Characteristics of query topics for the WikipediMM task at ImageCLEF 2008

Jana Kludas, UNIGE, Switzerland
Theodora Tsikrika, CWI, Netherlands
WikipediaMM Task

Description:
• ad-hoc image retrieval
• collection of Wikipedia images
  – large-scale
  – heterogeneous
  – user-generated annotations
  – availability of multi-lingual data
• diverse multimedia information needs

Aim:
• investigate mono-media and cross-media retrieval approaches
• focus on fusion/combination of evidence from different modalities
• attract researchers from both text and visual retrieval communities
• support participation through provision of appropriate resources
WikipediaMM collection

• 151,590 images
  – wide variety
  – global scope
  – JPEG, PNG formats

• Annotations
  – user-generated
  – highly heterogeneous
  – varying length
  – noisy
  – semi-structured
  – monolingual (English)

• Used in INEX MM 2006 - 2007
MM task topics @INEX 2006/2007

2006:

<table>
<thead>
<tr>
<th></th>
<th>MMfragments</th>
<th>MMimages</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Topics</td>
<td>9</td>
<td>13</td>
<td>22</td>
</tr>
<tr>
<td>Avg. num. terms in &lt;title&gt;</td>
<td>2.7</td>
<td>2.4</td>
<td>2.6</td>
</tr>
<tr>
<td>Number of topics with src:</td>
<td>1</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>Number of topics with concept:</td>
<td>0</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

2007:

<table>
<thead>
<tr>
<th></th>
<th>MMfragments</th>
<th>MMimages</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of topics</td>
<td>19</td>
<td>20</td>
<td>39</td>
</tr>
<tr>
<td>Average number of terms in &lt;title&gt;</td>
<td>3.21</td>
<td>2.35</td>
<td>2.77</td>
</tr>
<tr>
<td>Number of topics with &lt;mmtitle&gt;</td>
<td>6</td>
<td>10</td>
<td>16</td>
</tr>
<tr>
<td>Number of topics with src:</td>
<td>2</td>
<td>7</td>
<td>9</td>
</tr>
<tr>
<td>Number of topics with concept:</td>
<td>4</td>
<td>6</td>
<td>10</td>
</tr>
<tr>
<td>Number of topics with both src: and concept:</td>
<td>0</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>

- NEXI format: XML based
- small number of topics
- not many multimedia hints
- text-based runs always best
Topic format

descriptions of multimedia information needs

- keywords
- optional: one or several visual and conceptual evidences

```xml
<topic>
  <number> 62 </number>
  <title> cities by night </title>
  <concept> building </concept>
  <image> http://www.bushland.de/hksky2.jpg </image>

  <narrative> I am decorating my flat and as I like photos of cities at night, I would like to find some that I could possibly print into posters. Photos of cities (or the earth) from space are not relevant. I would like to find photos of skylines or photos that contain parts of a city at night (including streets and buildings). </narrative>

</topic>
```
Topic development 2008

candidate topic pool:
(I) topics previously used in INEX 2006-2007 MM task
(II) topics submitted by this year's task participants
  - initial topic statement
  - exploration phase with assessment of top 25 results
  - feedback search with assessment of top 100
  - write <narrative>, optionally add <image> and <concept>
  - finalize topic

goal: diverse set of topics with
  - different characteristics (visual, semantic …)
  - different amount of multimedia resources
  - different domains: narrow/broad
Topic classification

**visual:**
- topics with visual, highly discriminating properties e.g. ‘blue flower’
- CBIR systems are likely to solve them

**textual:**
- topics containing proper nouns of persons, buildings, locations …
  e.g. ‘Da Vinci paintings’
- correctly annotated images easily found with text-only approaches

**semantic:**
- topics with a complex set of constraints, need world knowledge or contain ambiguous terms e.g. ‘labor demonstrations’, ‘plant’
- most likely no modality alone is effective

How to determine topics for each class?
Topics in 2008

intuitive classification of candidate topics according to class definition

- **5 visual**: blue flower, red ferrari, white cat …
- **35 textual**: oak tree, daily show, George W Bush, Golden gate bridge, can or bottle of beer …
- **35 semantic**: mountains under sky, winter landscape, people riding bicycles, famous buildings of Paris, plant …

topic statistics

<table>
<thead>
<tr>
<th></th>
<th>all</th>
<th>visual</th>
<th>textual</th>
<th>semantic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of topics</td>
<td>75</td>
<td>5</td>
<td>35</td>
<td>35</td>
</tr>
<tr>
<td>Average number of terms in title</td>
<td>2.64</td>
<td>2.2</td>
<td>2.3</td>
<td>2.9</td>
</tr>
<tr>
<td>Number of topics with image(s)</td>
<td>43</td>
<td>3</td>
<td>22</td>
<td>18</td>
</tr>
<tr>
<td>Number of topics with concept(s)</td>
<td>45</td>
<td>4</td>
<td>16</td>
<td>25</td>
</tr>
<tr>
<td>Number of topics with both image and concept</td>
<td>28</td>
<td>3</td>
<td>11</td>
<td>14</td>
</tr>
<tr>
<td>Number of topics with text only</td>
<td>15</td>
<td>1</td>
<td>8</td>
<td>6</td>
</tr>
</tbody>
</table>
Task submissions

