Large-scale News Image Analysis

with MapReduce based LSH and VisualRank

Hao Li Myeong Lee

Outline

- Image in News
- VisualRank
- LSH and Weighted PageRank
- Result Analysis

Outline

- Image in News
- VisualRank
- LSH and Weighted PageRank
- Result Analysis

Images in News

A picture is worth a thousand words

Images in News

The more important the news, the more times it is repeated in different news sources

460 × 345

Boston Marathon bombing kills 3, injures over 140
bigstory.ap.org › Boston
Apr 16, 2013 – BOSTON (AP) — Two bombs exploded in the crowded streets near the finish line of the Boston Marathon on Monday, killing at least three ...

460 × 276

Boston Marathon blasts: three dead and more than 100 injured – as ...
www.guardian.co.uk › World news › Boston Marathon bombing
Apr 15, 2013 – Both Vice President Joe Biden and Boston Marathon officials referred to " bombing" or "bombs." Boston police walked back an initial assertion ...

350 × 232

Carlos Arredondo, **Boston Marathon** Hero in a Cowboy Hat, on the ...
www.thedailybeast.com/.../carlos-arredondo-boston-marathon-hero-i...
Apr 16, 2013 – Carlos Arredondo tells Michael Daly about saving a man with his legs blown off by the explosion.

665 × 498

Boston marathon bombing suspect's pals texted him, tried to protect ...
www.nola.com/crime/.../boston_marathon_bombing_suspec_4.html
May 1, 2013 – Dias Kadyrbayev was driving back to his apartment when he got a call from a college buddy. A clearly anxious Robel Phillipos told him ...

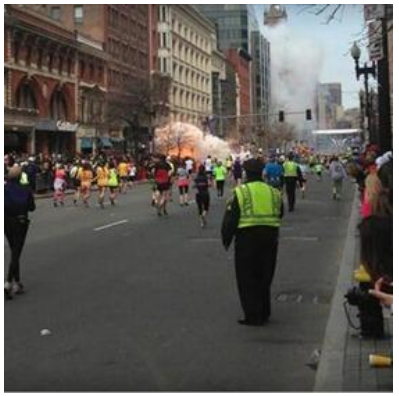
350 × 232

Boston Marathon Explosion - The Daily Beast
www.thedailybeast.com/features/.../boston-marathon-explosion.html
4 days ago – Authorities are widening their gaze in the Boston Marathon bombing case to the Tsarnaev brothers' closest associates, including Tamerlan's ...

Images in News

News Images about the same event are most near-duplicate with slight modification or similar to each other







Police offic
the finish l
Boston, M

A plume of smoke rises from a fertilizer plant fire near Waco, Texas

reme Court Patenting Genes

Investigators from the FBI inspect the boat where Boston Marathon bombing suspect Dzhokhar Tsarnaev was found hiding

NewsStand: a news aggregation system

Indexed 10,000+ RSS news feeds: CNN, FOX, NY Times, BBC, CBS….

extracted 5K+ news images/day

150K+ news images/month

1.8 M+ images/year

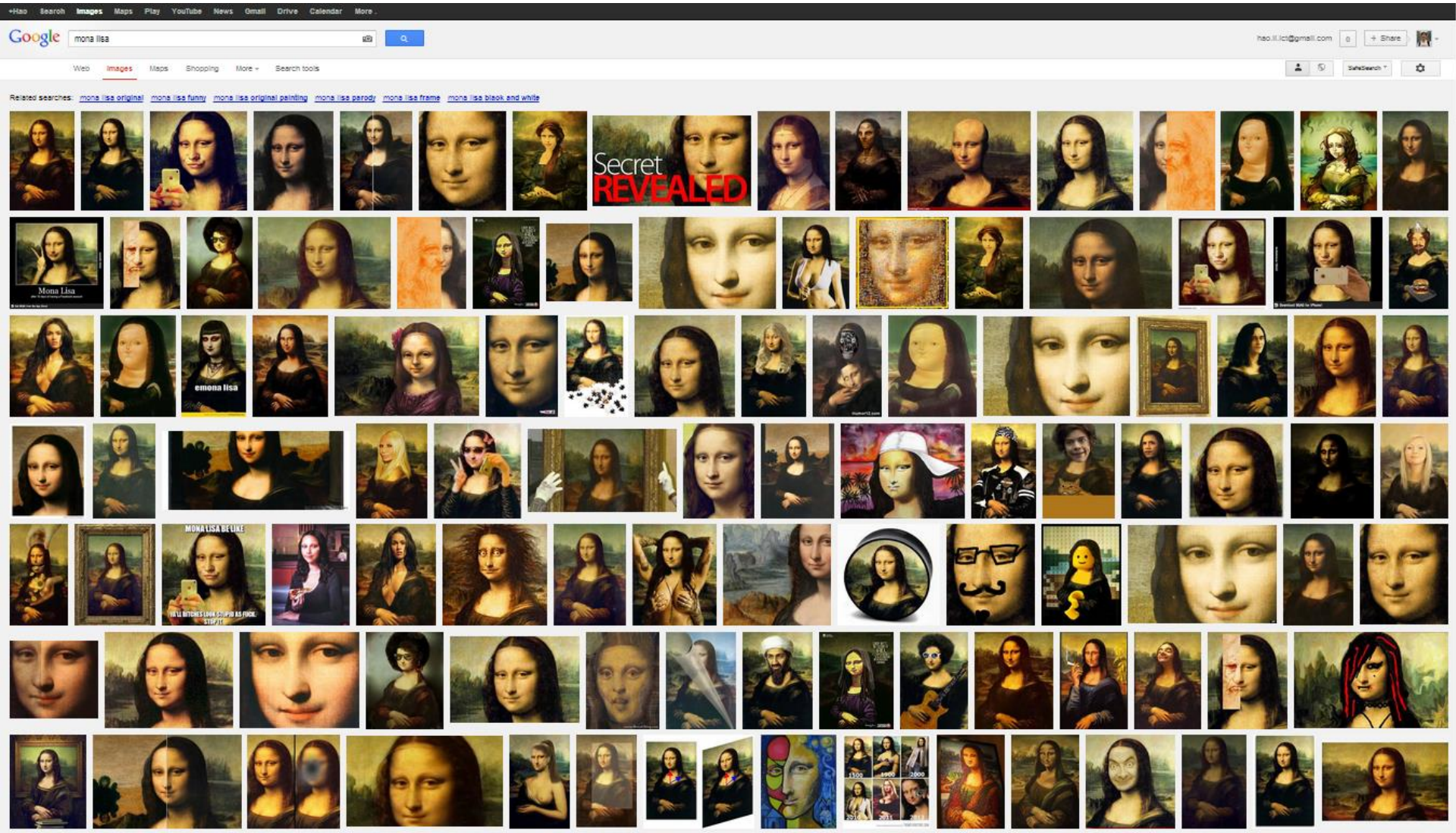
How to identify the most representative/valuable news images given a large image corpora?

