Harnessing Engagement for Knowledge Creation Acceleration in Collaborative Q&A Systems

Jie Yang, Alessandro Bozzon, and Geert-Jan Houben

Delft University of Technology, Mekelweg 4, 2628 CD Delft, the Netherlands
{j.yang-3, a.bozzon, g.j.p.m.houben}@tudelft.nl

Abstract. Thanks to reputation and gamification mechanisms, collaborative question answering systems coordinate the process of topical knowledge creation of thousands of users. While successful, these systems face many challenges: on one hand, the volume of submitted questions overgrows the amount of new users willing, and capable, of answering them. On the other hand, existing users need to be retained and optimally allocated. Previous work demonstrates the positive effects that two important aspects, namely engagement and expertise valorisation, can have on user quality and quantity of participation. The magnitude of their effect can greatly vary across users and across topics. In this paper we advocate for a more in-depth study of the interplay that exists between user engagement factors in question answering systems. Our working hypothesis is that the process of knowledge creation can be accelerated by better understanding and exploiting the combined effects of the interests and expertise of users, with their intrinsic and extrinsic motivations. We perform a study over 6 years of data from the `StackOverflow` platform. By defining metrics of expertise and (intrinsic and extrinsic) motivations, we show how they distribute and correlate across platform's users and topics. By means of an off-line question routing experiment, we show how topic-specific combinations of motivations and expertise can help accelerating the knowledge creation process.

1 Introduction

Collaborative Question Answering (CQA) systems (e.g. StackOverflow, Quora, Yahoo Answers) are an important class of Web knowledge repositories [1]. They coordinate practitioners with varying levels of expertise in the creation of evolving, crowdsourced, and peer-assessed knowledge bases, often in a reliable, quick and detailed fashion.

In CQAs users (askers) create questions, counting on topically-defined communities to provide an answer to their needs. Community members can browse existing questions, and decide whether or not to contribute to ongoing discussions. Such decisions are influenced by a multitude of factors, including time constraints, quality and difficulty of the question, and the knowledge of the answerer. Previous work [2, 17] shows how engagement elements such as gamifications mechanisms, and expertise valorisation can provide users with the right incentive for participation and collaborative knowledge creation.

While successful, such factors cannot prevent CQAs from facing several sustainability challenges. The volume of submitted questions overgrows the amount of new users willing, and capable, of answering them; a large portion of questions do not receive good (up-voted) answers, and even well-posed and relevant questions might wait for a long time before receiving a good answer [15, 16]. Recent studies have proposed to solve these problems with acceleration mechanisms such as: automatic detection of poorly formulated questions, question editing suggestion [15, 16], or question routing [8, 14, 5, 18].

To maximise the effectiveness of such mechanisms, a better comprehension of the mechanisms of knowledge creation in CQAs is needed. Recent research [9, 17] shows that engagement and topical expertise are complementary user properties. In this paper we advocate for a more in-depth understanding of the interplay that exists between them, and we aim at demonstrating how they can be used to accelerate knowledge creation in CQA systems.

Our working hypothesis is that the process of knowledge creation is topically dependent, and that is driven by a mix of *intrinsic* motivations, *extrinsic* motivations, and *topical expertise* of CQAs users. We suggests that different topic-specific knowledge needs demand for different types of contributor: intuitively, to generate the best answer, some questions may require active answerers engaged in discussion; others may only need one expert user to directly provide the right answer. To test our hypothesis we focus on `StackOverflow`, a question answering system specialised in programming-related issues. The paper provides the following original contributions:

1. A study, focusing on the relation that exist between intrinsic motivations (e.g. interest), extrinsic motivations (e.g. reward), and expertise in topically-centred communities;
2. An off-line question routing experiment, aimed at verifying the impact of (intrinsic and extrinsic) motivations and expertise in user modelling for question recommendation.

Our work provides novel insights on the mechanisms that regulates knowledge creation in CQA systems. Although the study and the experiment focus on `StackOverflow` data, we believe that our results are of general interest. The study shows the relevant impact that different topics exercise on (intrinsic and extrinsic) motivation and expertise: the results can be used to devise novel engagement and retention mechanisms, aimed at accelerating knowledge creation by maximising the effectiveness of contributors. The experiment presented in the paper provides empirical evidences of how existing CQAs can profit from the adoption of question routing mechanisms that include topical interest, motivations, and expertise as user modelling properties.

The remainder of the paper is organised as follows: Section 2 briefly introduces engagement dimension in CQAs. Section 3 analyses (intrinsic and extrinsic) motivations and expertise in `StackOverflow`, while Section 4 shows how they can improve question routing performance. Section 5 describes related work, before Section 6 presents our conclusions.

