WSSE: A Web Service Search Engine for Large Scale Web Service Discovery based on the Probabilistic Topic Modeling and Clustering
Afnan Bukhari, aab2790@rit.edu
Supervised by: Dr. Xumin Liu, Rochester Institute of Technology

Introduction

- A few number of free web service search engines are available such as Programmableweb.com. Most of them are keyword-based search engines which suffer from the low precision and recall rating.
- The solution is to develop an efficient search engine that can retrieve the most relevant web services in a short time.
- Our search engine is a semantic-based that integrated the probabilistic topic modeling and clustering technique to increase the precision and recall of web service discovery and provide a good performance in expressing semantic concepts.
- Users don’t need to specify the concise and correct keywords of the query in order to retrieve satisfied results.

Methodology

- Our experiments are performed on the programmableweb dataset. 7560 web services (WSDL and Restful) from the largest 20 categories are parsed and preprocessed to obtain the BOW for each service. Web services (BOWs) modeled as Word vectors by TF/IDF model as 7560 WS X 7751 words = 296.3 MB.
- The LDA reduces the dimensionality of web service matrix significantly. It captures the semantic concepts of web service descriptions. Each web service modeled as a Topic vector by LDA as 7560 WS X 100 topics = 22 MB.
- Using topic vectors to feed the clustering algorithms such as Kmeans and agglomerative to obtain best clusters. The clustering algorithms group the web services based on their related topics.
- The user query modeled as a topic vector.
- Q = [T1 T2 T3 ... Tn]
- Find the closest similar cluster.
- Calculate the cosine similarity between Q and WS.
  Sim(Q,WS) = Dot product(Q,WS) / ||Q|| ||WS||
- Rank the results based on the highest similarity scores.

LDA

- The optimal number of topics is 100 based on the perplexity result.

Evaluation

- Clustering methods with Word vectors and Topic vectors are compared based on the accuracy metric.
- Using vector-based keywords and topic vectors.
- Word vectors: Kmeans on word vectors Hierarchical on word vectors
  - Accuracy: 47% 36.7%
- Topic vectors: Kmeans on Topic vectors, Topics = 20
  - Accuracy: 53%
- Kmeans on Topic vectors, Topics = 100
  - Accuracy: 59%
- Hierarchical on Topic vectors
  - Accuracy: 46%
- LDA (clusters = the highest weight) Topics = 20
  - Accuracy: 39%
- LDA (clusters = the highest weight) Topics = 100
  - Accuracy: 56%

- Comparing the performance of word vectors and topic vectors:
  - Word vectors: Word vectors: Kmeans on word vectors Hierarchical on word vectors
    - Accuracy: 47% 36.7%
  - Topic vectors: Kmeans on Topic vectors, Topics = 100
    - Accuracy: 59%
  - Hierarchical on Topic vectors
    - Accuracy: 46%
  - LDA (clusters = the highest weight) Topics = 20
    - Accuracy: 39%
  - LDA (clusters = the highest weight) Topics = 100
    - Accuracy: 56%

Framework

- Data Preprocessing
- Web service descriptions
- Pre-Processing
  - Extract words
  - Tokenization
  - Splitting
  - Removing stop words
  - Stemming
- Topic modeling (LDA model)

- Service Discovery
- Query
  - Topic Vector
- Search Engine
  - Indexing
- Response
  - Cluster 1
  - Cluster 2
  - Cluster n

Results

- Our search system approach WSSE is compared with the available keyword-based discovery approaches:
  - Programmableweb.com.
  - Text-search system.
- Different 20 queries include 3 types of queries are generated and tested.
  - One keyword query e.g. “Government”.
  - Two-three words query e.g. “city weather”.
  - Sentence query e.g. “API for editing images and videos”.

- Other 40 queries are tested. The query is the mashup description that uses at least one API. The number of APIs of each mashup is used as the ground truth that we compared our results.

Conclusion

- Using the probabilistic topic modeling improves the accuracy of clustering results and integrating them improves the quality and efficiency, and accuracy of web service discovery by reducing the time and space overheads and providing relevant results.