Weekly reading report

My imported papers cover a variety of topics, including entity, latent search, MRF, n-gram and so on. Since the focus is on QAC (Query auto-completion), I read many papers on QAC. Since last week we watched a talk related to Susan Dumais, I am also interested by her biggest contribution: Using latent-based methods to do indexing. I also read [7] whose title is very appealing to me because it seems related to structured data search (but actually not), [7] proposes a entity TREC that is judged by crowd-source, to help search structural RDF (Resource description formate) WEB or "Semantic WEB", I also read a few papers on entities.

1 QAC

[1] gives a good summary on different auto-completion methods and their relevance performance. [2] is also a good starting point (a survey book).

For particular methods, [3] improve systems that can only recommend queries for prefixes that have been previously seen by the search engine with adequate frequency. It explores ranking signals that are appropriate for both types of candidates based on n-gram statistics and a convolutional latent semantic model (CLSM). Specifically, it uses suffixes that are popular n-grams, appends it to user's query prefix, QAC generates synthetic suggestion candidates that have never been observed in the history log. [4] Proposes a language model that employs neural networks to resolve the dimensionality problem with a distributed representation of text, using RNN. It achieves "significant" results compared to n-gram for both seen and unseen prefixes. [5] is suggesting QAC to be context-sensitive at entity level. These QAC papers are usually comparing results with MostPopularCompletion (MPC) method which relies on the popularity of query sequences rather than just the popularity of individual queries. The search engine suggests to the user the completions that have been most popular among users in the past (we call this algorithm MostPopularCompletion). MPC is the most popluar and standard QAC algorithm which is used by many QAC papers as baseline.

2 Latent-related

The latent semantic structure method in [6] uses the derived factor representation to process the query, represent a query (or "pseudo-document") as the weighted sum of its component term vectors from SVD.

3 Entity-related

[8] applies LDA to query log data. The topic model is constructed by a novel and general learning method referred as WS-LDA. The problem it tries to solve is Named Entity Recognition in Query (NERQ), e.g. harry potter walkthrough is a Game instead of Book because walkthrough strongly indicates that harry potter here is more likely to mean the Harry Potter game. According to [8], 71% of search queries contain named entities. In LDA model, classes of the named entity are represented as topics of the model. It proposes a probabilistic approach to address NERQ problem assuming each named entity can only belong to one class. In [8] model, If equals 0, WS-LDA learning will degenerate to LDA learning.

4 Reference

[1] Di Santo et al., Comparing Approaches for Query Autocompletion.
[2] Cai and de Rijke, A Survey of Query Auto Completion in Information Retrieval.
[3] Mitra and Craswell, Query Auto-Completion for Rare Prefixes.
[4] Park and Chiba, A Neural Language Model for Query Auto-Completion.
[5] Schmidt et al., Context-Sensitive Auto-Completion for Searching with Entities and Categories.
[6] Deerwester et al., Indexing by Latent Semantic Analysis.
[7] Blanco et al., Entity Search Evaluation over Structured Web Data.
[8] Guo et al., Named Entity Recognition in Query.