2 Engagement Dimensions In CQA Systems

User engagement is defined as "the emotional, cognitive and behavioural connection that exists, at any point in time and possibly over time, between a user and a resource" [3]. Among the attributes that characterise engagement (e.g. aesthetics, endurability, novelty, reputation), *user context* embeds a combination of user- and context-dependent factors that profoundly influence and affect the relation between CQAs and their users. In this work we focus on two factors, namely: the *motivation*s driving users' activities; and users' *expertise*, as assessed by their peers, in a given topic of interest.

Motivation, is a precondition for action. To foster user engagement, system designers must understand the reasons why users take a particular action. The Self-Determination Theory [6] differentiates between *intrinsic* and *extrinsic* motivation.
Intrinsic motivations lead individuals to perform an activity because of their personal *interest* in it; or because its execution gives some form of *satisfaction*. Users of CQAs are often intrinsically motivated [12]; they decide to interact with systems and their communities:*a*) To look for existing solutions to their issues; this involves browsing the CQA content, in search for the right formulation of knowledge need and, the answer(s) to it. *b*) To post a new question to the community, when no existing solution can be found. Or, *c*) to get satisfaction from the sense of efficacy perceived when, convinced to possess the skills and competence required to contribute to an ongoing discussion, they provide a new answer, or they comment/vote existing questions and answers.
When *extrinsically motivated*, individuals perform an activity for an outcome different from the activity itself, e.g. to obtain *external rewards*. A typical example of an engagement mechanism that exploits extrinsic motivation is *gamification* [2]. CQA systems often adopt two forms of external rewards: 1) a public *reputation score*, calculated by summing the number of votes obtained by all the posted questions and answers; and 2) a set of *badges*, assigned after achieving pre-defined goals (e.g. complete at least one review task, achieve a score of 100 or more for an answer).

Expertise. An expert can be defined as someone who is recognised to be skilful and/or knowledgable in some specific field [7], according to the judgment of the public, or of peers. In CQAs, social judgement is critical for expert identification. A question is usually answered by a set of users, whose answers are voted up or down by other members of the platform, thus reflecting the a user's capability of applying knowledge to solve problems. Hence, voting from other users can be seen as an unbiased, cyber simulation of social judgement for the answerers' expertise level [17]. Expertise can be seen as an example of intrinsic motivation related to competence. However, we stress the fundamental difference that exists between one's perception of competence (which is self-established, and often biased), and social judgement: by being externally attributed, the latter might not set off the same type of intrinsic triggers. Simply put: being perceived as an expert does not necessarily imply behaving like one. Next section elaborates on this behavioural difference, and provides quantitative support to our classification choice.

3 Analysing Extrinsic Motivations, Intrinsic Motivations, and Expertise in StackOverflow

The first part of our work studies how intrinsic motivations, extrinsic motivations, and expertise manifest themselves in topically-centred CQAs communities. We analyse `StackOverflow`, a popular CQA system launched in 2008 with the goal of becoming a very broad knowledge base for software developers. `StackOverflow` now features more than 2.7M users, 6.5M active questions, 11.5M answers, 26.1M comments, and 35.2K tags used by users to briefly characterise the subjects of the submitted questions. `StackOverflow` periodically releases a public version of the platform database, which can be accessed at `https://archive.org/details/stackexchange`. Our study is based on data created up until January 2014. Due to space limitations, the following sections report only part of the performed analysis and experiments. An extended description is available at `http://wis.ewi.tudelft.nl/umap2015`.

To investigate topical diversity, we categorise tags into 14 topics, shown in Table 1. Topics are identified by analysing the tag co-occurrence graph, using the approach described in [4].

Table 1: Topical categorisation of tags, with basic knowledge demand and contributors composition statistics.

	Knowledge Demand			Contributor Composition				
Topic	**Tags**	**#Q**	**#A**	**#CU**	**#AU**	**%(AU∩CU)**	**%(CU-AU)**	**%(AU-CU)**
.Net	c#, asp.net, .net, vb.net, wcf	571K	1222K	102K	119K	73.84%	6.40%	19.76%
Web	javascript, jquery, html, css	569K	1181K	146K	149K	85.43%	6.40%	8.17%
Java	android, java, eclipse	566K	1097K	136K	136K	85.84%	7.32%	6.84%
LAMP	php, mysql, arrays, apache	432K	927K	39K	128K	21.10%	6.98%	71.92%
C/C++	c, c++, windows, qt	269K	679K	78K	80K	87.45%	10.19%	12.36%
iOS	iphone, ios, objective-c	262K	441K	59K	57K	79.24%	11.98%	8.78%
Databases	sql, sql-server, database	177K	406K	74K	73K	79.37%	10.68%	9.95%
Python	python, django, list	186K	390K	55K	61K	67.59%	12.00%	20.41%
Ruby	ruby, ruby-on-rails	129K	226K	32K	39K	59.57%	12.89%	27.54%
String	regex, string, perl	99K	264K	47K	57K	57.26%	13.93%	28.81%
OOP	oop, image, performance, delphi	88K	212K	52K	61K	59.51%	14.20%	26.29%
MVC	asp.net-mvc, mvc	50K	98K	23K	29K	54.18%	15.06%	30.76%
Adobe	flex, flash, actionscript	39K	65K	18K	17K	73.98%	14.34%	11.68%
SCM	git, svn	34K	74K	21K	25K	44.31%	21.81%	33.88%

