Websearch using DSSM

Saniya Sharma(20162020)
Mansi Singh (20162034)
Yojana Birje (20162130)
Ramakrishna Kota

Background

- **TF-IDF**

  TF*IDF is a rough way of approximating how users value the relevance of a text match. This technique relies on two parameters :

- Term Frequency:How often does term occur in the document
- Inverse Document Frequency:Inverse document frequency (`1/df`) then measures how special the term is.

- **BM25**
- BM25 is better version of TF-IDF .Term frequency in BM25 dampens the impact of term frequency even further than traditional TF*IDF. The impact of term frequency is always increasing, but asymptotically approaches a value
- But it does word matching so it won’t match if we have words of similar meaning.

- **LSA**
- A better approach would allow users to retrieve information on the basis of a *conceptual topic* of a document.
- LSI assumes that there is some underlying or latent structure in word usage .This data is more robust indicator of meaning than individual terms.

Deep Structured Semantic Model

Clickthrough Data:

- Clickthrough data in search engines can be thought of as triplets (q,r,c) consisting of the query q, the ranking r presented to the user, and the set c of links the user clicked on.
- While clickthrough data is typically noisy and clicks are not “perfect” relevance judgments, the clicks are likely to convey some information.

DSSM

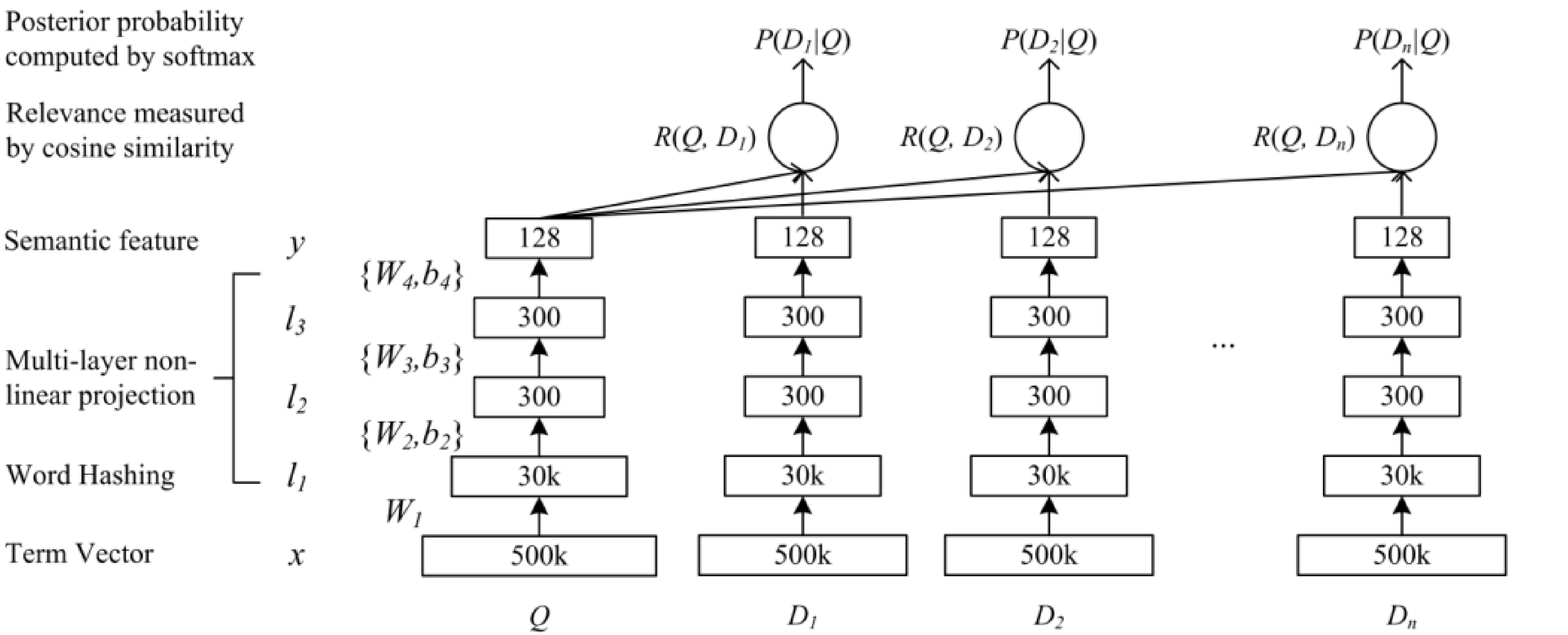


Figure 1: Illustration of the DSSM. It uses a DNN to map high-dimensional sparse text features into low-dimensional dense features in a semantic space. The first hidden layer, with 30k units, accomplishes word hashing. The word-hashed features are then projected through multiple layers of non-linear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

