Knowledge-Based Systems 000 (2018) 1–11

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Learning a unified embedding space of web search from large-scale query log

Lidong Bing[a,*], Zheng-Yu Niu[b], Piji Li[c], Wai Lam[c], Haifeng Wang[b]

[a] Tencent AI Lab, Shenzhen, China
[b] Baidu Inc, Beijing, China
[c] Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Hong Kong

ARTICLE INFO

Article history:
Received 24 September 2017
Revised 18 January 2018
Accepted 24 February 2018
Available online xxx

Keywords:
Web search
Query representation
Embedding space
Session analysis

ABSTRACT

In the procedure of Web search, a user first comes up with an information need and a query is issued with the need as guidance. After that, some URLs are clicked and other queries may be issued if those URLs do not meet his need well. We advocate that Web search is governed by a unified hidden space, and each involved element such as query and URL has its inborn position, i.e., projected as a vector, in this space. Each of above actions in the search procedure, i.e. issuing queries or clicking URLs, is an interaction result of those elements in the space. In this paper, we aim at uncovering such a unified hidden space of Web search that uniformly captures the hidden semantics of search queries, URLs and other involved elements in Web search. We learn the semantic space with search session data, because a search session can be regarded as an instantiation of users' information need on a particular semantic topic and it keeps the interaction information of queries and URLs. We use a set of session graphs to represent search sessions, and the space learning task is cast as a vector learning problem for the graph vertices by maximizing the log-likelihood of a training session data set. Specifically, we developed the well-known Word2vec to perform the learning procedure. Experiments on the query log data of a commercial search engine are conducted to examine the efficacy of learnt vectors, and the results show that our framework is helpful for different finer tasks in Web search.

1. Introduction

Web search is the fundamental tool for us to quickly retrieve the needed Web pages. To improve the retrieval performance, researchers have conducted extensive studies in different topics, such as query reformulation [8,9,50], query suggestion [2,34], query difficulty (or quality) prediction [21,22,29], query intent analysis [3,12,41], page quality evaluation and spam detection [7,11,20], and search result ranking [15,28,38]. In this paper, we exploit another paradigm which aims at mining distributed representation of Web search elements, such as terms, queries, pages/URLs, and websites. We advocate that Web search is governed by a unified hidden space, and each element can be embedded as a vector in the space. Fig. 1 depicts an example to show the intuition of this idea. The user has an information need (e.g. "I wanna repair my iPhone screen.") in mind which can be semantically represented as a vector of particular dimensions, 4 in this example, and each dimension indicates the relevance of his need with a particular hidden semantic topic. Under the guidance of this information need, three queries are issued. Although the user intends to formulate queries conveying his need on the third dimension, say "repair screen", the first two queries have a large semantic relatedness with the first dimension, say "Apple". Consequently, they retrieve pages mainly from the website of Apple. After browsing a few pages (e.g. u_1, u_2, and u_3) and feeling unsatisfied (maybe because of the high price), the user issues the last query with a hidden semantic representation well matching with his need, and accordingly, the returned URLs satisfy him better. To generalize the example, websites and query terms could also be involved and represented as vectors in the same space. Obviously, uncovering such a space governing the search procedure can be very useful for different finer tasks of Web search.

Researchers had observed the potential of generating semantic vectors for queries and URLs/pages, and conducted some pioneer investigations [19,25,46]. Our work is different from them in a few aspects. First, these works only learn vectors for queries and URLs, while our work also learns vectors for websites and terms. Therefore, the learnt vectors in our work can be applied to different tasks, not limited to result ranking. Second, our model can be eas-

* Corresponding author.
E-mail addresses: lyndonbing@tencent.com, binglidong@gmail.com (L. Bing), niuzhengyu@baidu.com (Z.-Y. Niu), pjli@se.cuhk.edu.hk (P. Li), wlam@se.cuhk.edu.hk (W. Lam), wanghaifeng@baidu.com (H. Wang).

https://doi.org/10.1016/j.knosys.2018.02.037
0950-7051/

ily extended to incorporate other types of elements, such as users. Third, the learnt term vectors enable our model to tackle new data such as unseen queries, while these previous works do not have such capability.

We conduct the semantic space learning from search session data since a search session can be regarded as an instantiation of a particular information need. To achieve the goal of learning continuous vector representations for different elements involved in search sessions (e.g., queries, URLs, and websites), a few research questions are raised, namely (a) How to represent the information from session data for vector learning? (b) What kind of models should one use to learn vector representations on session data? and (c) How to augment the model to deal with unseen queries from new sessions?