3.1 Topical Influence on Extrinsic and Intrinsic Motivated Actions

Table 1 reports topical knowledge demand statistics. For each topic, we include: the number of submitted questions $\#Q$, as a measure of knowledge demand popularity; the number of answers $\#A$ and the number of comments C, as a measure of community participation. Comments or answers to self-created questions are not considered as extra contributors to the topic. Results highlight great topical diversity for both popularity and participation. It also emerges a topic-dependent distribution of answers and comments, which underlines differences in the type of activities performed by contributors.

The difference is more evident when observing the right-hand side of Table 1, which analyses communities' composition. $\#AU$ and $\#CU$ respectively indicate the number answerers and the number of commenters; $\#AU \cap \#CU$ shows the percentage of contributors who are both commenters and answerers; $\#CU - \#AU$ reports the percentage of contributors that are only commenters, and $\#AU - \#AU$ the percentage of contributors that are only answerers.

The distribution of contributors across topics greatly varies. We observe a general trend towards communities where the number of users acting exclusively as answerers is higher; together with an absolute higher number of answers, these figures suggest a preference for rewarded actions. In the `LAMP` topic the trend is more evident. `iOS` and `Adobe` are exceptions, as the percentage of users that exclusively comment is higher, and the absolute numbers of comments and answers is comparable. We observe no trend related to topics's popularity or participation. For instance, `Web`, `Java`, and `Databases` have a very similar number of commenters and answerers, which are mostly overlapping. Other topics like `.Net`, `Python`, and `Ruby` show a slight predominance of answerers of comments, which is reflected in the uneven composition of contributors.

3.2 Measures of Motivations and Expertise in `StackOverflow`

Given its multi-faceted nature, user engagement has been measured in different ways: from subjective (e.g. user questionnaires) to objective (e.g. subjective perception of time) metrics, each measure is characterised by its own cost of acquisition, generalisation capabilities, and bias.

In this work we consider several objective measures. As common in related literature, we define such measures over the set of `StackOverflow` users' activities available in the public dataset. Namely: posting new questions, answers, or comments; and voting existing question and answers (although votes are only available as aggregates, not as individual actions). Our metrics focus on the 3 engagement factors described in Section 2.

Intrinsic motivations metric. The `StackOverflow` dataset does not provide page-access (view) data about individual users, thus making the task of measuring intrinsic motivations more challenging. To account for the missing data, we focus on *comments*, i.e. the only type of activity not rewarded by the scoring mechanism of `StackOverflow`. By being unrewarded, we assume commenting actions to be performed only for personal interest in a question, in its topic, or in the community. Figure 1 (a) plots the distribution of comments and answers for each user participating in the `.Net` topic. As typical in `StackOverflow` [17], user activeness brings a strong bias (0.9 correlation, $p < .01$), as most active users are also more likely to engage in discussions, or provide minor help and criticisms. To compensate for the activeness bias, we use as intrinsic motivation measure $IM_u = \frac{\#C_u}{\#A_u}$, defined as the ratio between the number of comments and the number of answers provided by a user for a given set of topics. Intuitively, IN_u provides a measure of intrinsic motivation by quantifying the self-driven likelihood of a user to contribute to an ongoing discussion. Figure 1 (b) plots the distribution of IN_u, showing its independence from user activeness (0.0 correlation, $p < .01$).

