Topic Shifts in StackOverflow: Ask it Like Socrates

Toni Gruetze(✉), Ralf Krestel, and Felix Naumann

Hasso Plattner Institute, Potsdam, Germany
{toni.gruetze,ralf.krestel,felix.naumann}@hpi.de

Abstract. Community based question-and-answer (Q&A) sites rely on well-posed and appropriately tagged questions. However, most platforms have only limited capabilities to support their users in finding the right tags. In this paper, we propose a temporal recommendation model to support users in tagging new questions and thus improve their acceptance in the community. To underline the necessity of temporal awareness of such a model, we first investigate the changes in tag usage and show different types of collective attention in StackOverflow, a community-driven Q&A website for computer programming topics. Furthermore, we examine the changes over time in the correlation between question terms and topics. Our results show that temporal awareness is indeed important for recommending tags in Q&A communities.

1 Tags in Q&A Communities

During the last two decades, various popular question-and-answer (Q&A) platforms, such as Ask.com, Experts-Exchange, Quora, and Yahoo! Answers have emerged. These platforms provide their users with the opportunity to ask questions and challenge other users of the community to share their knowledge based on these questions. Since 2008 StackOverflow (SO) is the de facto standard question-and-answer website for topics in computer science and programming. In SO, users earn reputation through active participation in the community, e.g., by posing well-received questions or providing helpful answers. After achieving a high reputation level, a user is allowed to up-vote or comment on questions and answers or even edit posts of other users.

Vasilescu et al. showed that experienced programmers (i.e., active GitHub committers) ask fewer questions and give more answers [9]. Anderson et al. investigated the correlation between delay of and reputation for a given answer [1]. Wang et al. show that the majority of questioners in SO asks only one question during their membership [10]. Only little research has been done on how to support these "newbies" on posing well-received questions. A qualitative study discusses efficient ways on how to give code examples for well-received questions [7]. Bazelli et al. establish a connection between extroversion and openness used in the questions and high reputation points of users [2].

Another means of improving a question is by categorizing the post appropriately and help a skilled responder to find it. To this end SO supports tags to

E. Métais et al. (Eds.): NLDB 2016, LNCS 9612, pp. 213–221, 2016.
DOI: 10.1007/978-3-319-41754-7_18

categorize questions. Questioners are forced to choose at least one computer science topic for their new post so that other users, who subscribed topics of their interest and profession, have the chance to find the post. Hence, an appropriately tagged question has a higher probability of getting useful answers.

Depending on the social platform, tags are used with different temporal characteristics. In platforms like Twitter many topics, such as sport events or disasters, are discussed in a very narrow time span (e.g., the Superbowl final: #SuperBowl or the Paris attacks of November 2015: #PrayForParis). Lehmann et al. study different classes of collective attention on Twitter [5] and Gruetze et al. evaluate the influence of such temporal changes for the tag recommendation task [4]. Based on the example of Quora, Maity et al. show that the topics discussed in general-purpose Q&A platforms underly strong temporal dynamics (e.g., political topics) [6]. The topics posted on SO (i.e., programming languages, databases, or operating systems) seem to be more static and lead to the assumption that they slowly evolve over time.

In contrast to this first intuition, we show that tag usage in SO indeed underlies strong temporal effects. In the following, we discuss four different types of temporal topic popularity patterns in SO. We further show how the likelihood for question terms can strongly change over time for SO topics. Finally, we show that recommender systems that incorporate these temporal changes outperform their static counterparts and thus improve the support for SO users in asking well-received questions to receive appropriate answers and thus earning the Socratic badge, which is awarded for asking a well-received questions.