News Image Summarization

Outline

- Image in News
- VisualRank
- LSH and Weighted PageRank
- Result Analysis

Text-based Image Search



VisualRank

Y. Jing, S. Baluja PageRank for Product Image Search. In WWW, 2008

Visual Link

- The most representative image has a lot of visual links to other modified images
  - SIFT feature

original

modified

Feature Extraction

- SIFT feature is quite storage consuming
  - a image may contain 500 descriptors
  - space for storing a 1M images would be

$$\frac{10^6 \times 500 \times 128 \times 4\ byte}{10^9 byte/G} = 256G$$! Even larger than the image set

- Gist Descriptor
  - 512 dimensions, a compact representation
  - Widely used in scene recognition

Outline

- Image in News
- VisualRank
- LSH and Weighted PageRank
- Result Analysis

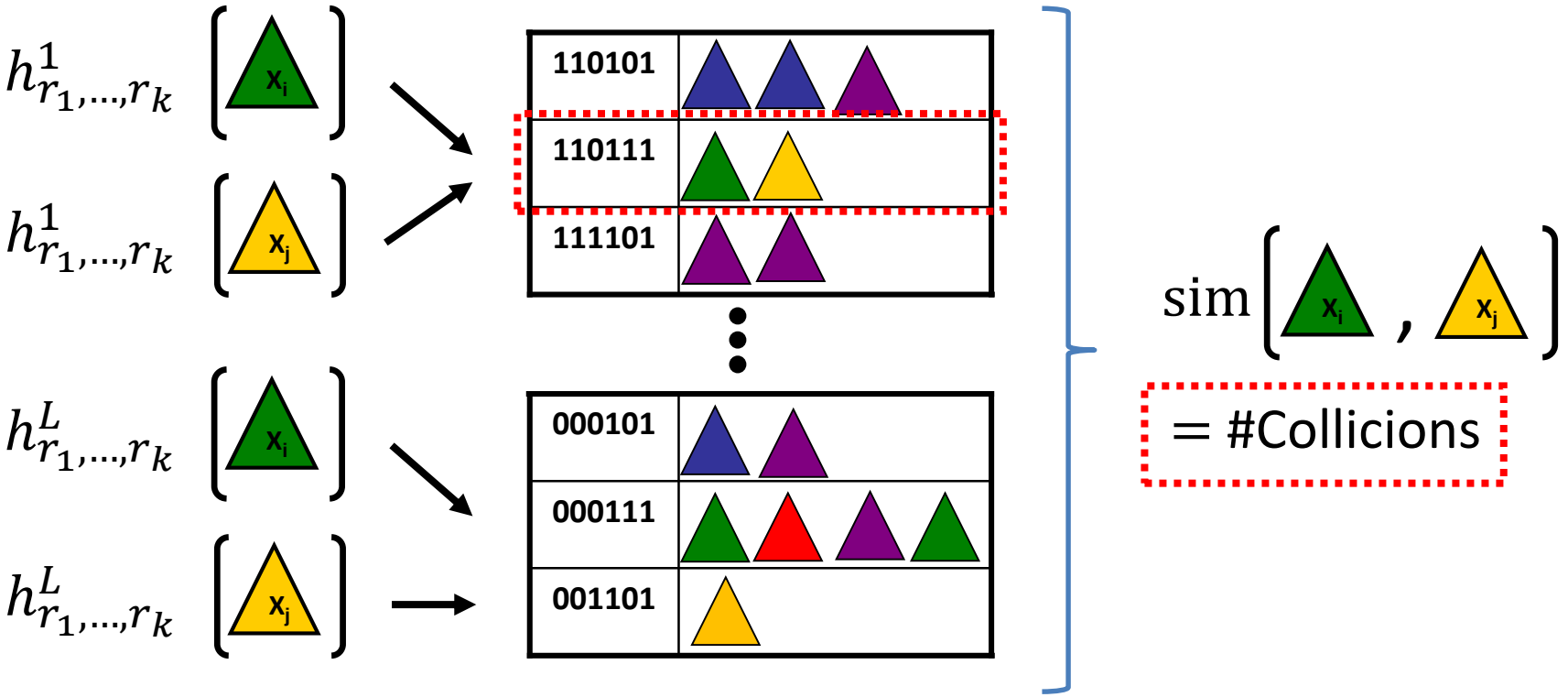
Similarity Graph Building

- Computational infeasible for Full K-NN Graph
  - $\binom{1000,000}{2}$ pairs for 1M documents.
  - If it takes a 0.001s to compute the similarity of two signatures, it takes 6 days to compute all the similarities

- Use LSH for K-NN Graph Approximation
  - The similarity can be approximated by counting the number of collisions for each pair of signatures.

Locality Sensitive Hashing

Hash functions $h^i_{r_1,\ldots,r_k}$ for i th Hash table



Job 1

Job 2

MapReduce Implementation

class MAPPER
method SETUP (K,L,W,dim)
build L LSH functions $\{\mathbf{h}_i\}_{i=1}^{L}$;
