6-18-2022

EFFECTS OF LABEL USAGE ON QUESTION LIFECYCLE IN Q&A COMMUNITY

Alyssa Shuang Sha
Australian National University, alyssa.sha@anu.edu.au

Armin Haller
The Australian National University, armin.haller@anu.edu.au

Yingnan Shi
NU A, yingnan.shi@anu.edu.au

EFFECTS OF LABEL USAGE ON QUESTION LIFECYCLES IN Q&A COMMUNITY

Research Paper

Alyssa Shuang Sha, Australian National University, Canberra, Australia, alyssa.sha@anu.edu.au

Yingnan Shi, Australian National University, Canberra, Australia, yingnan.shi@anu.edu.au

Armin Haller, Australian National University, Canberra, Australia, armin.haller@anu.edu.au

Abstract

Community question answering (CQA) sites have developed into vast collections of valuable knowledge. Questions, as CQA's central component, go through several phases after they are posted, which are often referred to as the questions' lifecycle or questions' lifespan. Different questions have different lifecycles, which are closely linked to the topics of the questions that can be determined by their attached labels. We conduct an empirical analysis based on the dynamic panel data of a Q&A website and propose a framework for explaining the time sensitivity of topic labels. By applying a Discrete Fourier Transform and a Knee point detection method, we demonstrate the existence of three broad label clusters based on their recurring features and four common question lifecycle patterns. We further prove that the lifecycles of questions in disparate clusters vary significantly. The findings support our hypothesis that questions with more time-sensitive labels are more likely to hit their saturation point sooner than questions with less time-sensitive labels. The research results could be applied for better CQA interface design and more efficient digital resources management.

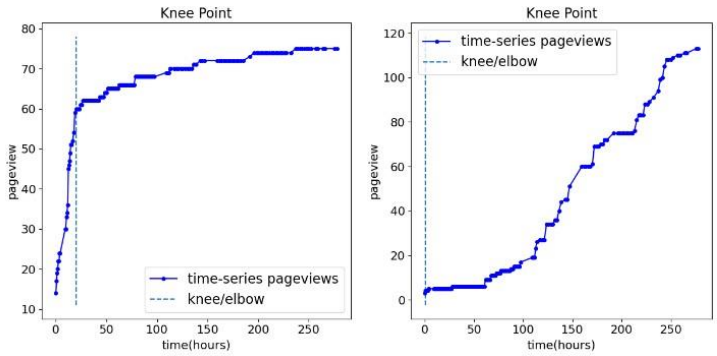
Keywords: Question-answering community, question lifecycle, topic classification, topic recurring pattern.

1 Introduction

After a question is posted on a Q&A site, how many phases will it go through? This is a well-known open research question pertaining to Community Question Answering (CQA) and collectively to online behaviour (Maity *et al.*, 2015; Yu *et al.*, 2015). CQAs, such as Yahoo! Answers, Stack Overflow, Quora and Zhihu, are forums for users to pose and answer questions. They have become a significant source of knowledge for online knowledge seekers. CQAs not only provide a platform for experts to share their insights and receive recognition, but they also assist new users in successfully solving a specific problem or get an answer to a question (DeVaro *et al.*, 2018; Huna *et al.*, 2016; Liu *et al.*, 2011; Pedro and Karatzoglou, 2014; Roy *et al.*, 2018; Yao *et al.*, 2015). The information provided in such sharing environments goes through several lifecycles over time. According to early CQA research, each question has a lifecycle as follows; it begins in an "open" state where it receives responses and users' attention; followed by a "closed" stage where no further responses are received (Agichtein *et al.*, 2008). Anderson et al. (2012) defined a two-phased question lifecycle on Stack Overflow that includes a growth phase during which most of the responding and voting occurs, followed by a saturation phase where the number of responses plateau over a long period of time. It is also claimed that the relatively stable plateau can potentially provide a useful public resource for future would-be questioners. The lifecycles of individual questions, however, can differ significantly depending on the topic of the question.

As an illustration of the different lifecycles of questions in a CQA system, we give an example of the short-term question lifecycles of two different questions on Zhihu, the largest CQA platform in China (Figure1), which shows the remarkable differences in the pageview trends of questions under different

topics. Figure 1 presents two different patterns respectively: a concave increasing curve and a convex increasing curve.



(a) (b)

Figure 1. The lifecycle of different questions on Zhihu: (a) The concave pageview trend of a policy changing question (Question ID = 447782565). (b) The convex pageview trend of a premenstrual syndrome question (Question ID = 447782641).

To investigate the lifecycles of questions under various topics, we use the labels attached by the questioner to capture and classify the question topics. Topic labels play a pivotal role in CQA sites, including but not limited to the following aspects (Nie *et al.*, 2020): (1) Question routing. In addition to the unidirectional user-follower relations, in CQA sites, users can also follow the topics of interest. As such, CQA sites can put the questions into the feeds of associated topic followers to draw more attention from potential answerers, and thus receive quicker and more accurate answers. (2) Topic labels can be leveraged to benefit index, search, navigation, and organization. Therefore, question tagging in CQA sites deserves researchers' attention.

When a user asks a question in a CQA site, the user must normally select a category label from a predefined hierarchy of categories. As a result, each question in a CQA archive is assigned one or more category labels, and questions in CQA sites are categorised into hierarchies (Cao *et al.*, 2009). For example, in Zhihu, topic labels are organised into a directed acyclic graph (DAG) by experienced users and hired experts, as shown in Figure 2. The DAG can be converted to a tree structure except that some nodes have multiple parents. From the root-to-leaf nodes, topic tags tend to be more specific. A question can be annotated by either the leaf or the internal nodes at the same time.

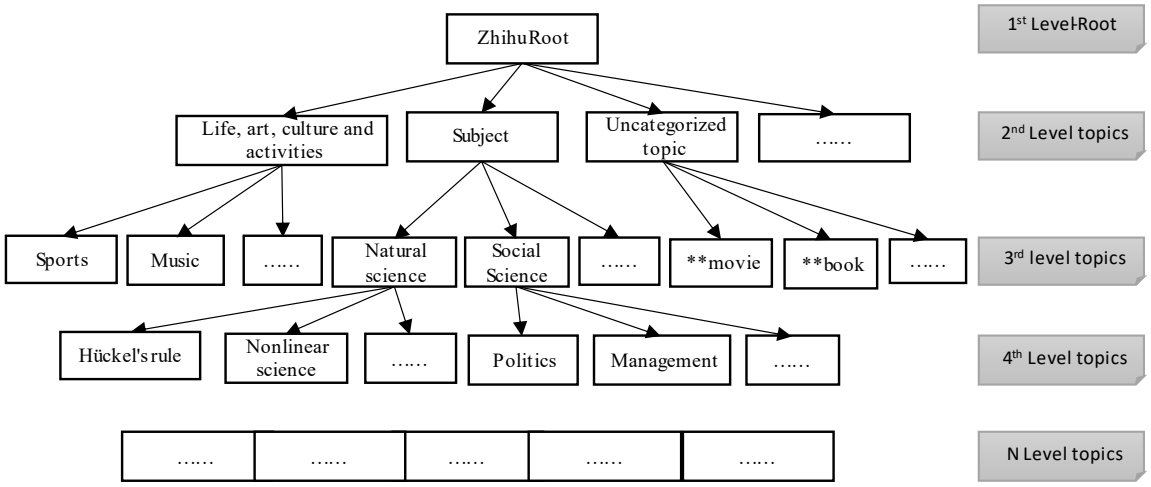


Figure 2. Category structure in Zhihu

The explanation why questions' lifecycles have different patterns is an intriguing topic that is related to users' engagement. However, very few studies have illustrated the relationship between time sensitivity of the label and the lifecycle of questions that are posted under this label. In this paper, we focus on studying the effect of one crucial aspect of user behaviour - i.e., the effect of a label on a questions' lifecycle. As has been discovered in Zhihu, the effect of label usage can represent the effect of the topic on a question's lifecycle. By demonstrating the relationship between label usage and the questions' exposure rate (traffic growth rate), we then elaborate on the indirect influence of topic classification on the questions' lifecycles.