2 Tags in StackOverflow

As the basis for the following analysis, we use the Stack Exchange dataset containing approximately 10 million questions posted between August 1^{st} 2008 and July 31^{st} 2015 on StackOverflow. The questions were posted by over 4.5 million users and categorized using over 40 thousand unique tags. The dataset is licensed under Attribution-ShareAlike 3.0 and available online.[1]

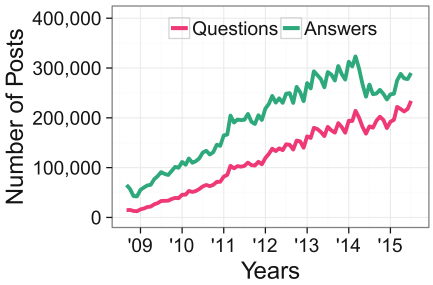
Figure 1b shows the number of questions and answers provided by the SO community. The steady growth of posts emphasizes the increase of popularity and spread of this Q&A platform. We presume that the large 2014 drop in the number of answers and questions originates from the Mighty Mjölnir community update that allows experienced users to close duplicated questions.

The temporal dynamics of discussed topics (i.e., applied tags) is shown in Fig. 1c. While the number of distinct topics per month increases nearly linearly over time, the number of new topics, i.e., tags that were not used before, does not. During the first two years, the number of new topics is relatively high but strongly decreasing. This shows that the platform is gaining coverage of the relevant topics for developers. As of mid 2010 the number of new topics is relatively

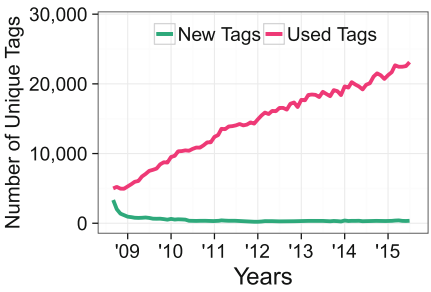
[1] https://archive.org/details/stackexchange.

Number of ...	Value
Questions	9,970,064
Answers	16,502,856
Users	4,551,130
Tags	29,497,960
Unique Tags	41,719

