Sustainable Questions

Determining the Expiration Date of Answers

Bart de Goede
bartdegoede@uva.nl

Anne Schuth
anne.schuth@uva.nl

Maarten de Rijke
derijke@uva.nl

Intelligent Systems Lab Amsterdam, University of Amsterdam

ABSTRACT

Community question answering platforms have large repositories of previously answered questions. Reusing the answers for new questions is tempting. However, not all stored answers will still be relevant. We define a new and challenging problem concerning the sustainability of questions, and present metrics aimed at distinguishing between sustainable and non-sustainable questions. We find that an intuitive approach to sustainability of questions is not sufficient, but that simple properties can already distinguish between sustainable and non-sustainable questions.

1. INTRODUCTION

Question-answering communities (or CQA; community question answering), such as StackOverflow[1] and Yahoo! Answers,[2] enable users to pose questions in natural language. These platforms typically have large volumes of previously asked questions available. Retrieval in CQA operates from the notion that, given the large amount of questions available, many of the posted questions are in some (semantically similar) form already available. If questions relevant to the information need of the user could be retrieved, that need would be satisfied quicker and strain on the community—not having to repeat the same answers—could be relieved.

A crucial factor in determining whether the answer to a question similar to the question a user poses is an answer to that question, is whether the answer is still relevant. For example, a question about last nights' soccer game loses its relevance very quickly, whereas other questions remain valid for longer periods of time (*'who is the current prime minister of the UK?'*) or will even remain valid forever (*'who designed the Eiffel Tower?'*).

The sustainability of a question depends on the time it takes before answers that satisfy the information need would change. Therefore, the sustainability of a question can be deduced from answers to questions that are similar. We define question *similarity* as 'semantic similarity,' with which we mean that if questions would have been asked at the same time, they would address to the same information need. Similarity is *not* the same as sustainability. Questions are defined to be *sustainable* when the answer to that question is independent of the point in time the question is asked. In this study, we operationalize sustainability in terms of semantic similarity over time. In this study, we investigate the patterns exhibited by sustainable questions, and whether these patterns can be used to distinguish sustainable from non-sustainable questions.

In order to be able to do so, we first explore three approaches to finding similar questions. We find that our best performing approach based on latent semantic analysis does not perform satisfactory causing us to create a manually labeled set of similar questions which we use for further analysis.

We consider two properties of questions to determine sustainability: the rate at which answers to similar questions change over time, and the time between asking a question and receiving answers.

We find that (*a*) clustering questions with similar semantics is a daunting task; (*b*) an intuitive approach to sustainability in questions by modeling change in their answers is not sufficient; and (*c*) very simple question properties can already distinguish sustainable questions from others in a reasonable manner. Our contributions are (*a*) the definition of this new problem concerning sustainability of questions; (*b*) our conclusion that the new problem is not trivially solved; (*c*) insight in factors that play a role by the definition of the problem; and (*d*) metrics aimed at distinguishing between sustainable and non-sustainable questions and answers.

2. RELATED WORK

Much of the research into CQA retrieval aims to find answers to questions of a user in the data already available in a question-answer repository. Although this touches the field of (traditional) information retrieval, where the similarity of the probability distributions of terms in query and documents is (traditionally) emphasized as a measure for relevance, CQA retrieval faces the lexical chasm. In addition to being very short in general, questions often contain different terms than their (relevant) answers (words as *how* or *why* will typically occur more in questions than in answers). Berger et al. [1] state that questions and their answers are linked on a semantical level, rather than just lexically, and should therefor be linked using 'intermediate concepts.' Query expansion, statistical translation models and latent variable models are proposed as means to create such links, as we will now discuss.

Jeon et al. [11] propose a translation-based retrieval model for finding semantically similar questions from a large Q&A-archive. Using a statistical machine translation approach, questions in the archive are ranked based on the probability that the users question translates into that question. In order to train the translation model, semantically similar questions in the repository are found by comparing their answers; questions with similar answers were used as training data. Xue et al. [19] extend this approach by not only estimating question-to-question probabilities, but incorporating

[1] `http://stackoverflow.com`
[2] `http://answers.yahoo.com`

TAIA workshop at SIGIR 2012, Portland, Oregon, USA

question-answer translation probabilities as well. Additionally, a novel translation-based language model is presented, essentially introducing Dirichlet smoothing to IBM alignment model 1, as well as a control mechanism for the impact either the translation model or the language model component has.

Although both approaches yield interesting results, these are not viable for finding similar questions with the intention to estimate sustainability; these approaches assume questions are similar when their answers are similar, while we assume that answers to similar questions can change over time. We are thus interested in clusters of similar questions regardless of whether their answers are similar.

