SocialTrove: A Self-summarizing Storage Service for Social Sensing

Tanvir Amin, Shen Li, Muntasir Rahman, Panindra Seetharamu, Shiguang Wang, Tarek Abdelzaher, Indranil Gupta
University of Illinois at Urbana Champaign

Mudhakar Srivatsa, Raghu Ganti
IBM Research

Reaz Ahmed
University of Waterloo

Hieu Le
TCG Corp

ICAC 2015
Social Sensing

Egypt Unrest  Fukushima Disaster  Sandy Gas Outage  Crimea Annexation  Syria Chemical Attack

Twitter  Facebook  Google+  Instagram  Flickr

Credibility Estimation  Anomaly Detection  Timeline Reconstruction  ...
60 Seconds of Social Media*

- 2.66 Million Google searches
- 433 Thousand Tweets
- 67 Thousand Instagram photo uploads
- 293 Thousand Facebook Status updates
- 277 Thousand Snaps over SnapChat
- 20 Million Photo views on Flickr
- 11 Thousand LinkedIn searches
- 17 Thousand Walmart transactions
- 20 Thousand Tumblr new photos
- 139 Million emails sent

* http://blog.qmee.com/qmee-online-in-60-seconds/
* http://blog.qmee.com/online-in-60-seconds-infographic-a-year-later/
Observations and Goals

- Social Sensing Applications
  - User facing – real-time updates
  - Many users – many queries
- Social Sensing Data Streams
  - Real-time information overload
  - Large amounts of redundant data
- Obtain a representative sampling
  - in a content agnostic manner
  - based on application specified distance metric
  - at a configurable granularity
  - with high query throughput
  - and low access latency
Real-time Social Sensing Stack

Social Sensing Applications

Credibility Estimation

Anomaly Detection

InfoMax

SocialTrove

Summarization Infrastructure

Back End

Instagram

Twitter Firehose

HDFS

Spark

Memcached
Customization Interface

- SocialTrove agnostic to stream content
- Application provides two callbacks
  - **Vectorize**($u$)
    - Convert stream content $u$ to a feature vector
  - **Distance**($u$.FeatureVector, $v$.FeatureVector)
    - Compute distance between two feature vectors
    - Should be a metric, obey Triangle Inequality
      
      $\text{distance}(u, v) + \text{distance}(v, w) \geq \text{distance}(u, w)$
Summary Query Interface

• Transform application specific query.
  • Degree of granularity for summary retrieval.
• Transform SocialTrove query response to application specific results.
• Example:
  • Rank the matching summaries and retrieve top 10.
  • Retrieve matching summaries as a list.
  • Retrieve matching objects and additional metadata.
    • Source ID, Source relations etc.
  • Retrieve summary clusters and member counts.
SocialTrove Runtime

• Maintains a **summary model in memory**
  • Nearest neighbor data structure based on application defined distance metric.
  • Encapsulates distance information.
  • Hierarchical organization summarizes data at arbitrary granularity.
  • In-memory model improves throughput and latency.

• **Updates limited to once per batching interval**
  • Local insertions without synchronizing model.
  • Improves throughput and latency.
  • Temporal locality for social sensing content.
  • Eventual consistency through gossip.
System Design

Distributed Summary Model

(Interval t)

Summary Model for Interval t

Summary Model for Interval t

Offline Job

Adjust Summaries

(Interval t-1)

Distributed Summary Model

(Interval t)

Input Proxy

Input Daemon

Summary Model Cache

Input Daemon

Summary Model Cache

Input Daemon

Summary Model Cache

External Source(s)

Online Summaries

(Interval t)

