Expert Systems With Applications 50 (2016) 89–99

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Search clicks analysis for discovering temporally anchored questions in community Question Answering

Alejandro Figueroa[a,b,c,*], Carlos Gómez-Pantoja[a], Ignacio Herrera[b,c]

[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
[c] Escuela de Ingeniería Informática y Telecomunicaciones, Universidad Diego Portales, Santiago, Chile

ARTICLE INFO

Keywords:
Community Question-Answering
Question analysis
Temporality
Question classification
User search activity
User generated content

ABSTRACT

Nowadays, community Question-Answering (cQA) sites are massive repositories for user-generated content, where members prompt questions expecting satisfactory answers from other members. However, in this dynamic, there is an intrinsic delay between the moment questions are posted to the arrival of acceptable responses. Therefore, cQA platforms have the pressing need for promoting unresolved questions to potential answerers and for taking advantage of resolved questions contained in their archives, whenever possible.

This paper studies cQA services from the viewpoint of the time frame where their questions attract the interest of their community members. By drawing a parallel with temporal patterns of user interests in web search activity, we are able to define three main types of temporally anchored questions: trend or bursty, periodic and permanent. Then, by analyzing user click distributions to Yahoo! Answers pages across Yahoo! Search logs, we automatically acquired a set of 35,000 cQA questions labeled with one of these three temporal anchors. Accordingly, we show the practicality of this approach by means of human assessments; and by using this automatically acquired corpus for studying several classification models.

Essentially, the proposed method was found to correlate well with these human judgements, and proven to be effective in building systems that automatically identify the temporal anchor of unseen cQA questions. In substance, our outcomes indicate that some contexts are strongly related to a particular temporal anchor. We believe that these anchors will contribute to the discrimination of resolved questions that are capable of being revitalized, as well as to foster the opportune participation in questions that generate enthusiasm only for a short time.

1. Introduction

Community question-answering (cQA) services offer a flexible on-line framework, whereby their members seek to fulfill their information needs by posting questions that are expected to be answered by other users of the platform. In order to connect open questions with potential answerers, askers are required to assign their questions in the most pertinent category available in a taxonomy provided by the site (e.g., sports, health and travel). This is a way of facilitating the process of linking new questions with other members that evince enthusiasm for the same topic or have an expertise in the matter. In result, members not only can answer open questions prompted by other participants of the community, but also they are able to browse and rate questions or answers embodied in the archives of the system.

A decisive advantage over traditional web search is that cQA members can fetch answers to questions that are not readily found across web documents. Take for instance, members produce answers that require the synthesis of several sources and/or highly specific knowledge (e.g., "*should UK leave the EU. How will UK pay its crippling pension bill and combat its aging population?*"). Oftentimes, these answers must be tailored to fit special needs or unique situations that users are placed in (e.g., "*Can I TAG for UC Merced in Spring term 2016 and TAG for another UC for Fall 2016?*"), or they just aim simply at expressing personal opinions or giving pieces of advice on particular subjects (e.g., "*friend totally ignoring me! don't know what to do please help!?*"). In practical terms, asking

* Corresponding author at: Departamento de Ciencias de la Ingeniería, Facultad de Ingeniería, Universidad Andres Bello, Antonio Varas 880, Santiago, Chile. Tel.: +56 2 27703795.
E-mail addresses: afiguero@yahoo-inc.com, alejandro.figueroa@unab.cl, alejandro.figueroa@mail.udp.cl (A. Figueroa), carlos.gomez.pantoja@unab.cl (C. Gómez-Pantoja), ignacio.herrera@mail.udp.cl (I. Herrera).

http://dx.doi.org/10.1016/j.eswa.2015.12.016
0957-4174/

other individuals is sometimes more convenient or appropriate in meeting the needs of the users than producing their own answers by visiting and reading a set of web documents.

All these features make cQA services, such as Stack Exchange and Yahoo! Answers, very popular. However, in this dynamic, an inherent delay time exists between the moment askers post new questions to the arrival of an acceptable response. For this reason, cQA platforms capitalize on traditional information retrieval approaches such that community members can profit from a search box for navigating their archives until they find an old question (if any) pertaining to their current need. This sort of strategy assists cQA platforms in re-using and revitalizing past questions and answers indexed in their archives. In like manner, web search engines benefit from this facility for enhancing user experience, whenever they detect that question-like search queries are submitted. As a matter of fact, web engines return hits found by browsing these archives at the top positions of their rank, displaying not only links to strongly related questions, but also producing their snippets from the best answers contained therein. However, the success of this type of technique depends on the usefulness of the resolved questions in the archives with respect to the current need of the web user or community member.