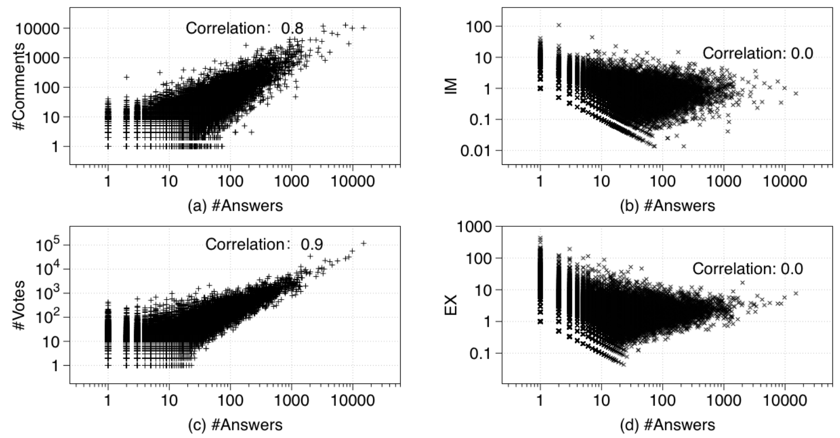

Fig. 1: Distribution of number of comments and votes in the `.Net` without – (a) and (c) – and with – (b) and (d) – activeness correction.

Extrinsic motivations. External rewards such as reputation score and badges are strictly correlated with user activeness [17] which, in turns, is linearly correlated with the number of provided answers. We therefore use as a measure of extrinsic motivation $EM_u = \#A_u$, i.e. the number of answers provided by a user for questions about a a given set of topics, in a given time frame.

Expertise. In `StackOverflow`, social judgment in expressed in terms of votes assigned to questions or answers provided by users. The number of votes received by other users can be used as a measure of expertise. As for intrinsic motivations, most active users are also more likely to receive more votes for their contributions, as can be seen from Figure 1 (c) ($p < 0.01$). To normalise for user activeness, we use as expertise metric $EX_u = \frac{\#V_u}{\#A_u}$, i.e. the average number of votes received for each answer. Figure 1 (d) plots the distribution of EX_u ($p < 0.01$) for each user in the dataset.

3.3 Topical Relation Of Extrinsic Motivation, Extrinsic Motivation, and Expertise

Table 2 reports, for each topic and engagement metric defined in Section 3.2, the mean value (μ), standard deviation (σ) and skewness (γ) of their distributions.

The distributions of users' expertise (EX_u), intrinsic (IM_u) and extrinsic (EM_u) motivations is topically diverse. With metrics of motivation, a general trend can be observed: the averaged EM_u value for very popular topics (e.g., `.Net`) is higher than less popular ones (e.g., `SCM`), while the averaged value of IM_u is lower, meaning that users of these topics are, on average, active in providing answers to gain reputation while less self-interested participating to unrewarded activities. All metrics features very skewed distributions, especially EM_u: this indicates a general trend towards the identification of a small group of users possessing high motivation and/or expertise.

Table 2: Distribution and correlation of IM_u, EM_u, and EX_u values across topics.

Topic	Basic Statistics ($\mu \pm \sigma, \gamma$) EX_u	IM_u	EM_u	Pearson Correlation EX_u-IM_u	EX_u-EM_u	IM_u-EM_u
.Net	1.65±4.73, 27.25	0.36±0.99, 18.57	10.23±84.17, 86.58	.03($p < .01$)	.02($p < .01$)	.05($p < .01$)
Web	2.06±9.55, 47.37	0.37±9.02, 7.01	7.90±52.35, 40.82	.02($p < .01$)	.00($p = .06$)	.08($p < .01$)
Java	2.30±9.69, 39.92	0.37±1.08, 21.85	8.10±64.61, 64.95	.01($p < .01$)	.00($p = .24$)	.01($p < .01$)
LAMP	1.59±6.09, 53.08	0.41±1.00, 10.42	7.25±44.05, 41.56	.02($p < .01$)	.01($p < .01$)	.08($p < .01$)
C/C++	2.01±6.76, 40.35	0.47±1.17, 9.30	8.41±58.29, 28.47	.03($p < .01$)	.02($p < .01$)	.09($p < .01$)
iOS	2.44±7.81, 18.98	0.38±1.08, 12.65	7.70±39.38, 19.11	.01($p < .01$)	.00($p = .35$)	.05($p < .01$)
Databases	1.69±7.11, 52.15	0.43±0.99, 6.57	5.53±40.94, 41.82	.01($p < .01$)	.01($p = .02$)	.05($p < .01$)
Python	2.43±6.68, 21.68	0.45±1.02, 6.52	6.44±75.77, 48.15	.03($p < .01$)	.02($p < .01$)	.06($p < .01$)
Ruby	2.68±8.44, 24.67	0.37±0.95, 8.27	5.88±27.23, 21.73	.02($p < .01$)	.01($p = .19$)	.06($p < .01$)
String	2.32±10.69, 51.44	0.57±1.21, 6.52	4.65±25.14, 31.61	.01($p < .01$)	.01($p < .01$)	.06($p < .01$)
OOP	2.18±7.54, 30.51	0.54±1.23, 8.18	3.47±16.43, 50.01	.04($p < .01$)	.02($p < .01$)	.07($p < .01$)
MVC	2.08±6.13, 22.34	0.40±0.95, 5.60	3.79±34.27, 132.39	.02($p < .01$)	.01($p = .19$)	.02($p < .01$)
Adobe	1.28±6.88, 77.62	0.24±0.71, 6.41	3.68±19.45, 36.33	.02($p = .03$)	.00($p = .97$)	.07($p < .01$)
SCM	5.51±28.48, 22.99	0.41±0.98, 6.01	2.99±23.64, 87.91	.01($p = .06$)	.00($p = .60$)	.03($p < .01$)

