ClusterEval

Introduction

ClusterEval is the counterpart of TrecEval (Link) to evaluate the diversity of a result set. ClusterEval is written in Java and is provided with 2 JAR-files (ClusterEval.jar and ImageClef.jar).

Usage

ClusterEval needs three input files in order to evaluate the diversity of a result set. First needed file is the standard QREL file, with the relevance judgments of the assessors. Second file is the cluster assessment file. Third and last file is the run file from a arbitrary retrieval (in TREC format). Before you use ClusterEval please make sure that all your input files follow the format guidelines (see below).

Instructions

Add the two Jar-Files ClusterEval.jar and ImageClef.jar to your classpath (set classpath).

Start evaluation process by :
```
java uk.ac.sheffield.is.ClustEval [yourQREL] [yourClusterAssessmentFile] [yourRun]
(e.g: java uk.ac.sheffield.is.ClustEval qrel cAss testrun)
```

In case you haven't set the Classpath you can alternatively invoke ClustEval by:
```
java -cp [PathtoJar1];[PathtoJar2] uk.ac.sheffield.is.ClustEval [yourQREL]
[yourClusterAssessmentFile] [yourRun]

(e.g Windows: java -cp C:\Cluster\ClusterEval.jar;C:\Cluster\ImageClef.jar
uk.ac.sheffield.is.ClustEval C:\Cluster\qrel C:\Cluster\cAss C:\Cluster\testrun)

(e.g Linux: java -cp
/home/thomas/workspace/ClusterEval/lib/ClusterEval.jar:/home/thomas/workspace/ClusterEval/li
b/ImageClef.jar uk.ac.sheffield.is.ClustEval
/home/thomas/workspace/ClusterEval/testRuns/qrel
/home/thomas/workspace/ClusterEval/testRuns/cAss
/home/thomas/workspace/ClusterEval/testRuns/dcu)
```

Attention : In Windows “;” (semi-colon) is used to separate JAR-Files, whereas in Linux “:” (colon) is used!

Limitations

At the this release version of September 2008 following limitations are in force:

- Multi Clustering is not implemented at the moment. An image can belong to more than one cluster (very seldom for the IARP TC12 collection). The current functionality just takes the last instance.
- It is only possible to calculate the cluster recall at the moment. It is currently not possible to get the cluster precision

Source code

Source Code is available here.

Format guidelines

QREL File

The Relevance Judgment files must be in following format, using a *tab* as the delimiter between columns.

```
...
3  0  04/4533  1
3  0  30/30011 1
3  0  00/28    1
3  0  02/2369  1
... etc
```

where:

1. The first column is the topic number
2. The second column is a dummy entry and is not used in ImageCLEF 2008 and should be set to 0.
3. The third column is the official document number of the retrieved document. This will take the form of: directory/filename, e.g. "15/15001" where the filename has the extension removed.
4. The fourth column is relevance assessment. This is done in two levels: "1" represents "relevant" and "0" represents "non relevant"

Cluster Assessment File

The Cluster Assessment files must be in following format, using a *space* as the delimiter between columns.

```
...
2 1 37/37393
2 1 37/37394
2 2 37/37169
2 2 37/37194
... etc
```

where:

1. The first column is the topic number
2. The second column is the subtopic id. It declares to which cluster a relevant images belongs to
3. The third column is the official document number of the retrieved document. This will take the form of: directory/filename, e.g. "15/15001" where the filename has the extension removed.

In order for better human understanding the ImageCLEFPhoto Cluster Assessment have been augmented with descriptions like

```
# T2 2 - Saint Petersburg (5)
```

which has to be interpreted that for topic 2 (T2) cluster id 2 stands for Saint Petersburg and contains 5 relevant images in its cluster. All lines beginning with a “#” are ignored by ClusterEval.

Run File (Submission file)

Run files must follow the standard TREC format and the relevant section from the general TREC guidelines have been reproduced almost verbatim below. Note in ImageCLEF, the document reference refers to both the image and caption.

The format to use when submitting results is as follows, using a *space* as the delimiter between columns. The width of the columns in the format is not important, but it is important to include all columns and have at least one space between the columns.

```
...
25 1 00/99  0 4238 xyzT10af5
25 1 00/11  1 4223 xyzT10af5
25 1 16/16998  2 4207 xyzT10af5
```

```
25 1 15/15001  3 4194 xyzT10af5
... etc
```

where:

1. The first column is the topic number -- these will be numbered 1-60 in 2008.
2. The second column is the query number within that topic and these allow for variation between the translations. This field is not used in ImageCLEF 2008 and should be set to 1.
3. The third column is the official document number of the retrieved document. This will take the form of: directory/filename, e.g. "15/15001" where the filename has the extension removed.
4. The fourth column is rank position (**starting from 0**).
5. The fifth column shows the score (integer or floating point) that generated the ranking. This score MUST be in descending (non-increasing) order and is important to include so that we can handle tied scores (for a given run) in a uniform fashion (the evaluation routines rank documents from these scores, not from your ranks).
6. The sixth column is called the "run tag" and should be a unique identifier for your group AND for the method used. That is, each run should have a different tag that identifies the group and the method that produced the run.