In order to overcome this problem, our approach to clustering similar questions—which we regard as a preprocessing step to our sustainability analysis—is based solely on the questions. We investigate three approaches. Two methods are based on semantic similarity: *latent semantic analysis* (LSA) [5] and *latent Dirichlet allocation* (LDA) [2]. An advantage of using LSA or LDA in our use case can be found in the dimensionality reduction of the vector space, as well as in the topic modeling that is conducted; (*i*) questions tend to be very short, and their representations can be somewhat expanded when semantically related concepts are considered as well; and (*ii*) Yahoo! Answers users tend to spell their questions poorly, which could be (partially) accounted for by matching of the distribution of other terms co-occuring in the same question. Successful use of LSA and LDA in measuring coherence between texts [8], document retrieval [6, 17], and even cross-language retrieval [4, 7] has been reported. Our third approach to clustering similar questions uses Locality Sensitive Hashing (LSH) [3], which is successfully applied to near duplicate detection [15, 16]. The approaches discussed above all rely on representations of an inherent property of questions asked on web-based question-answering communities: text. Non-textual properties such as ratings and comments, votes, clicks, the amount of answers, questioner and answerer activity levels, etcetera, are recorded by the providers of (most) question-answering communities as well. Jeon et al. [12] find that such properties can improve retrieval results. We find that such meta properties are helpful in deciding on the sustainability of questions.

3. MEASURING SUSTAINABILITY

Our definition of *sustainable* questions, as described in the introduction, implies that we first have to identify *similar* questions; we want to tell sustainable questions apart from questions that are just similar. We find that straightforward approaches to clustering questions do not yield satisfactory results. In order to still find properties that describe the sustainability of questions, we manually assess the clusters that are the output of our best performing clustering method; the details can be found in the experimental setup.

3.1 Change rate of answers

For each cluster of questions we create a tf-idf vector space of the answers labeled as '*best answer*' by either the question asker or the community. Subsequently, we fit a linear function on the cumulative cosine distances between the answers (as shown in Fig. 1), as well as the cumulative cosine distances between answers over time (shown in Fig. 2). Fig. 1 shows a rather constant change in answers over time, whereas Fig. 2 displays differences in the speed of change between different answers over time, suggesting that time might be an important factor in determining the evolution in answers to questions over time. Fig. 1 and Fig. 2 show the same cluster.

The idea behind this approach is that the slope of this linear function will provide an indication to how fast the answers to a set of similar questions change. Also, the sum of squared errors for this function given the dataset might provide clues to periodicity; if the answers to similar questions exhibit large amounts of change in short periods of time, that might indicate that the subject of these questions are subject to periodic changes (for example, *'who is the world champion soccer'* is expected to change suddenly at periodic time intervals). Additionally, we compute the standard deviation of the set of distances.

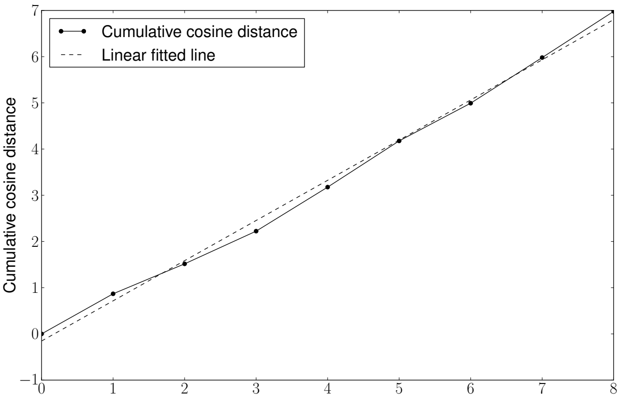

Figure 1: Cumulative cosine distance between vector representations of answers with a linear fitted line for a single cluster. For the 9 best answers in this cluster, the theoretical maximum of the cumulative distance is 8. Each step on the x-axis represents a question-answer pair.

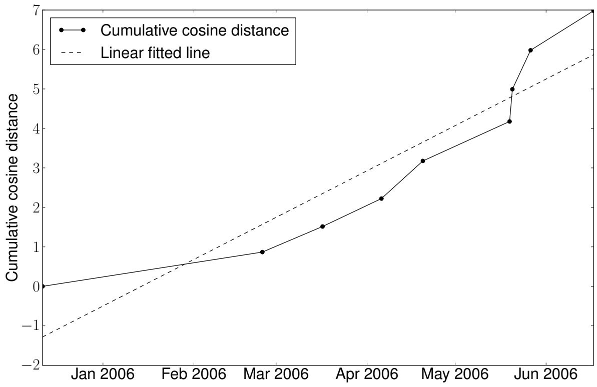

Figure 2: As in Fig. 1, the cumulative cosine distance between vector representations of answers with linear fitted line for a single cluster. However, here the timing of the answers is taken in to account.

