On the Measurement and Prediction of Web Content Utility: A Review

Yuan Yao
Department of Computer Science and Technology
Nanjing University, China
y.yao@nju.edu.cn

Hanghang Tong
School of Computing, Informatics and Decision Systems Engineering
Arizona State University, USA
hanghang.tong@asu.edu

Feng Xu, Jian Lu
Department of Computer Science and Technology
Nanjing University, China
{xf,lj}@nju.edu.cn

ABSTRACT

Nowadays, various types and large amount of content are available on the Web. Characterizing the Web content and predicting its inherent usefulness become important problems that may benefit many applications such as information filtering and content recommendation. In this article, we present a brief review of the existing measurements and the corresponding prediction methods for Web content utility. Specially, we focus on three close and widely studied tasks, i.e., content popularity prediction, content quality prediction, and scientific article impact prediction. While reviewing the existing work in each of the above three tasks, we mainly aim to answer the following two fundamental questions: how to measure the Web content utility, and how to make the predictions under the measurement. We find that while the three tasks are closely related, they bear subtle differences in terms of prediction urgency, feature extraction, and algorithm design. After that, we discuss some future directions in measuring and predicting Web content utility.

1. INTRODUCTION

In recent years, measuring and predicting Web content utility (e.g., popularity, quality, impact, etc.) have attracted much attention and have potential usefulness in many applications. For example, by measuring and predicting the popularity of Web content, we can design more profitable advertising strategies over the social media; by measuring and predicting the quality of Web content, we can more easily identify informative messages from the huge amount of noisy information; by measuring and predicting the future impact of research articles, we can better envision the research trends and plan our research roadmaps.

In this article, we briefly review the existing measurements and the corresponding prediction methods for Web content utility. Specially, we focus on the following three widely studied tasks.

- *Content popularity prediction.* Content popularity prediction is based on the current state of the content, and the goal is to predict its future popularity after a relatively long period. Popularity metrics (e.g., the number of views) are usually available in the social platforms that host the content. For this task, we summarize the existing methods on general Web content including text, videos, and images.
- *Content quality prediction.* The input of content quality prediction is similar to that of the content popularity prediction. For the output, content quality itself is a vague concept and different people may have different definitions on it over different social platforms. Therefore, most existing work either adopts human labeling or uses approximate metrics to define quality. Another difference from content popularity prediction (which estimates the future popularity) is that content quality prediction estimates the current quality of the content. In this article, we mainly summarize several mainstream content quality prediction problems including the helpfulness of product reviews, the credibility of microblog messages, and the quality of question/answers in question answering sites.
- *Scientific article impact prediction.* In addition the above to two general tasks, we review the impact prediction task for a specific type of content, i.e., the scientific articles. Scientific impact is of special interest to researchers. Similar to content popularity prediction, the impact prediction of scientific articles also aims to predict the future impact based on the current state, although scientific articles usually have a longer lifecycle than the other types of Web content.

While reviewing the existing work in each of the above three tasks, we mainly aim at answering two fundamental questions: how to measure the Web content utility, and how to make the predictions under the measurement. For the prediction of Web content utility, we focus on the following three aspects.

- *Prediction time.* The first aspect is about when to make the prediction. Based on the prediction time, existing work can be divided into three classes, i.e., before publication, at publication, and after publication. Before publication means that the prediction is made even before the content has been created and published; at prediction means that the prediction is made at the moment when the content is published; after publication means that the prediction is made after a short period of the publication time.
- *Features.* Depending on the prediction time, several types of features/factors may be indicative for the prediction task. For example, when the prediction time

is before publication, the content creator is the main factor; when the prediction time is at publication, the content itself is an indicative factor; when the prediction time is after publication, more context factors may play important roles.

- *Algorithms*. The third aspect is about the detailed algorithms used to finish the prediction. Many off-the-shell data mining algorithms have been used for the prediction tasks. Additionally, some researchers propose designed algorithms to better leverage the characteristics of the underlying social platforms.

The rest of this article is organized as follows. Section 2, Section 3, and Section 4 present the brief review of content popularity prediction, content quality prediction, and scientific article impact prediction, respectively. Section 5 summarizes the main findings and insights, and discusses some future directions in measuring and predicting Web content utility. Section 6 concludes the article.

2. CONTENT POPULARITY PREDICTION

In this section, we briefly review the existing literature for the popularity measurement and prediction of Web content.

2.1 Content Popularity Measurement

For content popularity, the measurement is relatively agreed. Most literature uses the view count (i.e., the number of views by the community users on the content) to measure popularity [74]. Since the view count may exhibit a power-law distribution in many social platforms, it is usually normalized by existing methods (e.g., [43; 19]). Compared to predicting view count, some researchers argue that identifying highly popular content is more meaningful. They find that a small amount of highly popular content usually dominates the major popularity views, and thus formulate a classification problem to identify such content (e.g., [45; 26]).
In addition to view count, some other metrics such as the number of votes/comments are also used to measure popularity with the assumption that the view count information is not always available. In this article, we treat popularity as view count as it is largely agreed by existing literature, and leave the discussions about other metrics in Section 3.

2.2 Content Popularity Prediction

Popularity prediction has been studied for many different types of Web content including videos [26], news [47], microblogs [39], reviews [79], images [19], codes [84], mobile apps [96], etc. Based on literature, popularity prediction is a challenging task due to the following two reasons. First, many factors are intuitively known to influence content popularity, and some of these factors are difficult to quantify (such as the quality of the content, the complex interactions between users and content, etc.). Second, content popularity can be affected by many external factors such as a shocking event in the physical world.
Next, we summarize the existing efforts for content popularity prediction in terms of prediction time, factors/features, and algorithms.