This research provides multiple significant theoretical and practical contributions to the study on question lifecycle and label clustering. First and foremost, this research addresses the ambiguous relationship between label usage and the lifespan of questions. Aside from the two patterns proposed by Crane & Scrnette (2008), we propose other recurrent online content lifecycle patterns and extend them to CQA to prove and further replenish this model. Furthermore, we propose a broad label clustering framework according to labels' recurring periods and recurring strengths. We present the time-sensitivity of labels using the Discrete Fourier Transform (DFT). DFT can be adopted to understand different dynamic types of question topics and perform deeper analysis on various research domains about online content lifecycles. The findings of this study can also be used to predict the long-term value of online content and user behaviour on various social media websites. Specifically, the outcomes of this study can assist the CQA platform operators to optimise the user experience by developing an immediate feedback feature that suggests the predicted content lifespan based on the labels attached to the post. Furthermore, as a critical component of Information Lifecycle Management (ILM), digital information classification has a number of difficulties that need to be addressed. This study tackling the information classification problem in CQAs by providing specific question classification method according to their lifecycles and topic time sensitiveness. The research findings could help platform management improve the data retention strategy and better manage the ageing, archival, and disposal of information to release more digital resources.

The remainder of this paper is structured as follows: Section 2 summarises related work and develops our key hypotheses. Section 3 provides an overview of our dataset, data collection process, and data processing methodology. Section 4 and 5 present our findings and contributions. Section 6 summarises the study and addresses its limitations and potential directions for future works.

2 Related Work & Hypotheses development

2.1 Information lifecycle management

The information in organizations including CQA platforms has been rapidly growing as a result of increased computerised systems, archiving and privacy rules, and customer support applications (Gantz and Reinsel, 2007). This brings upon many undesired consequences, for instance, limited digital resources and information overload tend to obscure useful information from users. Therefore, conservation of Web resources by archiving data to lower the cost of information storage is becoming a critical issue for digital platforms (Pennock, 2007; Waddington *et al.*, 2012). The rising demand for data protection and retention, as well as the prevention of data explosion brought by information growth, has pushed many digital organizations to deploy sophisticated information management strategies (Arutyunov, 2012; Hayes, 2008; Sabah, 2008).

Recently, Information Lifecycle Management (ILM), which is a cost-effective strategy for preserving information assets, has garnered significant attention as a solution to the data overload problem (Al-Fedaghi, 2013). The objective of ILM is to achieve digital resource allocation by matching the organization's storage capacity with active processing such as access times and frequency, which are highly related to the page traffic in CQAs.

The important aspect of ILM is "valuing" information, which refers to evaluating the relative importance of information. However, information on websites has varying values; certain information may be more valuable than others and decisions have to be made as to which information to preserve and which to

discard (Kraemer et al., 2009). For example, if a piece of information is evaluated to be more important than it is, platform maintenance expenses may be wasted. On the other hand, it is also important not to under-evaluate information, since information designated as an asset should be classified and protected according to its value and importance to the organization (Bergström and Åhlfeldt, 2014). Hence, ILM has to be implemented with caution.

To tackle this information classification problem, it is critical to assign time-dependent values to different types of contents according to their status in their lifecycles (Al-Fedaghi, 2013). Information might be in multiple stages, such as in use or archived, with different values at different times. Because information classification can change over time, reclassifying information to maintain an up-to-date classification is another issue raised by previous studies (Bunker, 2012; Fibikova and Müller, 2011; Virtanen, 2001). Thus, from the ILM standpoint, it is critical to concentrate on dynamic information classification tasks that can be implemented in different types of online communities.

An early study has already pointed out that more real-word ILM examples are needed in this research field (Bergström and Åhlfeldt, 2014). Nevertheless, much of the existing literature is focused on theoretical work, for instance, the information flow model (Sabah, 2008), information security policies (Bergström et al., 2019), and national models (Oscarson and Karlsson, 2009). According to the results of a survey on the underlying approaches of information classification practices, information classification policies need to be more precise and provide more actionable recommendations on how information lifecycle management is implemented in practice (Bergström and Anteryd, 2018). Some empirical studies have attempted to address the information classification task from an ILM perspective (Bergström et al., 2021; Büsch et al., 2017), but they were limited to intra-organizational information, and did not investigate online communities.

In the context of CQAs, there is no relevant research that addresses the information classification task to improve the effectiveness of ILM. An effective method for dynamic information classification in CQAs remains unknown. To bridge this research gap, our study aims to incorporate an ILM perspective to CQAs-related research and classify different types of content according to their lifecycle status, which can be utilised to optimise data retention policies and digital resource management.

2.2 Question Lifecycle

Previous research has discussed the value of utilizing a lifecycle strategy to assist managing digital information: different content types have varied lifespans, which are best managed by understanding their lifespans (or lifecycles) first (Rusbridge *et al.*, 2005; Waddington *et al.*, 2012). The typical social media lifecycle is strikingly comparable to the lifecycle in classic product adoption and lifecycle theory. The earliest systematic product lifecycle theory can be traced back to the four-stage lifecycle that Levitt (1965) put forward. He divided the product lifecycle into four stages: market development, growth, maturity and decline, and mentioned that the lifecycle concept can be effectively employed in the strategy of both existing and new products. Online content, as a product of websites or platforms, undoubtedly has its own lifecycle in today's highly established network media landscape.

The 90/90 data-use theory (Efraim *et al.*, 2017) states that the vast majority of stored data, up to 90%, is rarely accessed after 90 days. In other words, data loses a significant amount of its value after being available for more than three months. Social media content (e.g., posts, video, and articles) composed of numerous data follows the same principle. Thus, we pose the following hypothesis:

H1: Posts on Q&A sites adhere to the 90/90 data-use theory, which states that questions receive the majority of users' attention during a brief period after being posted.

Other social media models that observe online content popularity over time include power-law precursory growth and power-law relaxations (Crane and Sornette, 2008). Castillo et al. (2014) proposed a similar "80:10:10" rule based on the study of News articles that stated that for the first 12 hours, traffic to 80% of articles decreases monotonically, traffic to 10% of articles does not decrease, and traffic to the remaining 10% of articles decreases initially but then recovers.

The half-life of articles is shown to be power law spread over a wide range, with a mean of 36 hours (Dezsö *et al.*, 2006). Yu et al. (2015) enhanced this model by categorizing various videos into multiple

phases of popularity growth or decline over a specific period. Other research about media content lifecycle mentions that user activities in a social information network have a highly skewed distribution. This is referred to as the 90-9-1 law of participation discrimination (Xie and Sundaram, 2012). In most online groups, 90% of users are lurkers who never contribute, 9% of users contribute a bit, and 1% of users account for almost all activities (Nielsen, 2006). Proportion distribution varies across social media, but the basic principle applies to most. The same principle is explained by long-tail theory in CQA studies (Coelho and Mendes, 2019; Gu *et al.*, 2013; Han *et al.*, 2019; Taeuscher, 2019; Tucker and Zhang, 2007). These unbalanced user behaviors lead to different traffic stages after content is published.

A Q&A site's central components are the questions. The actions users perform on the Zhihu platform after reading the questions include browsing, posting answers, adding comments and following the questions. All of them can be classified as the questions' digital popularity (Sha *et al.*, 2020). By recording the questions' digital popularity after it is posted, we can track the stages that the question is going through which is considered as the questions' lifespan or lifecycle in our study. In this study, question lifecycle refers to the process of answering, browsing and subscribing to new questions from the point in time when they are first uploaded.

On Yahoo! Answers, either the asker or other users voted on the "best answer"; until the best answer was chosen, the question was considered "resolved" (Agichtein *et al.*, 2008). A general question lifecycle is similar to the product lifecycle and consists of four stages: an introduction period, a growth period, a maturity period, and a recession period (Liu *et al.*, 2020). Anderson et al. (2012) proposed a two-phase question lifecycle in CQAs: a "fast" phase during which the question receives responses and votes, and a "slow" phase during which members of the group indicate the question's longer-term value. According to their analysis conducted on Stack Overflow, the majority of responses and votes on both questions and answers occur during the first day after the question is posted. State-of-the-art studies on a questions' lifespan have been primarily concerned with a single broad pattern. It is still unknown how questions classified under different label clusters drive users' attention differently. Thus, we extend on previous research and raise the following hypotheses:

H2a: There exist multiple common types of question lifecycles in CQA.