To investigate the relation between engagement factors, for each topic we consider the list of contributing users; we calculate their topical IM_u, EM_u, and EX_u values, and evaluate the pairwise Pearson correlation. Results are reported at the right-hand side of Table 2. Correlation is generally very low, mostly at high level of significance ($p < .01$). Overall, this result validates our choice of measures: in the reference dataset, the three engagement factors are independently observable. $IM_u - EM_u$ correlation is more evident, although still very diverse across topics (e.g. 0.09 in `C/C++`, 0.01 in `Java`). Interestingly, the (low) correlation between extrinsic motivation and expertise is highly not significant in 5 topics. We interpret such lack of statistical support as the result of more homogeneous expertise distributions among very active community members. The phenomenon affects topics at varying levels of popularity and participation, so community size doesn't appear to be a relevant factor. Further investigations are left to future work.

4 Exploiting Extrinsic Motivations, Intrinsic Motivations, and Expertise for Question Routing Optimisation

In this section we provide empirical evidence of how (intrinsic and extrinsic) motivations and expertise can be exploited to improve the knowledge creation process. We employ question routing (i.e. recommendation of questions to the most suitable answerers) as knowledge acceleration mechanism, and compare the performance of different routing model configurations.

4.1 Data Preprocessing and Analysis

We split the dataset into two partitions. We build user profiles by considering actions executed up to Dec 31, 2012; we refer to this data partitions as the

`Training` partition. Routing performance on question-answering are assessed of the `Testing` partition, which includes 1 year worth of user actions (from Jan 19^{th} 2014). To avoid cold-start problems, we consider only users that performed at least one action in both partitions. As our assessment includes a comparison of answerers rankings, we include in our experiment only questions with at least two answerers. Table 3 reports the resulting dataset figures.

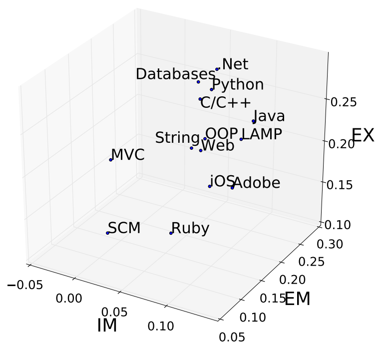

Fig. 2: Pearson correlation of (intrinsic and extrinsic) motivations and expertise w.r.t. answer quality across topics.

Table 3: Users and questions distributions in the `Training`, `Validation`, `Testing` dataset partitions.

	`Train`	`Test`	`Valid`	
Topic	**#U**	**#Q**	**#U**	**#Q**
.Net	10,118	37,641	29,357	156,512
Web	13,877	51,267	34,034	180,230
Java	11,679	46,568	40,287	197,688
LAMP	11,305	35,079	35,070	149,487
C/C++	6,114	31,255	19,248	94,409
iOS	4,218	14,508	13,725	70,114
Databases	4,794	16,011	17,488	53,489
Python	4,988	18,380	15,227	55,546
Ruby	2,477	6,640	8,802	30,390
String	4,898	12,805	16,526	39,074
OOP	3,256	4,500	14,059	21,127
MVC	1,435	2,077	5,622	10,613
Adobe	182	267	1,649	4,703
SCM	814	1,546	3,871	7,275

Our working hypothesis is that, by properly weighting different answerers according to their likelihood of being relevant to a given questions, the accuracy of question routing can be optimised. We test how the application of engagement factors in such weighting can lead to better question recommendation performance. To support our hypothesis, we first conduct the following experiment. For each question in the `Testing` set, we order answerers according to the number of votes they received from the community, and evaluate their intrinsic motivations, extrinsic motivations, and expertise measures over the `Training` set. We then calculate the Pearson rank correlations between the answering quality $AQ = \#votes$ and each of the three engagement factors (IM, EM, EX) : results are depicted in Figure 2. Each dot is a topic, and its coordinate indicate the respective correlations.