- 12 groups submitted 77 runs
- many text-only runs, **but** same amount of fusion runs!
- type of methods:

<table>
<thead>
<tr>
<th>Modality</th>
<th>MAP</th>
<th>P@20</th>
</tr>
</thead>
<tbody>
<tr>
<td>textual</td>
<td>0.3444</td>
<td>0.3794</td>
</tr>
<tr>
<td>visual</td>
<td>0.2735</td>
<td>0.3225</td>
</tr>
<tr>
<td>concept</td>
<td>0.2700</td>
<td>0.3075</td>
</tr>
<tr>
<td>textual/visual</td>
<td>0.3444</td>
<td>0.3794</td>
</tr>
<tr>
<td>textual/concept</td>
<td>0.2735</td>
<td>0.3225</td>
</tr>
<tr>
<td>textual/visual/concept</td>
<td>0.2700</td>
<td>0.3075</td>
</tr>
</tbody>
</table>

- best run: still text-only

<table>
<thead>
<tr>
<th>group</th>
<th>runID</th>
<th>Modality</th>
<th>MAP</th>
<th>P@20</th>
</tr>
</thead>
<tbody>
<tr>
<td>upeking</td>
<td>zzhou3</td>
<td>TXT</td>
<td>0.3444</td>
<td>0.3794</td>
</tr>
<tr>
<td>cea</td>
<td>ceaTxtCon</td>
<td>TXTCON</td>
<td>0.2735</td>
<td>0.3225</td>
</tr>
<tr>
<td>ualicante</td>
<td>IRnNoCamel</td>
<td>TXT</td>
<td>0.2700</td>
<td>0.3075</td>
</tr>
</tbody>
</table>
Evaluation on task results

- use only top 75% of the runs to eliminate noise and buggy runs

**topic difficulty**

- easy: $\text{MAP} > 0.4$
- medium: $0.2 < \text{MAP} \leq 0.4$
- hard: $0.1 < \text{MAP} \leq 0.2$
- very hard: $\text{MAP} \leq 0.1$
Difficulty vs topic classification

- visual and semantic topics hardest to solve
- too many hard and very hard topics
  (? due to many narrow topics: 29 with less than 25 relevant docs)
best method: maximum average MAP over runs that use the same resources

<table>
<thead>
<tr>
<th></th>
<th>TxtImg</th>
<th>Txt</th>
<th>TxtCon</th>
<th>TxtConImg</th>
</tr>
</thead>
<tbody>
<tr>
<td>visual</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>textual</td>
<td>5</td>
<td>3</td>
<td>16</td>
<td>11</td>
</tr>
<tr>
<td>semantic</td>
<td>6</td>
<td>4</td>
<td>13</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>8</td>
<td>31</td>
<td>24</td>
</tr>
</tbody>
</table>

- initial classification not accurate: 1/5, 19/35, 25/35
- most of the topics best solved with fusion methods: 67/75 - need for efficient fusion methods
Topic resources vs best method

- topic images and concepts can turn out as not useful
- characteristic of topic not depending on resources
- topic resources and resources of best runs not related – due to query expansion/feedback methods??

<table>
<thead>
<tr>
<th></th>
<th>TxtImg</th>
<th>Txt</th>
<th>TxtCon</th>
<th>TxtConImg</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Images</td>
<td>4</td>
<td>6</td>
<td>15</td>
<td>18</td>
<td>43</td>
</tr>
<tr>
<td>Concepts</td>
<td>6</td>
<td>6</td>
<td>19</td>
<td>14</td>
<td>45</td>
</tr>
<tr>
<td>Images/Concepts</td>
<td>2</td>
<td>4</td>
<td>10</td>
<td>12</td>
<td>28</td>
</tr>
<tr>
<td>Text-only</td>
<td>4</td>
<td>0</td>
<td>7</td>
<td>4</td>
<td>15</td>
</tr>
</tbody>
</table>

ImageCLEF pre-workshop
16.9.2008, Aarhus Denmark
Conclusions

(I) hard to determine a topic’s class visual/textual/semantic, depends on:
   – content of the topic
   – relevant results in the collection (quality of annotation, images …)
   – (a bit on) topic resources (relevance of example image)

(II) most topics best solved with fusion approach, also textual ones
Next year’s topics

• same amount of visual, textual and semantic topics based on this year’s experience

• avoid too hard topics (MAP~0)

• groups of topics
  – ‘mountains’, ‘mountains under sky’, ‘mountains under sky with snow’
  – ‘bridges’, ‘bridges at night’, ‘bridges at daylight’
  – see how constraints influence the retrieval result, which types of approaches perform best