Mining Tweets for Tag Recommendation on Social Media

Denzil Correa and Ashish Sureka
IIIT - Delhi, India

Tags ?
Carnival On The Harbor
Carnival On The Harbor

Carnival Ride, Wheel Ferris, Amusement Park, Roller Coaster, River, Lake
Steffi Graf Marriage Proposal
Steffi Graf Marriage Proposal

Tennis, Humor, Joke, Celebrity, Comedy
• Description
• Description

• Additional context and semantic information
• Description

• Additional context and semantic information

• Discover relevant content
Multimedia Search
• Multimedia processing expensive
• Multimedia processing expensive

• Relies on the surrounding text
• Multimedia processing expensive

• Relies on the surrounding text

• Tags help
Aids Information Retrieval on Multimedia
• At least 20% of Flickr photos have no tags [Garg ‘08]
• At least 20% of Flickr photos have no tags [Garg ‘08]

• 64% Flickr photos have 1-3 tags [Sigurbjöörnsson ’08]
• At least 20% of Flickr photos have no tags [Garg ‘08]
• 64% Flickr photos have 1-3 tags [Sigurbjörgnsson ‘08]
• Imprecise & unrelated to content [Zhuang ‘11][Kennedy ‘06]
• At least 20% of Flickr photos have no tags [Garg ‘08]
• 64% Flickr photos have 1-3 tags [Sigurbjörnsson ’08]
• Imprecise & unrelated to content [Zhuang ’11][Kennedy ’06]
• Incomplete, noisy & incorrect [Liu ’09] [Moxley ’10]
Tag Quality Issues
Automatic Tag Recommendation
Coffee w/ a view #seattle
http://flic.kr/p/acAbs2
1 hour ago via Flickr  ☆ Favorite  ⇢ Retweet  ↳ Reply
A beautiful collection of Seashell & Sea Glass Crafts http://flic.kr/s/aHsjvHwxyE
Twitter?
✓ Hashtags

✓ Complete Tweet including Hashtags
Tweets

YouTube
Dailymotion
Flickr
Photobucket
SoundCloud

Phase A

Filtering Module

Whitespace
Tokenizer

Phase B
Stopwords, @Mention and URL Filtering

Whitespace Tokenizer

Phase A

Phase B

Filtering Module

YouTube
Dailymotion
Flickr
Photobucket
SoundCloud
Stopwords, @Mention and URL Filtering

Whitespace Tokenizer

Hashtag Extractor

Phase A

Filtering Module

Phase B

Tweets

YouTube
Dailymotion
Flickr
Photobucket
SoundCloud
Stopwords, @Mention and URL Filtering

Whitespace Tokenizer

Hashtag Extractor

Tweets with Hashtags

Tweets without Hashtags

Phase A

Phase B
Stopwords, @Mention and URL Filtering

Whitespace Tokenizer

Hashtag Extractor

Tweets with Hashtags

Tweets without Hashtags

Phase A

Filtering Module

Phase B

URL Parser

Media Web 2.0 API
YouTube
Dailymotion
Flickr
Photobucket
SoundCloud

Phase A

Filtering Module

Stopwords,
@Mention and
URL Filtering

Whitespace
Tokenizer

Hashtag
Extractor

Tweets with
Hashtags

Tweets without
Hashtags

Phase B

URL Parser

Media Web
2.0 API

Media Information
Phase A

Filtering Module
- Stopwords, @Mention and URL Filtering
- Whitespace Tokenizer
- Hashtag Extractor

Tweets with Hashtags

Tweets without Hashtags

Phase B

URL Parser

Media Web 2.0 API

Media with Tags

Media without Tags
Filtering Module

- Stopwords, @Mention and URL Filtering
- Whitespace Tokenizer
- Hashtag Extractor

Evaluation Module

Phase A

- Tweets
- YouTube
- Dailymotion
- Flickr
- Photobucket
- SoundCloud

Phase B

- URL Parser
- Media Web 2.0 API
- Media Information
- Media with Tags
- Media without Tags

Tweets with Hashtags
Tweets without Hashtags
Data
<table>
<thead>
<tr>
<th>Service</th>
<th>Number of Tweets</th>
</tr>
</thead>
<tbody>
<tr>
<td>flickr</td>
<td>2833</td>
</tr>
<tr>
<td>YouTube</td>
<td>2280</td>
</tr>
<tr>
<td>photobucket</td>
<td>1487</td>
</tr>
<tr>
<td>DailyMotion</td>
<td>290</td>
</tr>
<tr>
<td>SoundCloud</td>
<td>5796</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>12736</strong></td>
</tr>
</tbody>
</table>
Evaluation
\[ R_i = \begin{cases} 
1 & \text{if content of Tweet}_i \text{ is marked relevant by annotators} \\
0 & \text{otherwise} 
\end{cases} \]

Relevance Score \[ = \frac{\sum_{i=1}^{N} R_i}{N} \]
Relevance Score
#NYC Cool street shot from New York city
#Manhattan #photography
http://flic.kr/p/aa7LGf

11 Aug via Flickr  Favorite  Retweet  Reply
#NYC Cool street shot from New York city
#Manhattan #photography
http://flic.kr/p/aa7LGf
11 Aug via Flickr  Favorite  Retweet  Reply