2.2.1 Prediction Time

For the prediction time, most existing work makes the prediction after a period of the publication time. The basic assumption is that the early popularity is a strong indicator for

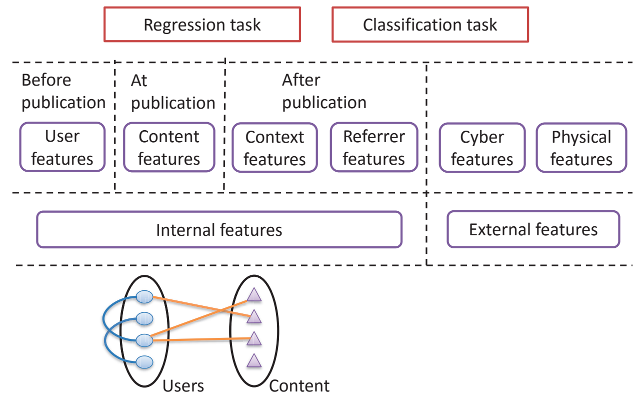

Figure 1: The features for content popularity prediction.

the future popularity. For example, an early practice finds a strong correlation between the logarithmically transformed past popularity and current popularity, and proposes to use the number of views that the Web content receives early after publication to predict its future popularity [73]. Within these after publication methods, some of them require fast predictions as users' interest may quickly fade (e.g., news), and some may afford several days to make predictions (e.g., videos).
In addition to most existing methods that make the prediction after publication, some methods focus on a harder problem to predict the popularity at publication time [6]. For example, Khosla et al. [43] use social features (e.g., the number of friends) and content features both of which are available at publication time to predict the popularity of images.
An even harder problem is to predict the popularity before publication, i.e., when the content is not published yet. This problem is seldom studied by existing researchers due to the fact that complex factors (e.g., content features) may affect the content popularity.

2.2.2 Features

The commonly used features for content popularity prediction are summarized in Fig. 1. Nowadays, Web content is typically published in social platforms where users can interact with the content and with each other. As shown in Fig. 1, the popularity of such content can be affected by both **internal features** and **external features**. Here, internal features indicate those features that can be extracted from the social platform where the content is published. These features are based on the interactions between users and content as well as the interactions between users. As shown in Fig. 1, we summarize the following four types of internal features.

- **User features**. User features indicate the impact of the content creator. Node-based user features (e.g., the account age) are relatively less visited as they are intuitively less indicative. In contrast, link-based user features (a.k.a. social features) have been widely explored [26; 27; 43]. Such features include the number of friends/followers/subscribers, social influence, etc.

- **Content features**. Content features are directly extracted from the content, and they are content-specific.

For example, language models can be used for text, representations from deep neural networks can be used for images [43], and abstract syntax tree features can be used for codes [84]. Some content features are easy to extract (e.g., the related tags, the time when the content is published, the location relevance in the content [10]), while others need careful design. For example, existing research has shown that emotion is one of the most important drivers for online audience [7; 33]; however, analyzing the emotion in Web content is a non-trivial task. Other non-trivial content features include quality, structure, readability, etc.

- **Context features**. Context features refer to the community users' early responses to the content after it is published. The early popularity is a typical context feature. In fact, Borghol et al. [9] have studied the popularity differences among videos that have essentially the same content, and find that the early popularity is one of the most indicative factor for videos. Actually, as we will later show in the next subsection, many existing methods use only this type of features to predict content popularity. Other used context features include the number of views on each monitored day, the popularity evolution patterns, the number of comments, the number of favorites, etc.

- **Referrer features**. Referrer features are related to the internal recommender systems, the internal search engine, or the internal social sharing tools of the given social platform. For example, the popularity can be dramatically increased if the content is recommended in the front page [27].

As to external features, they are extracted from outside the content's social platform. We divide these features into two classes.

- **Cyber features**. Cyber features are from cyber world, such as the number of incoming links from other social platforms, the recommendation from general search engines, etc. Castillo et al. [12] have shown that the features about a news article on social networks (e.g., the number of Facebook shares) are effective in predicting the popularity of the same article on a news site.

- **Physical features**. Physical features are real-world features with the assumption that the popularity of some Web content is strongly correlated with real-world events. However, it is challenging to transfer information from the physical world to the cyber world. An existing attempt is by Tsagkias et al. [77] who use the weather conditions to predict the number of comments for news articles.

For the above features, user features can be extracted before publication, content features can be extracted at publication time, and context features and referrer features can be extracted after publication. For external features (cyber features and physical features), they could be extracted at any time point. For example, a news article reporting a hot event may be popular, and a hot event may also make an old content popular.

2.2.3 *Algorithms*

Depending on the target measurements, many off-the-shell regression (e.g., SVR, linear regression, etc.) and classification (e.g., SVM, random forest, etc.) methods have been used for popularity prediction. Some other algorithms such as the stochastic process [38; 47], the extremely randomized ensemble trees [31; 26], and the two-step methods [4; 92] are also considered by existing literature. Overall, existing algorithms for content popularity prediction are not designed to fit the underlying social platforms.

2.3 Representatives

Here, we describe some representative popularity prediction methods.

As a pioneering practice, Szabo and Huberman [73] find a strong correlation between the logarithmically transformed past popularity and current popularity, and propose to use only the popularity value at prediction time as input. Following this work, Pinto et al. [59] find that different videos (and Web content in general) may have different temporal patterns in terms of popularity evolution, and they propose to use the view count in each monitored day before prediction time as features. Further, Ahmed et al. [4] propose to identify the temporal patterns of popularity evolution, and use these patterns to predict the future popularity. Specially, they take a two-step method, where the first step is to cluster the patterns based on the similarities of features, and the second step is to model the temporal evolution of popularity over the clusters. Finally, Yu et al. [92] define the concept of phase as a continuous time period in which a video's popularity has a salient rising or falling trend. Adding these phase features into the method by Pinto et al. [59], they achieve further accuracy improvement.

The above four pieces of work use only the context features (more specifically, the view count from publication time to prediction time) as input, and they all take a regression setting to predict the future view count of the content. However, as indicated by Castillo et al. [12], social features and other context features received in the first ten minutes after publication can achieve the same performance with the view count in the first three hours. Therefore, various other types of features that are useful for fast and accurate popularity prediction are considered by existing work. For example, Figueiredo [26] considers social features and content features in addition to context features, and formulates a classification problem to predict different types of popularity trends. Similarly, Vasconcelos et al. [79] consider social features, content features, and context features to predict the popularity levels[1] of micro-reviews (i.e., Foursquare tips); they formulate both regression and classification problems.

Speaking of social features, existing methods have some interesting findings that reveal their importance. For example, Khosla et al. [43] use both social features and image content features to predict the normalized view count. One of their findings is that popularity is difficult to precisely predict based on image content alone, and social features have a great influence on the popularity of images. Their methods do not use context features and referrer features, and thus can be used at publication time.

Intuitively, there is a tradeoff between the prediction time

[1]This work actually predicts the number of likes of Foursquare tips.

and prediction accuracy. Recently, Figueiredo et al. [28] explicitly address this tradeoff, and argue that accurate predictions should be made as early as possible (e.g., before users' interest has exhausted). They take a two-step method of trend extraction and trend prediction, and formulate multi-class classification problem for popularity trend prediction.

