Global Semantic Expansion for Hierarchical Query Intent Identification

Wei Song, Wenbin Xu, Lizhen Liu, Hanshi Wang

College of Information Engineering

Capital Normal University, China

Introduction

Understanding user intent is important for interactive and personalized information retrieval. User intent space actually forms a hierarchical, coarse to fine, top down architecture . Here, the **Hierarchical Query Intent** is defined as a two level structure.

● Query Senses. Each query sense represents a semantic sense of the given query, which usually refers to a specific object in reality.

● Query Subtopics. Each query subtopic represents an aspect or attribute of a specific query sense.

Formally, a query is noted as ***q***, a query aspect phrase is noted as ***a***, and a query subtopic candidate is represented as ***q+a***. For example, suppose *q*= ”xiaomi”, *a*=“company”, a query subtopic candidate is “xiaomi company”.

We propose a method based on global semantic expansion. The main contributions include:

● We use word semantic vectors and propose a query dependent semantic composition method for representing query aspect phrases.

● We expand query subtopics by introducing new words according to global semantic relatedness and cluster these words for query sense induction.

We seek to answer the following research questions.

● Whether word vectors are better representations compared with traditional ones for query subtopic mining?

● Whether global semantic expansion benefits the query sense induction?

Aspect Phrase Extraction

In this work, we focus on clustering query aspect phrase clustering and query sense induction. We didn’t pay much attention on aspect phrases extraction and only used the candidates provided by the NTCIR organizer. However, query logs and other resources could be exploited for extracting more query aspect phrases.

Aspect Phrase Representation

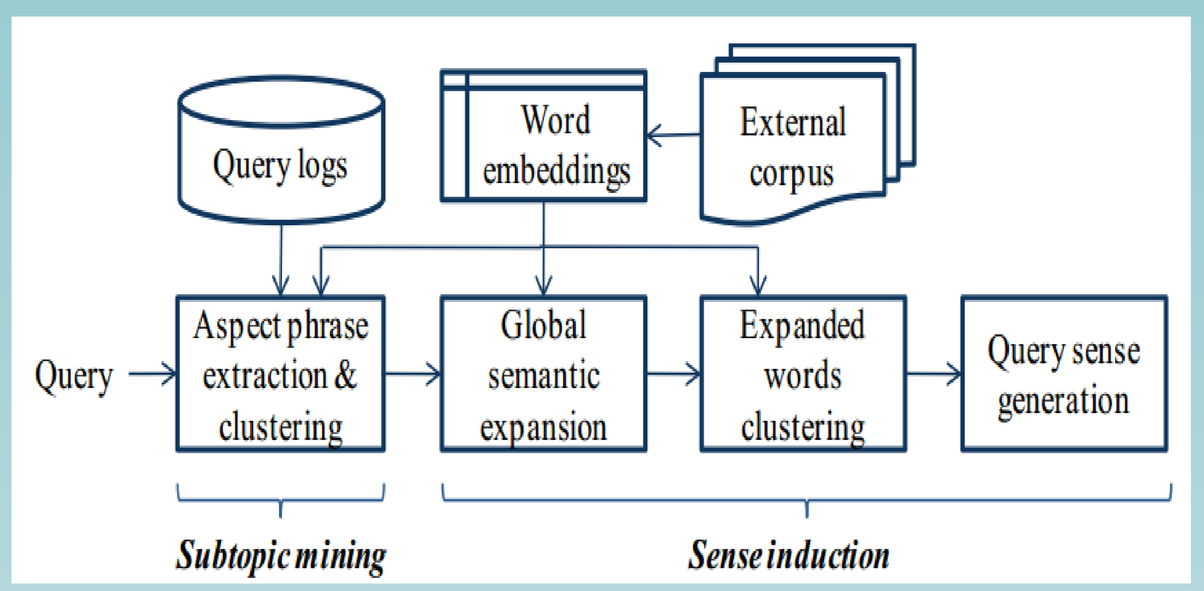
Word Embedding: each word within an aspect phrase is represented by a distributed continuous vectors which is learned using the word2vec toolkit on a crawled Baidu Baike dataset.

Aspect Phrase Embedding: If an aspect phrase contains multiple words, it is represented as a vector as well. We use a linear weighted composition method for constructing the vectors based on individual word embeddings.

$$phrVec(w_1, ..., w_n) = \sum_{i=1}^{n} \alpha_i vec(w_i)$$

The weighting parameters are determined based on their co-occurrence frequencies with the query in query logs.

Our System



Query Subtopic Mining

Query aspect phrases are clustered based on K-means clustering algorithm. *K* is set to 20, and we further remove clusters with few aspect phrases. The remaining clusters forms query subtopics.

Query Sense Induction

Given a set of fine-grained subtopics, we propose a semantic expansion strategy for query sense induction.

● Step 1: For each query subtopic, we view every aspect phrase *a* as a seed, and expand a list of most similar words based on the learned word embeddings.

● Step 2: We cluster these expanded words with Affinity Propagation which could determine the number of clusters automatically. Word clusters are viewed as sense candidates.

●Step 3: We assign each query subtopic to one word cluster according to the distance between the subtopic centroid and the word cluster centroid. The word clusters having at least one query subtopic are used to represent query senses.

Conclusion

According to our experiments, we found that

●Word embedding representation benefit query aspect phrase clustering compared with traditional representations including: bag of words and co-occurrence in search results.

●Global semantic expansion benefit for inducing query senses compared to clustering query subtopics directly. It is useful for separating different senses for ambiguous queries, such as “Xiaomi” .