T-61.6030 MULTIMEDIA RETRIEVAL

SYSTEM EVALUATION

Pauli Ruonala
pruonala@niksula.hut.fi
25.4.2008

Contents

Retrieve test material

Researchers test their new methods with:

- Own holiday pictures
- Historical images
- School yearbooks
- Corel Photo-CD
- TRECVID data sets
- Web images and videos - http://vaunut.org

- There are no standardized test data sets in general use

- Data sets are big, few hundred images is not large enough, and should contain query tasks and correct answers (*ground truth set*).

- Efforts towards standard test collections:

- TRECVID: 133 hrs news and documentary video with standard query types and answers

- IAPR TC-12 (International Association fo Pattern Recognition Technical Committee 12) uses still images

- ImageCLEF (Cross-Language Evaluation Forum) combines text and image retrieval. Historical photographs and image captions. Also medical image collection (X-rays, CT-scans, etc.) with captions.

- INEX (Initiative for the Evaluation of XML Retrieval) studies XML retrieval systems.

- Everyone uses his own data sets ⇒ generalization of findings is restricted

- Image type determines best metric ⇒ no general search algorithm.
  · One must learn to search. Example: Try to find a tumor in X-ray-image

- In same data set samples are not equal; true random subject samples are rare

Sources of retrieval errors

Main multimedia retrieval problems are selection of features and metrics for the features.

Night and day by Tom of Finland

What it is exactly that your feature detector is detecting?

How to combine several feature detectors optimally?

- Negative indication of Cheshire Cat fur does not indicate that Cheshire Cat is absent
- This is a decision theory problem and ex. Charles Dodgson (alias Lewis Carroll) did not find optimal solution

Traditional evaluation methods

Precision

$$\text{precision} P = \frac{\left|\text{relevant items}\right| \cap \left|\text{retrieved items}\right|}{\left|\text{retrieved items}\right|} = \frac{\text{no. of relevant items retrieved}}{\text{total no. items retrieved}}$$

- proportion of items relevant within a result set
- P = 100% means that all items retrieved are relevant.
- Because *ground truth set* is essentially a random selection (or hard to replicate) 100% result is random too
- Also called 'average precision' when 'precision' is given when 10, 20, 30, etc. items are retrieved.

- Idea is to give measures for different types of users. At 'precision at 10' is user that is satisfied with any relevant object.
- The least stable of evaluation measures; total number of items in set has a strong influence on precision at x.

Recall

$$\text{recall R} = \frac{|\text{relevant items}| \cap |\text{retrieved items}|}{|\text{relevant items}|} = \frac{\text{no. of relevant items retrieved}}{\text{total no. relevant items in set}}$$

- Proportion of relevant items that have been retrieved
- 100% = all relevant items are retrieved

Average precision

$$\text{average precision } AP = \left\{ \sum_{i=1}^{k} \frac{i}{r_i} \right\} \frac{1}{N_r}$$

k = no of items system has found
N_r = total no. relevant items in set
r_i = rank that system gave to item i

- Several algorithms have been suggested
- Ex. a query produces list of five items.
  - Put item that system thinks is most relevant first in list
  - First item gets rank number 1
  - If only three of the items, 1st, 3rd and 5th in list, are relevant so

