Expert Systems With Applications 77 (2017) 11–19

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Automatically generating effective search queries directly from community question-answering questions for finding related questions[☆]

Alejandro Figueroa[a,b]

[a] Departamento de Ciencias de la Ingeniería, Facultad de Ingeniería, Universidad Andres Bello, Antonio Varas 880, Santiago, Chile
[b] Yahoo! Research Latin America, Blanco Encalada 2120, Santiago, Chile

ARTICLE INFO

Article history:
Received 24 May 2016
Revised 7 December 2016
Accepted 25 January 2017
Available online 1 February 2017

Keywords:
Real-time intelligent automation
Expert systems
Knowledge bases
Knowledge processing
Natural language processing
Community question answering
Question analysis

ABSTRACT

Community Question-Answering platforms are massive knowledge bases of questions and answers pairs produced by their members. In other to provide a vibrant service, they are compelled to provide answers to new posted questions as soon as possible. However, since their dynamic requires their own users to answer questions, there is an inherent delay between posting time and the arrival of good answers. In fact, many of these new questions might be already asked and satisfactorily answered in the past. Ergo, one of the pressing needs of these services is capitalizing on good answers given to related resolved questions across their large-scale knowledge base. To that end, current approaches have studied the effectiveness of human-generated web queries across search logs in fetching related questions and potential good answers from these community archives. However, this kind of strategy is not suitable for questions without click-through data, in particular those recently posted, limiting their capability of providing them with real-time answers.

In this paper, we propose an approach to find related questions across the cQA knowledge base, which automatically generate effective search strings directly from question titles and bodies. In so doing, we automatically construct a massive corpus of related questions on top of the relationships yielded by their click-through graph, and generated candidate queries by inspecting dependency paths across the title and body of each question afterwards. Then, we utilize this corpus for automatically annotating the retrieval power of each of these candidates. With this labelled corpus, we study the effectiveness of several learning to rank models enriched with assorted linguistically-motivated properties. Thus deducing the linguistic structure of automatically generated search strings that are effective in finding related questions. Since these models are inferred solely from each question itself, they can be used when search log data (i.e., web queries) is unavailable.

Overall, our experiments underline the effectiveness of our approach, in particular our outcomes indicate that named entity recognition is instrumental in structuring and recognizing 2–5 terms effective queries. Furthermore, we carry out experiments considering and ignoring question bodies, and we show that profiting only from question titles is more promising, but most effective queries are harder to detect. Conversely, adding question bodies makes the retrieval of past related questions noisier, but their content helps to generalize models capable of identifying more effective candidates.

1. Introduction

Nowadays, the largest community Question Answering services, including Yahoo! Answers and Stack Exchange, maintain over 100 million answered questions, making them a massive and valuable repository of knowledge in the form of questions and answers pairs (Savenkov, Lu, Dalton, & Agichtein, 2015; Zhao, Wang, Li, Liu, & Guan, 2011). In effect, there was an 889% growth in the visits to online cQA platforms from the United States between 2006 and 2008 (Choi, Kitzie, & Shah, 2012). Basically, the dynamic of these platforms consists of members posting questions at any time that receive several responses from multiple members during a limited time window (e.g., four to eight days). These questions and answers are posed in natural language and resulting in

[☆] This piece of research was carried out when the author worked for Yahoo! Research Latin America.

E-mail address: alejandro.figueroa@unab.cl

http://dx.doi.org/10.1016/j.eswa.2017.01.041

personalized content produced by means of the collective wisdom of the members of the community. However, there is an intrinsic delay between the moment a new question is published and the arrival of its first and good answers. For this reason, cQA sites have the pressing need for automatically detecting, reusing and revitalizing past resolved questions, whenever they are found to be related to new posted ones, this way enhancing user experience by prompting to community members past questions and answers in real-time. By benefiting from the knowledge stored in their repositories, cQA sites are additionally able to mitigate duplicate-asking (Figueroa, Gmez-Pantoja, & Herrera, 2016; Pal, Margatan, & Konstan, 2012).

Due to the magnitude of these knowledge bases and the fact that cQA questions are conveyed in natural language, searching for related past questions and answers is a complex task. In order to overcome these difficulties, and hence to profit from the content stored in these archives, recent studies have focused on how they can obtain effective search queries for a particular cQA question. In so doing, they have examined human-generated web queries submitted to web search engines (e.g., Google, Bing and Yahoo!) that are connected to cQA pages via user search clicks. To be more exact, they have examined their effectiveness in fetching past related questions and in identifying and ranking past best answers (Carmel, Mejer, Pinter, & Szpektor, 2014; Figueroa & Neumann, 2013; 2014). Although, this kind of strategy is promising as it allows to collect a large amount of human-generated candidate queries, it suffers from two key drawbacks:

(a) There is a linguistic gap between web queries and documents (Barr, Jones, & Regelson, 2008), i.e., web queries are dominated by nouns and proper nouns, while verbs are much more prominent across documents. Moreover, contrary to most edited web documents, the usage of uppercase is inconsistent across web queries.
(b) Web queries for many, especially recently posted questions, are not available (via click-through), thus in these cases, it is not possible to obtain web-queries for searching the cQA archives.