H2b: Questions on Zhihu follow a rule similar to the 80:10:10 rule, with the majority of questions following a 2-phase lifecycle, while the remaining questions obey some other lifecycle patterns.

2.3 Online content topic classification

Previous research about online user behaviour in CQAs (e.g., visiting, voting, and sharing) has regularly established diverse groups of temporal trends. These groups can be broadly classified according to the presence or lack of distinct "peaks" of activity and the sum of activity immediately preceding and after the peak (Crane and Sornette, 2008; Lehmann *et al.*, 2012).

Crane & Sornette (2008) apply an epidemic spreading model on YouTube and divide the burst activities into 'exogenous' and 'endogenous' groups. They also define groups of online video visitation patterns and propose models that are associated with social network dissemination phenomena. Lehmann et al. (2012) expand these groups by demonstrating that for Twitter "hashtags" (user-defined topics), the distributions of behaviour through time intervals (before/during/after) generate distinct clusters of activities that can be understood in terms of a hashtag's semantics. Romero et al. (2011) discuss the relationship between manually allocated groups of hashtags and the various shapes of the exposure curve.

Multiple types of media content may provide a range of different exposure curves. Yang and Leskovec (2011) define six distinct types of temporal attention patterns. Attention is quantified in terms of the number of times a given expression appears in relation to an occurrence. The trends illustrate how traffic is distributed over time, as well as the order in different media content (professional blogs, news, etc.).

In general, prior research has demonstrated that the popularity evolution of various online objects is class-dependent (Figueiredo *et al.*, 2011; Yu *et al.*, 2015). The findings of Gharan and Wang (2010) indicate that some topics, such as politics and finance, are more time-sensitive than others. Mason (2011) extends this study by describing how articles' lifespans are affected by their different time-sensitive

topics. For example, business-related contents have a longer half-life on average, while posts about politics/celebrities/entertainment have a shorter half-life. Castillo et al. (2014) adopt a hybrid observation method to characterise distinct classes of articles and describe their different visit patterns. According to the studies above, the implementation of online content classification is heavily weighted toward forecasting potential user behaviour. Thus, we make the following further hypotheses:

H3a: Questions under different time-sensitive label groups have different lifespans.

H3b: Questions belonging to more time-sensitive topic categories are likely to reach their saturation points sooner than those under less time-sensitive topic categories.

2.4 Label clustering

For CQA sites, the most common way to capture questions' topics is to focus on tags/hashtags which are called labels in our study. There has been a diverse array of academic work that studies tags, most of them aiming to find relevant information by using tags to predict the popularity or information flow, or to develop clustering methods of online content (Cha *et al.*, 2010; Hong *et al.*, 2011; Naveed *et al.*, 2011; Romero *et al.*, 2011; Suh *et al.*, 2010). Other researchers investigated tags themselves, trying to analyse their dynamics, popularity, semantics and engagement (Cha *et al.*, 2010; Crane and Sornette, 2008; Lehmann *et al.*, 2012; Lin *et al.*, 2013; Shamma *et al.*, 2011; Yang and Leskovec, 2011). Our study emphasises recurring periods and recurring strengths of labels' digital popularity as a critical factor in clustering labels. The main findings from Bhat et al. (2015) indicate that tag-related factors in CQA, such as their "popularity" (how often the tag is used) and "subscribers" (how many users will answer questions containing the tag), provide significantly stronger effects on user engagement than non-tag-related factors. Thus, by referring tag-related features that were defined by prior studies, topics' digital popularity in Zhihu can be measured by counting the number of questions and views under topic labels.

Previous research has examined the dynamics of label usage and discovered that there are numerous types. According to some research, there are at least three distinct categories of dynamics: continuous activity, periodic activity, and activity clustered around a single time domain (Hsu *et al.*, 2010; Lehmann *et al.*, 2012). These studies, however, did not provide a systematic structure for demonstrating the existence of these classes or for illustrating how to classify and index tags within each class. Several other research have concentrated on defining and studying a single type of temporal pattern. "Peaky" events, such as news have been analysed and divided into up to six distinct categories (Cha *et al.*, 2010; Lin *et al.*, 2013; Shamma *et al.*, 2011). We therefore propose the following hypothesis for labels in CQA:

H4: Labels can be classified as high-time-sensitive or low-time-sensitive with varying proportions.

In conclusion, we build upon previous works on the classification of online content by examining the lifecycle of questions under different time-sensitive topics on a Q&A platform.

3 Research Methods

3.1 Dataset

Using an API (api.zhihu.com) and the Python zhihu-oauth package (Pypi, 2019), this study randomly collects question-related and their tag-related panel data for the period from 7th April, 2021 to 21th May, 2021. Questions on Zhihu can be retrieved using a Question ID (QID) as a key. Based on our observation, we found that the question IDs are always eight to nine digits long and the absolute values of the QIDs monotonically increase with time. We hence traced the QID for a set of questions that were just issued, then we adopted zhihu-oauth to track the real-time digital popularity data which includes the questions' pageviews, followers, answers, and comments every hour. For the recorded question dataset, we generated their labels' ID and plug the label-based dataset into another similar label crawler to track labels' popularity panel data.

We applied time normalization to all the panel data and calculated the standard deviation of their growth rate during each crawling interval. Note that the majority of the traffic arrived within a very short period (12 hours) after the questions were posted in CQA (Anderson *et al.*, 2012). Hence, we only focus on the

questions' lifecycles from a short-term perspective (within weeks instead of months or years) in this study. The dataset contains in total 1,575 labels (approximately 5% of the total number of existing labels on this website) and the digital popularity of the labels were traced 1,440 times for one week period. The traffic of the labelled questions were traced 2,946 times over a 16-day period.

3.2 Finding Periodic Labels

Users employ labels as a form of social annotation, to define a shared context for a specific topic. We analyse the record of Zhihu activity and find that different labels can be categorised based on the evolution of its popularity over time. Furthermore, it is crucial to detect recurring periods and recurring strengths of labels' popularity in order to measure a label's time sensitiveness. To achieve this, we performed a Discrete Fourier Transform (DFT) on the temporal evolution data of different label popularities which are tracked over time on Zhihu. This technique allows us to detect all the possible recurring frequencies of a label's popularity. The significance of the peaks in the frequency domain are then evaluated based on their relative intensity with respect to a three-sigma baseline level. Finally, we perform a categorization of the labels based on the significant recurring frequencies of their popularity.

Fourier Transform-based methods are commonly applied to a time-domain signal to identify potential periodicity in the signal (Bluestein, 1970). DFT is a variant of Fourier Transform that is used on discrete time-domain signal. The DFT technique has been commonly employed for periodicity detection in time-series panel data (Cook *et al.*, 2013; DeMasi *et al.*, 2016; Vlachos *et al.*, 2005). The DFT technique transforms a discrete sequence in time-domain x_n, where $n = 0, 1, \ldots N - 1$, into another discrete sequence in the frequency domain, X_k:

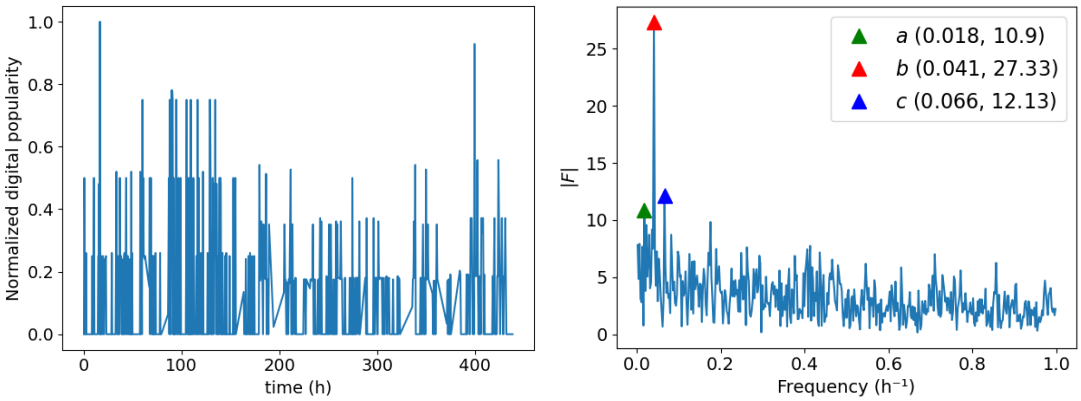
$$X_k = \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} x_n e^{-\frac{j2\pi kn}{N}}, \qquad k = 0, 1 \ldots\ldots N-1$$

where N is the sequence length, and the frequency captured by each Fourier coefficient is given by k/N. In our analysis, we sample the digital popularity of different topic labels on Zhihu as our discrete time-series x_n at a non-uniform sampling time interval T_s of approximately 30 minutes. We define the digital popularity of a particular topic label, collected at the n^{th} time-step, x_n, as the rate of increase of the number of questions tagged with the label over the sampling time interval:

$$x_n = \frac{Q_n - Q_{n-1}}{T_s};\ T_s = t_n - t_{n-1}$$

where Q_n denotes the total number of questions tagged with the label and t_n represents the real time at the n^{th} time-step, respectively.

The intensity of the entire digital popularity time-series is then normalised to between 0 and 1. Before performing a DFT on the time-series data. We also perform an interpolation (SciPy.org, 2021a) between the time-series data points to standardise the non-uniform sampling interval to 30 minutes. After performing DFT on the digital popularity time-series data (Figure 3a) with a Fast Fourier Transform package in Python (SciPy.org, 2021b), we extract the absolute magnitude of the complex Fourier coefficients $|F|$ as a function of frequency f, as shown in Figure 3(b). The $|F|$ value at a particular frequency ξ indicates the degree of periodicity of the label popularity with a period of $1/\xi$. These signals could be recurring hourly, daily, or weekly, which holds important information about the usage pattern of a particular label on Zhihu.



(a) *(b)*

Figure 3: (a) An example of the label's digital popularity time-series data (label ID = 19698303, label's translated name = "Desktop configuration"). (b) Significant peaks detection in the $|F|$ plot of the same label's digital popularity. The maximum resolvable frequency is limited by the Nyquist theorem to $1h^{-1}$, which is half of the standardised sampling rate of $2h^{-1}$.

For most of the label popularity data, the corresponding Fourier-transformed signal contains multiple frequency components. To differentiate significant signals from noise in our data, we detect all the signal peaks on the $|F|$ plot with a peak detection algorithm (SciPy.org, 2021c) and record the outliers (the peaks above the baseline $\bar{x} + 3\sigma$). For example, the Fourier coefficient plot for the label popularity, as shown in Figure 3(b) (label ID = 19698303), contains three significant peaks, labelled as peak *a*, *b* and *c* in the plot. Thus, we identify the most significant recurring signal to be peak *b*, with a recurring period of 1/0.041 ≈ 24.4 hours, which shows that the label can be classified as a daily recurring label and the topic tagged by this label can be identified as a daily chatter topic. The valid peaks as the significant signals are grouped as different clusters according to their recurring periods and recurring strengths.

As for the clustering methods, this study employs both K-Means and DBSCAN to classify question labels on a Q&A platform based on the repeating periods and strengths of their digital popularity. K-means clustering as a vector quantization technique that originated in signal processing usually used to split n observations into k clusters, with each observation belonging to the cluster with the closest mean (cluster centres or cluster centroid), which serves as the cluster prototype (MacQueen, 1967). By adopting K-means, this study aims to define multiple label clusters according to labels' recurring period. Density-based spatial clustering of applications with noise (DBSCAN) is a density-based algorithm; it assumes dense areas are clustered. It does not require that each point be assigned to a cluster, and so does not divide the data; rather, it extracts the 'dense' clusters and leaves the sparse background as 'noise' (Ho *et al.*, 2020; Leland McInnes, John Healy, 2016). DBSCAN is often used in conjunction with agglomerative clustering. This study aims to extract strong recurring signals of labels by utilizing DBSCAN clustering method.

3.3 Knee point detection

Incoming traffic is usually aggregated into flows, where a flow score is the cumulated sum of its abnormality level in every subspace. A knee point in the curve indicates a sudden change in flow scores and therefore, in flows degree of abnormality (Dromard *et al.*, 2017). We used the "Kneedle" method to determine the saturation/knee point of each question's lifecycle under different label clusters. We can measure the average reaching time of the second stage in the lifecycle of a question based on the coordination location of their knee points. As such, we can determine if questions in various label clusters hit the second stage of their lifecycle at different points.

Satopää et al. (2011) presented "Kneedle", a general approach to online and offline knee detection that is applicable to a wide range of systems. The knee definition comes from Salvador and Chan (2004): the knee of a curve is loosely defined as the point of maximum curvature. Kneedle is based on the notion that the points of maximum curvature in a data set, i.e., the knees, are approximately the set of points in

a curve that are local maxima if the curve is rotated θ degrees clockwise about (x_{min},y_{min}) through the line formed by the points (x_{min},y_{min}) and (x_{max},y_{max}). Put simply, knees occur when a curve becomes more "flat," indicating a decrease in curvature. For a given monotonically increasing function f(x), a knee-point is a point with maximum curvature. The curvature at each point x of the function f(x) is defined as $C_{f(x)}$, hence a knee-point can be formulated as of the equation below (Ghafoori *et al.*, 2016; Satopää *et al.*, 2011):

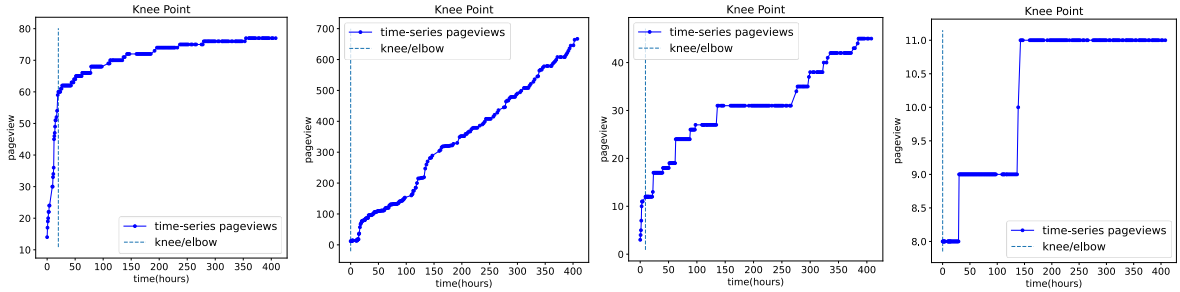
$$x_{C_f}^{max} = argmax C_f(x), \text{where } C_f(x) = \frac{f''(x)}{(1+f'(x)^2)^{0.5}}$$

Figure 4 illustrates examples of a knee point detection. Figure 4(a) exhibits the question pageview curve based on the original time series data which shows a concave and increasing pattern (QID = 447782565). The intersection of the vertical dotted line and the lifespan trendline is the detected knee point (21.1, 60).

4 Results

4.1 Four types of question lifecycles

We classified questions' lifecycles into four distinct patterns using the needle approach, which proves our hypothesis 2a: there exist multiple common types of question lifecycles in CQA. The most common lifecycle is a two-stage lifecycle (Figure 4-a), which consists of a "growth" phase in which the question receives most of the traffic and a "saturation" phase where the pageviews plateau over a long period of time. Certain questions' lifecycles exhibit a periodic recurring pattern (Figure 4-c), and their traffic typically grows in phases. Additionally, some questions can exhibit a consistent linear pattern of traffic increase (Figure 4-b); in this instance, we classify them as linear growing questions. Furthermore, not all questions can generate a sufficient number of pageviews throughout our observation period; for example, Figure 4-d depicts a question that received only three pageviews after being posted for almost 300 hours. In most circumstances, these questions will remain silent and stay at the bottom of the longtail; thus, they are referred to as dead questions.



