

Question Temporality: Identification and Uses

Aditya Pal, James Margatan, Joseph A. Konstan
Department of Computer Science & Engineering
University of Minnesota, Minneapolis, MN 55455, USA
{apal,james,konstan}@cs.umn.edu

ABSTRACT

In this paper, we introduce the concept of *question temporality* as a measure of the usefulness of the answers provided on the questions asked in the Question Answering sites (QA). We define question temporality based on when the answers provided on the questions would expire. We use classification methods to show that the question temporality can be assessed automatically. Our regression analysis highlights features that predict temporality of the questions. Our research can be instructive for interface designers to design temporality-aware interfaces and influence selection of questions and answers for display.

Author Keywords

Question Temporality, Question Answering Community, Social Q&A, Machine Classification

ACM Classification Keywords

H.3.1 Content Analysis and Indexing: General

General Terms

Experimentation; Human Factors

INTRODUCTION

In this paper, we introduce the concept of *question temporality* as a measure of the usefulness of the answers provided on the questions asked in the Question Answering sites (QA) and propose Machine Learning methods to automatically identify question temporality. We define a question's temporality as how soon the answers to that question are likely to expire or become obsolete. We illustrate this point using an example.

EXAMPLE 1. Consider the question-answer exchange presented in Figure 1, where a user is inquiring about the day and the date of occurrence of Thanksgiving. Even though the question asker has not specifically mentioned the year, it can be inferred that the asker is interested in knowing the date of Thanksgiving in the year 2007 based on the date the question was asked. The answer that was judged as the best answer

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or to publish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

CSCW'12, February 11–15, 2012, Seattle, Washington, USA.

Copyright 2012 ACM 978-1-4503-1086-4/12/02...\$10.00.



Figure 1. A question-answer thread taken from Yahoo Answers.

rightly points out that Thanksgiving would occur on November 22nd, which was correct in 2007. Even though the answer is labeled as the best answer, it is no longer correct in the year 2011. This is because the date of Thanksgiving changes every year. The question is temporal and its answers thus become outdated after some time, a few months in this case. We can envision scenarios in which the outdated information can lead to serious negative consequences when a person is acting on the advice provided on a question like - *whether a particular company's stock is overvalued, where to travel to see the next eclipse, or what hours a particular store is open.*

Applications

Identification of question temporality can improve the use of question-answering sites in several ways. First, it can help sites and search engines in weeding out outdated content. A major purpose of QA sites is to provide information that would not otherwise be available online. While internal mechanisms such as "best answer" awards can help identify whether the information provided was good at the time, temporality can help determine whether older answers are likely to have retained their value months or years later¹. An understanding of how long a question's answers are likely to be valid could also enable different archiving strategies of question-answer threads [4].

Second, QA sites could better handle the phenomenon of duplicate question-asking if they were temporality-aware. When a user presents a question that has been asked and answered before, many QA sites treat this as a user error and

¹It would also be useful to be able to identify the temporality of each answer to a question; while such work would be an interesting follow-on study, it is beyond the scope of this work.

proceed to present prior answers to the user. For many questions, this behavior may be correct, but for questions with short-lived answers, it provides an annoying and unwanted experience. Consider the real-world example of questions regarding immigration and visa rules. Old answers are almost always unsatisfactory and misleading due to frequent changes in visa rules. QA communities can benefit by automatically flagging old answers, requesting updated information, or providing tools to help the community do so.

Third, temporal analysis of questions could result in more useful lists of open and answered questions. Recent work by Kulkarni et al. [5] suggests that some web queries are also strongly influenced by time. As a result it is more appropriate to alter search result rankings based on time in order to present higher quality ranked lists. We draw a parallel in the QA domain where question presentation can be altered based on the temporal nature of the questions. For e.g. a question whose answer expires everyday (such as *How is the weather in Seattle?*) can be displayed on the main page more frequently than questions which have a longer life. Answers on some questions expire faster than others, e.g., *How is the weather in Seattle?* vs *When is Thanksgiving?* We don't propose to know the right strategy for factoring temporality into QA displays, but only by first establishing temporality can these interfaces be explored.

