
<td>52,259</td>
</tr>
<tr>
<td></td>
<td>Videos</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>Gamareana</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>
First Order Social Media Analytic
Where are the check-in data? (Distribution in Singapore)
Generation of First Order Analytics
(Map of UGCs on “Mother’s Day”)
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
(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)
• Second Order Social Media Analytic
Large Scale Image Search

- 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

Visual + Text + User-log

Image Search Engines

Text + User-log

Visual
One limitation:
The feature must be sparse.
Our Framework

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 $1 + C_{32}^{1} + C_{32}^{2}$.  

```
0 1 1 0 0 1 1 1 1 1 1 1
```

0 1

1
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)
- Multimedia query suggestion (Zha et al. 2009)
- Related Sample Feedback (Yuan et al. TMM’11)
- Attribute Feedback (Zhang et al. MM’12)
- Related Social Sense Media Search (Cui et al 2012)
Attribute Feedback – Sample Results
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.

Travel patterns of users
Social communities of users
Continuous query system
• Mining structure of data from multiple UGC sources
  – Including: Wikipedia, cQA, Forum and Twitters, or equivalent
• OneSearch: Multimedia/multilingual question-answering
Differential News and Social TV

- Differential news portal
- Social TV
- Jiku: Live video capturing and sharing platform
CONTENTS

- The Environment
- My Research Path
- A Personal Reflection
My Reflection

• Lucks in 3 areas

• NUS provides an excellent environment that I can’t focus on my research with no “life” pressure

• My choice of research areas:
  – Content analysis
  – Text and media
  – + user community → social media
  – Social media applications → NExT project with Tsinghua U

• Early involvement in international professional activities and large-scale evaluation activities

• A smooth journey:
  Learning → Catching up → Consolidating → Contributing
Post-Doc Journey

- A post-PhD training/ A beginning of research
- But a Post-Doc is not a Senior PhD Student !!

- Roles
  - Research:
    - Focus on 1 area
    - Work towards a second area
    - Demonstrate strong knowledge, competence, vision and with passion
  - Leadership
    - Provide research leadership; Guide PhD students; Lead & plan projects

- Reflection: academic or industry
- Pave ways for impactful work

- Needs luck: But hardworking people tend to have better luck
Final Reflection

• I wish I had a Post-Doc journey with a good mentor to help me quickly acquire:
  – Good foundation
  – Good work ethics and principle
  – Good attitude to life and work

• Life is not an obstacle

All the Best to your journey
THANKS

Q & A