2.4 Summary

Content popularity may play important roles in many applications such as media advertising, content caching, and traffic management. Therefore, understanding what makes a piece of content popular and being able to predict its future popularity have attracted a lot of research interest in the past few years. Typically, the view count is used as the popularity measurement, and the main focus of existing work is on the feature extraction aspect. Overall, we summarize the existing features into internal features and external features, where internal features are more commonly used. We further divide internal features into four types, in which user features, content features, and context features are mostly used.

Although not covered in this article, many other methods are related to the popularity prediction of Web content. For example, Anderson et al. [5] propose to predict the long-lasting value which is defined as the pageviews of a question-answer thread in community question answering sites; Roy et al. [65] use transfer learning from the Twitter domain to the video domain to detect popularity burst; Cheng et al. [17; 18] propose to predict the phenomenon of large resharing behaviors in social network services; Cha et al. [15] find that the presence of multiple versions of the same content tends to limit the popularity of each individual copy; Lerman et al. [47] propose to predict which articles will get on the front page. A more thorough survey on these topics can be found in [74].

3. CONTENT QUALITY PREDICTION

In this section, we briefly review the existing work on the measurement and prediction of Web content quality.

3.1 Content Quality Measurement

In general, Web content quality is a vague concept, and it is related to a wide range of concepts like popularity, readability, conciseness, trustworthiness, helpfulness, etc. Additionally, content quality is a subjective concept and it depends on the factors that can be different among different social platforms. In other words, quality measurement tends to differ based on the underlying social platforms. In this article, to address the vagueness issue, we put our focus on the quality measurements that either are labeled by human annotators or can reflect the community users' overall opinions; to address the subjectiveness issue, we summarize the quality measurements on several typical types of social platforms as follows.

- **Collaborative content quality**. Collaborative content is the content that can be collaboratively edited by the registered users. Wikipedia article is a typical example. In Wikipedia, the quality of collaborative content is also collaboratively defined. That is, users can evaluate an article to several quality levels (e.g., Featured Article, A-Class, and B-Class) provided by Wikipedia, and these quality levels serve as the quality measurement [20].
- **Microblog credibility**. Rumors or non-credible messages may spread over microblogs like Twitter and Weibo, and evaluating the credibility of microblog messages becomes an important problem. To this end, human efforts are usually involved to judge if a microblog message corresponds to a newsworthy event [13; 14]. The official services for identifying rumors may also be used [88].
- **Question and answer quality**. Community question answering sites such as Yahoo! Answers and Stack Overflow contain rich knowledge, and thus have attracted much recent attention. There are several types of measurements for question/answer quality in these sites. The first one resorts to human annotators to manually label the question/answer quality [42; 35; 72]; the second one is the so-called questioner satisfaction, indicating whether an answer is chosen as the accepted/best answer by the questioner [53; 82; 66; 76]; the third one is based on the votes on the questions/answers [63; 89; 90; 85].
- **Review/comment helpfulness**. Reviews in many product review sites may contain noisy and misleading information. Therefore, identifying the helpfulness of the reviews becomes an important task in these sites. Similarly, it is also important to identify useful comments in social platforms like YouTube, Flickr, and Digg. To measure the review/comment helpfulness, the voting information is usually employed. Typically, the helpfulness is defined as the number of positive votes divided by the number of total votes [44; 69; 32; 40].

3.2 Content Quality Prediction

As we can see from the above four typical types of content quality problems, the quality is mainly reflected by either the community users' responses or the manual labeling. In the former case, it usually needs a relatively long time to accumulate sufficient user responses to accurately assess the quality. However, it would be more helpful if we can evaluate the content quality as soon as it is published. In the latter case, manual labeling is costly and it cannot scale to large data sets. Consequently, predicting the content quality as soon as it is published, and predicting with the input of a small amount of labeling information become necessary problems.

3.2.1 Prediction Time

The prediction time of quality prediction usually happens at the publication time. Methods in this category believe that many quality factors are inherent in the content itself. Also, these methods usually use human annotators to label quality. When the community's overall response is used as quality measurement, the prediction is usually made after publication. This is due to the fact that social platforms that allow users to evaluate the content quality may also provide other mechanisms where the content quality can be implicitly implied. For after publication methods in this category, the prediction should be made as soon as possible [90]. Although predicting content quality before publication is also feasible, seldom researchers consider this setting.

3.2.2 Features

Basically, the features used for content quality prediction resemble the features for popularity prediction as shown in Fig. 1. For example, Kim et al. [44] explore content features including structural features (e.g., the length of reviews), lexical features (e.g., unigrams), syntactic features (e.g., percentage of verbs and nouns), and semantic features (e.g., product feature mentions) for review helpfulness prediction. Based on these content features, Lu et al. [54] further incorporate the features from the reviewers. They exploit user features like the number of past reviews and social features like the out-degree of the reviewer in the social network.

In addition to the common features, some platform-specific features have been considered by existing content quality prediction methods. For example, when evaluating the quality of collaborative content, review features which are extracted from the review history of each article can be considered [20]. These features (e.g., the number of revisions) are indicative for the maturity and stability of an article. Additionally, since collaborative content involves many citations between each other, features can also be extracted from the citation network. These two types of features are actually complementary. That is, a mature article should be stable; however, it may be stable because no one is interested in it (e.g., in-degree) due to its poor quality. For microblogs, since there are extensive re-sharing behaviors, the propagation features about these re-sharing paths can be considered. Examples include the depth of the re-tweet tree, or the number of initial tweets of a topic [14]. For community question answering sites, complex relationships among questioners, answerers, questions, and answers have been considered [3]. The features like whether the service is free [35] and authors' offline reputation [75] are also considered.

3.2.3 Algorithms

Similar to popularity prediction, many off-the-shell regression and classification methods (e.g., SVR [41], SVM [69], neural networks [46]) are used by existing quality prediction methods.

Based on the characteristics of the underlying social platforms, some advanced algorithms are designed to finish the prediction task. For example, to predict microblog message credibility, Gupta et al. [34] propose to propagate credibility on a constructed network consisting of events, tweets, and users. One of their basic assumptions is that credible users tend to provide credible tweets. Similarly, to predict the quality of questions/answers, Bian et al. [8] propose to propagate the labels through user-question-answer graph with the assumption that reputable users tend to give high-quality questions/answers. These methods are also designed to tackle the sparsity problem where only a small number of messages/questions/answers are labeled. In review helpfulness prediction, Lu et al. [54] study two methods for incorporating social context into the prediction process: either as features, or as regularization constraints. The proposed regularization framework may also be used when a small amount of data is labeled.