3.2 Speed of response

Another property that might be indicative to the sustainability of a cluster of questions, is the time it takes for a question to be answered. The intuition here is that the probability of a sustainable question soliciting answers over a longer period of time would be higher, as the question would still be relevant.

For each cluster, we computed the average time for a question to be resolved in days (i.e., the time between the posting of a question and the posting of the best answer). Also, we computed the standard deviation in answering time, as well as the total amount of days questions in a cluster had to 'wait' for their best answer.

In addition, we computed the average time in days between the posting of a question, and the last answer it received. The intuition is that sustainable questions are more likely to solicit answers longer after they were posted than non-sustainable questions; many questions are answered straightaway and disappear in the timeline quickly, whereas some questions keep getting attention, and are therefore not expired (yet).

Table 1: Accuracy of several question clustering methods. Missing values represent experiments that never terminated.

	sample size		
algorithm	10K	100K	all
LDA	0.435	0.500	-
LSA	0.706	0.638	-
$\text{LSH}_{\text{16bits}}$	0.472	0.484	0.500
$\text{LSH}_{\text{24bits}}$	0.465	0.502	0.495
$\text{LSH}_{\text{32bits}}$	0.512	0.514	0.509
$\text{LSH}_{\text{40bits}}$	0.523	0.537	0.542

4. EXPERIMENTS

Our experiments are aimed at answering the following questions. What are the distinguishing properties of sustainable questions? Can we measure these properties of sustainability? Can we tell sustainable and non-sustainable questions apart based on these properties?

4.1 Experimental Setup

All our experiments are run on the Yahoo! Answers Comprehensive Questions and Answers version 1.0[3] dataset. This data set consists of 3.4M questions with often multiple answers. We perform case and accent folding and employ simple tokenization on both the text of the question and the answers.

We view the clustering of questions as a preprocessing step and therefore take it as part of the experimental setup. We explore three approaches to finding similar questions: latent semantic analysis [5], latent Dirichlet allocation [2] and locality senstive hashing [3].[4]

From the output of each clustering method on the 10K dataset, we sampled 559 pairs of questions and manually judged 205 as rightfully clustered together and 354 as wrongly clustered together. We used the combined set of judgements (randomly sampling 205 questions from the wrongly-clustered set) to arrive at the accuracy results in Table 1; for each judged pair of questions we observe whether the algorithm was correct in either putting both questions in the same cluster or keeping them separate. Based on these accuracy results we decided on using LSA as our clustering approach for the remainder of our experiments. We also decided on taking the sample of 10K documents as the basis for our analysis.

While we consider clustering of similar questions as a preprocessing step for our approaches to sustainability, we can not ignore the fact that obtaining a reasonable clustering performance is important for our sustainability estimation. We do not consider an accuracy of 0.706 to be good enough. Therefore, we opt to manually label data for further investigation.We labeled the 904 clusters in the output of our LSA clustering approach on the before mentioned subset of 10K questions with one of three classes: 752 *all* clusters, 144 clusters with *similar* questions and 8 clusters with *sustainable* questions. The clusters in the *similar* class are only required to have similar questions—questions asking for the same information—regardless of the answers; these clusters can thus be sustainable and non-sustainable. Additionally, the clusters in the *sustainable* class are required to have answers that do not change over time. Note that this definition implies that the *sustainable* class is a subset of the *similar* class which is a subset of the *all* class.

Subsequently, for each cluster we compute cosine distances between chronologically sorted best answers, as described in Section 3.1. For each set of distances (per cluster), we compute the average, standard deviation, average change per day, standard deviation on the cumulative distances, plus the slope and sum of squared errors of a linearly fitted function on the cumulative distances. Also, we compute for each cluster the time between the moment a question was posted and the answer labeled as best answer, and the time between the last answer that question received, as described in Section 3.2. For each set of these distances in time, we compute the average, standard deviation, standard deviation on the cumulative distances in time, plus the slope and sum of squared errors of a linearly fitted function on the cumulative distances in days.

