Improving Document Ranking for Long Queries with Nested Query Segmentation

Rishiraj Saha Roy
Max Planck Institute for Informatics
rishiraj@mpi-inf.mpg.de

Anusha Suresh
Indian Institute of Technology Kharagpur
anusha.suresh@cse.iitkgp.ernet.in

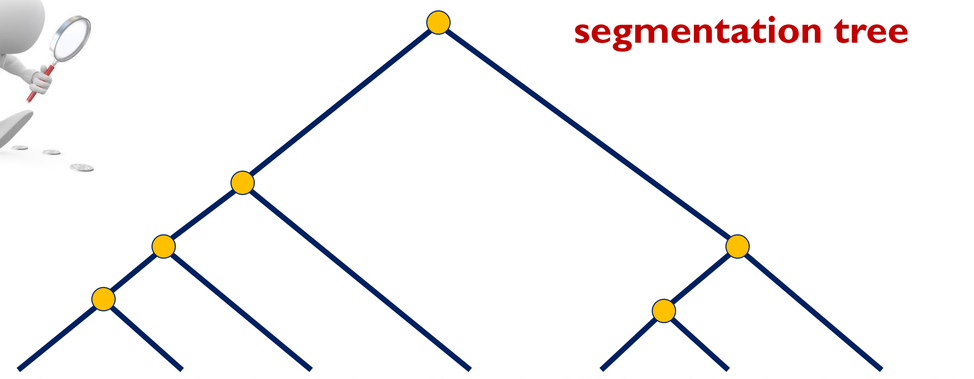
Niloy Ganguly
Indian Institute of Technology Kharagpur
niloy@cse.iitkgp.ernet.in

Monojit Choudhury
Microsoft Research India
monojitc@microsoft.com

MOTIVATION

- Long queries in the tail are the biggest challenge to commercial search engines!!
- Long queries benefit from advanced syntactic processing
- Flat query segmentation fails to capture relationships *within and between segments*
- Can be discovered if we allow nesting of segments inside bigger segments
- Information can be leveraged for re-ranking documents using a term proximity model

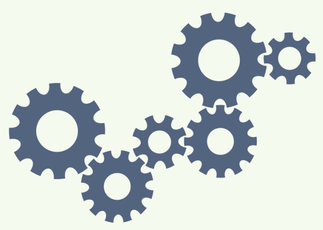
A nested segmentation tree



Nested query segmentation (((windows xp) home) edition) ((hd video) playback)

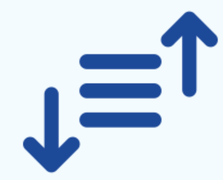
Flat query segmentation windows xp home edition | hd video | playback

ALGORITHM



- Split flat segments recursively using lower order *n*-gram word association scores
- Greedy approach works well during grouping
- Join flat segments using relative orders of scores of bigrams straddling flat segment boundaries
- Prioritize determiners, conjunctions and prepositions while joining
- All association scores learnt from **query logs only**

DOCUMENT RE-RANKING



- Pair of words that have low tree distance should not have a high document distance
- Document fragment: **… windows 7 compatibility of office xp home edition…** ➡ High penalty [TD = 2, QD = 1]
- Document fragment: **… windows xp home edition has various tools for hd video …** ➡ Lower penalty [TD = 5, QD = 1]
- Define (accumulated inverse) document distance rewarding multiple occurrences of term pairs

$$AIDD(a,b;D)_{a\neq b} = \frac{1}{d_1} + \frac{1}{d_2} + \frac{1}{d_3} + \ldots + \frac{1}{d_k}$$

- Score every pair of matched query words, **scaling document distance by tree distance**

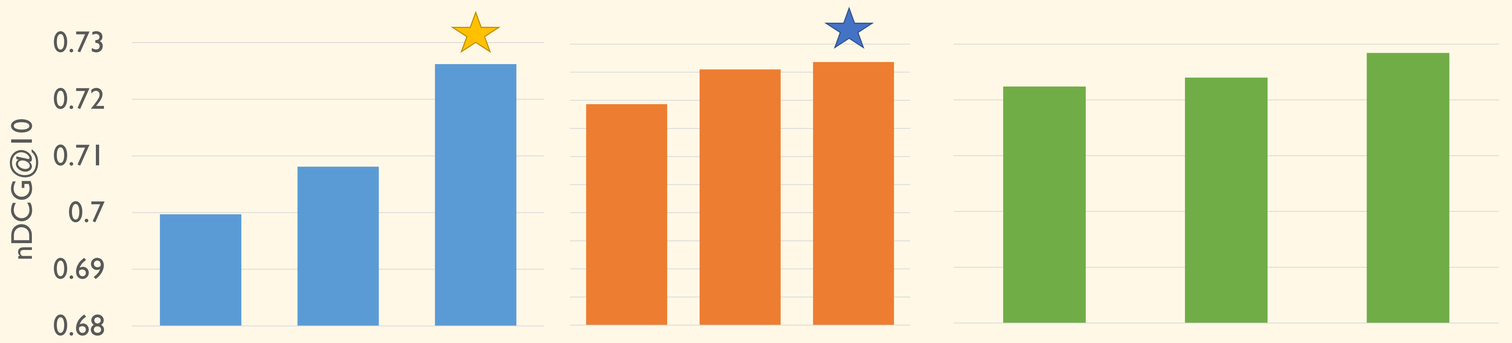
$$RrSV_D = \sum_{\substack{t_i,t_j \in q \cap D \\ t_i \neq t_j}} \frac{AIDD(t_i,t_j;D)}{td(t_i,t_j;n(q))}$$

- Score documents by aggregating scores for every pair of matched query words, then **re-rank** them
- Fuse original rank (unsegmented query) and new rank

EVALUATION

- 17 Million **Bing queries** from Australia (May 2010) used for generating nested segmentations
- **Baselines:** Flat segmentation (Hagen et al. 2011), scaling with query distance/no scaling, Huang et al. (2010)
- **Datasets:** SGCL12 with 500 *long* queries (5 – 8 words), ~14000 documents, ~30 judgments per query; TREC-Web Track (2009 - 2012)

Segmentation	Segmented Query
Flat	garden city shopping centre | brisbane | qld
Nested	((garden city) (shopping centre)) (brisbane qld)
Flat	the chronicles of riddick | dark athena
Nested	(the ((chronicles of) riddick)) (dark athena)
Flat	sega superstars tennis | nintendo ds game
Nested	((sega superstars) tennis) ((nintendo ds) game)



Nested query segmentation can be viewed as the first step towards query parsing, and can lead to a generalized query grammar!