MAP(*id, feature f*)
for $i = 1 : L$ **do**
  EMIT($(i, \mathbf{h}_i(f))$, id);
end
class REDUCER
method REDUCE($(i, \mathbf{h}_i(f))$, $[id_1, id_2...]$)
EMIT($(i, \mathbf{h}_i(f))$, $[id_1, id_2...]$);

Algorithm 1: Building hashing table

class MAPPER
method MAP(*hashcode* $(i, \mathbf{h})$, *ids* $[id_1, id_2, ...]$)
for $i = 1 : |[id_1, id_2, ...]| - 1$ **do**
  for $j = i + 1 : |[id_1, id_2, ...]|$ **do**
    EMIT $((id_i, id_j), 1)$;
    EMIT $((id_j, id_i), 1)$;
  end
end
class REDUCER
method REDUCE((id_i, id_j), $[1, 1...]$)
$sum \leftarrow 0$;
for $i = 1 : |[1, 1...]|$ **do**
  $sum \leftarrow sum + 1$
end
EMIT $((id_i, id_j), sum)$;

Algorithm 2: Occurrence matrix computation

Weighted PageRank

- PageRank concerns the weight of edges
- Edge weight = Image Similarity

PageRank

$$PR(x) = \alpha(\frac{1}{N}) + (1-\alpha)\sum_{i=1}^{n}\frac{PR(t_i)}{C(t_i)}$$

Weighted PageRank:

$$PR(x) = \alpha(\frac{1}{N}) + (1-\alpha)\sum_{i=1}^{n}PR(t_i)\frac{w(t_i,x)}{\sum w(t_i,*)}$$

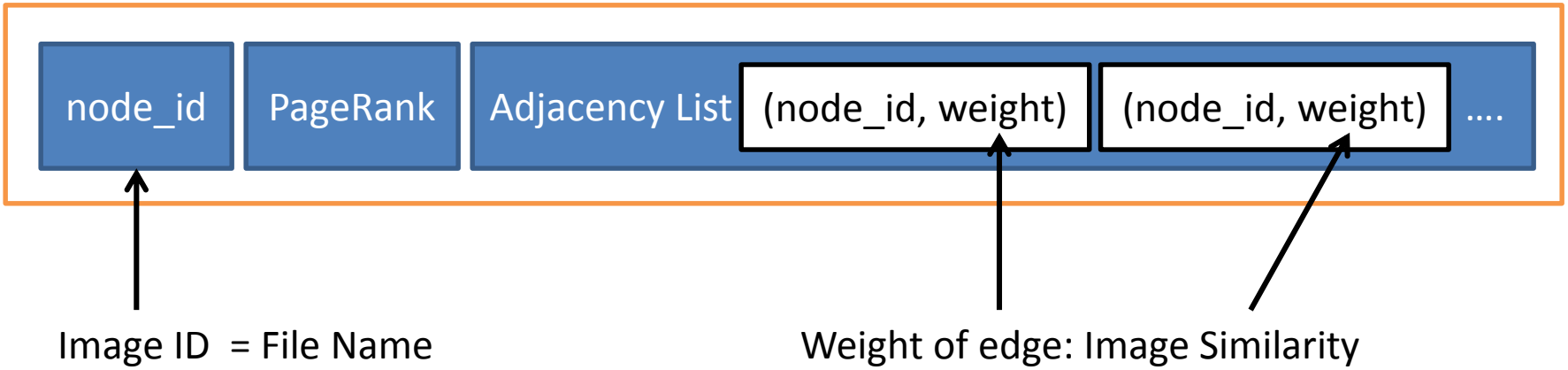
MapReduce Implementation

```
class MAPPER
method MAP(nid n, node N)
EMIT (nid n, N);
for nodeidm ∈ N.AdjencyList do
    p ←N.PageRank * sim(n, m)/ ∑ sim(n, *);
    EMIT (nid m, p);
end
class REDUCER
method REDUCE(nid m, [p1, p2...])
M ← 0;
for p ∈ [p1, p2...] do
    if IsNode(p) then
        M ← p;
    else
        s ← s + p;
    end
end
M.PageRank ← s EMIT (nid m, node M);
```