Finally, identification of question temporality can provide deeper insights into how questions can be altered to increase their longevity. Improving the usefulness of questions answered is a major goal of ours. Consider again the Thanksgiving example. The way this question has been posed, the answers are likely to expire within a year. If we modify the question slightly to read: *What day/date is Thanksgiving in the year 2007?*, correct answers never expire. The example highlights that a question's temporality can be altered by binding it with specific time reference; resulting in greater archivability (and potentially greater usefulness in documenting history). E.g. *Who won last Thursday on American Idol?*, can be improved automatically by adding date binding to read *Who won May 10, 2011 in American Idol?* We're not proposing simplistically adding time bindings to all questions (many would produce little value and simply be annoying to users), but for questions with potential long-term value, the approach seems promising.

Our Approach

We propose 4 temporal categories based on how soon the answers are likely to expire on any given question. These temporal categories range from short-duration (questions whose answers can expire within weeks) to permanent (questions whose answers never expire). To separate questions where temporality doesn't apply, we also propose an *other* category. We use the 5 category taxonomy to label a set of 100 questions taken randomly from Yahoo Answers. The 5-category taxonomy allows us the flexibility to group categories together depending on the objective of our classification. We then identified and extracted the key features that we thought would help in identifying their temporality. We use the extracted features to run Machine Learning models to show that

the temporality of the questions can be estimated automatically with promising precision. We further run regression analysis to identify the features that are most effective in identifying question temporality.

QUESTION TEMPORALITY

We define *question temporality* as a measure of how long the answers provided on a question are expected to be valuable. The Thanksgiving example shows that the answers on it would be valid for up to a year, depending on when the question was asked. On the other hand consider the question: *Which is the largest country in the world?*. The answers for this question can be expected to be valid for several years, though likely not permanently. In contrast, answers to a question that asks about the weather of a certain place, e.g., *Is it raining in Seattle?*, are expected to expire within a few hours. The examples provided here are representative of the factual questions on actual QA sites.

Category 1: Permanent

The answers on these questions do not expire. Generally these questions directly or indirectly specify a specific time frame or they are permanent factual questions. A factual question can be of the kind *What is the chemical symbol for Gold?* In a time-specified frame question, the time frame can simply be a year, or a past dated event. E.g. *Who was the US president in 1957?*, *When did World War II happen?* We can modify our Thanksgiving example such that it can belong to this category by adding a time reference: *What day/date is Thanksgiving in year 2007?*

Category 2: Long-Duration

Long-duration questions are those whose answers do not change often. The answers are expected to change every few years or decades. Questions that fit this category are: *Which is the largest country in the world?*, *Which is the oldest mammal found by archaeologists?*, *How many planets are there in the solar system?*

Category 3: Medium-Duration

Answers to medium-duration questions are expected to become invalid within a few months to a few years. The Thanksgiving question presented above fits this category. Other examples of questions in this category are: *Who's the mayor of Washington DC?*, *What is the fastest CPU in the market to buy?*

Category 4: Short-Duration

Short-duration questions are those whose answers are expected to expire in less than few months. Some examples of questions that belong to this category are: *How is the weather in Seattle?*, *What is the most recent House episode?*

Additionally, we introduce *Category 0: Other* to cover all questions where temporality does not apply.

Category 0: Other

Questions that do not fit any other category such as non-factual questions, opinion questions, advice questions are put in this category. Questions such as *Do you like Brad Pitt?*,

What is better Coke or Pepsi? belong to this category. Additionally, future directed questions that are open for speculation belong to this category, e.g., *Who will be nominated as the best actress in Oscar 2012?*

With only 100 questions, most of our analysis could not make full use of our category set. Based on our goal of identifying questions where the answers are not likely to remain useful for long, we make a binary distinction between Categories 3 and 4 (short- and medium-duration temporal questions) and all the other categories. Most of our results report on our efforts to identify these Category 3 and 4 (strongly temporal) questions.

EXPERIMENTAL DESIGN

Dataset

We sampled 100 questions randomly from the main page of Yahoo Answers over a period of one week. The sampled questions contained 10.59 words on average and a total of 448 words. The word frequency distribution follows a power law [1], indicating that the selected dataset captures the characteristics that occur naturally in online social systems.

Survey Design

The first screen of the survey provided details of the 5 categories along with example questions as described above. Subsequent screens would show 100 questions² one by one and provide check box to select only of the above 5 categories. A user could also select a “not sure” button and move to the next question. The order in which the 100 questions were presented was also randomized.

