A Relation Pattern-Driven Probability Model for Related Entity Retrieval

Peng Jiang, Beijing Institute of Technology, China
Qing Yang, Beijing Institute of Technology, China
Chunxia Zhang, Beijing Institute of Technology, China
Zhendong Niu, Beijing Institute of Technology, China
Hongping Fu, Beijing Institute of Technology, China

ABSTRACT
As the Web is becoming the largest knowledge repository which contains various entities and their relations, the task of related entity retrieval excites interest in the field of information retrieval. This challenging task is introduced in TREC 2009 Entity Track. In this task, given an entity and the type of the target entity, a retrieval system is required to return a ranked list of related entities extracted from a given large corpus. It means that entity ranking goes beyond entity relevance and integrates the judgment of relation into the evaluation of the retrieved entities. This paper proposes a probability model using relation patterns to address the task of related entity retrieval. This model takes into account both relevance and relation between entities. The authors focus on using relation patterns to measure the level of relations matching between entities, and then to estimate the probability of occurrence of relation between two entities. In addition, the authors represent entity by its context language model and measure the relevance between two entities by a language model. Experimental results on TREC Entity Track dataset show that the proposed model significantly improves retrieval performances over baseline. The comparison with other approaches also reveals the effectiveness of the model.

Keywords: Information Retrieval, Language Model, Related Entity Retrieval, Relation Pattern

1. INTRODUCTION
In recent years, the rapidly increasing scale and wide spread of the Web has rendered it an immense knowledge repository, which is a rich information source of entities and their relations. The need to find appropriate retrieval techniques to track these entities and their relations raises some challenging problems in the field of information retrieval (IR). Related entity retrieval is a task to solve the challenges and serve the growing interest in IR.

Different from the traditional retrieval task in which the retrieval unit is document,
the retrieval unit of related entity retrieval is a kind of entity with a fixed type such as person, organization, product, location, etc. We need to extract these entities from relevant document in advance. In addition, related entity retrieval also differs from traditional entity search which does not consider the relations between entities. In traditional entity search, typical information need is like “find me a list of experts whose interesting research area is natural language processing (NLP)”, where the retrieved results are persons who are relevant to NLP. Typical information needs in related entity retrieval include “find me a list of experts who are the students of Claire Cardie”. The retrieved persons are not only relevant to “Claire Cardie”, but also have the relation “students of” with “Claire Cardie”. Another example is “find me a list of airlines that currently use Boeing 747”. The retrieved airlines must use “Boeing 747” planes currently. There is a relation between airlines and “Boeing 747”. By resorting to traditional search engine, we are returned a list with massive amount of information. It is exhausted to select related entities from that list manually.

TREC 2009 Entity Track highlights the information needs about relations between entities in the Web, whereby Related Entity Retrieval task is introduced. It aims at finding entities (target entities) that have a given relation with a given entity (source entity). Some requirements are introduced by this task: 1) types of target entities are not fixed to one compared with that of expert finding and time search. Besides, the retrieval domain is open. It requires the retrieval approach to be domain-independent and applicable to varied entity types. 2) The retrieved entities must have a given relation with the source entity. It means that the task goes beyond entity relevance and must integrate the judgment of relation into the evaluation of the retrieved entities. However, the relation is usually described in a short and free text. It is hard to represent the relation information in the retrieval process. In TREC 2009 Entity Track, the typical approach (Balog, Vries, Serdyukov, & Thomas, 2009) to the related entity retrieval is to gather snippets for the source entity, followed by extracting co-occurring entities from these snippets using named entity taggers. Wikipedia and other external resources are applied to improve named entity recognition. However, most approaches do not effectively make use of the relations specified in topics.

In this paper, we introduce a probability model to formalize and accomplish the related entity retrieval task. This model considers both relevance and relations between two entities, so that it is effective to evaluate the levels of relevance and relation matching of a target entity. In order to measure the relevance between two entities, we represent entity by its context language model which can be estimated by pseudo-relevance feedback. Thus, we can employ the language model approach to measure the relevance. For the relation judgment, we use an approach based on relation pattern. For the given relation defined in each topic, the relation patterns are learned automatically to measure the level of relation matching between two entities. Although relation patterns are based on rules, we integrate them into our proposed probability model by the statistical information of relation patterns.

The proposed probability model is general and has many potential applications. Specifically, the model can be applied to the Factoid Question Answering (QA). For factoid QA, the answers can be considered as target entities, while the association between answers and question can be considered as a relation. Besides, the experiments are conducted on TREC dataset which contains about 50 million English-language web pages covering a wide range of topics.

The remainder of the paper is organized as follows: In Section 2, we make a brief introduction of the related works in the field. The problem is defined in Section 3. The whole approach is described in Section 4. The experiments and result analysis are presented in Section 5. Finally we conclude the paper and discuss the future work in Section 6.
Adaptation and Dependability and Their Key Role in Modern Software Engineering
www.igi-global.com/chapter/adaptation-dependability-their-key-role/63579?camid=4v1a