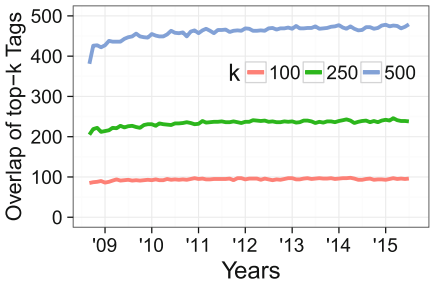
(a) Dataset statistics



(b) Monthly questions & answers



(c) Monthly tag usage



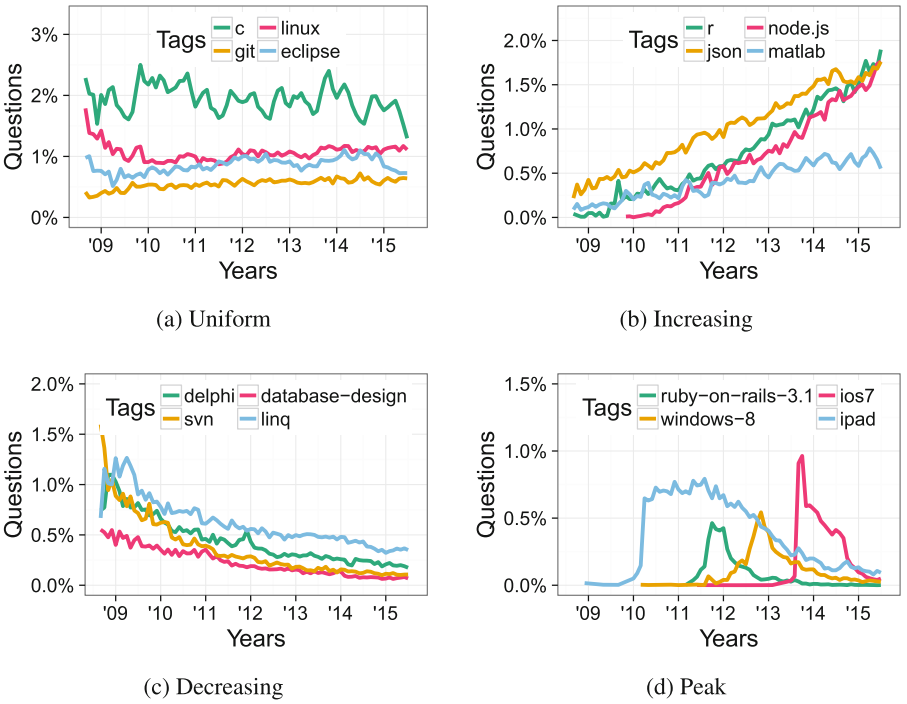
(d) Stability of top-k tags

Fig. 1. SO dataset statistics including monthly usage charts (Color figure online)

stable between 200 and 400 per month. We believe that these topics are real new topics, i.e., topics that emerge at this point in time, such as a new programming languages (e.g. Dart (2011)) or new frameworks (e.g. Apache Spark (2013)). The continuing increase of distinct tags per month originates from a gain in topic coverage and is asymptotic to the total number of topics in SO.

Figure 1d shows the stability of the most questioned topics, i.e., the top 100, 250, and 500 tags. The top-100 topics are very stable such that, starting from 2010, these topics overlap with the top-100 of the previous month by approximately 95 %. The number of overlapping top 250 and 500 topics is large too. Beginning with 2011, both values show a relatively constant overlap of about 92 %.

All three statistics show that the community, as well as the coverage of relevant topics for programmers, is constantly increasing. Because the community is growing faster than the newly emerging topics, we expect the number of distinct topics discussed per month to converge in the near future. However, we show that the topics discussed in SO — while making the impression to be well covered by a significant number of questions and answers — still underly significant temporal effects, such that the consideration of up-to-the-minute statistics significantly increase the understanding of the topics in SO.



(a) Uniform

(b) Increasing

(c) Decreasing

(d) Peak

Fig. 2. Examples for the four most common temporal tag attention pattern. Depicted in each figure, the percentage of questions that were assigned to a specific topic. (Color figure online)

3 Temporal Tagging Behavior in StackOverflow

In comparison to other social media platforms, such as Twitter and Facebook, SO users exhibit a differing temporal tagging behavior.

Lehmann et al. show different classes of collective attention for tags in Twitter that span a period of one day to one week [5]. Similarly, we are able to classify most **types of attention** in SO into four classes, namely uniform, increasing, decreasing, and peak. Examples for all four classes are shown in Fig. 2 and are next discussed in more detail:

Uniform: Figure 2a shows a uniform usage pattern of the topics with a relatively stable expected monthly usage value. Topics with the uniform usage ratio pattern are growing as fast as the platform (Fig. 1b). While showing a stronger variance of monthly percentages, the topic C still provides a constant expected value of around 2 %. The other topics stay relatively constant over the seven year time span with around 1 % (Linux and Eclipse) or 0.5 % (Git).

Increasing: Figure 2b depicts topics with an increasing attention on the SO platform. The two programming languages R and MATLAB gain in popularity.

Both languages, which have a strong focus on mathematics and statistics, are commonly used in the area of data science and data analytics. Until mid 2009 the number of MATLAB questions was higher than R related posts. Afterwards, the number of R questions follows a stronger growth. Notably, the number of questions concerning R is higher than the ones concerning C at the end of the data period.

Decreasing: Figure 2c depicts the decrease in spreading of the topics (halved popularity). For instance, the number of questions about the proprietary and expensively licensed programming framework Delphi decreases. Simultaneous, the number of questions regarding open source alternatives, such as .NET and Java, increases. The decreasing spread of question regarding SVN shows that more and more software development projects switch to other version control solutions (e.g., Git).

