Learning deep structured semantic models for web search using clickthrough data

Lee ji soo

Department of Statistics
Seoul National University

2017.12.26

Overview

1. Latent semantic models
   - keyword-based matching(LSA) often fails
2. Proposed deep structured model
   - trained by maximizing the conditional likelihood of the clicked documents given a query using the clickthrough data

latent semantic model

- A query and a document, represented as two vectors in the lower-dimentional semantic space, can still have a high similarity score even if they do not share any term
- the model address the language discrepency between Web document and search queries by grouping different terms that occur in a similar context into the same semantic cluster.

Bi-lingual Topic Model(BLTM)-latent semantic model

- generative model that requires that a query and its clicked documents not only share the same distribution over topics but also contain similar factions of words assigned to each topic
- hierarchically models documents by treating each document as a set of segments, e.g. sections

Bi-lingual Topic Model(BLTM)

- clickthrough data consists of a list of queries and their clicked documents
- used for semantic modeling so as to bridge the language discrepancy between search queries and Web documents
- the training is to maximize a conditional likelihood of the clicked documents given a query using the clickthrough data

Deep Structered Semantic method

- Salakhutdinov and Hinton extended the semantic modeling using deep auto-encoders
- using the hierarchical semantic structure embedded in the query and the document can be extracted via deep learning
- a non-linear projection is performed to map the query and the documents to a common semantic space
- the relevance of each document given the query is calculated as the cosine similarity between their vectors in that semantic space

Deep Structured Semantic method

- popular models can be grouped into two categories, linear projection models and generative topic models
- By using the singular value decomposition of a document-term matrix, a document can be mapped to a low-dimensional concept vector $\hat{D} = A^T D$, where A is the projection matrix

$$sim_A(Q, D) = \frac{\hat{Q}^T \hat{D}}{\|\hat{Q}\|\|\hat{D}\|} \tag{1}$$

- The relevance score between a query and a document, represented respectively by term vectors Q and D, is assumed to be proportional to their cosine similarity score of $\hat{Q}$ and $\hat{D}$

Deep Structered Semantic method

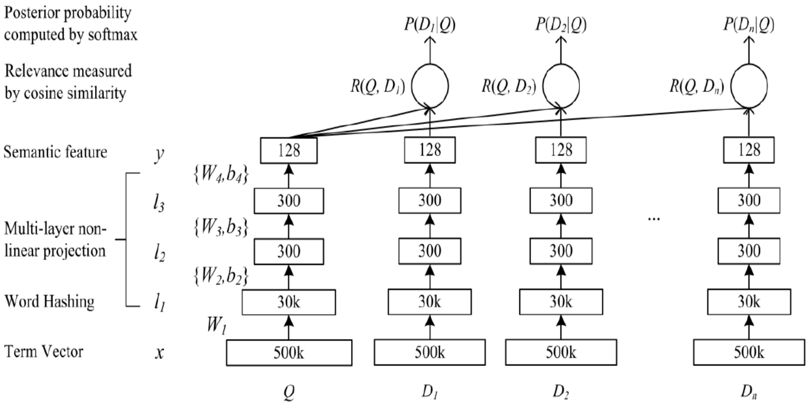

Figure 1: Illustration of the DSSM. It uses a DNN to map high-dimensional sparse text features into low-dimensional dense features in a semantic space. The first hidden layer, with 30k units, accomplishes word hashing. The word-hashed features are then projected through multiple layers of non-linear projections. The final layer's neural activities in this DNN form the feature in the semantic space.

Deep Structured Semantic method

- map term vectors to their corresponding semantic concept vectors
- compute the relevance score between a document and a query as cosine similarity of their corresponding semantic concept vectors
- we donate x as the input term vector, y as the output vector, l_i as the intermediate hidden vector, W_i as the i-th weight matrix, b_i as the i-th bias term

Use of clickthrough data

- y_Q and y_D are the concept vectors of the query and the document

$$l_1 = W_1 x \tag{2}$$

$$l_i = f(W_i l_{i-1} + b_i), i = 2, ..., N-1 \tag{3}$$

$$y = f(W_N l_{N-1} + b_N) \tag{4}$$

- where we use the tanh as the activation function at the output layer and hidden layers l_i

$$R(Q, D) = cosine(y_q, y_d) \tag{5}$$

word hashing

- reduce the dimensionality of the bag-of-words term vectors
- add word starting and ending marks and break the word into n-grams
- the word is represented using a vector of letter n-grams
- while the number of English words can be unlimited, but the number of letter n-grams in English is often limited
- it allows to scale up the DNN solution when extremely large vocabularies are desirable

Conclusion

- we make use of the clickthrough data to optimize the parameters of all version of the models by directly targeting the goal of document ranking
- the deep architectures adopted have further enhanced modeling capacity so that more sophisticated semantic structures in queries and documents can be captured and represented
- a letter n-gram based word hashing scale up the taining so that very large vocabularies can be used in realistic web search

The End