3.3 Representatives

Here, we present some representatives.

Collaborative content quality. Dalip et al. [20] use SVR to estimate the quality of Wikipedia articles. They use content features and revision features. The content features include the length, style, readability, and structure of the article, and the citations between articles. Take structure as an example. The basic intuition is that a good article must be organized clearly, and it should provide necessary references to additional material. Therefore, features like section count and reference count are extracted. For revision features which are extracted from the revision history, they consider features like the discussion count, the review count, and the number of revisions in the last three months. The quality measurement is based on users' ratings, and user ratings are based on the quality criteria defined by Wikipedia which considers completeness, neutral point-of-view, good organization, factual accuracy, and provision of references to relevant sources of information.

Microblog credibility. Castillo et al. [13; 14] propose to assess the credibility of tweets. They first find some tweets that are about current hot topics, and then classify them as credible or non-credible by a group of human annotators. Next, they formulate a classification problem by considering a set of features. The considered features include content features (e.g., twitter-independent features like the length of a message and the number of positive/negative sentiment words, and Twitter-dependent features like if the tweet contains a hashtag and if the message is a re-tweet), user features (e.g., the accout age, the number of followers, and the number of followees), as well as propagation features as we previously mentioned. They show that the way microblog messages propagate (i.e., propagation features) is indicative for message credibility.

Question and answer quality. Jeon et al. [42] propose to use non-textual features to predict answer quality and incorporate it to improve retrieval performance. The used features include answerer's acceptance ratio, answerer's activity level, and answerer's category specialty. Kernel density estimation and the maximum entropy approach are used to handle different types of features, and a stochastic process is built to make the prediction. The quality labels are annotated by humans.

Agichtein et al. [3] propose to predict the quality of both questions and answers. They use some common features including content features and user features. They also extract several platform-specific features from the user-question-answer graph. Example platform-specific features include the average number of answers with references (URLs) given by the asker of the current question, the average number of positive votes received by answers written by the asker of the current question, etc. Finally, they lable the quality of questions and answers by human annotators, and use stochastic gradient boosted trees [29] to classify high-quality questions/answers.

Review/Comment helpfulness. Ghose and Ipeirotis [32] use content features and user features to predict the review helpfulness. The content features include subjectivity aspects and readability aspects. They find that reviews that have a mixture of objective and subjective sentences are rated more helpful. The user features consider the users' performance in previous reviews. They first use positive votes and negative votes to define quality, and then formulate a classification problem by choosing a threshold based on two human annotators. They adopt the random forest-based classifiers.

As to comment quality/helpfulness, Momeni et al. [56] propose to classify useful comments on YouTube and Flickr with logistic regression and Naive Bayes. They use content features and user features. User features describe the users' posting and social behaviors (e.g., the size of the user's network). Content features include syntactic, semantic, and topical features. By human labeling, they show that semantic and topical features are very important for accurate classification for both Flickr and YouTube.

3.4 Summary

Measuring and predicting the content quality can help to identify some valuable information from the large amount of content in social platforms. In this review, we mainly focus on four types of quality measurement and prediction settings, i.e., the quality of collaborative articles, the credibility of microblog messages, the quality of questions and answers, and the helpfulness of reviews and comments. Different from popularity prediction in the previous section, content quality is a vague concept and many existing methods use human annotators to label the quality. However, human labeling is costly and not scalable. Therefore, community users' responses are also widely used as quality. As to the prediction, some platform-specific features and algorithms are designed by existing methods.

In this section, we only summarize some typical and mainstream social platforms. Within the covered social platforms, other types of measurements that are related to content quality have also been studied. In community question answering sites, examples include the usefulness of questions (i.e., the possibility that a question would be repeated by other people) [71], the arrival speed of the best answer [48], whether a question will be answered [25], site searcher satisfaction (i.e., whether or not the answer satisfies the information searcher) [52], and the subjectiveness of questions and answers [49; 95].

4. SCIENTIFIC IMPACT PREDICTION

In this section, we put our focus on the measurement and prediction of the impact of scientific articles.

4.1 Impact Measurement for Scientific Articles

Typically, the impact of a research article is closely related to the citations of this article, and the study of research article citations can date back to the 1960s and 1970s [21; 60]. Recently, it attracts attention after the citation count prediction competition in KDD Cup 2003 [30].

In literature, there are several ways to measure the impact of scientific articles, which are summarized as follows.

The first measurement is the number of articles that cite the given article (i.e., citation count). Citation count is a most straightforward measurement for article impact, and it is adopted by most existing work (e.g., [86; 81; 16; 70]). However, as noticed by several researchers, citation count has several disadvantages. For example, it tends to favor those aging articles, and it may be quite different in different fields [57].

The second measurement neglects the actual number of citations, but focuses on whether the citation count of a given scientific article is above the average (or above a certain percentile). For example, Newman [58; 57] proposes to predict highly cited articles, and argues that pure citation count may not be a good indicator for article impact because it is

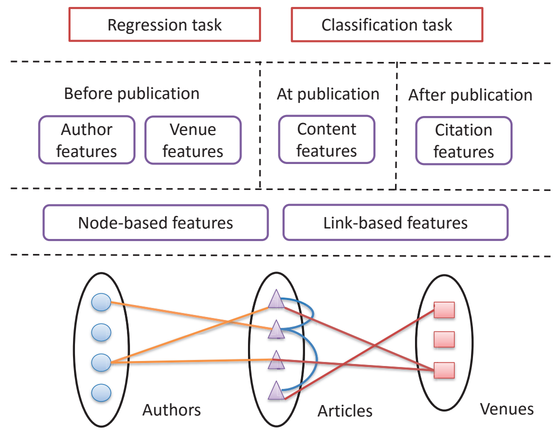

Figure 2: The features for scientific article impact prediction.

highly related to publication dates. Compared to citation count which can be viewed as a regression problem, this measurement can be viewed as a classification problem.

In addition to the above two measurements, Dong et al. recently propose to measure the impact of an article based on whether this article will increase its authors' h-indices in the future [22; 23]. This measurement can be seen as a personalized article impact as it depends on the authors of the article.

In addition to the above measurements in literature, other measurements for article impact may also be further studied. For example, we can consider the downloads of an article, and ratio between the citations and the downloads, the citation differences among different fields, etc.

4.2 Impact Prediction for Scientific Articles

Next, we discuss the prediction time, features, and algorithms for scientific article impact prediction.