(a) 2-stage lifecycle (qid=447782565) (b) Linearly growing lifecycle (qid=447782156) (c) Periodically recurring lifecycle (qid=447782734) (d) Dead question's lifecycle (qid=447783005)

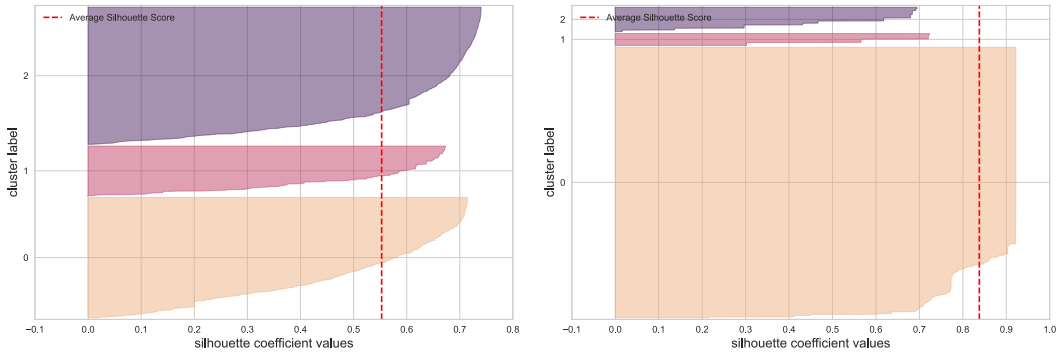
Figure 4. Four patterns of question lifecycles

Our data statistics show that 84.87% of all questions had a two-stage lifecycle, which supports our hypothesis 2b: the majority of questions on Zhihu follow the typical two-phase lifecycle. 10.76% of questions are linearly growing questions. However, only 2% of the questions are periodically recurring questions and 2.37% of the questions are dead at the bottom of the longtail. We thus put forward a '85:11:2:2' lifecycle rule to summarise the findings above for the questions on Zhihu.

For periodically recurring questions with multiple saturation (knee) points, we captured the most significant ones while determining their saturation phase. We excluded linear and dead questions from the calculation of the knee point because they do not have any knees. In aggregate, 2-stage questions and periodically recurring questions have an average knee point of 47.3637, indicating the average time it takes for the majority of questions to reach their second lifecycle stage after being posted is almost two days. This finding supports our hypothesis 1: questions on Zhihu obtain a majority of user attention after being posted only for a brief period of time.

4.2 Three Label clusters

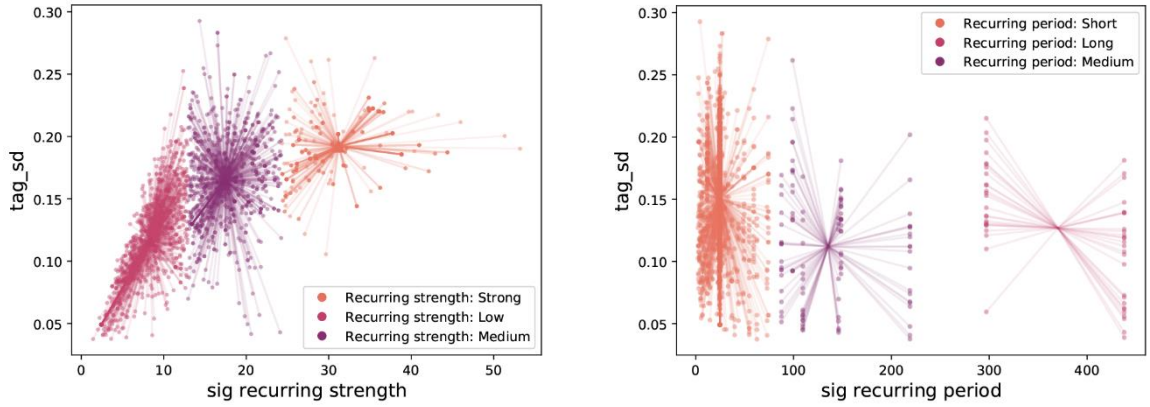
The most notable peak and its prominence in DFT are referred to as the most significant repeating period and its relative strength for each label. We choose to classify labels using K-means based on their recurring periods and strengths. Silhouette analysis examines the separation distance between the resulting clusters which can be used to choose an optimal value for cluster numbers in K-means clustering (Rousseeuw, 1987). The value of the silhouette ranges between [1, -1], where a high value indicates that the object is well matched to its own cluster and poorly matched to neighbouring clusters. In order to determine the number of clusters, we calculate the silhouette scores for 2, 3, 4 and 5 clusters. The silhouette analysis for K-means clustering on recurring strengths and recurring periods are shown in Figure 5. The results reveal that most data points can achieve above-average silhouette coefficient values with three clusters, outperforming 2, 4, and 5 clusters, demonstrating that a 3-cluster configuration is appropriate for continuing the k-means analysis.



(a) *(b)*

Figure 5. Silhouette plots of K-means clustering of Labels' recurring strengths(a) and recurring periods(b).

By using the K-means clustering approach, we can visualise the three primary clusters regardless of the standard deviation of the tags. Figure 6ab presents the connected scatter plots for different label clusters.



(a) *(b)*

Figure 6. K-means clusters of Labels' recurring strengths(a) and recurring periods(b).

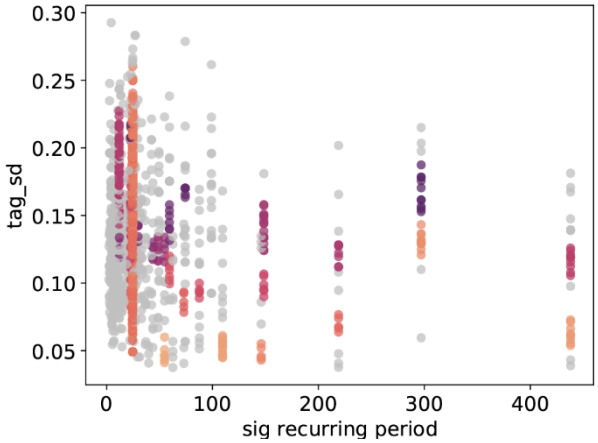


Figure 7. DBSCAN clusters of Labels' recurring period

The three clusters can be labelled as a 'weak recurring group', 'medium recurring group', and 'strong recurring group' based on the recurring strengths of the labels (Figure 6a). The weak group contains labels with signal strengths ranging from 0 to 13 (less time-sensitive). Labels in the medium category have signals with a prominence of 13-24. The strong group includes more time-sensitive labels expressing signals with a prominence greater than 24. The data statistics show that only 14.64% of the labels are high-time-sensitive labels, 32.56% of them are medium-time-sensitive labels and 52.8% of the labels are low-time-sensitive labels. The presented K-means clustering analysis of signal recurring strengths proves our hypothesis 4: Labels can be classified as high-time-sensitive or low-time-sensitive with varying proportions.

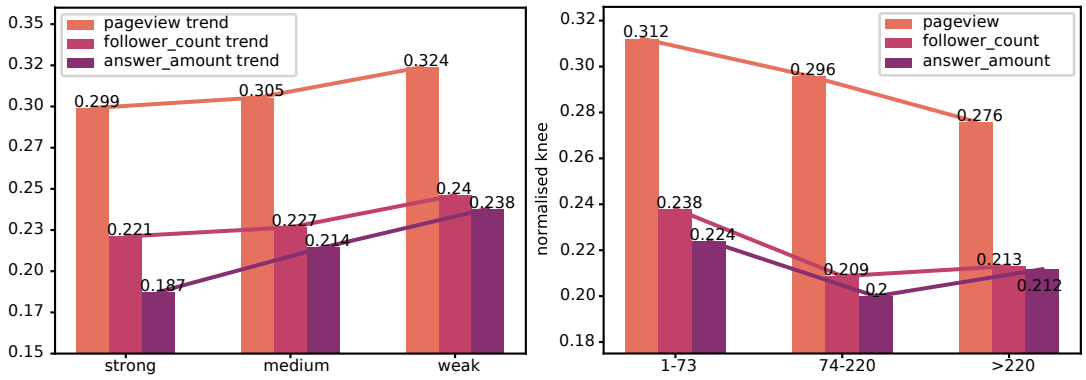
Similarly, as shown in Figure 6b, by employing the K-means clustering method, recurrent periods of labels can be categorised into three clusters (recurring every 0-73 hours; recurring every 74-220 hours; and recurring over every 220 hours), regardless of how their followers or subordinate inquiries change.