Algorithm 3: Weighted PageRank

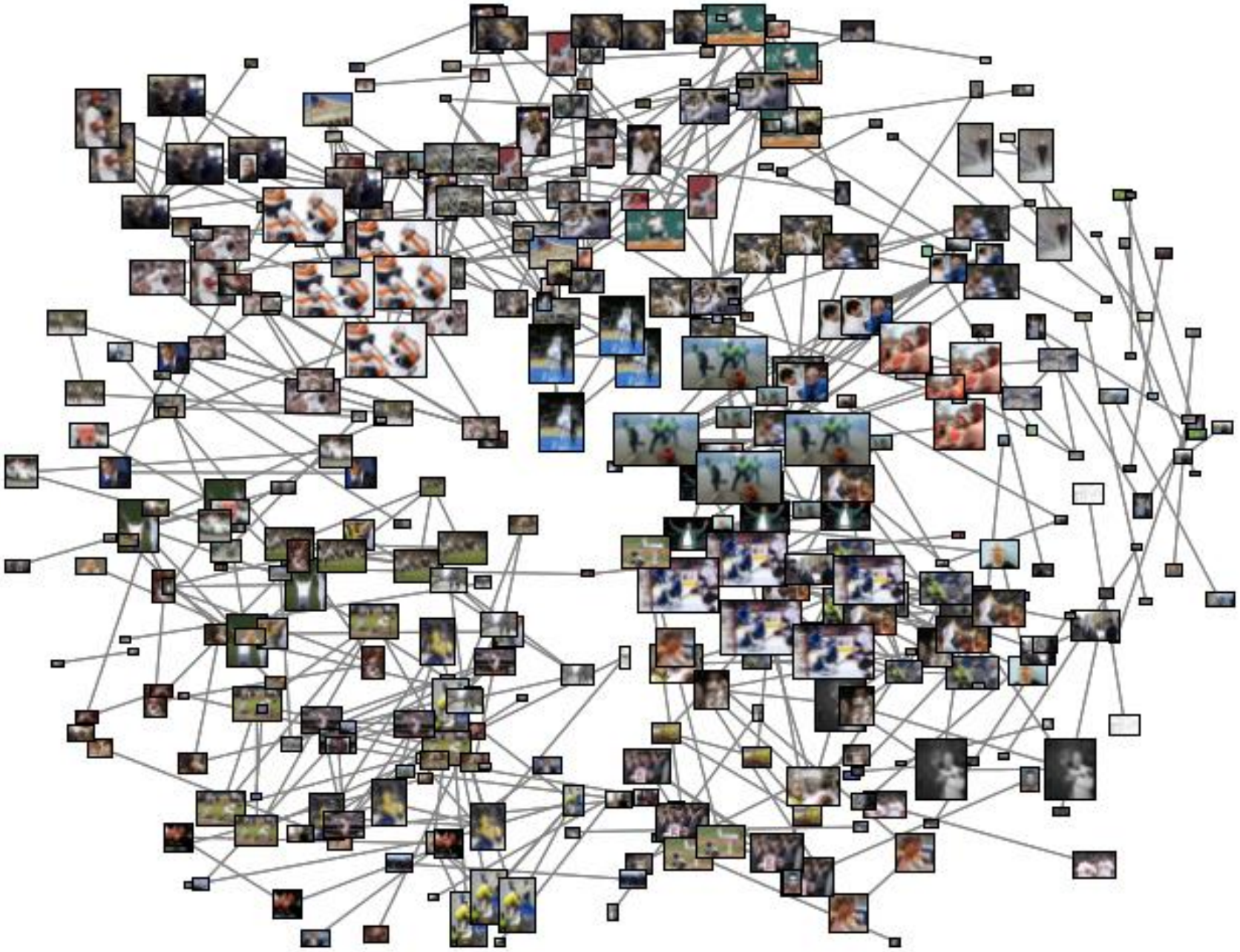
MapReduce Implementation

- PageRankNode Structure



Outline

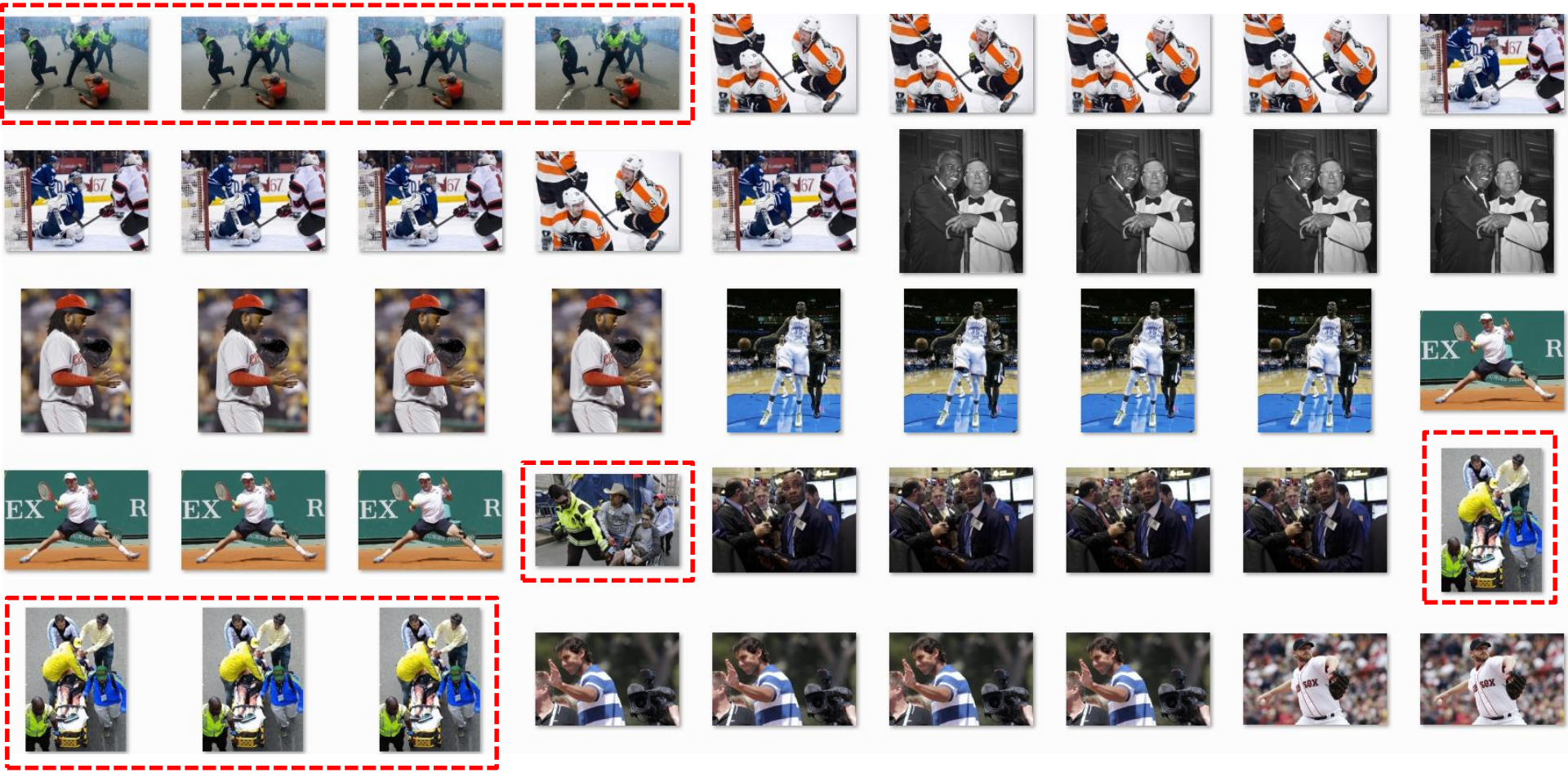
- Image in News
- VisualRank
- LSH and Weighted PageRank
- Result Analysis