4.2.1 *Prediction Time*

Most existing work makes the prediction after publication time. For scientific articles, predicting their impact is not as urgent as the popularity/quality prediction of Web content, as scientific articles have a longer lifecycle than the general content on social media. For example, Wang et al. [81] make the citation prediction for an article five or even ten years after publication.

Next, since the research community is more stable and controlled, the research impact is more predictable than the content popularity. For example, some existing methods find that accurate predictions can be made when the prediction time is at publication or even before publication (e.g. [87; 86]).

4.2.2 *Features*

The features for article impact prediction are summarized in Fig. 2. First, the features are extracted from the heterogeneous network among authors, articles, and venues. In this heterogeneous network, authors write articles, articles are published on venues, and articles can cite each other. Although not explicitly drawn in the figure, authors are also related to authors (by collaboration or citation) and venues

(by publishing on the venues).
Rich features can be extracted from the heterogeneous network. We divide these features into node-based features and link-based features. Authors, articles, and venues all have some attributes from which node-based features can be extracted. For example, we can extract the total number of citations of authors, the topics of articles, and the h-indices of venues. For link-based features, we can apply PageRank-like algorithms on the network to achieve the authority of authors, centrality of venues, etc.
Based on the two types of node-based and link-based features, we further divide these features into four classes, i.e., author features, venue features, content features, and citation features.

- **Author features**. Author features reflect how well the authors have performed before the publication of the current article. Intuitively, these features may indicate the future performance of the authors. Typical example features include author productivity (e.g., the number of authors' previous publications), author influence (e.g., the number of authors' citations/h-indices), author sociality (e.g., the number of co-authors), author authority (e.g., PageRank value on author-author citation network), and authors diversity (e.g., the research fields that the author publishes).
- **Venue features**. Venue features indicate the impact of venues. High-impact venues may attract much attention from related scholars, and thus have a better chance to be cited. Similar to author features, venue features can be extracted before publication. Typical example features include venue influence (e.g., its h-index and citations), venue centrality (e.g., the PageRank value on venue-venue citation network), and venue diversity (e.g., the diversity of its articles).
- **Content features**. Intuitively, the content of the scientific article is a major factor that can affect its impact. To extract content features, topic modeling is widely used in existing literature. Typical content features include content novelty (e.g., the similarity between the article and its references), content diversity (e.g., the topical breadth of the article), and content popularity (e.g., the popularity of content topic).
- **Citation features**. Finally, when the prediction time can be lagged after publication, citation features may be very indicative for the impact prediction of scientific articles. The intuition behind is the richer-get-richer phenomenon. Here, the citation features are extracted based on the citations in a time window after publication. In fact, several researchers propose to predict the future citations of an article by fitting a function on the citations during the time window [81; 67]. Other citation features such as the total citations in the time window and the number of countries citing the current article in the time window have also been considered [93].

In addition to the above four classes of features, some researchers claim that they have considered the temporal features [87; 86]. Basically, these temporal features are applied on a recent short-term publication data, and the extracted features can also be categorized into the above four classes.

4.2.3 Algorithms

Based on the extracted features and the prediction target (i.e., impact measurement), the next component is the algorithm used to finish the prediction. Depending on the impact measurement, regression task and classification task can be formulated. For the regression task, methods such as linear regression, Gaussian Process regression [62], and SVR have been considered by existing literature. For the classification task, existing methods have considered logistic regression, decision trees, random forest, SVM, etc.
Unlike content quality prediction where platform-specific algorithms are designed, the characteristics of the underlying author-article-venue network is relatively ignored in terms of the algorithm level. We will discuss this more in Section 5.

4.3 Representatives

Here, we describe some representative methods for scientific impact prediction.
As an early practice, Castillo et al. [11] propose to extract both node-based and link-based author features as well as the number of citations after publication to predict the future citations. They state that their prediction can be done both before and after publication depending on whether to include the citation features. They formulate both regression and classification (top 10% cited articles as successful articles) problems, and use linear regression and C4.5 decision trees for them, respectively.
Although Castillo et al. [11] have considered a wide range of aspects, the extracted features are relatively simple. Yan et al. [87; 86] create a short-term data and extract content features, author features, and venue features from this short-term data as well as the full data. Since they do not use citation features, their prediction time is at publication or even before publication (if the content features are not used). As for prediction algorithm, they formulate a regression problem, and use both CART (classification and regression tree) and Gaussian Process regression to predict the citations after five and ten years.
The citation features are also widely used. For example, Chakraborty et al. [16] add first-year citation as features, and consider content features, author features, and venue features. The vary the time when the ground truth value of impact is computed, and observe the evolution of feature importance. One of their findings is that content features seem to have less significance in the initial few years, and become more effective in the later years. This result indicates that, to some extent, a high-quality article would eventually gain high impact irrespective of the reputation of the authors and the publication venue. Following Chakraborty et al. [16], Singh et al. [70] consider more citation features such as the number of times an article is cited within the same article and the number of words within the citation context. Both Chakraborty et al. [16] and Singh et al. [70] adopt a two-step regression formulation. That is, they first use SVM to divide articles into several predefined categories (based on citation patterns) and then use SVR to predict citations.
Taking only the citation information after publication as input is also studied. For example, Wang et al. [81] and Shen et al. [67] propose to mine the citation patterns of each article to predict its future citations. To fit the pattern function, these methods consider three factors, i.e., preferential

attachment (highly-cited articles are more visible and are more likely to be cited again), aging (article's novelty fades), and fitness (the inherent competitiveness against other articles).

Recently, Dong et al. [22; 23] propose to infer whether a given article will increase its authors' h-indices in the future. They use content features, author features, and venue features, and the prediction time is at publication. In content features, they further extract some reference features such as the average citations of the reference articles in the current article. They find that authors' authority on the topic and the publication venue are key factors for citations.

4.4 Summary

Measuring and predicting the impact of scientific articles have gained attention in recent years. Overall, we find three types of measurements for scientific article impact, and categorize four types of features for two types of tasks. A key difference from general content popularity/quality prediction is that article's scientific impact is more predictable and it can be accurately predicted at or even before publication. There are some interesting findings from existing literature. For example, authors' authority and publication venue are found to be the key factors for citations, while content features become more effective as time goes by.