Thus this kind of strategy does not help in presenting related past questions and answers in real-time, that is to say at the moment of being posted on the community.

In this study, we extend our previous works by extracting search strings directly from cQA questions instead of taking advantage of web queries. However, producing effective query strings from cQA questions poses additional challenges. The number of candidate strings increases exponentially in consonance with the number of unique terms embodied in the question. Another difficulty regards determining their length, and if query terms and features should be extracted solely from question titles, or also from their bodies. As a means of overcoming both challenges, we modelled each question by amalgamating the lexicalized dependency trees of all its sentences. By traversing this graph representation, we are able to automatically generate candidate queries for searching related questions across the cQA knowledge base. Effective candidates are subsequently detected by learning to rank (L2R) models equipped with assorted linguistically oriented attributes. In summary, the contribution of this paper to this task is the following:

1. A dependency-tree based strategy to produce search strings directly from a cQA question. In so doing, we fuse the tree representation of each sentence embedded in a question into one graph, and in a Bread-first fashion, we walk through this new graph constructing several candidate queries afterwards.
2. Since this graph representation can be built on top of question titles, or on top of question titles and bodies, we study the effectiveness of candidate queries harvested not only from titles, but also from combining titles and bodies. It is worth highlighting here that some questions do not contain a body, they just provide a title.
3. The effectiveness of candidate queries in retrieving related questions is then assessed by several L2R models (e.g., ListNet, RankNet and Ranking SVM) equipped with assorted linguistically-oriented features. Note here, since we are using question content to generate candidate queries, attributes were extracted not only from the query being ranked, but also taking into account the linguistic role its terms play across the whole question. Specifically, we studied properties distilled from a host of natural language processing tools such as part-of-speech tagging, named entity recognition and sentiment analysis.

In brief, experiments on a large-scale dataset show the promise of our strategy to extract and identify effective search queries from cQA questions for discovering their related questions across the cQA knowledge base. Additionally, our outcomes gives insight into the linguistic structure of effective, especially 2–5 terms, search queries. The reminder of this paper is organized as follows. Section 2 outlines the related work, next Section 3 dissects our method, Section 4 describe our experiments and findings. Eventually, Section 5 draws some conclusions and sketches future work.

2. Related work

In recent years, a variety of studies have focused their attention on profiting from user search activity for improving the search across cQA archives (Wu et al., 2014). In general, user click patterns have proven to supply valuable relevance feedback for several tasks (Radlinski, Szummer, & Craswell, 2010). To name a few, Ji et al. (2009) extracted relevance information from clicked and non-clicked documents within aggregated search sessions. They modeled sequences of clicks as a means of learning to globally rank the relative relevance of all documents with respect to a given query. Click-through data also assisted in predicting labels for improving the quality of training material for L2R approaches (Xu, Chen, Xu, Li, & Abib, 2010).

In effect, Zhao, Wang, and Liu (2010) pioneered the extraction of paraphrase patterns from search logs distilled from general-purpose web search engines. Their investigation revealed that queries hitting the same title are likely to be paraphrases of each other. And the other way around, paraphrases can be found among titles linked with the same query. In other words, query results observing similar click patterns are likely to share similar meanings (Wen, Nie, & Zhang, 2002). This was the germ of the idea of Zhao et al. (2011), who automatically generated questions from queries for cQA. Fundamentally, their method gathered query-to-question pairs from web search logs, whereby question generation templates were acquired. For new queries, they select the most similar templates, and used them for constructing questions afterwards.

Later, Zheng, Si, Chang, and Zhu (2011) proposed a keywords to question approach that accounts for both query history and user feedback. They employed an adaptive language model to describe the process of formulating questions, and also utilized automatically induced question templates to create unseen questions. Like Zhao et al. (2011), their sorting approach is tailored to prefer candidate questions which bear a higher quality of being a well-formulated human-like question.

Incidentally, recent research have extended the use of this notion of query-to-question relatedness to other tasks. Take for instance, search queries landing on cQA pages have proven to be useful for determining temporally-anchored questions (Figueroa et al., 2016), and also for categorizing question-like web search queries (Figueroa & Neumann, 2016). The work of Carmel et al. (2014) analyzed the relationship between question titles and