Information Processing and Management 54 (2018) 380–393

Contents lists available at ScienceDirect

Information Processing and Management

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

Unsupervised Latent Dirichlet Allocation for supervised question classification

Saeedeh Momtazi*

Department of Computer Engineering and Information Technology, Amirkabir University of Technology, Tehran, Iran

ARTICLE INFO

Keywords:
Community-based QA
Question classification
LDA

ABSTRACT

Question answering systems assist users in satisfying their information needs more precisely by providing focused responses to their questions. Among the various systems developed for such a purpose, community-based question answering has recently received researchers' attention due to the large amount of user-generated questions and answers in social question-and-answer platforms. Reusing such data sources requires an accurate information retrieval component enhanced by a question classifier. The question classification gives the system the possibility to have information about question categories to focus on questions and answers from relevant categories to the input question. In this paper, we propose a new method based on unsupervised Latent Dirichlet Allocation for classifying questions in community-based question answering. Our method first uses unsupervised topic modeling to extract topics from a large amount of unlabeled data. The learned topics are then used in the training phase to find their association with the available category labels in the training data. The category mixture of topics is finally used to predict the label of unseen data.

1. Introduction

Question answering systems aim at providing exact answers to natural language questions without overwhelming users with a large number of retrieved documents, which they must sort through to find the desired answer. Providing such advanced systems, however, requires more effort and complex linguistics analysis compared to search engines.

In the traditional architecture of open-domain question answering, in the first step, linguistic analyses are performed on the question. This encompasses syntactic parsing and named entity tagging. The information is used later for answer extraction. The expected type of answer; e.g., person, location, etc. is determined in a separate step. In the next step, a query is constructed from the input question and is run against document and passage retrieval on the input corpus. The retrieved texts are then passed to the answer extraction component to extract answers from them. Finally, the answer validation component re-ranks the list of candidate answers and selects the top item as the final answer (Momtazi, 2010).

Although many questions can be answered using the described pipeline, answering complex questions, such as why questions, how questions, and questions on opinions, is still challenging and cannot be addressed by returning a single phrase (Sutcliffe, Kruschwitz, & Mandl, 2010). Addressing such complex questions requires complex reasoning which is of major importance for employing question answering for various purposes (Gurevych, Bernhard, Ignatova, & Toprak, 2009). Since a large portion of users' questions have already been asked by other people in Internet FAQs (Frequently Asked Questions) or social question-and-answer platforms like Yahoo! Answers and are answered by experts or crowds, the available question-and-answer archives are

E-mail address: momtazi@aut.ac.ir.
* Principal corresponding author.

https://doi.org/10.1016/j.ipm.2018.01.001
Received 24 August 2015; Received in revised form 21 August 2017; Accepted 4 January 2018
Available online 04 February 2018

```
<Question>
    <QId>27571710</QId>
    <QSubject>Lake Garda Italian</QSubject>
    <QContent>Hi. We want to spend our vacation in Italy on Lake
    Garda. Of course we also want to do water skiing, but I have heard
    that the police prescribes that one is allowed to do water skiing if
    only 2 persons are on the boat. Is that correct?</QContent>
    <QCategories>
            <QCategory>Law</QCategory>
    </QCategories>
</Question>
```

Fig. 1. Example question with the category label.

valuable information sources that can be used to answer these complex questions and minimize human effort in searching through information. This line of research is known as community-based question answering (CQA). In this case, the task is done through three major steps: question classification which aims at finding the category of the question (Qu, Cong, Li, Sun, & Chen, 2012), text retrieval which searches through question-and-answer pairs to find similar questions to the new input question and ranking their answers (Bernhard & Gurevych, 2009; Burke et al., 1997; Duan, Cao, Lin, & Yu, 2008; Jijkoun & de Rijke, 2005), and summarization which combines information from several user-generated answers and produces a coherent and informative answer to the input question (Tomasoni & Huang, 2010).

In this paper, we focus on the first part of this pipeline and propose a new method based on Latent Dirichlet Allocation (LDA) (Blei, Ng, & Jordan, 2003) to perform the question category classification as a multi-label classification task.[1] A question category classifier can be used for two different purposes:

(1) By classifying all available questions in the question-and-answer archives as well as the input question, we can enhance the retrieval component (Cao, Cong, Cui, & Jensen, 2010; Cao, Cong, Cui, Jensen, & Zhang, 2009; Figueroaa & Neumann, 2014). In this case, the retrieval component uses the category information to prioritize the questions that are in the same category as the input question. For example, consider the question in Fig. 1, which asks about a law in Italy. As can be seen in the content of the question, the main words of the question are related to “Entertainment”, but the question is asking about a “Law”. Considering the question words, it is very likely that the retrieval component retrieves many question-and-answer pairs related to spending vacation in Lake Garda, but irrelevant to the aim of the input question. Using a question classifier, however, the system gives higher priority to questions related to “Law”.[2]
(2) Moreover, such a classifier can be used in the normal scenario of using social question-and-answer platforms. In the available platforms, the users are asked to provide one or more category label(s) for their question. The assigned category labels are used to organize questions and facilitate further browsing and searching. They can also be used to pass the question to specific experts who are registered in the forums. The users’ suggested labels, however, are normally too noisy and with different levels of granularity, which cannot be utilized in an appropriate way. An alternative solution provided in some forums is asking the users to choose a label from a predefined list of categories which can be too large. Running a question classifier in this case will help to fully or partially automate the process. The classifier can either find the category of the question and assign it to the question automatically or suggest a short list of relevant categories to the user to pick the best one(s).

Our proposed method uses an unsupervised LDA topic modeling approach to exploit latent semantics from text and then use this information to classify questions by estimating the association between LDA topics and labels. We compare our new model with the state-of-the-art multi-label classification algorithms while using well-known keyword-based features such as bag-of-words and *n*-grams as well as semantic features such as LDA topics (Blei et al., 2003) and Word2Vec embeddings (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013).

We run our experiments on crawled data from German and Persian social question-and-answer forums. For the German data, we use GuteFrage forum[3]. For the Persian data we used 140 different forums to gather data. Although the experiments are run on these two specific languages, the proposed method is language independent and can be used for any language. We also run our experiments without text pre-processing, except stop-word removal, to show that how the method can achieve good results while keeping it

[1] Multiclass classification means a classification task with more than two classes which is normally compared with binary classification. By multi-label classification mentioned in this paper, we mean not only using more than two classes, but also having the possibility of assigning more than one target label to each instance, since the questions can be about more than one topic at the same time; e.g., a question about health and sport.

[2] To make the examples understandable for all readers, all examples are translated from German to English.

[3] http://www.gutefrage.net/.