Peak: Figure 2d shows the spreading of discussion topics following a peak pattern, which behave like a mixture of increasing and decreasing pattern happening in a short time frame. The peak attention pattern is a typical scheme for topics covering particular versions of frameworks, platforms, etc., as shown for iOS7. However, the pattern can also appear for non-version topics such as iPad, which was hyped on SO in spring 2010, the time of the official release of the device. While being a major concern of developers also shortly after the release, the community lost interest, such that, 2 years and 9 month later the relative frequency of questions with this topic halved.

Topic Shifts in SO

Besides the changes of topic popularity, rather the community also produces **topic shifts**, i.e., temporal changes of the vocabulary used for questions associated with a specific topic.

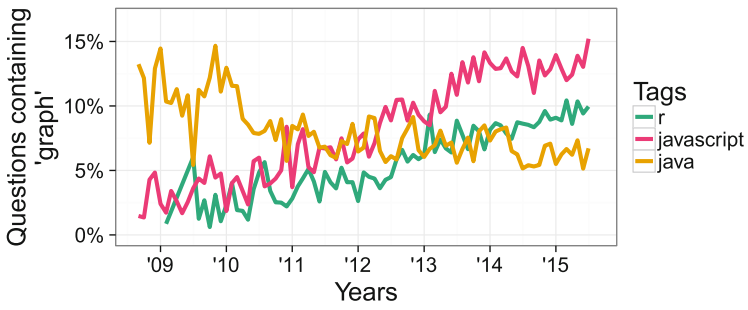


Fig. 3. Topic shift example for tags on questions containing the term 'graph' (Color figure online)

For instance, Fig. 3 contrasts three different topics in the context of posts containing the term 'graph' in the question text. Due to the ambiguity of the term,

different meanings might be covered, e.g., a chart or infographic, a structure connecting a set of objects by links (graph-theory), or a plot of a mathematical function. These meanings are expressed by the assignment of different tags. For example, the percentage of tag Java is decreasing over time, due to the fading importance of graph theoretic questions. In contrast, the number of questions about R, which is usually used for plotting of mathematical functions and calculating statistics, increase. The questions assigned to Javascript typically cover visualization of graph data.

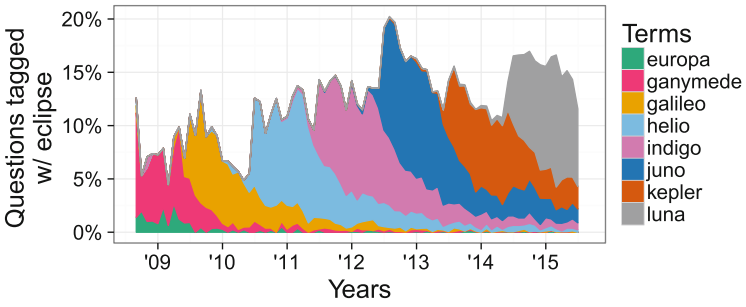


Fig. 4. Topic shift example for keywords on questions tagged with eclipse (Color figure online)

Figure 4 presents the frequency of selected terms in questions tagged with Eclipse. Specifically, the terms represent the version names of the IDE during the examined period. Evidently, the number of questions follow a temporal pattern that is strongly correlated with the yearly release cycle of the IDE. Similar effects can be observed for other topics, such as Android with version names from Cupcake to Lollipop.

4 Tagging Like Socrates

By assigning appropriate tags to a question, the likelihood to be found by the community is increased. The Socratic badge is awarded for asking a well-received question in SO. To model a tag recommender, we phrase the assignment of topics to a question as a posterior probability of topic parameters given a question: $P(\theta_{topic}|q)$. As shown in [4], such a formalization can be used to recommend the k-most probable topics for a given question by:

$$rec_k(q) = \underset{\{t_1..t_k\}\subset T}{\arg\max} \sum_{i=1}^{k} P(\theta_{t_i}) \cdot P(q|\theta_{t_i})$$

where T refers to the set of known topics. As shown in Fig. 2, the collective attention for topics in SO underlies strong temporal changes. In our experiments,