Survey Takers

We asked Computer Science graduate and undergraduate students to take the survey. Overall 8 students took the survey. The inter-rater agreement between the raters was 0.22 (Fleiss kappa, error = 0.009 $p < 0.01$ with 95% CI). The agreement between the raters was not accidental and is considered to be satisfactory agreement. Moreover, when computed based on our binary classification (strongly temporal vs. not), inter-rater agreement is 0.88 which is very strong agreement (and which illustrates that most disagreement was over which category within high-temporal or not should be used).

Survey Results

We need a measure to combine the eight user ratings to compute the categories of the questions. The most commonly used strategy of taking the mean does not work since the ratings are categorical. As a result we consider the majority rating strategy: As per this strategy, we assign a category to the question based upon the majority votes (out of 8). We break the ties randomly (5/100 had 2-way ties and only 1 question had a 3-way tie). Figure 2 shows the frequency distribution of question being assigned to each of the 5 categories. It shows that 47 question exhibit temporality and among those 47 questions, 17 are short-duration and medium-duration questions. It is these 17 strongly temporal questions that we attempt to identify automatically.

²We only showed the questions and not the answers provided on those questions.

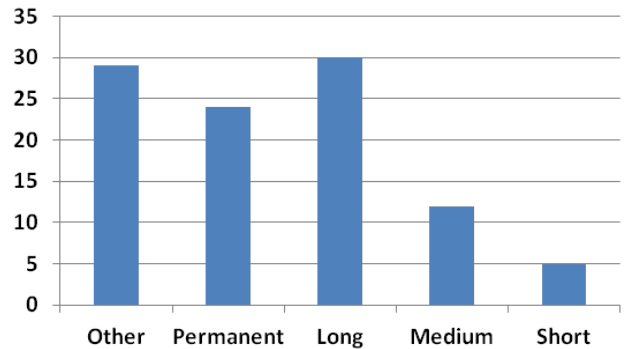


Figure 2. Frequency distribution of the questions per category.

CLASSIFICATION METHODOLOGY

We combine short- and medium-duration temporal questions as class 1 (17 questions) and all other questions as class 0 (83 questions). Note that the categories are combined as performing multi-way classification over a small dataset would not be effective. This binary separation enables us in answering the question: *Can we automatically separate questions which would likely expire “soon” from those more likely to remain valid for years or longer?*

Question Features

We consider a question as a vector of words (after removing stop words and special symbols like ?, !, *, etc). Then we computed the tf/idf score for the question vector, which were used as question features. Additionally, we considered several other features such as *Does the question reference a specific time (e.g. date, month)?, Is the referenced time fixed or relative (ago, after, before, today, tomorrow)?, Does it reference a date in past?, Does it reference a date in future?, Is the tense of question past tense?* These features along with the tf/idf scores were used to build a feature vector per question, which is then used for question classification.

Classification Method

We considered several state-of-art classification algorithms (such as Support Vector Machines, Decision Trees, Ada Boost, Naive Bayes, Gaussian discriminant Analysis). Ada Boost [3] which is an ensemble algorithm consistently performed better than other algorithms. We ran our experiments using 10-fold cross validation and report the aggregate accuracy over the 10 folds.

RESULTS

Identifying Strongly Temporal Questions

Table 1 shows the performance results of the model in identifying short- and medium-duration questions (class 1) from the rest of the questions. It shows that we can automatically

| | Precision | Recall | F1 |
|----------------------------|-----------|--------|------|
| Class 1 (Category 3, 4) | 0.70 | 0.41 | 0.52 |
| Class 0 (Category 0, 1, 2) | 0.89 | 0.96 | 0.92 |

Table 1. Performance of Ada Boost model in distinguishing strongly temporal questions from other types.

identify 41% of class 1 questions (7/17). It also signifies that of all the questions called out by the model as temporal, 70% are short- or medium-duration temporal. Overall the result shows promise for automatic detection of question temporality.

Identifying Features of Strongly Temporal Questions

We run ridge regression [6] using binary class as the output variable. The R^2 of the fitted distribution is 0.83 which indicates that the extracted features provide a good linear fit over the ratings ($error = 0.2$). Table 2 shows features with the highest and lowest weights found using ridge regression. The highest weight features are indicators of question temporality and alternately the lowest weight features are indicators of the non-temporality of the questions. We can see that words such as what, who, current, etc or presence of a *relative time* indicates that a question is strongly temporal. Indeed we found short-duration temporal questions such as *What is the current weather/temperature in San Diego, CA?*, *How much has the price of gas gone up in the past 5 months?* containing these features.