We cast this vector learning task with session data as a learning problem on a set of graphs, where each graph corresponds to a session, and the elements in a session are represented as vertices and related vertices are connected by edges. The use of graph serves as a suitable choice for session representation since it can capture the semantic interactions among the elements. Given the user's information need represented as a semantic vector, the probability of obtaining a session is jointly determined by the semantic meaning of involved elements, i.e., vertices of the session graph. Then we perform vector learning for vertices by maximizing the log-likelihood of a training session data set. With the learnt representation scheme, we can perform hidden semantic analysis on new session data, given queries or clicked URLs of the new session as source information.

Contributions. The main contributions of this work are as follows.

- We propose a framework of learning a unified semantic space of Web search, and different elements, such as queries, URLs, and terms, are embedded as vectors in this space. The vectors of different types of elements are directly comparable for similarity calculation. And our model also has good applicability on unseen data.
- We use graph structure for session data representation and develop an approach for vertex vector learning on a set of graphs. Our model can capture fine-grained structured information in click-through data. It is generic and naturally lends itself to extensions incorporating other types of elements from session data.
- Our model is trained on a large query log data generated by Baidu[1] search engine. Extensive experiments are conducted to examine the efficacy of the constructed semantic space, and the results show that the learnt vectors are helpful for different tasks.

This article substantially extends our previous work published as conference paper [10]. First, we elaborate on more technical details of the proposed model. Second, more experiments are conducted and more case studies are given, such as the experiments on entity recommendation in Section 6.6 and case studies in Section 5.2. Third, the differences between our work and previous works are discussed more thoroughly across different sections, such as Sections 1, 4, and 7.

The remainder of the article is organized as follows. We first give the problem definition in Section 2. After the first model is described in Section 3, an enhanced model is described in Section 4. The experimental dataset and case studies are given in Section 5, then the experimental settings and results are discussed in Section 6. After related works are reviewed in Section 7, we conclude the article and provide some possible directions for the future work in Section 8.

2. Problem formulation

We aim at uncovering a semantic space that governs the Web search procedure via projecting each involved element (such as search query and URLs) in Web search log as a vector of certain dimensions. The task is precisely defined as follows. Given a set of search sessions $S = \{s_i\}_{i=1}^{n}$ as training data, we aim at finding a unified semantic space to model Web search scenario so that the probability of observing the sessions in S is maximized. Let θ denote an instantiation of model parameters of the space. The log-likelihood objective function is expressed as follows:

$$\ell(\theta; S) = \sum_{s_i \in S} \log P(s_i; \theta), \tag{1}$$

where $P(s_i; \theta)$ denotes the probability that s_i is observed in the space with the parameters θ. Our goal is to find the best parameters by maximizing the above objective.

In our model, the parameters refer to the hidden vectors of involved elements in sessions. Let e_j denote an element such as a query or a URL in s_i, and $\mathbf{v}(e_j)$ denote the vector representation of e_j. Let $\mathbf{v}(s_i)$ denote the information need, represented as a vector, of the user corresponding to the session s_i. $\mathbf{v}(s_i)$ is also called session vector. $\mathbf{v}(s_i)$ and $\mathbf{v}(e_j)$ have the same dimensionality. Let $C(e_j)$ denote the context elements of e_j in s_i. We assume that the probability of e_j only depends on its context $C(e_j)$ and the user's information need, and it is denoted as $P(e_j; C(e_j), \mathbf{v}(s_i))$. Therefore, $P(s_i; \theta)$ can be calculated as:

$$P(s_i; \theta) = \prod_{e_j \in s_i} P(e_j; C(e_j), \mathbf{v}(s_i)). \tag{2}$$

$P(e_j; C(e_j), \mathbf{v}(s_i))$ is calculated with the vectors of elements in $C(e_j)$ and the vector $\mathbf{v}(s_i)$. The calculation will be described later. To summarize, our task is to learn vector representations of the elements in search sessions so that the objective function in Eq. (1) is maximized. To perform learning, we need to transform each session into training instances with the form $(e_j, C(e_j), s_i)$ for calculating $P(e_j; C(e_j), \mathbf{v}(s_i))$. To do so, a major task is to define the context $C(e_j)$ of the element e_j in the session s_i. For better capturing the structured information in click-through data, we introduce a graph representation of session data, which will be discussed in Section 3.

There are some existing studies conducting vector representation learning for words in Natural Language Processing and Speech Recognition [6,35,37,44]. However, they cannot be directly applied to our task due to the following reasons. Our training data is a set of sessions and each of them is represented as a session graph, while the training data of existing methods is a set of word sequences. In addition, a vector capturing the user's information need is incorporated into our learning procedure. Moreover, we intend to learn a unified space that embeds the elements of different granularities such as queries, URLs and terms.

3. Basic model for vector learning on session graphs

3.1. Session graph and training instances

In a search session, there are several types of involved elements. A user first issues a query, and some URLs are clicked in the result list. To obtain better results, she may issue more queries. When browsing a clicked page, the user may also browse other pages in the same website. And the corresponding website is also involved as an element of the session. Thus, a session involves three types

[1] http://www.baidu.com/.