A higher correlation implies that the corresponding measure is more predictive for answering quality. The plot shows how, in general, intrinsic motivation is a poor predictor of answer quality, while EX and EM are more correlated, although often in a complementary fashion (e.g. `iOS`, `Adobe`). We observe great topical variety in the predictive power of the three engagement features. For instance, expertise in `Java` are more predictive than in `iOS`, while intrinsic motivation for these two topics are similar. Such a diversity calls for a routing model that weights the contributions of engagement properties differently across topics.

4.2 Routing Model

We propose a linear model, defined as follows:

$$S(u, q) = \alpha^t_{IM} IM_u + \alpha^t_{EM} EM_u + \alpha^t_{EX} EX_u,$$

For each question q of topic t in the `Testing` partition, $S(u, q)$ scores the answerer u's answer quality. $\alpha^t_{IM}, \alpha^t_{EM}, \alpha^t_{EX}$ respectively model the topic-specific needs for intrinsic motivated, extrinsic motivated, and expert answerers.

The optimal, topic-specific values for the $\alpha^t_{IM}, \alpha^t_{EM}, \alpha^t_{EX}$ parameters are calculated as follows. We identify a third dataset partition, called `Validation`, defined over the original, unfiltered dataset, and containing two years worth (from Jan 01, 2011 to Dec 31, 2013). Table 3 provides a basic description of the `Validation` partition, where the number of questions higher due to the lack of filtering conditions. We use the Linear Ordinal Regression SVM with L_2 regularization to learn the parameters, such that the profiled users are optimally ranked in the `Validation` partition. To learn more accurate parameters, we exclude the answer pairs in the training phase if the difference of #votes to the answers is less than 2. Such parameters are then used in the routing model to recommend questions to users in the `Testing` partition.

4.3 Experimental Setup

Evaluation Metrics The routing performance are assessed with three metrics, commonly used in the evaluation of recommender systems: NDCG (normalized discounted cumulative gain) [10], Kendall Tau, and Pearson rank correlation coefficients. The goal is to measure the quality/correlation of the recommended list of potential answerers by comparing it to the ground truth.

The evaluation of NDCG is performed against the #votes received by an answerer in a question. We use **NDCG@1** to assess the quality of the best recommended answerer, while **NDCG** assess the overall quality of the answerer set ranking. Due to the presence of negative voted answers, we exclude from the evaluation questions where the sum of DCGs is negative. Pearson correlation is calculated against #votes to answers, while Kendall Tau only measures similarity in the relative order of answers. Correlation is calculated only for questions where at least one answerer has a unique number of votes in the answer set.

Experimental Configuration. We compare the performance of 5 routing configurations. In the `Rdm` configurations, we randomly order the original answerers in the tested question. This configuration provides a performance baseline, as it measure a purely casual recommendation strategy. In the `Exp`, `Int`, and `Ext` configurations we respectively configure the routing system to return answerers according to their EX, IM and EM scores. These configurations simulate a recommendation strategy based on a single feature of engagement. Finally, `Cmb` applies the routing model described in Section 4.2, using the topic-specific learned parameters. As a remark, we exclude content-based model (e.g., bag-of-words of user answers) since our preliminary experiment show that it is less effective than configurations (e.g., `Ext`) that measure user answering activities.

4.4 Results

Table 4 summaries the results of our experiment. As expected, the topic of interest is an important performance diversification element for all the considered engagement factors and evaluation metrics. W.r.t. the numbers reported in Table 4, it is important to highlight how the range of values for NDCG metrics is necessarily narrower than for Pearson and Kendall Tau correlations. This is due to the definition of the metric which, by considering the number of votes received by an answerer, compress results in a more compact spectrum of values[1]. This is also demonstrated by the considerably high performance obtained by the `Rdm` configuration. Therefore, minor variations in NDCG values entails relevant differences in the quality of the returned answerers list.

As expected, among the configurations of `Int`, `Exp` and `Ext`, `Int` is the one providing worse results, whereas `Exp` configuration usually performs better than the others. On the other hand, we observe that `Cmb` configuration has in general performance better than or comparable with `Exp`. Small improvements, however, can provide tangible impacts. For instance, in topics such as `Web` and `iOS`, `Cmb` achieves better rankings – for 833 and 167 questions respectively. For some topics `Exp` could give even slightly better result than `Cmb` configuration, e.g., `Ruby, OOP`. Results suggests that the `Cmb` configuration could leverage different user engagement factors for question routing; however, is many topics, expertise is the most important factor for recommendation quality.

As a final remark, we highlight how the routing performance of `Cmb`, `Exp` is generally higher than the ones reported in related literature [18]. This is despite the different targeted dataset, which is more extensive in our setting.