Interval t

Interval t-1

Interval 2

Interval 1

Distributed Storage

Interval t

Interval t-1

Interval 2

Interval 1

Interval t

Interval t-1

Interval 2

Interval 1

Interval t

Interval t-1

Interval 2

Interval 1

Interval t

Interval t-1

Interval 2

Interval 1

Interval t

Interval t-1

Interval 2

Interval 1

Interval t

Interval t-1

Interval 2

Interval 1

Interval t

Interval t-1

Interval 2

Interval 1

Interval t

Interval t-1

Interval 2

Interval 1

Interval t

Interval t-1

Interval 2

Interval 1

Interval t

Interval t-1

Interval 2

Interval 1

Interval t

Interval t-1

Interval 2

Interval 1

Interval t

Interval t-1

Interval 2

Interval 1

Interval t

Interval t-1

Interval 2

Interval 1

Interval t

Interval t-1

Interval 2

Interval 1

Interval t

Interval t-1

Interval 2

Interval 1

Interval t

Interval t-1

Interval 2

Interval 1

Interval t
Summary Model

- Represented as a binary tree
  - Constructed using a divide and conquer paradigm.
- Current set of vectors partitioned
  - Into two disjoint sets using 2-means clustering.
  - Centroids of the two sets become two children.
    - Each set contains points nearest to the centroid of that set, compared to the centroid of the other set.
  - Both sets scheduled for further division in parallel.
  - Distance function must satisfy Triangle Inequality.
- Generated once per batching interval
  - Updates summary model based on the incoming objects in the last interval.
Points to Summary Model
Insertion

• For example, consider two dimensional space and Euclidean distance
• Insert object u to appropriate cluster
Insertion

- u arrives at root, A
- u compared to B and C
- u pushed to C
Insertion

- u compared to F and G
- u pushed to G
Insertion

- u compared to J and K
- u pushed to J
- u assigned to cluster J
Twitter Application

- **Twitter based Fact-finder application**
  - Tweets crawled using search api during events.
    - Crimea, Sandy, Fukushima, Egypt Revolution etc.
  - More than 4 million tweets.

- **Customization Interface**
  - Tweets converted to high dimensional vectors of tokens and frequencies.
  - Used Tanimoto distance function.

\[
d(u, v) = 1 - \frac{u \cdot v}{u^2 + v^2 - u \cdot v}
\]

- **Summary Query Interface**
  - Queries SocialTrove using keywords.
  - Ranks matching summaries according to credibility.
Implementation

• Summary Model
  • Written in Java
  • MapReduce job to generate model runs on Spark.
  • Broadcasts serialized model every interval.

• Client Proxy
  • Java / Python – based on different client applications.
  • Serves matching objects from distributed memcache

• Input Proxy
  • Python – periodically polls twitter via search api

• Distributed Storage
  • HDFS (Hadoop Distributed File System)
SocialTrove Testbed

- Runs on UIUC Green Data Center
  - http://greendatacenters.web.engr.illinois.edu
  - Capacity: 40 large nodes, 40 small nodes

- Testbed Setup
  - 8 nodes
  - Intel Xeon E5-2620 – 12 logical cores
  - 16 GB Ram per node
  - 1 Gbps Ethernet – Star topology
  - 1 TB disk

- One driver machine
  - apollo3.cs.illinois.edu
  - Intel Xeon E5-2690 – 40 logical cores
  - 128 GB Ram
  - 16 TB Raid-5 disk
Evaluation

• SocialTrove compared to 4 schemes
  • **Baseline**
    • No summary model computed.
    • Load balanced assignment to storage.
  • **Indexing**
    • No summary model computed.
    • Tweets indexed by keywords.
  • **Summary Baseline**
    • Computes a flat summary model in advance.
  • **Summary Indexing**
    • Computes a keyword indexed summary model.
Response time is very low with on demand summarization.
Response Time with summary model