Similarity Graph of 3544 images of Apr. 15th 2013

Top Ranked Images

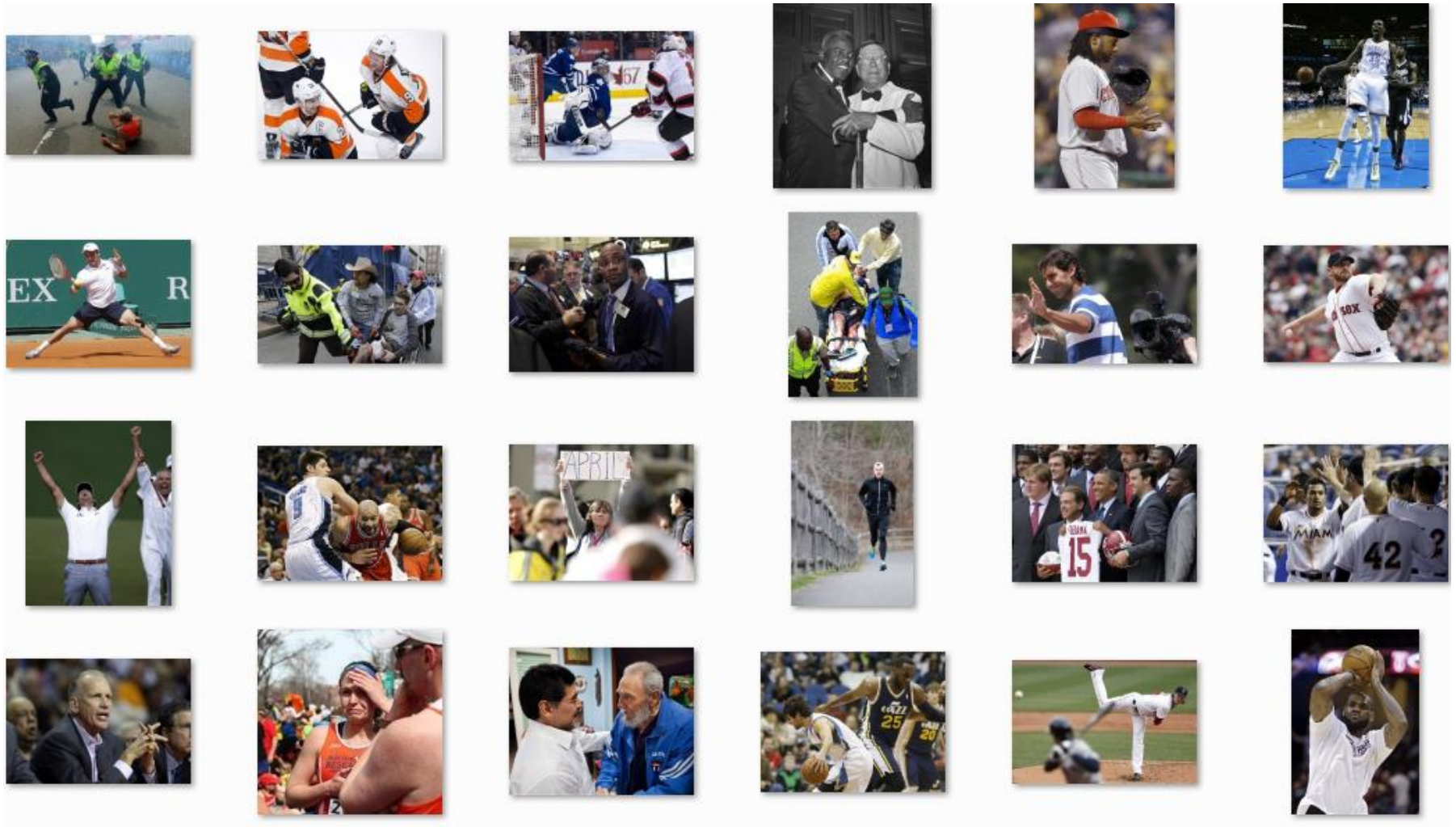
Rank 1 ⟶



Apr. 15th 2013

Top Ranked Images (No Duplicate)