In addition to scientific articles, some researchers propose to measure and predict the impact of research scholars [2; 24]. For example, h-index is proposed to measure both the productivity and impact of a research scholar [37]; later, the differences of disciplines and years are incorporated into h-index [61]; finally, the number of top-venue publications and the citations of big-hit publications are also used [24]. Some other researchers propose to measure the quality of conference/journal venues based on the overall citations of the venue [55] or the program committee characteristics [97]. Other related directions include recommending research articles for potential readers [80], recommending previous articles for citation [64], predicting citation relationships [94], ranking existing research articles [83], summarizing the research articles from multiple scholars viewpoints [1], mining the citation patterns for different articles [68], revealing the relationships between impact and combination of prior work [78], etc.

5. DISCUSSIONS AND FUTURE DIRECTIONS

In this section, we discuss some insights and present some future directions.

Prediction time v.s. prediction accuracy. Intuitively, there is a tradeoff between prediction time and prediction accuracy. An illustrative example is shown in Fig. 3. Time T1 in the figure means before publication. The accuracy is relatively low as only a small set of features are available before publication. Time T2 means at publication. At publication time, some useful features may be extracted from the published content. The curve means accuracy trend after publication. Intuitively, the prediction accuracy at time T4 would be higher than that at time T3, as more information is available. However, in many prediction tasks, the prediction should be made as early as possible, resulting a tradeoff between prediction accuracy and prediction time.

Based on our review of existing literature, seldom work has explicitly addressed this tradeoff. There are two reasons.

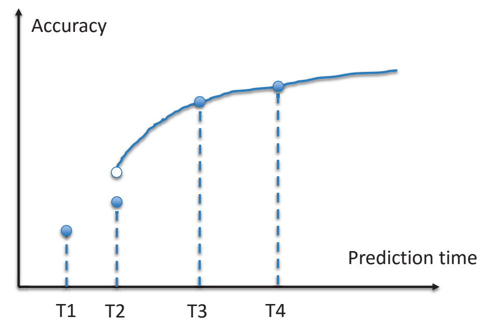

Figure 3: Prediction time v.s. prediction accuracy.

First, it is difficult to determine the time when the users' interest is exhausted. Second, experimenting over many time points would be costly. A possible future direction would be determining a suitable prediction time for different tasks. Take popularity prediction as an example. We may find the time when at least a certain (significant) amount of attention or a certain percentage of attention has not been received, especially for those highly popular content.

Learn from each other. The reviewed three tasks (i.e., content popularity prediction, content quality prediction, and scientific article impact prediction) bear some subtle differences, and they can borrow experiences and ideas from each other.

First, referrer features and external features have been considered by popularity prediction. Such features can also be used for the other prediction tasks. Take the impact prediction of scientific articles as an example. Whether the article is freely available on the Web and whether the authors have released their codes may also be helpful for the impact prediction. Second, some platform-specific features and algorithms are designed for content quality prediction. These ideas may also work for the other two tasks. Third, in scientific article impact prediction, the importance of features over time has been studied. Based on this study, some interesting findings have been observed. Similar study can be applied on the other tasks.

New measurements. For scientific impact, new measurements can be defined. The first direction is to make use of the popularity information. As mentioned above, we can consider the ratio between the number of citations and the number of downloads. The second direction is to add weights on the citation articles (i.e., the articles that cite the given article). Instead of pure citation count, we may define impact as weighted sum/average of the citation articles where the weight is based on the impact of the citation article itself. The intuition behind is that a citation from a high-impact article is more meaningful than that from a low-impact article. The third direction is to consider pathway measurements [51]. That is, instead of predicting the citation count at a given future time point, we can predict the citation numbers over several years. Such prediction is capable of visioning the impact trends of articles.

New algorithms. In the algorithm aspect, many existing methods apply off-the-shell machine learning algorithms to make the predictions. Here, some advanced algorithms that take the characteristics of the underlying social platforms into account can be designed. Existing examples in-

clude Agichtein et al. [3] who construct user-question-answer graph for the question/answer quality prediction task. However, their goal is to facilitate the label sparsity problem. More general methods for improving the prediction accuracy when we have enough labels are still needed. Recently, some efforts have been made towards this direction. For example, in question answering sites, the question-answer correlation has been considered to improve prediction accuracy [91]; in scientific research filed, the differences between different research domains have been considered [50].
Another future direction is to view the problem as a prediction problem on a data stream. Web content can be generated in a rapid speed. Therefore, a future direction is to design online algorithms that can efficiently make the predictions.
Additionally, some authors argue that the relative ranking rather than the actual value is more important [36]. However, they still use actual value as the prediction target. Here, pair-wised prediction target such as the learning to rank methods can be considered.
Beyond Web content. In this article, we focus on the measurement and prediction of Web content utility, where we define Web content as any individual item (in the form of text, image, or video) available on a web site in which a certain level of interest can be reflects by the community users. In addition to measuring and predicting on an individual item, we may also consider to predict the popularity or the research impact of a certain topic. Considering the fact the advertising usually choose a set of nodes, predicting the popularity of a given set of content nodes is also potentially useful.

6. CONCLUSIONS

In this article, we have reviewed the existing literature for the measurement and prediction of Web content. Specially, we focus on three tasks, i.e., content popularity prediction, content quality prediction, and scientific article impact prediction. We analyze the measurement definition, the prediction time, the extracted features, and the employed algorithms for each task. Finally, we identify the connections and differences between these three tasks, and discuss some future possible directions.

Acknowledgements

This work is supported by the National Key Research and Development Program of China (No. 2016YFB1000802), National 863 Program of China (No. 2015AA01A203), National Natural Science Foundation of China (No. 61702252, 61672274, 61321491), the Fundamental Research Funds for the Central Universities (No. 020214380033), and the Collaborative Innovation Center of Novel Software Technology and Industrialization. Hanghang Tong is partially supported by DTRA under the grant number HDTRA1-16-0017, by Army Research Office under the contract number W911NF-16-1-0168, by National Institutes of Health under the grant number R01LM011986, Region II University Transportation Center under the project number 49997-33 25 and a Baidu gift.

7. REFERENCES

[1] A. Abu-Jbara and D. Radev. Coherent citation-based summarization of scientific papers. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 500–509. Association for Computational Linguistics, 2011.

[2] D. E. Acuna, S. Allesina, and K. P. Kording. Future impact: Predicting scientific success. *Nature*, 489(7415):201–202, 2012.

[3] E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne. Finding high-quality content in social media. In *WSDM*, pages 183–194. ACM, 2008.

[4] M. Ahmed, S. Spagna, F. Huici, and S. Niccolini. A peek into the future: predicting the evolution of popularity in user generated content. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 607–616. ACM, 2013.