We discover several very strong signals when examining the recurring periods of label clusters. We can extract these strong signals by utilizing the Density-based spatial clustering of applications with noise (DBSCAN). As seen in Figure 7, outlier points have been shaded grey, while the colourful scatters remain inside represent the high-density zones. There are two particularly powerful and frequent recurrent periods among those small clusters: every 12 hours and every 24 hours. According to Alexa's Audience Geography statistics, approximately 92.7 % of all visitors of Zhihu came from the same time zone (Beijing time zone) from 13th May to 13th June 2021 (Alexa, 2021). Thus, we may infer that they are stable half-day/daily recurring labels. Additionally, other bright regions on the image include 438hs (18days), 298hs (12days), and 149hs (6days), which are all relatively common signals.

4.3 Questions' knee points occurrence under label groups

We standardise the knee points and conduct bivariate correlation analysis on two-stage and regularly recurring questions. The results from our correlation analysis reveal that both the recurring periods and strengths of labels have a negative correlation with the knee points of respective questions. Labels' recurring strengths and normalised knee have a significant negative correlation $r(N) = -.042^*$, $p = .038$ (N =1945), which proves hypothesis 3a and 3b.

We adopted questions' digital popularity data which includes pageview, answer amount and follower count to track each question's lifecycle and the knee point of its lifecycle. The statistical results can also be visualised in each label cluster. Figure 8a shows that as time passes, the questions with more time-sensitive (strong periodically recurring) labels reach the saturation point of their lifecycle earlier than the questions with less time-sensitive labels. From a short-term viewpoint, Figure 8b demonstrates that questions with longer repeating period labels attain saturation earlier than questions with shorter recurring period labels.



(a) *(b)*

Figure 8. Normalised knee value grouped by label recurring strengths(a) and recurring periods(b)

We can then take the following conclusion based on the findings above to demonstrate our Hypothesis 3a and 3b: Questions have different visit patterns under different label groups. Questions with more time-sensitive labels or longer recurring periods typically reach saturation sooner.

5 Discussion

The findings of the study addressed the hypotheses that have been proposed, which are summarised in Table 1.

Index	Hypothesis Description	Result
H1	Posts on Q&A sites adhere to the 90/90 data-use theory, which states that questions receive the majority of users' attention during a brief period after being posted.	Supported
H2a	There exist multiple common types of question lifecycles in CQA.	Supported
H2b	Questions on Zhihu follow a rule similar to the 80:10:10 rule, with the majority of questions following a 2-phase lifecycle, while the remaining questions obey some other lifecycle patterns.	Supported
H3a	Questions under different time-sensitive label groups have different lifespans.	Supported
H3b	Questions belonging to more time-sensitive topic categories are likely to reach their saturation points sooner than those under less time-sensitive topic categories.	Supported
H4	Labels can be classified as high-time-sensitive or low-time-sensitive with varying proportions	Supported

Table1. Summary table of hypotheses

This research provides some theoretical contributions. Firstly, this study has proven the 90/90 data-use rule on CQAs by visualising various question lifecycles. It also addresses the information classification problem that exists in existing ILM studies by providing precise topic categorisation methods based on the recurring strengths and periods of labels, as well as quantitative real-world examples. This is also the first time the concept of ILM has been introduced to knowledge sharing community management. Future studies can employ the proposed question lifecycle classification framework and the "85:11:2:2" question lifecycle rule as valuable takeaways to develop relevant functionalities in CQAs.

The practical implications of this study are twofold. The question lifespan exploration can aid in improving platform design, which benefits both the users and CQA platform providers. The findings and methods of this study can be applied to develop toolsets that allow users to visualise their expected question lifecycles by selecting different labels, which can potentially predict the traffic saturation status of their posted content more accurately. For instance, based on the selected question labels, the toolset can suggest the estimated timeframe to obtain the best answer to the users' questions. Users who asked questions on more time-sensitive topics (e.g., entertainment news) should expect to receive the attention of the majority of users in a shorter amount of time than users who asked questions about less time-sensitive topics (e.g., a city). Furthermore, the findings may benefit CQA platform management in

gaining a deeper understanding of the traffic trends behind various types of topics, which may help them in creating effective label recommendation systems that are capable of predicting content lifecycles.

The management of digital resources is another practical contribution of this study. As many systems struggle with limited digital resources and high data maintenance costs, data archival, retention, and deletion are critical operations in ILM to address the information overload problem for users. As a result, knowing when to "retire" data from systems in online communities is becoming increasingly important to utilise system resources more efficiently. The results of this research could be utilised to build algorithms that help CQAs in developing better performing data retention policies. Questions that have already reached saturation points, for example, can be safely archived and not promoted. It might even be worth considering deleting "dead" questions. The 2% dead questions in CQAs can be considered as noise, therefore, eliminating or relocating them can help reduce costs. Moreover, this study has established a relationship between label usage and questions lifecycles, which provides possibilities for platforms to apply alterative data preservation rules for questions with varied time-sensitive topics.

6 Conclusion

This article examines the relationship between temporal label usage and the lifespan of questions using real-world dynamic panel data from one of the largest CQA platforms. We used a Fourier Transformation to identify and quantify question labels' prominent recurring signals. We then used a K-means and DBSCAN method to classify the labels into different clusters and determine the knee values of questions within each cluster based on their recurring periods and strengths. We proposed a novel representation, *saturation point*, for describing the lifecycle of questions on Q&A sites. We summarise the questions' lifecycles into four common patterns (2-stage; periodically recurring; linearly growing and dead questions) and put forward a '85:11:2:2' question lifecycle rule to further illustrate each pattern's corresponding proportion. We used the 'Needle' method to determine the saturation point of all two-stage and periodically recurring questions and then mapped them to the associated label clusters to analyse the corresponding effects.

To summarise, we demonstrate that the majority of questions follow a standard two-stage lifespan, where most of users' attention is being received shortly after being posted (around 47.36 hours in our study). Label clusters have a significant effect on the lifecycles of questions, as evidenced by the following more detailed finding: questions with more time-sensitive (strong periodically recurring) labels reach the second stage of their lifespans earlier than questions with less time-sensitive labels. This work closes a theoretical research gap by establishing a connection between label usage and the lifecycles of questions to address the information classification issue in Information Lifecycle Management (ILM). Practically speaking, the research findings can be used by CQA platforms to forecast the progression of posted questions from users based on the features of the labels that they attach to their post. The lifecycle visualisation feature can help users predict when they will likely receive the majority of the traffic for their posted questions. The findings can also assist platform provider in optimizing the data retention rules to save system resources.

This study has certain limitations and elicits several insightful future research possibilities. Firstly, when summarizing the DFT data, we chose only the most significant peaks. Given that some of the less significant peaks may influence the label clusters, a more accurate peak detection method is required. Besides the topics, there might exist other features that can influence a question's lifecycle (e.g., askers' reputation, semantic features, etc.) which is worth further exploring in the future. In this study, we focus exclusively on the short-term lifecycle of questions, with the longest observation time being 400 hours (about 2 and a half weeks). However, from a long-term perspective, certain two-stage questions may become periodically recurring or even linearly expanding, which is another subject we intend to explore in the future. While this study utilised data from Zhihu, a domain-general CQA, future research may try using a domain-specific platform, such as Stack Overflow, to conduct a comparative analysis and determine whether the research findings vary across different types of question-answering communities.

7 References

Agichtein, E., Castillo, C., Donato, D., Gionis, A. and Mishne, G. (2008), "Finding High-quality Content in Social MediaAgichtein, Eugene Castillo, Carlos Donato, Debora Gionis, Aristides Mishne, Gilard", *International Conference on Web Search and Data Mining*, pp. 183–193.