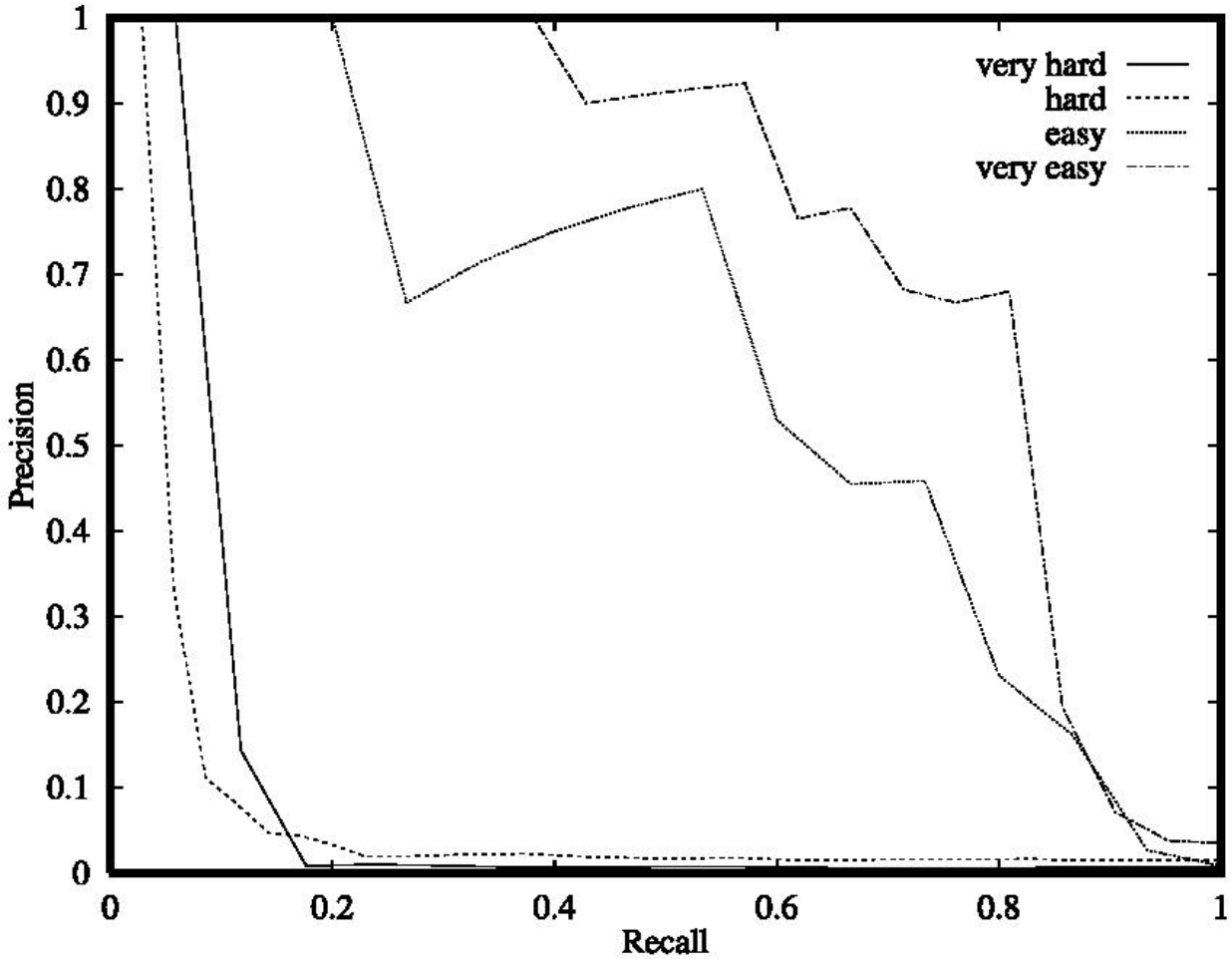
$$AP = \left(\frac{1}{1} + \frac{2}{3} + \frac{3}{5} \right) \cdot \frac{1}{3} \approx 0{,}76$$

- Items at top of the list have more weight ⇒ precision is sensitive to small change of ordering of items ⇒ generalization performance suffers

- Some authors call this Mean Average Precision.

Mean Average Precision (MAP)

Overall performance measure. Mean of all average precisions measured after all queries are done.



Precision/Recall graph

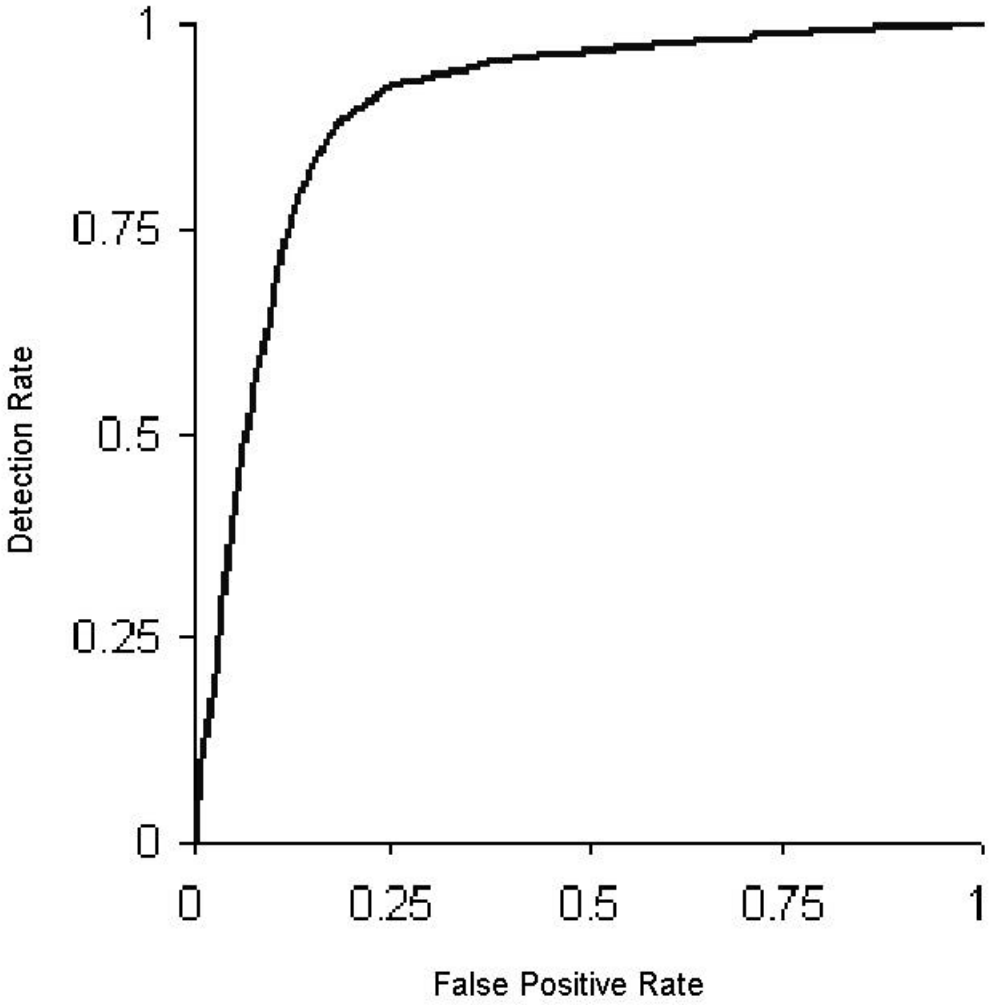
- Usually P/R-graph has sudden jumps
- Graph gives indication of how many items it is best to retrieve