[5] A. Anderson, D. Huttenlocher, J. Kleinberg, and J. Leskovec. Discovering value from community activity on focused question answering sites: a case study of stack overflow. In *KDD*, pages 850–858. ACM, 2012.

[6] I. Arapakis, B. B. Cambazoglu, and M. Lalmas. On the feasibility of predicting news popularity at cold start. In *International Conference on Social Informatics*, pages 290–299. Springer, 2014.

[7] J. Berger and K. L. Milkman. What makes online content viral? *Journal of marketing research*, 49(2):192–205, 2012.

[8] J. Bian, Y. Liu, D. Zhou, E. Agichtein, and H. Zha. Learning to recognize reliable users and content in social media with coupled mutual reinforcement. In *WWW*, pages 51–60. ACM, 2009.

[9] Y. Borghol, S. Ardon, N. Carlsson, D. Eager, and A. Mahanti. The untold story of the clones: content-agnostic factors that impact youtube video popularity. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1186–1194. ACM, 2012.

[10] A. Brodersen, S. Scellato, and M. Wattenhofer. Youtube around the world: geographic popularity of videos. In *Proceedings of the 21st international conference on World Wide Web*, pages 241–250. ACM, 2012.

[11] C. Castillo, D. Donato, and A. Gionis. Estimating number of citations using author reputation. In *International Symposium on String Processing and Information Retrieval*, pages 107–117. Springer, 2007.

[12] C. Castillo, M. El-Haddad, J. Pfeffer, and M. Stempeck. Characterizing the life cycle of online news stories using social media reactions. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 211–223. ACM, 2014.

[13] C. Castillo, M. Mendoza, and B. Poblete. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web*, pages 675–684. ACM, 2011.

[14] C. Castillo, M. Mendoza, and B. Poblete. Predicting information credibility in time-sensitive social media. *Internet Research*, 23(5):560–588, 2013.

[15] M. Cha, H. Kwak, P. Rodriguez, Y.-Y. Ahn, and S. Moon. Analyzing the video popularity characteristics of large-scale user generated content systems. *IEEE/ACM Transactions on Networking (TON)*, 17(5):1357–1370, 2009.

[16] T. Chakraborty, S. Kumar, P. Goyal, N. Ganguly, and A. Mukherjee. Towards a stratified learning approach to predict future citation counts. In *Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 351–360. IEEE Press, 2014.

[17] J. Cheng, L. Adamic, P. A. Dow, J. M. Kleinberg, and J. Leskovec. Can cascades be predicted? In *Proceedings of the 23rd international conference on World wide web*, pages 925–936. ACM, 2014.

[18] J. Cheng, L. A. Adamic, J. M. Kleinberg, and J. Leskovec. Do cascades recur? In *Proceedings of the 25th International Conference on World Wide Web*, pages 671–681. International World Wide Web Conferences Steering Committee, 2016.

[19] A. Choudhury and S. Nagaswamy. A bayesian approach to identify photos likely to be more popular in social media. *Journal of Computer and Communications*, 3(11):198, 2015.

[20] D. H. Dalip, M. A. Gonçalves, M. Cristo, and P. Calado. Automatic assessment of document quality in web collaborative digital libraries. *Journal of Data and Information Quality (JDIQ)*, 2(3):14, 2011.

[21] D. J. de Solla Price. Networks of scientific papers. *Science*, 149(3683):510–515, 1965.

[22] Y. Dong, R. A. Johnson, and N. V. Chawla. Will this paper increase your h-index?: Scientific impact prediction. In *WSDM*, pages 149–158. ACM, 2015.

[23] Y. Dong, R. A. Johnson, and N. V. Chawla. Can scientific impact be predicted? *IEEE Transactions on Big Data*, 2(1):18–30, 2016.

[24] Y. Dong, R. A. Johnson, Y. Yang, and N. V. Chawla. Collaboration signatures reveal scientific impact. In *2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 480–487. IEEE, 2015.

[25] G. Dror, Y. Maarek, and I. Szpektor. Will my question be answered? predicting question answerability in community question-answering sites. In *ECML/PKDD*, pages 499–514, 2013.

[26] F. Figueiredo. On the prediction of popularity of trends and hits for user generated videos. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 741–746. ACM, 2013.

[27] F. Figueiredo, J. M. Almeida, M. A. Gonçalves, and F. Benevenuto. On the dynamics of social media popularity: A youtube case study. *ACM Transactions on Internet Technology (TOIT)*, 14(4):24, 2014.

[28] F. Figueiredo, J. M. Almeida, M. A. Gonçalves, and F. Benevenuto. Trendlearner: Early prediction of popularity trends of user generated content. *Information Sciences*, 349:172–187, 2016.

[29] J. H. Friedman. Stochastic gradient boosting. *Computational Statistics & Data Analysis*, 38(4):367–378, 2002.

[30] J. Gehrke, P. Ginsparg, and J. Kleinberg. Overview of the 2003 kdd cup. *ACM SIGKDD Explorations Newsletter*, 5(2):149–151, 2003.

[31] P. Geurts, D. Ernst, and L. Wehenkel. Extremely randomized trees. *Machine learning*, 63(1):3–42, 2006.

[32] A. Ghose and P. G. Ipeirotis. Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. *IEEE Transactions on Knowledge and Data Engineering*, 23(10):1498–1512, 2011.

[33] R. E. Guadagno, D. M. Rempala, S. Murphy, and B. M. Okdie. What makes a video go viral? an analysis of emotional contagion and internet memes. *Computers in Human Behavior*, 29(6):2312–2319, 2013.

[34] M. Gupta, P. Zhao, and J. Han. Evaluating event credibility on twitter. In *SDM*, pages 153–164. SIAM, 2012.

[35] F. Harper, D. Raban, S. Rafaeli, and J. Konstan. Predictors of answer quality in online q&a sites. In *CHI*, pages 865–874. ACM, 2008.

[36] X. He, M. Gao, M.-Y. Kan, Y. Liu, and K. Sugiyama. Predicting the popularity of web 2.0 items based on user comments. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, pages 233–242. ACM, 2014.

[37] J. E. Hirsch. An index to quantify an individual's scientific research output. *Proceedings of the National academy of Sciences of the United States of America*, pages 16569–16572, 2005.

[38] T. Hogg and K. Lerman. Stochastic models of user-contributory web sites. In *Third International AAAI Conference on Weblogs and Social Media*, 2009.

