Expert Systems with Applications 41 (2014) 4730–4742

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Category-specific models for ranking effective paraphrases in community Question Answering

Alejandro Figueroa [a,b], Günter Neumann [c,*]

[a] Yahoo! Research Latin America, Blanco Encalada 2120, Santiago, Chile
[b] Escuela de Ingeniería Informática, Universidad Diego Portales, Santiago, Chile
[c] DFKI GmbH, Stuhlsatzenhausweg 3, Campus D3_2, D-66123 Saarbrücken, Germany

ARTICLE INFO

Keywords:
Community-based Question Answering
Learning to rank
Question paraphrases
Question categories

ABSTRACT

Platforms for community-based Question Answering (cQA) are playing an increasing role in the synergy of information-seeking and social networks. Being able to categorize user questions is very important, since these categories are good predictors for the underlying question goal, viz. informational or subjective. Furthermore, an effective cQA platform should be capable of detecting similar past questions and relevant answers, because it is known that a high number of best answers are reusable. Therefore, question paraphrasing is not only a useful but also an essential ingredient for effective search in cQA. However, the generated paraphrases do not necessarily lead to the same answer set, and might differ in their expected quality of retrieval, for example, in their power of identifying and ranking best answers higher.

We propose a novel category-specific learning to rank approach for effectively ranking paraphrases for cQA. We describe a number of different large-scale experiments using logs from Yahoo! Search and Yahoo! Answers, and demonstrate that the subjective and objective nature of cQA questions dramatically affect the recall and ranking of past answers, when fine-grained category information is put into its place. Then, category-specific models are able to adapt well to the different degree of objectivity and subjectivity of each category, and the more specific the models are, the better the results, especially when benefiting from effective semantic and syntactic features.

1. Introduction

Web browsing has become a de facto standard for information seeking in our daily life. Search engines play a key role here in bridging the gap between the information seekers and the massive collection of web data. Understanding web queries for guiding the search effectively is a difficult task, since distinct users do not only formulate their queries with different terminologies, intents, and linguistic patterns, but they also exhibit assorted browsing behaviors. This challenging nature together with the goal of enhancing and personalizing search experience encourage developers of web search engines to investigate more intelligent algorithms for understanding and satisfying the requests of their users.

The advances made by search engines, i.e., offering more powerful services, have given their users the chance of reaching more specific and ambitious goals, and actually, have caused them to become more audacious when prompting queries. With the advent of social media, users are now more and more likely to enter complex and complete questions instead of few keywords, especially when they are targeting at precise information needs. Nonetheless, answers to these complex questions are hardly found in short text fragments within web pages or across full documents, because they require the analysis, understanding, and synthesis of several documents and world knowledge. For example, complex questions aim at current events (e.g. "*Who will win this Australian Open?*"), finding sentiments of the general public about something or someone (e.g. "*What is the coziest Starbucks in Manhattan?*"), at subjective opinions regarding particular topics (e.g., contrasting different products), which, at the moment of searching, do not necessarily exist on the web in the form of conventional web documents (e.g., "*How do you envision tablets in the year 2020?*").

Since these kinds of information needs are difficult to fulfill by means of traditional information retrieval techniques, web users take advantage of community Question Answering (cQA) services for getting help from other individuals, who know or can readily produce satisfactory precise answers, or like in many cases, can

* Corresponding author. Tel.: +49 (0) 681/85775 0.
E-mail addresses: afiguero@yahoo-inc.com, alejandro.figueroa@mail.udp.cl (A. Figueroa), neumann@dfki.de (G. Neumann).

http://dx.doi.org/10.1016/j.eswa.2014.02.004
0957-4174/

provide help by conducting opinion polls and surveys. In a nutshell, these platforms (e.g., Yahoo! Answers) are the synergy of a information-seeking and a social network (Rechavi & Rafaeli, 2013), where members can post any kind of question, either simple, complex or detailed, or questions about opinions. In a similar way, posted questions can receive several responses from multiple members, which can not only be supplementary or complementary to each other, but also reflect different sentiments and aspects. When taking part in this network, members additionally provide social capital: rate the answers' quality (via positive/negative votes, thumbs-up/thumbs-down, etc.) and post comments. In summary, the information-seeking perspective of a cQA provides arbitrary members with content, motivating them to take part in asking and responding questions, especially when the experience of social interactions is positive; while the social network perspective causes members to engage in social activities (Rechavi & Rafaeli, 2013).

Through these social interactions, members share their knowledge so as to construct a valuable, rapidly growing and massive archive of questions and answers. Notably, one attractive part of these repositories yields a large quantity of diverse word-of-mouth tips (e.g., "*How to get rid of eye strain?*" and "*Teach my cat to use the toilet*"), insights and solutions to many common questions and daily problems that people may face (e.g., "*Removing cooked on grease from pans?*"). CQA services are usually organized in categories, which are selected by members when submitting new questions. These categories are later utilized for locating contents on topics of interest. In a category to which only social activity is attached, fewer members respond to questions, resulting in a small average number of answers per question causing a low rate of user satisfaction. This is in contrast to a category where social activities and information-seeking activities co-occur: the amount of answers is average or above (Rechavi & Rafaeli, 2013).