we model the prior probability $P(\theta_{t_i})$ as frequency of posts with this topic from the past, including all recent posts. The likelihood $P(q|\theta_{t_i})$ is estimated based on nouns found in the question texts for a given topic (e.g., such as the version names for Eclipse questions shown in Fig. 4). Note, we could have used more word classes or even all words, however, tests showed a comparable recommendation quality for models based on nouns as for models with all textual information while minimizing the vocabulary size. To further facilitate the probability estimation, we make the assumption of independence between the probabilities of terms (as the Naïve Bayes model). This plain model greatly facilitates the update of its parameters and thus enables the efficient incorporation of current topic changes, such that the model parameters can be updated for every newly posed question. In contrast, more complex methods, such as dynamic topic models, are based on more expensive parameter inference methods. In particular, they require a temporal partitioning of the data (i.e., the parameter estimation can be executed only for time slices, e.g. yearly) [3].

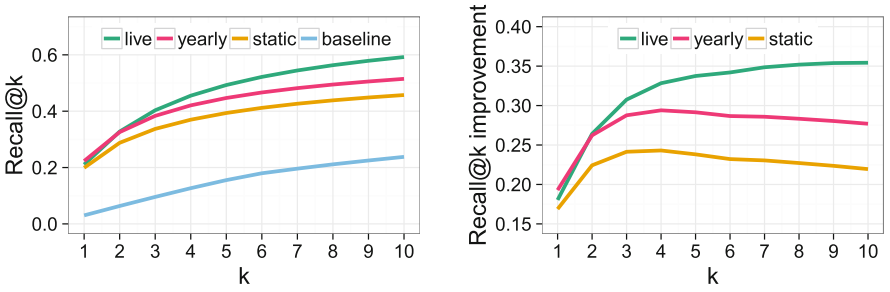
Note, this recommender definition is rather simple and does not include features like tag co-occurrence, code snippet information, or user preferences. As shown in previous research, such features yield a significant improvement of the recommendation quality [8,11]. However, due to the focus on the temporal aspects, we did not use this knowledge for the experiments and do not discuss these features here and rather focus on the improvements due to temporal patterns.

To measure the influence of topic shifts in SO for tag recommendations, we evaluated four different configurations of the recommender: the first model follows a static learn once strategy, which is based on the topic knowledge derived from all posts created in the year 2010. The second model is updated annually, this yearly model recommends based on the posts of the previous calendar year (e.g., recommendations in 2011 are based on the posts of 2010). Third, the live model is continuously updated based on the latest SO posts. This model is based on all posts that were created in the one year period before a question q was submitted. Finally, we define the simple baseline that is only based on the prior probability $P(\theta_{t_i})$ and thus always recommends the k most commonly used tags for all posts independent from the question text ($\forall t_i, t_j \in T : P(q|\theta_{t_i}) = P(q|\theta_{t_j})$). To compare the performance of all four models, we computed top-k recommendations for all questions Q from the years 2011–2015. Based on these recommendations, we measured recall in comparison to the actual used tags assigned to the question.

$$recall@k = \frac{1}{|Q|} \sum_{q \in Q} \frac{|rec_k(q) \cap act_{Tags}(q)|}{|act_{Tags}(q)|}$$

Results

Figure 5a shows the recall as the function of k for all four approaches. As expected, the baseline model performs worst. This is due to the ignorance of question contents. For $k = 5$ the static model was already able to recommend 39 % of the actually used tags correctly. The yearly model improved this performance



(a) Recall@k (b) Recall@k improvements over baseline

Fig. 5. Recall@k statistics for different recommendation models (Color figure online)

by 5 percent points, whereas the live model doubled these improvements and achieved a recall of 49 %. This is a respectable performance given that the static baseline is already able to correctly identify 2 out of 5 tags given the set possible set of over $40k$ tags. The significant increase of ten percentage points is based on the additional temporal information and shows the potential even for a seemingly static corpus such as SO. Considering the top-10 recommendations, the live model is able to improve on recall by 10 percentage points (up to 59 %), whereas the gap to the static (and yearly) model is increased by 6 (respectively 5) percentage points.

