Information Sciences 330 (2016) 403–423

Contents lists available at ScienceDirect

Information Sciences

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

Exploring heterogeneous features for query-focused summarization of categorized community answers

Wei Wei [a,d], ZhaoYan Ming [c], Liqiang Nie [b], Guohui Li [a,*], Jianjun Li [a], Feida Zhu [d], Tianfeng Shang [d], Changyin Luo [a]

[a] School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan, China
[b] Department of Computer Science, National University of Singapore, Singapore
[c] Department of Computer Science, Digipen Institute of Technology, Singapore
[d] School of Information Systems, Singapore Management University of Singapore, Singapore

ARTICLE INFO

Article history:
Received 19 January 2015
Revised 28 September 2015
Accepted 15 October 2015
Available online 23 October 2015

Keywords:
Summarization
Community-based question answering
Graph-based ranking

ABSTRACT

Community-based question answering (cQA) is a popular type of *online* knowledge-sharing web service where users ask questions and obtain answers contributed by others. To enhance knowledge sharing, cQA also provides users with a retrieval function to access the historical question-answer pairs (QAs). However, it is still ineffective in that the retrieval result is typically a ranking list of potentially relevant QAs, rather than a succinct and informative answer. To alleviate the problem, this paper proposes a three-level scheme, which aims to generate a query-focused *summary-style* answer in terms of two factors, i.e., *novelty* and *redundancy*. Specifically, we first retrieve a set of QAs to the given query, and then develop a smoothed Naive Bayes model to identify the topics of answers, by exploiting their associated category information. Next, to compute the global ranking scores of answers, we first propose a parameterized graph-based method to model a Markov random walk on a graph that is parameterized by the heterogeneous features of answers, and then combine the ranking scores with the relevance scores of answers. Based on the computed global ranking scores, we utilize two different strategies to construct top-K candidate answer set, and finally solve a constrained optimization problem on the sentence set of top-K answers to generate a summary towards a user's query. Experiments on real-world data demonstrate the effectiveness of our proposed approach as compared to the baselines.

© 2015 Elsevier Inc. All rights reserved.

1. Introduction

Recently, the trend to advance the functionality of search engine to an expressive semantic level has attracted extensive attentions [6,28]. Under such circumstance, an extremely active knowledge-sharing web service, community-based question answering (cQA), has emerged as one of the most widely-adopted web services in acquiring online information. It enables users to post questions in natural language and directly returns actual answers provided by other participators. Over the years, cQA has successfully accumulated millions of question–answer pairs, i.e., *QAs*, which carry rich human intelligence in various domains, e.g., arts, business or health, and are effectively organized by cQAs for re-use [14,28,35] in providing users with a retrieval function.

* Corresponding author. fax.: +86 27 87543104.
E-mail address: guohuili@hust.edu.cn (G. Li).

http://dx.doi.org/10.1016/j.ins.2015.10.024
0020-0255/© 2015 Elsevier Inc. All rights reserved.

Table 1
Examples of cQA question-answer pairs from Yahoo! Answers (**Q** denotes *question*; **BA** indicates *best answer*; **OA** refers to *other answer*).

Category	Type	Content
Health\Diseases&Conditions\Cancer	$\mathbf{Q}_1$	How to prevent getting cancer if you know it runs in your family?
	$\mathbf{BA}_1$	You are asking a general question regarding ... Cancer is rarely hereditary and cancers like leukemia are not one of them.
Health\Diseases&Conditions\Cancer	$\mathbf{Q}_2$	Can Vitamin D help prevent/cure ovarian cancer?
	$\mathbf{BA}_2$	It can prevent :) but not cure! :(... http://www.webmd.com/content/article/116/112304
Sports\Handball	$\mathbf{Q}_3$	Is it true that playing Handball regularly can help prevent Prostate Cancer?
	$\mathbf{BA}_3$	no, sounds made up
	$\mathbf{OA}_3$	Its just having exercise, not particularly handball http://www.canceriscurablenow.com
Food&Drink\Other-Food&Drink	$\mathbf{Q}_4$	What are the best foods to eat to prevent cancer?
	$\mathbf{BA}_4$	whole grains, dark fruit, dark chocolate, leafy greens, green tea, beans
Education&Reference\Other-Education	$\mathbf{Q}_5$	How can i study for english?
	$\mathbf{OA}_5$	I would say read newspaper to improve your vocab.
Society&Culture\Languages	$\mathbf{Q}_6$	How can I learn english?
	$\mathbf{OA}_6$	... Try to find topics that INTEREST you, read online newspapers (Times, Guardian, Independent...),
Health\Diseases&Conditions\Cancer	$\mathbf{Q}_7$	How do I prevent cancer?
	$\mathbf{OA}_7$	we all have a cancer gene. we just don't know what sets it off.