Recent studies have unveiled that this synergy is also projected into the relationship between categories and question intents (Harper, Moy, & Konstan, 2009). More precisely, they revealed that categories are good predictors of question goals. Although the number of types of intents varies from one approach to another (Harper et al., 2009; Chen, Zhang, & Mark, 2012; Liu et al., 2008), most studies agree on two main types of ends (Harper et al., 2009): informational (i.e., objective or information-seeking) and subjective (i.e., social, opinions or conversational). The following Yahoo! Answers categories exemplify this contrast: "*Polls* & *Surveys*" and "*Religion* & *Spirituality*" embrace almost solely subjective questions, while this kind of intent covers 70% of "*Singles* & *Dating*", 27.27% of "*Health*" and 16.17% of "*Science* & *Mathematics*", only.

Due to several reasons (e.g., system saturation (Liu, Liu, & Yang, 2010) or bad question formulations (Yang et al., 2011)), it has been observed that about 15% of all incoming questions in English go unresolved, poorly answered or never satisfactorily resolved in Yahoo! Answers (Shtok, Dror, Maarek, & Szpektor, 2012). Thus, an effective cQA platform should be capable of detecting similar past questions and relevant answers. Practical solutions would involve asking members for rephrasing a question (Yang et al., 2011), suggesting alternative questions (Cao, Duan, Lin, Yu, & Hon, 2008), or offering past answers, since at least 78% of best answers are reusable (Liu et al., 2008; Bian, Liu, Agichtein, & Zha, 2008). However, the lexical gap between past and new questions is the main obstacle to reuse these best answers (e.g., "*Remove pimples?*" and "*How to get rid of acne*"), thus some strategies have tried to combine social and textual (e.g., semantic and syntactic) features as a means of tackling lexical mismatches, cf. (Bian et al., 2008; Blooma, Chua, & Goh, 2011; Blooma & Kurian, 2012; Wanga, Ming, & Seng Chua, 2009).

A promising approach to improve the effectiveness of search in cQA by means of automatic identification of question paraphrases has been proposed by Lin (2008). The core idea is to use the user generated questions of a cQA along with search engine query logs to automatically formulate effective questions or paraphrases in order to improve search in cQA. Zhao, Wang, Li, Liu, and Guan (2011) have further elaborated this idea into the direction of generation of new questions from queries. A major advantage of such a query-to-question expansion approach for cQA is that it can help to retrieve more related results from cQA archives and hence, can improve the recall.

The automatic generation of paraphrases is a useful means to improve the search for finding best answers in cQA. But the generated paraphrases (although they might "mean" the same) do not necessarily lead to the same answer set, and hence, it might be that they differ in the expected retrieval quality of identifying and ranking best answers high. Thus, it makes sense to rank the generated paraphrases, so as to provide evidence according to recall and the position of the best answer of a paraphrase, i.e., its mean reciprocal rank (MRR). This is the major motivation behind our approach of computing *effective paraphrases*. An effective paraphrase is a reformulation of the posted question that narrows the lexical gap the best, i.e., an alternative formulation of a user question that can retrieve more past answers to the new question, or can rank past answers higher within the fetched set (see examples in Table 1). In Figueroa and Neumann (2013), we presented a first learning to rank approach based on general-purpose models that is able to determine effective question paraphrases by exploiting search engine query logs and connections to cQA, however, without taking into account question category-specific information. This work extends our earlier work on several innovative aspects:

1. We empirically demonstrate that the subjective and objective nature of cQA questions dramatically affect the recall and ranking of past answers. Since categories and question intents are closely related, we construct category-specific learning to rank models (i.e., SVMRank) for paraphrase ranking, showing that the retrieval and ranking from social media can be improved when category information is put in place.
2. Since we carry out experiments on a large data-set of automatically annotated question paraphrases harvested from Yahoo! Answers and Yahoo! Search logs, we are able to conduct experiments not only on broad, but also on fine-grained question categories. Specifically, we consider the three levels of granularity supplied by the Yahoo! Answers question taxonomy.
3. In addition, we study the impact on our category-specific models of Natural Language Processing (NLP) information in two ways: (a) we show that enriching question categorization with Wh-question typification enhances the performance; and (b) our models are built largely on the basis of effective semantic and syntactic properties.

The core idea of our method is as follows (see Fig. 1). Given a huge collection of query logs from Yahoo! Search, we extract all pairs consisting of a query and a title, where at least one user click links the query with a title from Yahoo! Answers. Note that the title is the user entered question of the answer web page together with the category selected by the user.[1] We further cluster these pairs into groups, where each group consists of all query-title pairs with same title and category. We interpret each group (including the title) as a set of paraphrases of the same underlying question and category. Note that each title is associated with an answer web page, and so also its paraphrases. This way we obtain a huge collection of 32 million answer web pages and their associated

[1] Actually, it is mandatory, that if a user enters a question to Yahoo! Answers, he or she also selects a category from a given set.