Study of Techniques for the Indirect Enhancement of Training Data in the Training of Neural Networks used in Re-ranking of Web Search Results

Panagiotis Parchas, George Giannopoulos, Timos Sellis

National Technical University of Athens, Greece
pn parchas@gmail.com, {giann, timos}@dmlab.ece.ntua.gr

2011 ACM SIGMOD/PODS Conference Undergraduate Poster Competition
12-16 June 2011, Athens, Greece

1. Personalization Problem

2. Defining the problem

Q: How to achieve personalization in web search?
A: Train a learning model using user’s Preferences (relative judgments)

Q: How to train a reliable model when an average user only evaluates <10% of the web search results?
A: Expand user’s relative judgments to unjudged results

3. What is Expansion

4. How do we perform Expansion

- We choose a Vector Space and project the search results there. The Vector Space might be one of the following:
  a) the textual representation of results,
  b) the feature vectors which consist of features that quantify the quality of the results, w.r.t. the query,
  c) a hybrid combination of the two.

- Then we cluster the results based on their similarity in the related Vector Space.

We try to expand the user judgments to clustered results.

In our experiments, the relevance judgments take the values:
0 --> irrelevant
1 --> quite relevant
2 --> highly relevant

5. Expansion Algorithm

1. Simple:
   - Discard all clusters with opposite judgments
   - To the rest, generalize the most popular judgment

2. Partially Expanded:
   - Let d=diff. of opposite judgments and k1 a threshold.
   - If (d<k) then
     Maintain cluster and generalize most popular judgment else
     Discard cluster

3. Fully Expanded (accept clusters with all 3 judgments)
   - Let s_i=amount of results with judgment i, and k2 a threshold
   - If (s_1+s_2<s_3+k2) then
     Maintain cluster, generalize s_3
     Else: Discard cluster

6. Our Method’s Dataflow

7. Results

A comparison to the ideal training using the entire set of results:

A comparison between:
- the tuning of parameters in expansion (thresholds k1,k2),
- the quality of expansion (columns)
- and the precision of the derived model (green line).

8. Conclusions

1) Our method, with the best tuning, approaches closely the ideal training lacking only by a 4.4%

2) In the hybrid space there seems to be a notable stability, independent of the tuning

3) In the process of training, the wrong training data have bigger (negative) impact than the correct ones, regardless of the growth of the training set