Something Happened On The Way To Heaven
English riots in the Iranian press in Tehran #mongolrally
English riots in the Iranian press in Tehran #mongolrally
English riots in the Iranian press in Tehran #mongolrally
http://flic.kr/p/abTkLs
English riots in the Iranian press in Tehran #mongolrally
http://flic.kr/p/abTkLs
<table>
<thead>
<tr>
<th>Hashtag</th>
<th>Relevance Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>YouTube</td>
<td>85.71%</td>
</tr>
<tr>
<td>Dailymotion</td>
<td>40%</td>
</tr>
<tr>
<td>flickr</td>
<td>79%</td>
</tr>
<tr>
<td>Photobucket</td>
<td>40.91%</td>
</tr>
<tr>
<td>SoundCloud</td>
<td>57.84%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Complete Tweet</th>
<th>Relevance Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>84.48%</td>
<td></td>
</tr>
<tr>
<td>29.62%</td>
<td></td>
</tr>
<tr>
<td>53.86%</td>
<td></td>
</tr>
<tr>
<td>37.02%</td>
<td></td>
</tr>
<tr>
<td>62.59%</td>
<td></td>
</tr>
</tbody>
</table>
Limitations
• Noisy tags may get generated

• External media presence on Twitter
Future Work
• ⚫ Sophisticated Tag Selection Algorithm

• ⚫ Types of Tweets applicable to this approach
Summary
• Social Tag Recommendation

• Wisdom of Crowd approach using Twitter

• Approaches: Hashtags and Complete Tweet

• Using Twitter for Tag recommendation is effective and feasible
References


In more detail ...
Case I
Case 1

Tweets with HT + Media with Tags
Case I

Tweets with HT + Media with Tags

Case II
Case I

Tweets with HT + Media with Tags

Case II

Tweets + Media with Tags
Case I
Tweets with HT + Media with Tags

Case II
Tweets + Media with Tags

Automated
Case I

Tweets with HT + Media with Tags

Case II

Tweets + Media with Tags

Automated

+ string match
+ overlap count
Case I

Tweets with HT + Media with Tags

Automated

+ string match
+ overlap count

Case II

Tweets + Media with Tags

Case III
Case I

Tweets with HT + Media with Tags

Case II

Tweets + Media with Tags

Case III

Tweets with HT + Media

Automated

+ string match
+ overlap count
Case I
Tweets with HT + Media with Tags

Automated
+ string match
+ overlap count

Case II
Tweets + Media with Tags

Case III
Tweets with HT + Media

Case IV
Case I
Tweets with HT + Media with Tags

Automated
+ string match
+ overlap count

Case II
Tweets + Media with Tags

Case III
Tweets with HT + Media

Case IV
Tweets + Media
Case I
Tweets with HT + Media with Tags

Case II
Tweets + Media with Tags

Case III
Tweets with HT + Media

Case IV
Tweets + Media

Automated
+ string match
+ overlap count

Manual
Case I
Tweets with HT + Media with Tags

Case II
Tweets + Media with Tags

Case III
Tweets with HT + Media

Case IV
Tweets + Media

Automated
+ string match + overlap count

Manual
<table>
<thead>
<tr>
<th>Case</th>
<th>Tweets</th>
<th>Media</th>
<th>Evaluation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>+HT</td>
<td>-HT</td>
<td>+Tag</td>
</tr>
<tr>
<td>I</td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>II</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>III</td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>IV</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>
Metrics
$RT^j_i = \text{Set of Recommended Tags in Tweet}_i \text{ for } RQ_j$

$T_i = \text{Set of media tags corresponding to Tweet}_i$

$TT_i = \text{Set of tokens in Tweet}_i$

$N = \text{Total number of Tweets}$
Automated
Overlap Count = $\sum_{i=1}^{N} |R^j T_i \cap T_i|$

Average Overlap % = $\left( \frac{\sum_{i=1}^{N} \left( \frac{|R^j T_i \cap T_i|}{|T^T T_i|} \right)}{N} \right) \times 100$
<table>
<thead>
<tr>
<th>Service</th>
<th>OC/Tweets</th>
<th>AOP</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>YouTube</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RQ1</td>
<td>133/656</td>
<td>1.92</td>
</tr>
<tr>
<td>RQ2</td>
<td>1833/2222</td>
<td>5.17</td>
</tr>
<tr>
<td><strong>Dailymotion</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RQ1</td>
<td>12/65</td>
<td>1.88</td>
</tr>
<tr>
<td>RQ2</td>
<td>360/261</td>
<td>17.71</td>
</tr>
<tr>
<td><strong>flickr</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RQ1</td>
<td>420/613</td>
<td>7.78</td>
</tr>
<tr>
<td>RQ2</td>
<td>1580/2083</td>
<td>5.58</td>
</tr>
<tr>
<td><strong>photobucket</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RQ1</td>
<td>17/24</td>
<td>9.45</td>
</tr>
<tr>
<td>RQ2</td>
<td>530/593</td>
<td>17.14</td>
</tr>
<tr>
<td><strong>SoundCloud</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RQ1</td>
<td>18/189</td>
<td>0.89</td>
</tr>
<tr>
<td>RQ2</td>
<td>4508/3133</td>
<td>15.41</td>
</tr>
</tbody>
</table>