In spite of the great progress reported, the retrieval function of cQA is still imperfect as its results are usually a series of relevant QAs. It is apparently infeasible and tedious for users to 'click-and-view' each of the returned QAs. For example, Yahoo! Answer returns more than 182K QAs for question "*How can I lose weight?*". To alleviate the problem, several previous work [7,11,32] are proposed to improve the retrieval accuracy. However, most of the existing methods would suffer from several problems, such as the returned questions might contain *incomplete*, *redundant* or *irrelevant* answers (or even no answer), which will be elaborated with several examples given in Table 1.

- *Incompleteness*: The relevant answers to a query might come from different QAs, since a query typically involves different aspects. Hence, the use of any single answer (even the best answer) is not enough to fully answer a query [6,14,28]. For example, question Q_1 can be answered from different aspects, e.g., "Health", "Sports" and "Food", and it seems that the use of answers BA_2, OA_3 and BA_4 is more complete than the best answer BA_1 for answering Q_1.
- *Redundancy*: Questions from different categories are phrased differently however have semantically duplicated answers such as questions Q_5 and Q_6, their answers (i.e., OA_5 and OA_6) indicate the same meaning, namely, reading the newspaper is a good way in learning English.
- *Irrelevance*: Relevant questions might have irrelevant answers that usually cannot provide any useful information. For instance, answer OA_7 is an useless answer for question Q_7.

To help users digest numerous retrieved QAs, this paper formulates it as a *query*-focused answer summarization (QAS) problem.[1] For users, immediately providing an informative answer is benefit to avoid a lag time of waiting for response, and determine whether need to refine the query to focus on a more specific aspect of that question.

However, many traditional *query*-focused multi-document summarization (QMDS) methods cannot be applied to the QAS problem directly, such as cluster-based [30,39], graph-based [2,19,34], manifold-ranking based [31,37] and graphical model-based [27]. These methods usually assume that the relevant documents [3] contain sufficient statistical signals for comparison, whereas it might not be held in cQA domain, since cQA answers are typically short and contain many category-specific frequent words that are useless for comparing answers within a category. Therefore, the problem of query-focused answer summarization is more challenging.

Indeed, there exist several attempts on the QAS problem. For example: (i)clustering-based methods like [14], which clusters answers based on their own information for summarization; (ii) concept-based methods, e.g., [28], which employs the concept functions to capture the semantic overlaps of answers for summarization; and (iii) condition random field (CRF) based methods, such as [6], which exploits four contextual and textual features for summarization of multi-sentence questions. However, it remains largely unexplored to apply the cQA three-level structure features (as shown in Fig. 1), i.e., *category-level*, *question-level* and *answer-level*, to sufficiently model the query information (i.e., "*query relevance*" issue) for answer summarization especially category-level features. Without loss of generality, we take *category-level* information as example, intuitively the more relevant answers a category contains, the more likely that category is relevant to the query, which in turn favors answers in such category to be assigned a high relevance score for summarization. Moreover, [7] has also proven that the category-specific frequent words (words that appear relatively frequently in a category, e.g., "NBA" for category "Sport\Basketball\NBA") are more important in computing the relevance of answers to the query across categories, but less important in measuring the relevance of the answers to the query within a same category.

[1] Here, we mainly focus on open questions ([14], which are those asking for facts or methods), as most questions (up to 18–67%, [14]) belong to this category. Note that the answers with a strict sequence are usually expected to be answered with a linking resource because the summarization will lead to a useless summary due to missing the original order. Hence, we tend to find linking resources for such answers.

Download English Version:

https://daneshyari.com/en/article/392861

Download Persian Version:

https://daneshyari.com/article/392861

Daneshyari.com