Aksentijević, S., Tijan, E. and Agatić, A. (2011), "Information security as utilization tool of enterprise information capital", *2011 Proceedings of the 34th International Convention MIPRO*, IEEE, pp. 1391–1395.

Al-Fedaghi, S. (2013), "Information management and valuation", *International Journal of Engineering Business Management*, Vol. 5 No. 1, pp. 1–11.

Alexa. (2021), "zhihu.com Competitive Analysis, Marketing Mix and Traffic", available at: https://www.alexa.com/siteinfo/zhihu.com.

Anderson, A., Huttenlocher, D., Kleinberg, J. and Leskovec, J. (2012), "Discovering value from community activity on focused question answering sites: A case study of stack overflow", *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 850–858.

Arutyunov, V.. (2012), "Identification and authentication as the basis for information protection in computer systems", *Scientific and Technical Information Processing*, Springer, Vol. 39 No. 3, pp. 133–138.

Bergström, E. and Åhlfeldt, R.-M. (2014), "Information Classification Issues", in Bernsmed, K. and Fischer-Hübner, S. (Eds.), *Secure IT Systems*, Springer International Publishing, Cham, pp. 27–41.

Bergström, E. and Anteryd, F. (2018), "Information Classification Policies : An Exploratory Investigation".

Bergström, E., Karlsson, F. and Åhlfeldt, R.-M. (2021), "Developing an information classification method", *Information & Computer Security*, Emerald Publishing Limited, Vol. 29 No. 2, pp. 209–239.

Bergström, E., Lundgren, M. and Ericson, Å. (2019), "Revisiting information security risk management challenges: a practice perspective", *Information & Computer Security*, Emerald Publishing Limited.

Bhat, V., Gokhale, A., Jadhav, R., Pudipeddi, J. and Akoglu, L. (2015), "Effects of tag usage on question response time: Analysis and prediction in StackOverflow", *Social Network Analysis and Mining*, Springer-Verlag Wien, Vol. 5 No. 1, pp. 1–13.

Bluestein, L.I. (1970), "A Linear Filtering Approach to the Computation of Discrete Fourier Transform", *IEEE Transactions on Audio and Electroacoustics*, Vol. 18 No. 4, pp. 451–455.

Bunker, G. (2012), "Technology is not enough: Taking a holistic view for information assurance", *Information Security Technical Report*, Elsevier, Vol. 17 No. 1–2, pp. 19–25.

Büsch, S., Nissen, V. and Wünscher, A. (2017), "Automatic classification of data-warehouse-data for information lifecycle management using machine learning techniques", *Information Systems Frontiers*, Vol. 19 No. 5, pp. 1085–1099.

Cao, X., Cong, G., Cui, B., Jensen, C.S. and Zhang, C. (2009), "The use of categorization information in language models for question retrieval", *International Conference on Information and Knowledge Management, Proceedings*, pp. 265–274.

Castillo, C., El-Haddad, M., Pfeffer, J. and Stempeck, M. (2014), "Characterizing the life cycle of online news stories using social media reactions", *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW*, pp. 211–223.

Cha, M., Haddadi, H., Benevenuto, F. and Gummadi, K. (2010), "Measuring User Influence in Twitter: The Million Follower Fallacy", *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 4 No. 1, available at: https://ojs.aaai.org/index.php/ICWSM/article/view/14033.

Coelho, M.P. and Mendes, J.Z. (2019), "Digital music and the 'death of the long tail'", *Journal of Business Research*, Elsevier, Vol. 101 No. June 2018, pp. 454–460.

Cook, J., Kenthapadi, K. and Mishra, N. (2013), "Group chats on Twitter", *WWW 2013 - Proceedings of the 22nd International Conference on World Wide Web*, pp. 225–235.

Crane, R. and Sornette, D. (2008), "Robust dynamic classes revealed by measuring the response function of a social system", *Proceedings of the National Academy of Sciences*, Vol. 105 No. 41, pp. 15649–15653.

DeMasi, O., Mason, D. and Ma, J. (2016), "Understanding communities via hashtag engagement: A clustering based approach", *Proceedings of the 10th International Conference on Web and Social Media, ICWSM 2016*, No. Icwsm, pp. 102–111.

DeVaro, J., Kim, J.H., Wagman, L. and Wolff, R. (2018), "Motivation and performance of user-contributors: Evidence from a CQA forum", *Information Economics and Policy*, Elsevier B.V., Vol. 42, pp. 56–65.

Dezsö, Z., Almaas, E., Lukács, A., Rácz, B., Szakadát, I. and Barabási, A.-L. (2006), "Dynamics of information access on the web", *Phys. Rev. E*, American Physical Society, Vol. 73 No. 6, p. 66132.

Dromard, J., Roudière, G. and Owezarski, P. (2017), "Online and Scalable Unsupervised Network Anomaly Detection Method", *IEEE Transactions on Network and Service Management*, Vol. 14 No. 1, pp. 34–47.

Efraim, T., Linda, V. and Gregory R, W. (2017), *Information Technology for Management*, Wiley Direct.

Fibikova, L. and Müller, R. (2011), "A simplified approach for classifying applications", *ISSE 2010 Securing Electronic Business Processes*, Springer, pp. 39–49.

Figueiredo, F., Benevenuto, F. and Almeida, J.M. (2011), "The Tube over Time: Characterizing Popularity Growth of Youtube Videos", *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, Association for Computing Machinery, New York, NY, USA, pp. 745–754.

Gantz, J.F. and Reinsel, D. (2007), "A forecast of worldwide information growth through 2010", *The Expanding Digital Universe*, Vol. 4.

Ghafoori, Z., Rajasegarar, S., Erfani, S.M., Karunasekera, S. and Leckie, C.A. (2016), "Unsupervised Parameter Estimation for One-Class Support Vector Machines BT - Advances in Knowledge Discovery and Data Mining", in Bailey, J., Khan, L., Washio, T., Dobbie, G., Huang, J.Z. and Wang, R. (Eds.), , Springer International Publishing, Cham, pp. 183–195.

Gharan, S.O. and Wang, Y. (2010), "What Memes Say about the News Cycle ?"

Gu, B., Tang, Q. and Whinston, A.B. (2013), "The influence of online word-of-mouth on long tail formation", *Decision Support Systems*, Elsevier, Vol. 56, pp. 474–481.

Han, E.S. and goleman, daniel; boyatzis, Richard; Mckee, A. (2019), "Recommendation Networks and the Long Tail of Electronic Commerce", *Journal of Chemical Information and Modeling*, Vol. 53 No. 9, pp. 1689–1699.

Hayes, J. (2008), "Have data? Will travel [IT security threat]", *Engineering & Technology*, IET, Vol. 3 No. 15, pp. 60–61.

Ho, M.K., Tatinati, S. and Khong, A.W.H. (2020), "Distilling Essence of a Question : A Hierarchical Architecture for Question Quality in CQA Sites", *2020 International Joint Conference on Neural Networks*, Vol. 1, pp. 1–7.

Hong, L., Dan, O. and Davison, B.D. (2011), "Predicting Popular Messages in Twitter", *Proceedings of the 20th International Conference Companion on World Wide Web*, Association for Computing Machinery, New York, NY, USA, pp. 57–58.

Hsu, M.H., Chang, Y.H. and Chen, H.H. (2010), "Temporal Correlation between Social Tags and Emerging Long-Term Trend Detection", *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 4 No. 1 SE-Poster Papers, available at: https://ojs.aaai.org/index.php/ICWSM/article/view/14049.

Huna, A., Srba, I. and Bielikova, M. (2016), "Exploiting content quality and question difficulty in CQA reputation systems", *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Vol. 9564, pp. 68–81.

Kraemer, S., Carayon, P. and Clem, J. (2009), "Human and organizational factors in computer and information security: Pathways to vulnerabilities", *Computers & Security*, Elsevier, Vol. 28 No. 7, pp. 509–520.

Ku, C.-Y., Chang, Y.-W. and Yen, D.C. (2009), "National information security policy and its

implementation: A case study in Taiwan", *Telecommunications Policy*, Elsevier, Vol. 33 No. 7, pp. 371–384.