	NDCG@1					NDCG					Pearson					Kendall				
Topic	Rdm	Exp	Int	Ext	Cmb	Rdm	Exp	Int	Ext	Cmb	Rdm	Exp	Int	Ext	Cmb	Rdm	Exp	Int	Ext	Cmb
.Net	.572	.687	.589	.676	**.693**	.834	.882	.842	.877	**.884**	.015	.279	.055	.244	**.290**	.014	.266	.054	.231	**.275**
Web	.578	.679	.624	.679	**.689**	.838	.879	.857	.878	**.883**	-.004	.234	.104	.225	**.255**	-.003	.226	.100	.217	**.245**
Java	.572	.665	.602	.647	**.666**	.835	**.873**	.847	.865	**.873**	.007	**.220**	.067	.169	.219	.005	**.210**	.064	.162	.209
LAMP	.579	.675	.602	.664	**.677**	.839	.877	.848	.873	**.878**	-.004	.219	.044	.193	**.228**	-.004	.212	.043	.187	**.220**
C/C++	.568	**.673**	.589	.644	.663	.834	**.878**	.843	.865	.874	.013	**.256**	.056	.181	.233	.015	**.244**	.052	.174	.222
iOS	.569	.644	.605	.650	**.658**	.835	.867	.850	.868	**.871**	-.002	.175	.080	.173	**.204**	.000	.171	.075	.169	**.197**
Databases	.593	.700	.614	.694	**.704**	.847	.889	.855	.886	**.890**	.001	.254	.037	.232	**.259**	.002	.248	.036	.228	**.254**
Python	.582	.682	.605	.684	**.695**	.842	.882	.851	.882	**.887**	.005	.244	.054	.236	**.265**	.005	.235	.052	.229	**.255**
Ruby	.607	**.656**	.628	.651	.651	.853	**.872**	.861	.870	.871	.016	**.141**	.073	.119	.130	.015	**.138**	.071	.119	.130
String	.572	.660	.601	.656	**.663**	.837	.874	.850	.872	**.875**	-.013	.200	.056	.171	**.206**	-.013	.192	.058	.165	**.196**
OOP	.578	**.682**	.614	.672	.680	.840	**.883**	.855	.879	.881	-.011	**.231**	.067	.185	.228	-.005	**.224**	.065	.185	.220
MVC	.623	.692	.613	.697	**.699**	.860	.888	.857	.890	**.890**	.034	.194	-.027	.193	**.214**	.034	.193	-.020	.199	**.208**
Adobe	.60	.663	.654	.649	**.674**	.853	.873	.872	.872	**.879**	-.013	.174	.96	.113	**.184**	-.009	.174	.97	.115	**.186**
SCM	.598	**.663**	.621	.650	.647	.853	**.875**	.861	.873	.871	-.038	**.101**	.010	.078	.083	-.044	**.100**	.008	.079	.081

Table 4: Experiment results of question routing with different configurations. Numbers in bold are the highest among all configurations.

[1] **NDCG** = 1 entails a perfect recommendation.

5 Related Work

This section positions our paper in the context of previous work related to user engagement and knowledge creation acceleration in CQA systems.

Although both are factors of user engagement, user motivations and expertise in CQAs have been typically discussed in isolation. In a qualitative study based on interviews with CQA users, [12] finds that altruism, learning and competency are frequent motivations for participation. In addition, previous research shows gamification mechanisms can largely influence users' behaviours [4, 2]. Expertise, on the other hand, is mostly studied in the problem expertise identification. Related work typically adopts indicator-based methods such as Z_{score} [19], or graph-based methods such as the adapted PageRank method [11]. A recent study [17] shows how existing metrics of expertise can be heavily biased toward most active users. The normalisation of user activeness in our EX_u metric is inspired by such consideration. Combining expertise and motivation, [13] explores their effect in the specific task of expert finding. W.r.t. literature our work further the understanding of user engagement factors in CQAs, providing new insights about the interplay of user expertise, intrinsic motivation, and extrinsic motivation. We contribute an original and extensive analysis that shows the independent manifestation of these three engagement factors across topical communities.

Knowledge creation acceleration is a topic recently emerged in research related to CQA systems. Typical methods include automatically detection of question quality [15], editing suggestions for poorly formulated questions [16], and active routing of questions to potentially relevant answerers [8, 14, 5, 18]. The latter is the most popular technique, and is the inspiration for our experiment. Previous work typically considers only users' topical activeness[8] as user modelling feature. These works were extended by considering the problem of routing question to a user community, for collaborative problem solving [14, 5]. More recently, [18] proposes a question routing user model that includes expertise, providing empirical evidences of its contribution to performance improvement. To the best of our knowledge, our work is the first considering a broader spectrum of engagement factors, and we extensively demonstrate their applicability.