In cQA archives, questions pertaining to trend topics are likely to generate enthusiasm among its participants only for a short time (e.g, "*what sports bars are showing the Mayweather fight tomorrow in NC??*"), and accordingly, potential answerers must be encouraged to become involved during this brief period of enthusiasm in order to make the community more vibrant. When this sort of question is resolved or its peak passes, it normally has low archival value, unlike other resolved questions, which can be of interest at any time, and thus have permanent archival value (e.g, "*Is beer good for lactation?*"), or unlike questions which their focus of attention, and hence their archival value, varies seasonally (e.g, "*What CAN you eat on Yom Kippur?*"). Thus, identifying the temporality of questions can help to improve the retrieval of resolved questions pertinent to an information need. Furthermore, it can assist interface designers in enhancing user experience by devising temporality-aware displays, especially considering the limited space on mobile devices. Take for instance, questions on trend topics which catch much attention at one particular short period of time. This kind of question could pass unnoticed, and hence unsatisfactorily answered, if they are not brought to the attention of the community members in a timely manner. On the other hand, periodic questions can be anticipated in a sort of "check this, before asking" fashion, when a new season of heightened interest is approaching. In this way, temporality analysis on cQA services would be able to mitigate duplicate question-asking as by-product. Simply put, our work contributes to this area of research in the following aspects:

1. Inspired by recent studies into the temporal dynamics of web search activity (i.e., clicks and queries), we draw a parallel between search queries and cQA questions. This parallel led to the definition of a taxonomy comprising three classes of temporal anchors: trend and periodic as well as permanent questions.
2. By analyzing temporal patterns of user clicks to Yahoo! Answers pages during two years, we are able to automatically build a corpus of ca. 35,000 questions, each of them annotated with one of these three temporal anchors. Note that the two-years extent of our query log helps to identify cQA questions on topics that grasps the attention, at most, annually.
3. We validate our approach in two ways: by asking two human assessors to judge these automatic anchors; and by testing the predictions on unseen data of several supervised models trained with this material. More precisely, we executed experiments regarding seven state-of-the-art multi-class learners (e.g., SVM and MaxEnt) equipped with five different combinations of features.

In a nutshell, human judgments and the performance achieved on tagging unseen questions confirm the effectiveness of our method. In particular, the outcome reveals that these types of temporally anchored questions are conveyed in similar contexts, and ergo these can be automatically learned and recognized. The remainder of this paper is organized as follows. Section 2 outlines the related work, then Section 4 dissects our automatic corpus construction strategy, Section 5 describe our experiments and findings. Eventually, Section 6 draws some conclusions and renders future work.

2. Related work

To the best of our knowledge, our work pioneers the idea of profiting from user search activity, when automatically detecting the temporal anchor of cQA questions. Broadly speaking, our study is at a crossroads for three research topics: temporal analysis of user-clicks in web search, the interpretation of cQA questions and web user clicks for cQA.

2.1. Question analysis in cQA

By and large, questions posted on cQA sites are sharply different from their counterparts prompted at traditional question-answering systems. For example, cQA questions comprise a title and a body. In this format, the title typically bears, if not all, most of the fundamental aspects of the question (e.g., "*What happens if the Dow Jones falls to 5,000 points?*"), whereas the body gives its specifics (e.g., "*Is that considered a stock market crash? Will there be a Great Depression? Should we start preparing now?*"). Another important difference is that question titles are likely to be expressed in multiple-sentences, and the length of question bodies can vary from being long-winded to be left blank. Essentially, there are two central research branches in question analysis: question classification and the identification of similar questions. Our study falls under the former.

Early works categorized cQA questions, targeted at one answer, by extracting the most important sentence, while removing all noisy or unnecessary elements (Tamura, Takamura, & Okumura, 2005; 2006). From another standpoint, Li, Liu, Ram, Garcia, and Agichtein (2008) proposed a cost-efficient solution to check whether questions are objective or subjective built on top of an SVM trained with trigrams attributes. Similarly, Adar, Weld, Bershad, and Gribble (2007) utilized Bayesian Networks for labeling questions as informational or conversational. In the same vein, the study of Amiri, Zha, and Chua (2013) devised subjectivity classifiers for recognizing questions with an implicit opinionated intent. Along the same lines, Chen, Zhang, and Levene (2013a) designed language models combined with textual and metadata properties, to categorize the user intent of questions into objective, subjective, and social. By the same token, Chen, Zhang, and Mark (2012) used co-training for building an SVM approach that groups questions into these three categories.

Incidentally, Yang et al. (2011) revealed some patterns observed by unresolved questions, making it possible, in some cases, to encourage askers to rephrase their questions. Their analysis showed that short and long questions are quickly answered, while medium-length questions have a higher probability of remaining unresolved. In the same vein, (Chua & Banerjee, 2015) developed a conceptual framework aimed at explaining why some questions draw answers while others stay pending. As a result, they suggested that the probability of getting answers relies on the