4 million tweets
On demand summarization is not scalable
Lookup Throughput

with summary model

4 million tweets

Number of Summaries requested from 4 Million tweets
Model Generation Time

8 worker machines

<table>
<thead>
<tr>
<th>Number of Workers</th>
<th>Model Generation Time (4 million tweets)</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>337 minutes</td>
</tr>
<tr>
<td>8</td>
<td>239 minutes</td>
</tr>
<tr>
<td>11</td>
<td>195 Minutes</td>
</tr>
<tr>
<td>Protests – Real-time Summary - Thu 25 Jun 2015 07:19:02 GMT</td>
<td></td>
</tr>
<tr>
<td>---------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>OG: &quot;Chinese war veterans protest Beijing over missing benefits: <a href="http://t.co/T4x3mgXhB">http://t.co/T4x3mgXhB</a>&quot;</td>
<td></td>
</tr>
<tr>
<td>Massive protest by AAP Youth Wing against smriti Irani. @AamAadmiParty @DrKumarVishwas @dilipkandey @ArvindKejriwal <a href="http://t.co/EkMbK5MU2w">http://t.co/EkMbK5MU2w</a></td>
<td></td>
</tr>
<tr>
<td>Video: Activists protest Israeli killer drones at Paris Air Show <a href="http://t.co/q0fKmmYBzk">http://t.co/q0fKmmYBzk</a></td>
<td></td>
</tr>
<tr>
<td>On the day of the #SaveILF protest against cuts for the disabled, I cannot begin to describe how angry this makes me: <a href="http://t.co/RsRDhjjuvF">http://t.co/RsRDhjjuvF</a></td>
<td></td>
</tr>
<tr>
<td>Thousands of Buddhists protest to get their Religious Freedom #DalaiLama #Glastonbury [<a href="http://t.co/HJrXjGoxAs">http://t.co/HJrXjGoxAs</a> #UK](<a href="http://t.co/HJrXjGoxAs">http://t.co/HJrXjGoxAs</a> #UK)</td>
<td></td>
</tr>
</tbody>
</table>
Ranked Summary


#Cholera outbreak in #SouthSudan is wake up call to gov’t & humanitarian agencies [link]

MSG’ movie showcases the need for citizens to adopt various humanitarian causes and support welfare activities [link]

There is no humanitarian crisis in Calais just a bunch of thugs trying to enter the UK illegally. The law should deal with ALL criminals....

#MSGFilmDVDLaunched the purpose and motivation behind the film @MSGTheFilm is to promote Humanitarian Activities

Vayapam scam in MP, Lalit Gate in Rajasthan, Humanitarian scam in Delhi, fakes in debate...... good speed

Dear media, as soon as you are done with worshipping BBC, pls make sure to report the humanitarian crises in Karachi due to the Heatwave.

9 humanitarian INGOs refused registration. Ironically, the UN banned JuD was allowed to work in #Awaran #Balochistan [link]
## Crimean Crisis – Offline Summary – February 2014

<table>
<thead>
<tr>
<th>Event</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thousands at Moscow rally against Russian intervention in #Ukraine</td>
<td></td>
</tr>
<tr>
<td>Man in Ukraine plays the piano to help calm down a riot.</td>
<td><a href="http://t.co/fdNAc0cfJ2">http://t.co/fdNAc0cfJ2</a></td>
</tr>
<tr>
<td>For Crimea, Google Shows Different Borders Based on Your Location:</td>
<td>Russia’s Minister of Communications and Mass Media <a href="http://t.co/vIHGYlibOC">http://t.co/vIHGYlibOC</a></td>
</tr>
<tr>
<td>Militants in eastern #Ukraine were equipped with Russian weapons and</td>
<td>the same uniforms as those worn by Russian forces that invaded Crimea.</td>
</tr>
<tr>
<td>50,000 #Ukraine supporters march in Moscow protest Russia’s</td>
<td>intervention in #Crimea.</td>
</tr>
<tr>
<td>Some Russian tanks on Ukrainian border already painted with</td>
<td>‘peacekeeping’ slogans. How much longer until the ‘humanitarian</td>
</tr>
<tr>
<td>I’ve been speaking to @BarackObama about the situation in</td>
<td>intervention’?</td>
</tr>
<tr>
<td>Ukrainian Defense Ministry says its lone submarine has been</td>
<td><a href="http://t.co/7Rk2k8iOIK">http://t.co/7Rk2k8iOIK</a></td>
</tr>
<tr>
<td>Ukraine prepares armed response as city seized by pro-Russia forces</td>
<td></td>
</tr>
<tr>
<td>Ukraine crisis: Nato warns Russia against further intervention</td>
<td>BBC News</td>
</tr>
</tbody>
</table>
Conclusions

• Summarizes social content stream with arbitrary granularity, in real time.
• Caches a summary model in memory.
  • Scale-out input and client query proxies.
• High throughput and low response time.
  • Thanks to asynchronous updates to summary model, and temporal locality in social stream.
• Outperforms traditional indexing methods.
• Limitation: Evaluation performed only in the context of tweets.