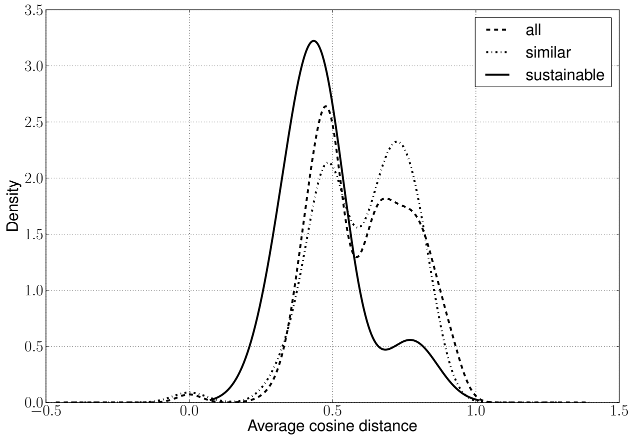

Figure 3: Kernel density estimation of the average cosine distance between answers labeled as best according to either the user or the community.

4.2 Results

Fig. 3 shows a kernel density estimation[5] plot of the average cosine distance between the best answers for each class of clusters. Although there seems to be some evidence for this metric to distinguish similar and sustainable clusters from regular clusters, it is not that strong. However, when we consider the time between the moment of posting a question and the moment that question receives its final answer, we see that questions we deem sustainable keep receiving answers far longer than 'regular' or even similar questions. Fig. 4 shows a kernel density estimation[6] plot for the time between the posting of a question and the reception of its last answer. It should be noted that the set of sustainable clusters is a subset of the set of similar clusters, and that the set of similar clusters is a subset of the set of all clusters. This explains the second local maximum in the 'all clusters' line.

4.3 Analysis

When comparing a kernel density estimation of the average cosine distance between the best answers to the questions in a cluster (shown in Fig. 3) with a kernel density estimation of the average time in days between posting a question and that question receiving its last answer (shown in Fig. 4) we see that the time between the posting of a question and receiving its last answer is very indicative in describing sustainability: the longer a question solicits answers, the higher the probability of said question to be sustainable.

When training a simple tree classifier[7] using the properties defined in Section 4.1 as featureswe find that the combination of

[3] http://webscope.sandbox.yahoo.com/catalog.php?datatype=l

[4] We use LSA and LDA to reduce the dimensionality (almost a million unique terms in the corpus) of the tf-idf representations of the questions. Then, we perform clustering by grouping questions with a cosine similarity greater or equal to 0.95. For LSH, we hash each question, and group questions with a Hamming distance [10] less than 3 together.

[5] We use kernel density estimation because it models the density of data points at a given value. In this way, a fairer comparison between the instances of our three classes can be made; we have far less *sustainable* than *similar* questions [18].

[6] This is why the plot covers negative values for time as well.

[7] We use the WEKA [9] implementation of C4.5 by Quinlan [14].

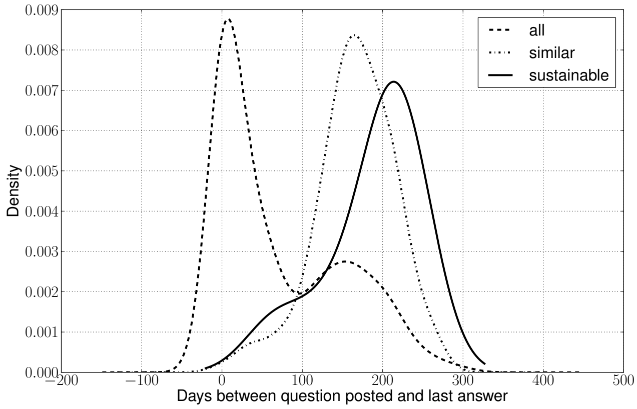

Figure 4: Kernel density estimation of time in days between posting of a question and the last answer a question received.

features is capable of obtaining a classification accuracy of 91.5%,[8] indicating that using very simple properties such as the time between a question and its last answer, and the cosine distance between the answers over time allow for a reasonable distinction between sustainable and non-sustainable questions. Similar comparisons were done for the change rate approach discussed in Section 3.1, but no meaningful distinctions between cluster types could be made. We attribute this to a large degree of the sparsity of the answer vectors and the relatively small size of most clusters; a linear line fitted on two data points is always perfect.

5. DISCUSSION AND CONCLUSION

Those parts of this work that are based on textual similarity measures—the change rate as well as the clustering of similar questions—are not performing as well as we projected. Questions and their answers generally tend to be very short, complicating the use of traditional document representations and similarity measures to distinguish between cluster categories. To overcome this representation problem, we made an exploratory attempt to expand our question and answer models by linking *n*grams in answers to Wikipedia pages (considering a page as a concept) using semantic linking [13], and use the concepts found as vectors to estimate similarity between documents. However, we found this approach did not yield the expected improvements.