To investigate the adaptation of the models with respect to topic shifts, Fig. 5b shows the recall improvement of the content based models (static, yearly, and live) in comparison to the baseline. The improvements are significant for all competitors, however, only the live model is able to constantly increase the improvements for increasing k values. This underlines the superior adaptation to topic shifts due to the fast update of the model parameters.

5 Conclusion

We studied the temporal topic changes in the Q&A platform StackOverflow. We examined four different types of popularity patterns of topics over time to show that indeed there are topic shifts. We showed that the likelihood for question terms can strongly change over time for SO topics. Finally, we showed that the tag recommender that are able to incorporate these temporal changes are able to significantly outperform their static counterparts.

In future work we will survey further Q&A platforms and compare the identified temporal effects. We will investigate whether other features, such as user preferences, code snippets, named entities, or tag co-occurrences, are changing over time and can help to improve tag recommendation for new questions. Another interesting application for our findings would be trend prediction in SO and thus for the computer programming field.

Acknowledgments. This research was funded by the German Research Society, DFG grant no. FOR 1306. We thank the StackOverflow community for sharing the valuable knowledge targeted in this work. Finally, we wish to acknowledge the anonymous reviewers for their detailed and helpful comments to the manuscript.

References

1. Anderson, A., Huttenlocher, D., Kleinberg, J., Leskovec, J.: Discovering value from community activity on focused question answering sites: a case study of Stack Overflow. In: Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 850–858 (2012)
2. Bazelli, B., Hindle, A., Stroulia, E.: On the personality traits of StackOverflow users. In: Proceedings of the International Conference on Software Maintenance (ICSM), pp. 460–463 (2013)
3. Blei, D.M., Lafferty, J.D.: Dynamic topic models. In: Proceedings of the International Conference on Machine Learning (ICML), pp. 113–120 (2006)
4. Gruetze, T., Yao, G., Krestel, R.: Learning temporal tagging behaviour. In: Proceedings of the International Conference on World Wide Web (WWW Companion), pp. 1333–1338 (2015)
5. Lehmann, J., Gonçalves, B., Ramasco, J.J., Cattuto, C.: Dynamical classes of collective attention in Twitter. In: Proceedings of the International Conference on World Wide Web (WWW), pp. 251–260 (2012)
6. Maity, S., Sahni, J.S.S., Mukherjee, A.: Analysis and prediction of question topic popularity in community Q&A sites: a case study of Quora. In: Proceedings of the International AAAI Conference on Web and Social Media (ICWSM) (2015)
7. Nasehi, S.M., Sillito, J., Maurer, F., Burns, C.: What makes a good code example? A study of programming Q&A in StackOverflow. In: Proceedings of the International Conference on Software Maintenance (ICSM), pp. 25–34 (2012)
8. Stanley, C., Byrne, M.D.: Predicting tags for StackOverflow posts. In: Proceedings of the IEEE International Conference on (ICCM), pp. 414–419 (2013)
9. Vasilescu, B., Filkov, V., Serebrenik, A.: StackOverflow and GitHub: associations between software development and crowdsourced knowledge. In: Proceedings of the International Conference on Social Computing (SocialCom), pp. 188–195 (2013)
10. Wang, S., Lo, D., Jiang, L.: An empirical study on developer interactions in StackOverflow. In: Proceedings of the Annual ACM Symposium on Applied Computing (SAC), pp. 1019–1024 (2013)
11. Wang, S., Lo, D., Vasilescu, B., Serebrenik, A.: EnTagRec: an enhanced tag recommendation system for software information sites. In: Proceedings of the International Conference on Software Maintenance and Evolution (ICSME), pp. 291–300 (2014)