Query Segmentation using only Query Logs

Nikita Mishra, Rishiraj Saha Roy, Niloy Ganguly (IIT Kharagpur)
Srivatsan Laxman and Monojit Choudhury (Microsoft Research India)

contact:
monojitc@microsoft.com
niloy@cse.iitkgp.ernet.in

---

**What?**

Query segmentation refers to identification of meaningful units in a query. For example, “australian open” “home page”. Query segmentation can help improve IR performance, query expansion and query suggestion. Past work has used a variety of resources such as content of Web documents, POS tagger and chunker, named entity lists, and human annotated data.

**Why only Query Logs?**

Queries have their own unique structure. Use of document text or human annotation for query segmentation increases the risk of projection of natural language structure onto queries. So the use of query logs as the only resource is the best and scalable approach to unravel units specific to queries.

---

1. **X:** observed counts of candidates.
   - Leonardo da vinci: 5
   - XP Vista: 2

2. Expected probability of candidate:
   \[ P_i = \frac{(\ell_i - n + 1) \times (\ell_i - n)!}{\ell_i!} \]
   \[ P(Leonardo da vinci) = \frac{1}{30} \]
   \[ P(XP Vista) = \frac{1}{6} \]

3. Expected counts
   - Leonardo da vinci: 7/30
   - XP Vista: 7/6

4. Prune the candidates by using Hoeffding’s inequality:
   \[ \text{Prob}[X \geq N] \leq \exp \left( -\frac{(N - E(X))^2}{k} \right) \]
   \[ p[Leonardo \ldots \geq 5] = 0.015 \]
   \[ p[XP Vista \geq 2] = 0.82 \]
   \[ XP Vista \rightarrow \text{not a segment} \]

5. Given a query, segment it such that the sum of the negative log-probabilities (obtained in step 4) of all the segments is maximized.

---

Leonardo da vinci artwork