Rank 1 ⟶



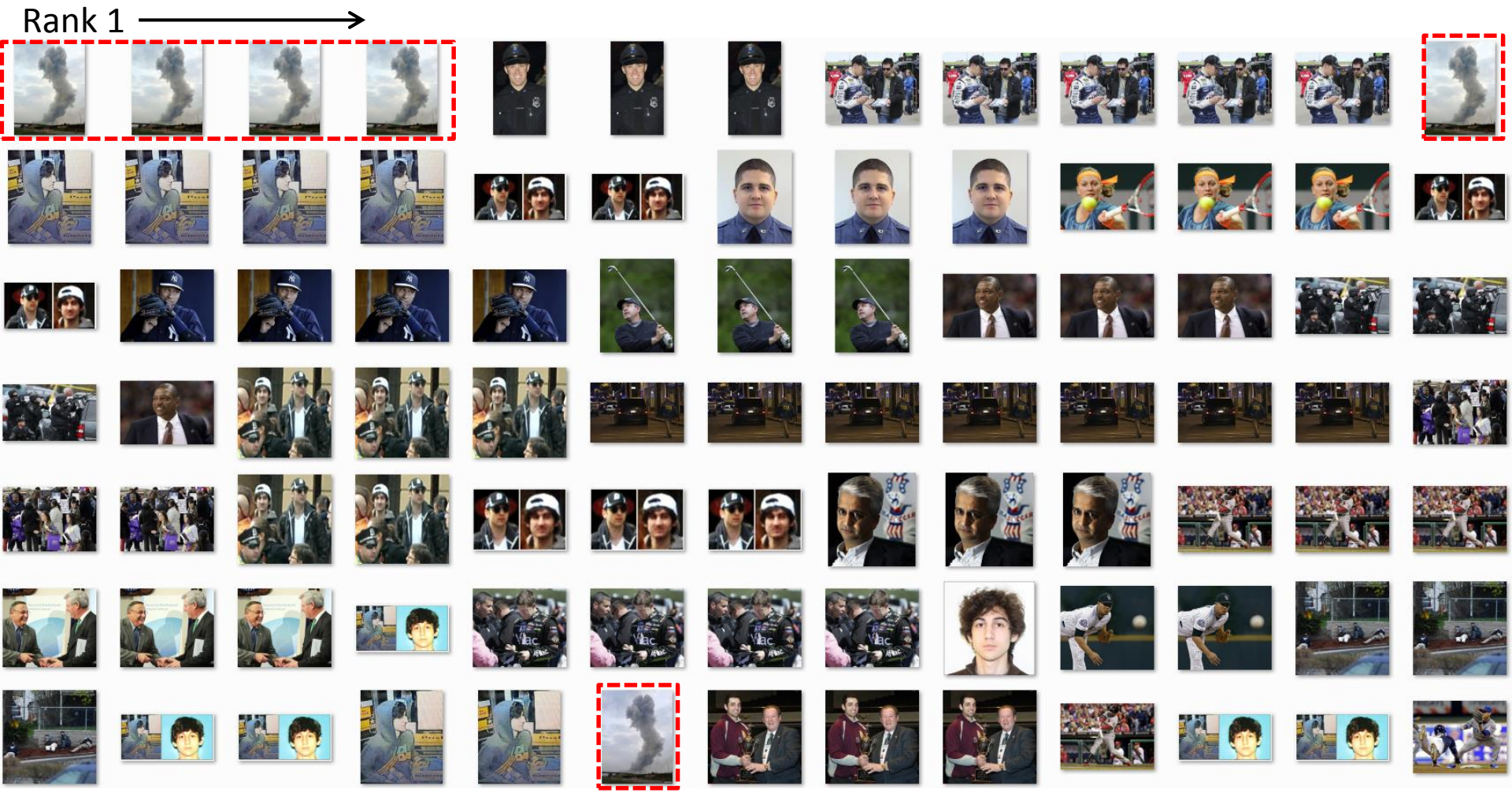
Apr. 15th 2013

Tag Cloud

Apr. 15th 2013

Similarity Graph of 4428 images of Apr. 19th 2013