In future experiments, expansion of the question and answer models in a meaningful way could improve measuring sustainability based on textual features. Our current experiments show that the textual content of questions and answers alone is not sufficient to make reasonable comparisons between questions. Also, we expect that scaling up the amount of questions considered may improve clustering results. Given the resources available, we were not able to cluster and annotate all questions. The sample considered is small compared to the available data set, and could therefore have yielded many meaningless clusters. Another open issue that remains is a robust approach to clustering questions such that they fit our definition of similar questions.

Nevertheless, we did find that very simple indicators such as the time that questions keep soliciting answers can already distinguish sustainable questions from others in a reasonable manner.

We conclude that (*a*) sustainable questions tend to be answered longer than regular questions, (*b*) given a robust clustering method, that property can be measured in a set of similar questions, and (*c*) in combination with some very simple properties, sustainable and non-sustainable questions can be distinguished.

[8] We resampled the data such that a random classifier obtaines an accuracy of 33.3%. We used stratified 10 fold cross-validation.

Acknowledgments. This research was supported by the European Union's ICT Policy Support Programme as part of the Competitiveness and Innovation Framework Programme, CIP ICT-PSP under grant agreement nr 250430, the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreements nr 258191 (PROMISE) and 288024 (LiMoSINe), the Netherlands Organisation for Scientific Research (NWO) under project nrs 612.061.814, 612.061.815, 640.004.802, 380-70-011, 727.011.-005, 612.001.116, the Center for Creation, Content and Technology (CCCT), the Hyperlocal Service Platform project funded by the Service Innovation & ICT program, the WAHSP and BILAND projects funded by the CLARIN-nl program, the Dutch national program COMMIT, and by the ESF Research Network Program ELIAS.

6. REFERENCES

[1] A. Berger, R. Caruana, D. Cohn, D. Freitag, and V. Mittal. Bridging the lexical chasm: statistical approaches to answer-finding. In *SIGIR '00*, pages 192–199, 2000.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet Allocation. *JMLR*, 3:993–1022, March 2003.

[3] M. Charikar. Similarity estimation techniques from rounding algorithms. In *STOC '02*, pages 380–388, 2002.

[4] P. Cimiano, A. Schultz, S. Sizov, P. Sorg, and S. Staab. Explicit versus latent concept models for cross-language information retrieval. In *IJCAI '09*, pages 1513–1518, 2009.

[5] S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshman. Indexing by latent semantic analysis. *JASIS*, 41 (6):391–407, 1990.

[6] S. Dumais. Enhancing performance in latent semantic indexing (LSI) retrieval, 1992.

[7] S. Dumais, T. Letsche, M. Littman, and T. Landauer. Automatic cross-language retrieval using latent semantic indexing. In *AAAI Spr. Symp. Cross-Lang. Text and Speech Retr.*, 1997.

[8] P. Foltz, W. Kintsch, and T. Landauer. The measurement of textual coherence with latent semantic analysis. *Discourse processes*, 25(2-3):285–307, 1998.

[9] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. Witten. The WEKA data mining software: an update. *SIGKDD*, 11(1):10–18, 2009.

[10] R. Hamming. Error detecting and error correcting codes. *Bell System technical journal*, 29(2):147–160, 1950.

[11] J. Jeon, W. Croft, and J. Lee. Finding similar questions in large question and answer archives. In *CIKM '05*, pages 84–90, 2005.

[12] J. Jeon, W. Croft, J. Lee, and S. Park. A framework to predict the quality of answers with non-textual features. In *SIGIR '06*, pages 228–235, 2006.

[13] E. Meij, W. Weerkamp, and M. de Rijke. Adding semantics to microblog posts. In *WSDM '12*, pages 563–572, 2012.

[14] J. Quinlan. *C4. 5: Programs for machine learning*. 1993.

[15] B. Stein. Principles of hash-based text retrieval. In *SIGIR '07*, pages 527–534, 2007.

[16] M. Theobald, J. Siddharth, and A. Paepcke. Spotsigs: robust and efficient near duplicate detection in large web collections. In *SIGIR '08*, pages 563–570, 2008.

[17] X. Wei and W. Croft. Lda-based document models for ad-hoc retrieval. In *SIGIR '06*, pages 178–185, 2006.

[18] W. Winner. Introduction to kernel density estimation. *NTIS, Springfield, Virginia (USA)*, 1985.

[19] X. Xue, J. Jeon, and W. Croft. Retrieval models for question and answer archives. In *SIGIR '08*, pages 475–482, 2008.