F-measure

- Weighted harmonic mean of precision (P) and recall (R)

$$F = \frac{2 \cdot P \cdot R}{P + R}$$

Receiver operating characteristic (ROC) curve



- ROC curve plots TPR vs. FPR
  - true positives (TPR = true positive rate)
  - false positives (FPR = false positive rate)
- Used also in medicine, radiology, data mining, etc.
- ROCArea is area at right side of the curve.
- Large ROCArea can be indication of good retrieval performance

Rank normalization

- Used to even out effects of different data set sizes in each query
- Several methods proposed but none has been taken into use

Evaluation and machine learning

- Retrieval results can be improved with machine learning
- But simple methods do not produce good results because
  - AP is computed from discrete values; ranking of items
  - Ex. 99% of items in dataset are non-relevant and learning algorithm soon learns the method that produces 99% correct results: classify all items as non-relevant
- Good precision is easy to get: find a good order if items
  - This does not lead to good retrieval results.
- Optimizing ROCArea has shown that big ROCArea is not necessarily proof of good precision.

- Using precision for automatic feature detector weight adjustment is difficult.
  - Recall might suffer
  - Gradient-based optimization methods can not be used
- A new method for optimizing average precision with software from http://www.yisongyue.com

Statistical methods

- Used for comparing performance of two systems
- Recommended methods:
  - Wilcoxon signed rank test
  - paired t-test
  - paired sign test

Example: Wilcoxon signed rank test

Make two queries A and B from same data set by using two different methods. Results are different. Is difference statistically significant?

- Formulate a null hypothesis = difference is not statistically significant.
- Take one result from A and one from B and make a pair out of them. Arrange rest in similar way.
- Compute absolute differences (D) between pairs.
- Create an ordered list of results - smallest value first.
- If several pairs have same D replace their ranks with mean of rank values. Remove pairs where D == 0.
- Replace differences with rank values - retain the sign.

Smallest D gets rank 1, R(D) is rank value.

$$W^- = \sum_{D<0} R(|D_i|) \qquad W^+ = \sum_{D>0} R(|D_i|)$$

- Compute W^+ and W^-. i is number of pairs.
- Take W^- or W^+, whichever is smallest, and use tables or computer to compute value 'p'.
  - Ex. if p is 0,01 then there is 1% possibility of error if we say that null hypotheses is valid = difference is not statistically significant.

Same with Matlab:

```
[p, h] = ranksum (a,b);
```

- Query results in vectors a and b

- If h == 1 it means rejection of null hypotheses at 5% significance level

- Wilcoxon test does not produce good results if many rank-sums tie with each other.

- Wilcoxon test does not care about distribution of errors or result values. Student t-test is more vulnerable to deviations from the normal distribution and can be used when number of pairs is larger than 50.

Conclusions

- Current evaluation methods are incompatible - there is no clear connection between ex. ROCArea and MAP
- You are free to add your own.
- Evaluation methods can help in selecting appropriate features and metrics for the features, but there are difficulties
- There are no solved problems - some niche problems have a working solution.
- Nobody has measured user satisfaction.
- Human similarity judgments do not follow any metric; they are user- and task-dependent.
- Humans attach meanings to pictures which can not yet be modelled.

References

Henning Müller, Wolfgang Müller, David McG. Squire, Stéphane Marchand-Maillet, Thierry Pun. Performance evaluation in content-based image retrieval: overview and proposals. Pattern recognition letters 22 (2001).

Alexander G. Hauptmann, Michael G. Christel. Successful Approaches in the TREC Video Retrieval Evaluations. In International Multimedia Conference New York, NY, USA 2004.

Michael S. Lew, Nicu Sebe, Chabane Djeraba, Ramesh Jain. Content-based Multimedia Information Retrieval: State of the Art anc Challenges. ACM Transactions on Multimedia Computing, Communications and Applications, Feb 2006.

Mathias Lux, Gisela Dösinger, Günter Beham. Empirical Studies in Multimedia Retrieval Evaluation. In BTW Workshops 2007: 199-217, Aachen, Germany.

Yisong Yue, Thomas Finley, Filip Radlinski, Thorsten Joachims. A Support Vector Method for Optimizing Average Precision. Proceedings of SIGIR 2007. http://www.yisongyue.com

Thank you!