Results: Top Ranked Images



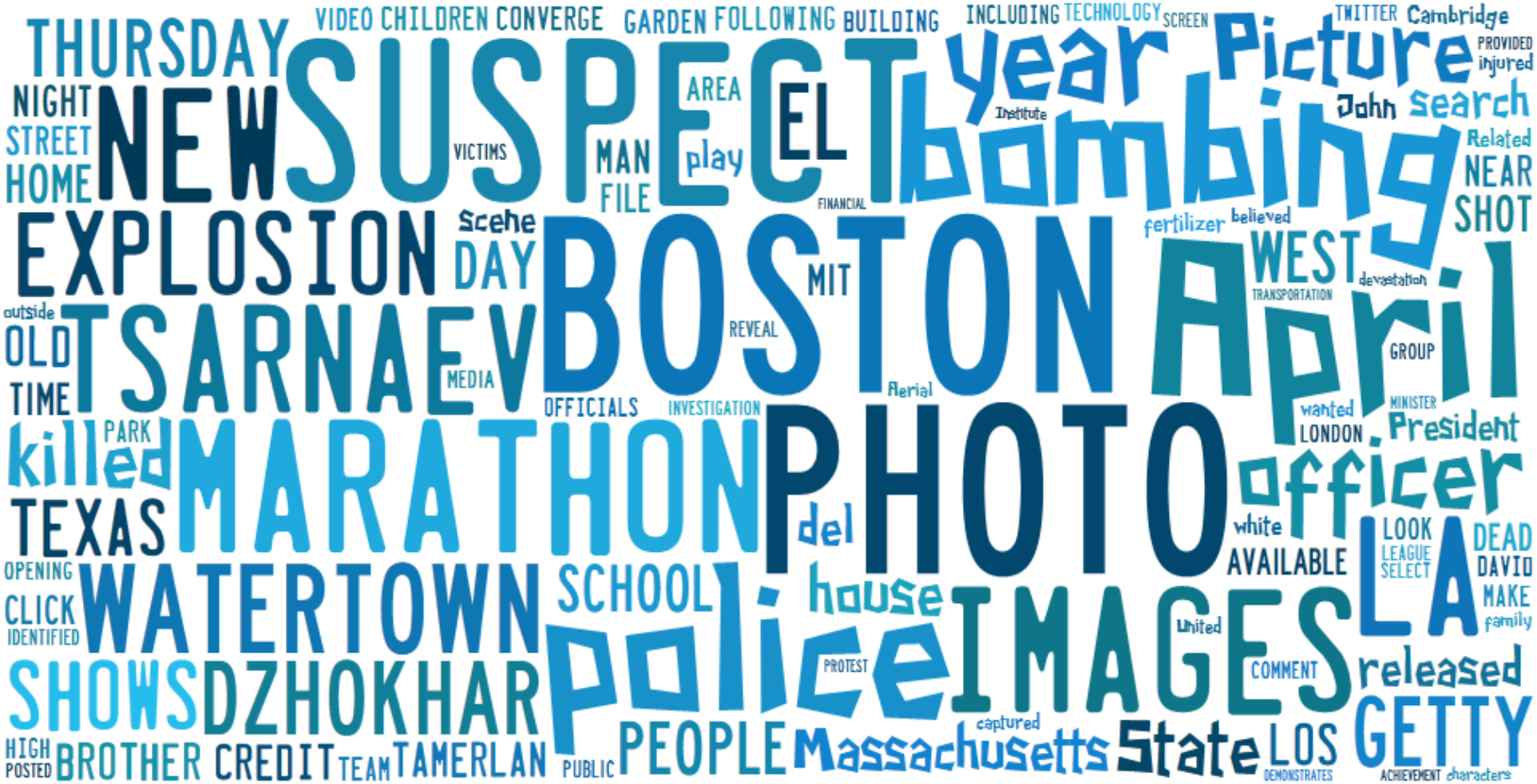
Apr. 19th 2013

Top Ranked Images (No Duplicate)

