An Algorithm in Determining Field Weight Distributions for Language Models and a New Language Model

Di Xu, Purdue University

ADVISORS: Prof. Marianne Winslett, Arash Termehchy

Background

- **Semi-structured (XML) data (SSD):** many fields store different information.
- **Ranker:** ranks retrieved answers to an unstructured query.
- **Language models (LM):** have been proposed for existing rankers, but are incompetent and error-prone.

$$P(Q|d) = \prod_{i=1}^{\text{terms}} \sum_{j=1}^{\text{fields}} \text{FWD} \cdot P(q_i|f_j),$$ where $\text{FWD}$ (field weight distribution) is the general knowledge of users’ intentions, and $P(q_i|f_j)$ is the query term likelihood. One such LM is PRMS: a Probabilistic Retrieval Model for SSD.

- **FWD:** is difficult to find; bad ones result in bad performance. That’s why PRMS is bad.

Research Plan

- **Goal:** Designing better approaches in finding FWD for existing rankers using LM (such as PRMS).
- **Challenges:**
  - There exists no optimal ranker for all kinds of queries and data.
  - Improvement in one query can sacrifice other queries of different targets.
  - Data are noisy and full of garbage.
  - Analysis on queries is tough work.
  - Time and space complexity.

Research Results

- **Found a mechanism that gives better FWD for each query term, which brings significant improvements.**
- **Designed a new ranker (KFILM) with decent performance, which outperforms PRMS and Multi-Style Ranker.**

Algorithm

We assume the dataset is storing realistic information and is reasonably diversified. Construct a universal FWD (UFWD) by inspecting a certain number of queries. UFWD provides knowledge about the general importance of the fields. This is done by training over 30 queries.

For each query term, we construct specific FWD for this term (UTFWD), based on UFWD. UTFWD provides knowledge about the general importance of the fields subject to a certain query term. As we can see, it is just a general idea that is not specific to a term or a document.

For each document, we constructed specific FWD for this query term in this document (DTFWD), which is essentially part of UTFWD. At that point, we obtained a specific distribution for the term “Titanic” over all candidate documents. Apparently, this eliminates lots of useless information that is irrelevant to “Titanic.” For example, in the document “Titanic: A Tale of Two Journeys (2005) (TV),” “titanic” appears in “Title” and “Keywords” only.

Related Involvement in Other Projects

- “Principled and Optimal Language Model for Keyword Search Over Databases,” Professor Marianne Winslett, Arash Termehchy, Ph.D, and Yodsawalai Chodpathumwan. DAIS Laboratory, Department of Computer Science, University of Illinois at Urbana-Champaign, May 2012.

Referenced Works

- “A Probabilistic Retrieval Mole for Semistructured Data,” Jimyoung Kim, Xiaobing Xue, and W. Bruce Croft., CIIR, Dept. Computer Science, UMA.
- “Multi-Style Language Model for Web Scale Information Retrieval,” Kuansan Wang, Xiaolong Li, and Jianfeng Gao, One Microsoft Way, Redmond, WA.