EFFICIENT DIVERSIFICATION OF SEARCH RESULTS USING QUERY LOGS

Gabriele Capannini, Franco Maria Nardini, Raffaele Perego & Fabrizio Silvestri

ISTI–CNR, Pisa, Italy

ABSTRACT

We study the problem of diversifying search results by exploiting the knowledge mined from query logs. Our proposal exploits the presence of different "specializations" of queries in query logs to detect the submission of "ambiguous" queries, and manage them by diversifying the search results returned in order to cover the different possible interpretations of the query. More details are available in [2].

We introduce the following novel contributions to result diversification:

- a methodology to detect the ambiguous queries that would benefit from diversification based on query log analysis;
- a methodology to efficiently and effectively devise the possible topics to include in the diversified list of results along with their probability distribution;
- an objective diversification usefulness-measure to assess how valuable a diversified results list is;
- a diversification algorithm that re-ranks the results list on the basis of the set of query refinements mined from the log along with the associated probabilities.

QUERY-LOG-BASED DIVERSIFICATION OF WEB SEARCH RESULTS

Let \( D \) be the collection of documents indexed by the WSE which returns, for any given query \( q \), an ordered list of documents \( R_q \subseteq D \). The rank of document \( d \) in \( D \) within \( R_q \) is indicated with \( \text{rank}(d, R_q) \). The distance function \( \delta : D \times D \rightarrow [0,1] \), having non-negative and symmetric properties is defined as \( \delta(d_1, d_2) = 1 - \text{cosine}(d_1, d_2) \), where \( \text{cosine}() \) denotes the cosine similarity function.

The utility function specified defined in Equation (1) denotes how good \( d \) in \( R_q \) is for satisfying a user intent that is better represented by specialization \( q' \in S_q \), which is the set of specializations of an ambiguous query \( q \) mined from the query log.

\[
U(d|R_q) = \sum_{d'\in R_q} \frac{1 - \delta(d, d')}{\text{rank}(d', R_q)}
\]

(1)

EXPERIMENTS

The experiments concerns the diversity task of the TREC’09 Web track. Two query logs, i.e. AOL and MSN, were preprocessed in order to devise the logical user sessions.

- Effectiveness: We evaluate our method versus two state-of-the-art competitors: xQUAD [3] and IASELECT [1] in diversifying the results retrieved using the DPH Divergence From Randomness model. OptSelect and xQUAD behave similarly in terms of effectiveness, while IASELECT performs always worse.

- Efficiency: The figure on the left shows the average time (msecs) required by the three algorithms to diversify the varying sized sets of documents for the 50 queries of the dataset.

REFERENCES