First Layer

- Term Vector – raw bag of words features
- (Q,D)
- D includes D+ and D-

Second Layer

- Word Hashing

  -tri-grams

  -reduce dimension

  -handle out-of-vocabulary problem

  -reduce collision

Cosine similarity

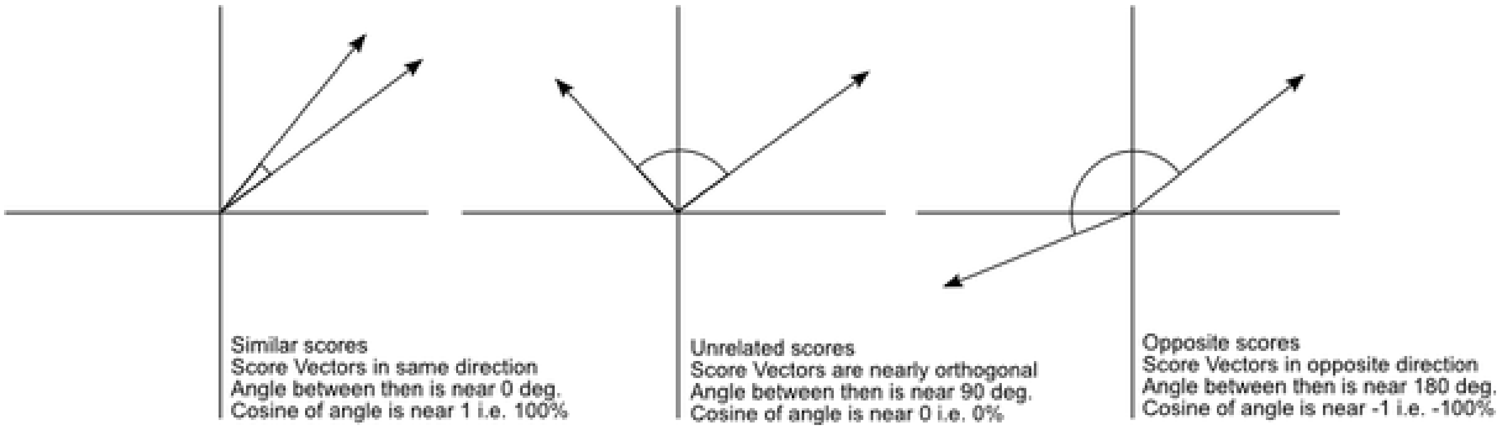
$$R(Q, D) = \text{cosine}(y_Q, y_D) = \frac{{y_Q}^T y_D}{\|y_Q\| \|y_D\|}$$

YQ= Query vector , YD=Document vector

- Better than Euclidean distance
- It is independent of document length.

Cosine similarity

- Cosine Similarity will generate a metric that says how related are two documents by looking at the angle instead of magnitude.



Hidden Layers

$$l_1 = W_1 x$$

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

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

- Wi is ith weight matrix
- Tanh activation function

$$f(x) = \frac{1 - e^{-2x}}{1 + e^{-2x}}$$

- we initialize the network weights with uniform distribution in the range between

Softmax function

Loss funct

$$P(D|Q) = \frac{\exp(\gamma R(Q,D))}{\sum_{D' \in \mathbf{D}} \exp(\gamma R(Q,D'))}$$

$$L(\Lambda) = -\log \prod_{(Q,D^+)} P(D^+|Q)$$

Gradient descent procedure

$$\Lambda_t = \Lambda_{t-1} - \epsilon_t \frac{\partial L(\Lambda)}{\partial \Lambda} |_{\Lambda = \Lambda_{t-1}}$$

RESULT

Results show that the deep structured semantic model is the best performer, beating other methods by a statistically significant margin.

WHY?

- Supervised learning on clickthrough data
- Word hashing allows us to use very large vocabularies for modeling.
- Using a deep architecture.

CONCLUSION

Our model contributes in three ways:

- Use Of clickthrough Data
- The deep architectures adopted have further enhanced the modeling capacity so that more sophisticated semantic structures in queries and documents can be captured and represented
- Third, we use a letter n-gram based word hashing technique

Applications of DSSM

1) **Contextual entity search**
-Given a user-highlighted text span representing an entity of interest
- search for supplementary document for the entity

2)**Automatic highlighting**
-given a document a user is reading
-discover the concepts/entities/topics that interest the user and highlighting the corresponding text span

3) **Document prefetching**
-given a document a user is reading
Prefetching a document that user will be interested in next

Thank you