Lehmann, J., Gonçalves, B., Ramasco, J.J. and Cattuto, C. (2012), "Dynamical Classes of Collective Attention in Twitter", *Proceedings of the 21st International Conference on World Wide Web*, Association for Computing Machinery, New York, NY, USA, pp. 251–260.

Leland McInnes, John Healy, S.A. (2016), "The hdbscan Clustering Library", available at: https://hdbscan.readthedocs.io/en/latest/index.html.

Levitt, T. (1965), "Exploit the product life cycle", Vol. 43 No. 6, pp. 1–33.

Lin, Y.-R., Margolin, D., Keegan, B., Baronchelli, A. and Lazer, D. (2013), "#Bigbirds Never Die: Understanding Social Dynamics of Emergent Hashtags", *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 7 No. 1 SE-Full Papers, pp. 370–379.

Liu, Q., Agichtein, E., Dror, G., Gabrilovich, E., Maarek, Y., Pelleg, D. and Szpektor, I. (2011), "Predicting web searcher satisfaction with existing community-based answers", *SIGIR'11 - Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 415–424.

Liu, Y., Tang, A., Sun, Z., Tang, W., Cai, F. and Wang, C. (2020), "An integrated retrieval framework for similar questions: Word-semantic embedded label clustering – LDA with question life cycle", *Information Sciences*, Elsevier Inc., Vol. 537, pp. 227–245.

MacQueen, J. (1967), "Some methods for classification and analysis of multivariate observations", in Lucien M, L.C. and Jerzy, N. (Eds.), *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, pp. 281–297.

Maity, S.K., Sahni, J.S.S. and Mukherjee, A. (2015), "Analysis and prediction of question topic popularity in community Q&A sites: A case study of Quora", *Proceedings of the 9th International Conference on Web and Social Media, ICWSM 2015*, pp. 238–247.

Mason, H. (2011), "You just shared a link. How long will people pay attention?", *Bitly Blog*.

Naveed, N., Gottron, T., Kunegis, J. and Alhadi, A.C. (2011), "Bad News Travel Fast: A Content-Based Analysis of Interestingness on Twitter", *Proceedings of the 3rd International Web Science Conference*, Association for Computing Machinery, New York, NY, USA, available at:https://doi.org/10.1145/2527031.2527052.

Nie, L., Li, Y., Feng, F., Song, X., Wang, M. and Wang, Y. (2020), "Large-Scale Question Tagging via Joint Question-Topic Embedding Learning", *ACM Transactions on Information Systems*, Vol. 38 No. 2, available at:https://doi.org/10.1145/3380954.

Nielsen, J. (2006), "Participation Inequality : Encouraging More Users to Contribute", available at: https://ci.nii.ac.jp/naid/10019960520/en/.

Oscarson, P. and Karlsson, F. (2009), "A national model for information classification", *AIS SIGSEC Workshop on Information Security & Privacy (WISP2009), Phoenix, USA, 2009.*

Pedro, J.S. and Karatzoglou, A. (2014), "Question recommendation for collaborative question answering systems with RankSLDA", *RecSys 2014 - Proceedings of the 8th ACM Conference on Recommender Systems*, pp. 193–200.

Pennock, M. (2007), "Digital curation: a life-cycle approach to managing and preserving usable digital information", *Library & Archives*, n, Vol. 1 No. 1, pp. 1–3.

Pypi. (2019), "zhihu-oauth 0.0.42", available at: https://pypi.org/project/zhihu-oauth/.

Romero, D.M., Meeder, B. and Kleinberg, J. (2011), "Differences in the Mechanics of Information Diffusion across Topics: Idioms, Political Hashtags, and Complex Contagion on Twitter", *Proceedings of the 20th International Conference on World Wide Web*, Association for Computing Machinery, New York, NY, USA, pp. 695–704.

Rousseeuw, P.J. (1987), "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis", *Journal of Computational and Applied Mathematics*, Vol. 20, pp. 53–65.

Roy, P.K., Ahmad, Z., Singh, J.P., Alryalat, M.A.A., Rana, N.P. and Dwivedi, Y.K. (2018), "Finding and Ranking High-Quality Answers in Community Question Answering Sites", *Global Journal of Flexible Systems Management*, Springer India, Vol. 19 No. 1, pp. 53–68.

Rusbridge, C., Burnhill, P., Ross, S., Buneman, P., Giaretta, D., Lyon, L. and Atkinson, M. (2005), "The digital curation centre: a vision for digital curation", *2005 IEEE International Symposium on*

Mass Storage Systems and Technology, IEEE, pp. 31–41.
Sabah, A.F. (2008), "On information lifecycle management", *Proceedings of the 3rd IEEE Asia-Pacific Services Computing Conference, APSCC 2008*, IEEE, pp. 335–342.
Salvador, S. and Chan, P. (2004), "Determining the Number of Clusters / Segments in Hierarchical Clustering / Segmentation Algorithms", No. Ictai.
Satopää, V., Albrecht, J., Irwin, D. and Raghavan, B. (2011), "Finding a 'kneedle' in a haystack: Detecting knee points in system behavior", *Proceedings - International Conference on Distributed Computing Systems*, pp. 166–171.
SciPy.org. (2021a), "Interpolation (scipy.interpolate)", available at: https://docs.scipy.org/doc/scipy/reference/interpolate.html.
SciPy.org. (2021b), "Fourier Transforms", available at: https://docs.scipy.org/doc/scipy/reference/tutorial/fft.html.
SciPy.org. (2021c), "scipy.signal.find_peaks", available at: https://docs.scipy.org/doc/scipy/reference/generated/scipy.signal.find_peaks.html.
Sha, A.S., Shi, Y. and Haller, A. (2020), *How Question Quality Drives Web Performance in Community Question Answering Sites, ArXiv*.
Shamma, D.A., Kennedy, L. and Churchill, E.F. (2011), "Peaks and Persistence: Modeling the Shape of Microblog Conversations", *Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work*, Association for Computing Machinery, New York, NY, USA, pp. 355–358.
Suh, B., Hong, L., Pirolli, P. and Chi, E.H. (2010), "Want to be Retweeted? Large Scale Analytics on Factors Impacting Retweet in Twitter Network", *2010 IEEE Second International Conference on Social Computing*, pp. 177–184.
Taeuscher, K. (2019), "Uncertainty kills the long tail: demand concentration in peer-to-peer marketplaces", *Electronic Markets*, Electronic Markets, Vol. 29 No. 4, pp. 649–660.
Tucker, C. and Zhang, J. (2007), "Long Tail or Steep Tail? A Field Investigation into How Online Popularity Information Affects the Distribution of Customer Choices", *Working Paper No. 4655-07*, No. 617, pp. 1–35.
Virtanen, T. (2001), "Design criteria to classified information systems numerically", *IFIP International Information Security Conference*, Springer, pp. 317–325.
Vlachos, M., Yu, P. and Castelli, V. (2005), "On periodicity detection and structural periodic similarity", *Proceedings of the 2005 SIAM International Conference on Data Mining, SDM 2005*, pp. 449–460.
Waddington, S., Green, R. and Awre, C. (2012), "CLIF: Moving repositories upstream in the content lifecycle", *Journal of Digital Information*, Vol. 13 No. 1.
Xie, L. and Sundaram, H. (2012), "Media lifecycle and content analysis in social media communities", *Proceedings - IEEE International Conference on Multimedia and Expo*, IEEE, Vol. 1 No. c, pp. 55–60.
Yang, J. and Leskovec, J. (2011), "Patterns of Temporal Variation in Online Media", *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, Association for Computing Machinery, New York, NY, USA, pp. 177–186.
Yao, Y., Tong, H., Xie, T., Akoglu, L., Xu, F. and Lu, J. (2015), "Detecting high-quality posts in community question answering sites", *Information Sciences*, Elsevier Inc., Vol. 302, pp. 70–82.
Yu, H., Xie, L. and Sanner, S. (2015), "The lifecyle of a YouTube video: Phases, content and popularity", *Proceedings of the 9th International Conference on Web and Social Media, ICWSM 2015*, pp. 533–542.