On the other hand questions from our dataset such as *When was Super Bowl XLV played in 2011?*, *Who was the first US president to visit the Soviet Union?* indicate that indeed the features such as when, was, fixed time, etc are indicator of permanence of the questions. The words such as *likely* are indicator of opinion questions (e.g. *who is likely to be Vikings qb this upcoming NFL season?*) and the questions containing it are considered non-temporal (category 0).

These features and their corresponding weights can be used to automatically detect question temporality class. The regression analysis can be used along with classification model to detect temporal questions.

| Type | Features |
|---------|---|
| Class 1 | what (0.12), who (0.07), current (0.08), recent (0.11), today (0.11), does (0.06), during (0.06), <i>Relative time</i> (0.11) |
| Class 0 | when (-0.09), was (-0.09), worst (-0.06), which (-0.06), this (-0.04), likely (-0.06), <i>Fixed time</i> (-0.06), <i>Ordinal Number</i> (-0.07) |

Table 2. Features (and their weights) that are most indicative of temporality or non-temporality of the questions.

RELATED WORK

Content temporality has been extensively studied in the context of web queries. Zhang et al. [7] analyzed query volume at Yahoo to detect time sensitive queries, *based on which queries are accompanied by year*, and proposed a ranking method for web results for the time sensitive queries. Dakka et al. [2] presented a framework for handling time sensitive queries and identifying the important time intervals that might be interesting for a given query. They ran their experiments on news articles to show that temporality aspect of a document is an important aspect during ranking of web results.

Recently, Kulkarni et al. [5] considered query popularity over time to re-rank web results and understand the user intent while keeping in perspective the temporally evolving web pages. Our work compliments prior work in several ways. We posit that question temporality can be useful in ranking of QA data. We also show how questions can be slightly modified to change the question’s temporality class. We demonstrate that question temporality can be automatically detected from simple features extracted from question text.

CONCLUSION

In this paper, we propose a novel concept of question temporality in QA communities. We argue that the detection of question temporality would benefit search engines and allows QA providers to explore interesting interface design choices towards providing better user experience. It also leads towards developing deeper insights into how questions can be altered to increase their longevity.

Our results show that question temporality can be satisfactorily detected automatically using question vocabulary and other simple features. Our results indicate that features such as *relative time* (e.g. today, tomorrow, ago, etc) or words such as what, who, etc can be used to successfully estimate the temporal questions. We propose using classification as well as regression models to build a confidence score about the temporality of the questions.

This paper focuses on introducing question temporality and showing that it can be estimated effectively. As part of future work, we would like to explore other features that could help improve the classification performance. We also aim at applying some of the ideas presented in this paper in a real world QA system and measuring the effectiveness of their application towards improving user interaction.

ACKNOWLEDGEMENT

We would like to thank the anonymous reviewers for providing their valuable feedback. This work was supported by the National Science Foundation, under grants IIS 08-08692 and 08-12148.

REFERENCES

1. Barabasi, A. L., and Albert, R. Emergence of scaling in random networks. *Science* 286, 5439 (1999), 509–512.
2. Dakka, W., Gravano, L., and Ipeirotis, P. G. Answering general time sensitive queries. In *CIKM* (2008), 1437–1438.
3. Freund, Y., and Schapire, R. E. A decision-theoretic generalization of on-line learning and an application to boosting. In *EuroCOLT*, vol. 904 of *Lecture Notes in Computer Science*, 23–37.
4. Harper, F. M., Moy, D., and Konstan, J. A. Facts or friends?: Distinguishing informational and conversational questions in social Q&A sites. In *CHI*, ACM (2009), 759–768.
5. Kulkarni, A., Teevan, J., Svore, K. M., and Dumais, S. T. Understanding temporal query dynamics. In *WSDM*, ACM (2011), 167–176.
6. Tychonoff, A. N. Solution of incorrectly formulated problems and the regularization method. In *Soviet Mathematics* 4 (1963), 1035–1038.
7. Zhang, R., Chang, Y., Zheng, Z., Metzler, D., and Yun Nie, J. Search engine adaptation by feedback control adjustment for time-sensitive query. In *HLT-NAACL, ACL* (2009), 165–168.