RUC @ INEX 2011 Data-Centric Track

Qiuyue Wang\textsuperscript{1,2}, Yantao Gan\textsuperscript{1}, and Yu Sun\textsuperscript{1}

\textsuperscript{1} School of Information, Renmin University of China
\textsuperscript{2} Key Lab of Data Engineering and Knowledge Engineering, MOE, Beijing 100872, P.R. China
qiuyuew@ruc.edu.cn, ganyantao19901018@163.com, yusun.aldrich@gmail.com

Abstract. We report our experiment results on the INEX 2011 Data-Centric Track. We participated in both the ad hoc and faceted search tasks. On the ad hoc search task, we employ language modeling approaches to do structured object retrieval, trying to capture both the structure in data and structure in query and unify the structured and unstructured information retrieval in a general framework. However, our initial experimental results using INEX test bed show that the unstructured retrieval model performs better than structured retrieval models. On the faceted search task, we propose a simple user-simulation model to evaluate the effectiveness of a faceted search system’s recommending facet-values. We implemented the evaluation system and conducted the evaluations for the track. We also tested basic redundancy and relevance based approaches for recommending facet-values. The results show that our basic approaches of recommending facet-values perform quite well.

1 Introduction

More and more data on the Web are structured, e.g. data in Deep Web and Semantic Web, while most end users prefer to search any data with a simple keyword query interface. There has been increasing interest structured data retrieval in recent years. INEX 2011 Data-Centric Track is one of the efforts to investigate retrieval techniques on highly structured XML data, where rich structure carries important semantic and relationship information about pieces of data. Basically, structural information can be exploited to improve ad hoc retrieval performance, and also can be used to help users navigate or explore a large set of results as in faceted search systems. We participated in both the ad hoc and faceted search tasks.

On the ad hoc search task, we study how to exploit structural information in data, and how to infer and exploit structural information implicit in queries to improve retrieval performance. We base our studies on language modeling approaches, and propose structured language models to automatically infer structural information from unstructured queries, and match structure in data and in query probabilistically. Compared with other structured language modeling approaches in information retrieval, our approach can capture both structure in data and in query and unify structured and unstructured data retrieval in a general framework. However, our experimental results on INEX show that the structured models are inferior to the unstructured ones.
On the faceted search task, to avoid expensive and non-repeatable user studies, we propose a cost-based metric and implement an evaluation system based on user simulations to evaluate all the submitted runs for the task. We also tested the basic redundancy and relevance based approaches for recommending facet-values. Among all the submitted runs based on the same reference result list, the basic redundancy based approach performs best.

2 Ad Hoc Search Task

Language modeling approach has a solid statistical foundation, and can be easily adapted to model various kinds of complex and special retrieval problems, such as structured document retrieval. In particular, mixture models [1] and hierarchical language models [2][3][4] were proposed to be applied in XML retrieval. On the ad hoc search task in INEX 2011 data-centric track, we employ the language modeling approach to do structured object retrieval as the IMDB data collection can be viewed as a set of structured objects, i.e. movies and persons. With the rich structural information in data, we intend to investigate how to capture the structural information in data as well as that in query in language models to retrieve more accurate results for an ad hoc information need.

In this section, we discuss different ways of adapting language modeling approach to structured object retrieval, and evaluate them on the IMDB data collection.

2.1 Unstructured Data, Unstructured Query

The basic idea of language modeling approach in IR is to estimate a language model for each document ($\theta_D$) and the query ($\theta_Q$), and then rank the document in one of the two ways: by estimating the probability of generating the query string with the document language model, i.e. $P(Q|\theta_D)$, as in Equation 1, or by computing the Kullback-Leibler divergence of the query language model from the document language model, i.e. $D(\theta_Q \parallel \theta_D)$, as in Equation 2.

$$P(Q|\theta_D) = \sum_{w\in Q} P(w|\theta_D)$$

(1)

$$-D(\theta_Q \parallel \theta_D) = -\sum_{w\in V} P(w|\theta_Q)\log\frac{P(w|\theta_Q)}{P(w|\theta_D)}$$

(2)

On the surface, the KL-divergence model appears to be quite different from the query likelihood method. However, it turns out that the KL-divergence model covers the query likelihood method as a special case when we use the empirical distribution to estimate the query language model, i.e. maximum-likelihood estimate.

In IMDB data collection, each document is a structured object, i.e. movie or person. Our first retrieval strategy is to ignore the structural information in each object, estimate a language model for each object based on its free-text content, and rank them using query likelihood.