Rank 1 ⟶

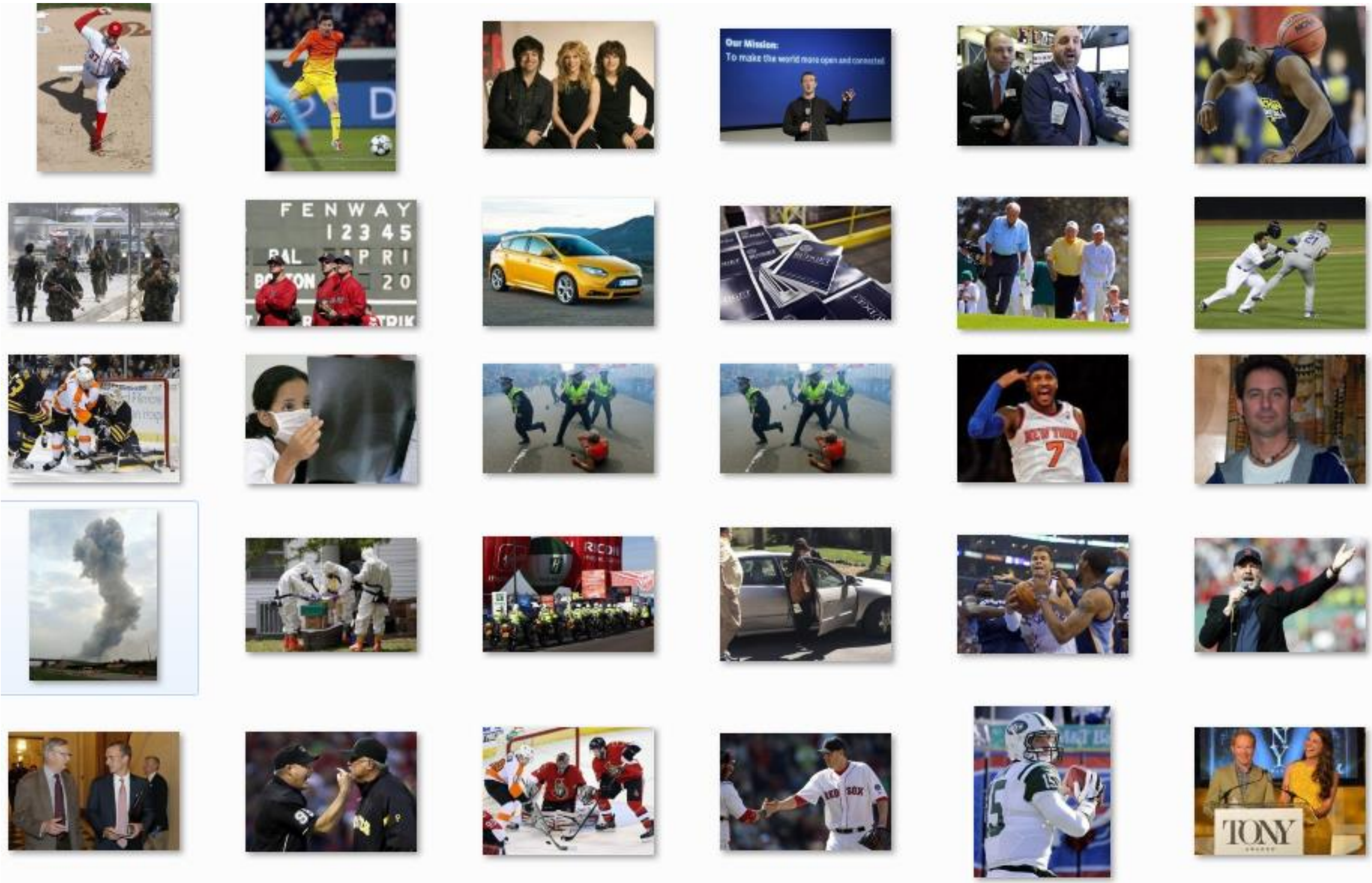
Apr. 19th 2013

Tag Cloud



Apr. 19th 2013

April's Daily Top News



What we have done?

- Implement MapReduce-based VisualRank
  - LSH for Similarity graph
  - Weighted PageRank
- Analysis on 150K news image of April, 2013.
  - Identifies the daily most important news images in April, e.g., Boston bombing and Texas explosion.

Future Work

- Use caption information
- News Classification
- Quantitative validation by user studies

Reference

- *A. Oliva, A. Torralba*. Modeling the shape of the scene: a holistic representation of the spatial envelope. IJCV 2001
- *M. Datar, N. Immorlica, P. Indyk, and V. S. Mirrokni*. Locality-sensitive hashing scheme based on p-stable distributions. In Symposium on Computational Geometry, 2004
- *Y. Jing and S. Baluja*. Visualrank: Applying pagerank to large-scale image search. In PAMI 2008
- *B. E. Teitler, M. D. Lieberman, D. Panozzo, J. Sankaranarayanan, H. Samet, and J. Sperling*. Newsstand: a new view on news. In SIGSPATIAL, 2008.