6 Conclusions

The main mechanisms that drive knowledge creation process in CQAs are still to be fully uncovered. In this paper we address the problem of characterising and measuring three engagement factors in `StackOverflow`. The rationale behind our work is simple: to drive participation, thus improving the quality and speed of knowledge creations, we need to better understand the driving forces behind user engagement. Inspired by engagement theory from literature, we focus on intrinsic motivations, extrinsic motivations, and expertise. We propose three metrics, defined over the set of actions available to `StackOverflow` users, and we show how *topic* plays a major role in influencing them. We investigate the relations that exist among these three engagement factors, and demonstrate their independent and decomposable nature. A question routing optimisation experiment confirms the relevant role that engagement can play in knowledge creation acceleration.

Acknowledgment. This publication was supported by the Dutch national program COMMIT. This work was carried out on the Dutch national e-infrastructure with the support of SURF Foundation.

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: KDD'12, pp. 850–858. ACM (2012)
2. Anderson, A., Huttenlocher, D., Kleinberg, J., Leskovec, J.: Steering User Behavior with Badges. In: WWW'13, pp. 95–106. ACM (2013)
3. Attfield, S., Kazai, G., Lalmas, M., Piwowarski, B.: Towards a Science of User Engagement. In: WSDM Workshop on User Modelling for Web Applications (2011)
4. Bosu, A., Corley, C.S., Heaton, D., Chatterji, D., Carver, J.C., Kraft, N.A.: Building Reputation in Stackoverflow: an Empirical Investigation. In: MSR'13, pp. 89–92. IEEE (2013)
5. Chang, S., Pal, A.: Routing Questions for Collaborative Answering in Community Question Answering. In: ASONAM'13, pp. 494–501. IEEE/ACM (2013)
6. Deci, E., Ryan, R.M.: Self-determination Theory. Handbook of Theories of Social Psychology, 1:416–433 (2008)
7. Ericsson, K.A., Charness, N., Feltovich, P.J., Hoffman, R.R.: The Cambridge Handbook of Expertise and Expert Performance. Cambridge University Press (2006)
8. Guo, J., Xu, S., Bao, S., Yu, Y.: Tapping on the Potential of Q&A Community by Recommending Answer Providers. In: CIKM'08, pp. 921–930. ACM (2008)
9. Guy, I., Avraham, U., Carmel, D., Ur, S., Jacovi, M., Ronen, I.: Mining Expertise and Interests from Social Media. In: WWW'13, pp. 515–526. ACM (2013)
10. Järvelin, K., Kekäläinen, J.: Cumulated Gain-based Evaluation of IR Techniques. ACM Transactions on Information Systems, 20(4):422–446 (2002)
11. Liu, J., Song, Y.-I., Lin, C.-Y.: Competition-based User Expertise Score Estimation. In: SIGIR'11, pp. 425–434. ACM (2011)
12. Nam, K.K., Ackerman, M.S., Adamic, L.A.: Questions in, Knowledge in?: a Study of Naver's Question Answering Community. In: SIGCHI'09, pp. 779–788. ACM (2009)
13. Pal, A., Farzan, R., Konstan, J.A., Kraut, R.E.: Early Detection of Potential Experts in Question Answering Communities. In: UMAP'11, pp. 231–242. Springer (2011)
14. Pal, A., Wang, F., Zhou, M.X., Nichols, J., Smith, B.A.: Question Routing to User Communities. In: CIKM'13, pp. 2357–2362. ACM (2013)
15. Ravi, S., Pang, B., Rastogi, V., Kumar, R.: Great Question! Question Quality in Community Q&A. In: ICWSM'14, pp. 426—435. AAAI (2014)
16. Yang, J., Hauff, C., Bozzon, A.., Houben, G.J.: Asking the Right Question in Collaborative Q&A Systems. Hypertext'14, pp. 179–189. ACM (2014)
17. Yang, J., Tao, K., Bozzon, A.., Houben, G.J.: Sparrows and Owls: Characterisation of Expert Behaviour in Stackoverflow. In: UMAP'14, pp. 266–277. Springer (2014)
18. Yang, L., Qiu, M., Gottipati, S., Zhu, F., Jiang, J., Sun, H., Chen, Z.: CQARank: Jointly Model Topics and Expertise in Community Question Answering. In: CIKM'13, pp. 99–108. ACM (2013)
19. Zhang, J., Ackerman, M.S., Adamic, L.: Expertise Networks in Online Communities: Structure and Algorithms. In: WWW'07, pp. 221–230. ACM (2007)