[39] L. Hong, O. Dan, and B. D. Davison. Predicting popular messages in twitter. In *Proceedings of the 20th international conference companion on World wide web*, pages 57–58. ACM, 2011.

[40] Y. Hong, J. Lu, J. Yao, Q. Zhu, and G. Zhou. What reviews are satisfactory: novel features for automatic helpfulness voting. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 495–504. ACM, 2012.

[41] C.-F. Hsu, E. Khabiri, and J. Caverlee. Ranking comments on the social web. In *International Conference on Computational Science and Engineering*, volume 4, pages 90–97. IEEE, 2009.

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

[43] A. Khosla, A. Das Sarma, and R. Hamid. What makes an image popular? In *Proceedings of the 23rd international conference on World wide web*, pages 867–876. ACM, 2014.

[44] S.-M. Kim, P. Pantel, T. Chklovski, and M. Pennacchiotti. Automatically assessing review helpfulness. In *Proceedings of the 2006 Conference on empirical methods in natural language processing*, pages 423–430. Association for Computational Linguistics, 2006.

[45] J. G. Lee, S. Moon, and K. Salamatian. Modeling and predicting the popularity of online contents with cox proportional hazard regression model. *Neurocomputing*, 76(1):134–145, 2012.

[46] S. Lee and J. Y. Choeh. Predicting the helpfulness of online reviews using multilayer perceptron neural networks. *Expert Systems with Applications*, 41(6):3041–3046, 2014.

[47] K. Lerman and T. Hogg. Using a model of social dynamics to predict popularity of news. In *Proceedings of the 19th international conference on World wide web*, pages 621–630. ACM, 2010.

[48] B. Li, T. Jin, M. R. Lyu, I. King, and B. Mak. Analyzing and predicting question quality in community question answering services. In *WWW*, pages 775–782. ACM, 2012.

[49] B. Li, Y. Liu, and E. Agichtein. Cocqa: co-training over questions and answers with an application to predicting question subjectivity orientation. In *EMNLP*, pages 937–946, 2008.

[50] L. Li and H. Tong. The child is father of the man: Foresee the success at the early stage. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 655–664. ACM, 2015.

[51] L. Li, H. Tong, J. Tang, and W. Fan. *iPath: Forecasting the Pathway to Impact*, pages 468–476. 2016.

[52] Q. Liu, E. Agichtein, G. Dror, E. Gabrilovich, Y. Maarek, D. Pelleg, and I. Szpektor. Predicting web searcher satisfaction with existing community-based answers. In *SIGIR*, pages 415–424, 2011.

[53] Y. Liu, J. Bian, and E. Agichtein. Predicting information seeker satisfaction in community question answering. In *SIGIR*, pages 483–490. ACM, 2008.

[54] Y. Lu, P. Tsaparas, A. Ntoulas, and L. Polanyi. Exploiting social context for review quality prediction. In *Proceedings of the 19th international conference on World wide web*, pages 691–700. ACM, 2010.

[55] W. S. Martins, M. A. Gonçalves, A. H. Laender, and N. Ziviani. Assessing the quality of scientific conferences based on bibliographic citations. *Scientometrics*, 83(1):133–155, 2010.

[56] E. Momeni, C. Cardie, and M. Ott. Properties, prediction, and prevalence of useful user-generated comments for descriptive annotation of social media objects. In *ICWSM*, 2013.

[57] M. Newman. Prediction of highly cited papers. *EPL (Europhysics Letters)*, 105(2):28002, 2014.

[58] M. E. Newman. The first-mover advantage in scientific publication. *EPL (Europhysics Letters)*, 86(6):68001, 2009.

[59] H. Pinto, J. M. Almeida, and M. A. Gonçalves. Using early view patterns to predict the popularity of youtube videos. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 365–374. ACM, 2013.

[60] D. d. S. Price. A general theory of bibliometric and other cumulative advantage processes. *Journal of the American society for Information science*, 27(5):292–306, 1976.

[61] F. Radicchi, S. Fortunato, and C. Castellano. Universality of citation distributions: Toward an objective measure of scientific impact. *Proceedings of the National Academy of Sciences*, 105(45):17268–17272, 2008.

[62] C. E. Rasmussen. Gaussian processes in machine learning. In *Advanced lectures on machine learning*, pages 63–71. Springer, 2004.

[63] S. Ravi, B. Pang, V. Rastogi, and R. Kumar. Great question! question quality in community q&a. In *ICWSM*, pages 426–435, 2014.

[64] X. Ren, J. Liu, X. Yu, U. Khandelwal, Q. Gu, L. Wang, and J. Han. Cluscite: Effective citation recommendation by information network-based clustering. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 821–830. ACM, 2014.

[65] S. D. Roy, T. Mei, W. Zeng, and S. Li. Towards cross-domain learning for social video popularity prediction. *IEEE Transactions on multimedia*, 15(6):1255–1267, 2013.

[66] C. Shah and J. Pomerantz. Evaluating and predicting answer quality in community qa. In *SIGIR*, pages 411–418, 2010.

[67] H. Shen, D. Wang, C. Song, and A.-L. Barabási. Modeling and predicting popularity dynamics via reinforced poisson processes. In *AAAI*, pages 291–297. AAAI Press, 2014.

[68] X. Shi, J. Leskovec, and D. A. McFarland. Citing for high impact. In *Proceedings of the 10th annual joint conference on Digital libraries*, pages 49–58. ACM, 2010.

[69] S. Siersdorfer, S. Chelaru, W. Nejdl, and J. San Pedro. How useful are your comments?: analyzing and predicting youtube comments and comment ratings. In *Proceedings of the 19th international conference on World wide web*, pages 891–900. ACM, 2010.

[70] M. Singh, V. Patidar, S. Kumar, T. Chakraborty, A. Mukherjee, and P. Goyal. The role of citation context in predicting long-term citation profiles: An experimental study based on a massive bibliographic text dataset. In *CIKM*, pages 1271–1280. ACM, 2015.

[71] Y. Song, C. Lin, Y. Cao, and H. Rim. Question utility: A novel static ranking of question search. In *AAAI*, pages 1231–1236, 2008.

[72] M. Suryanto, E. Lim, A. Sun, and R. Chiang. Quality-aware collaborative question answering: methods and evaluation. In *WSDM*, pages 142–151. ACM, 2009.

[73] G. Szabo and B. A. Huberman. Predicting the popularity of online content. *Communications of the ACM*, 53(8):80–88, 2010.

[74] A. Tatar, M. D. de Amorim, S. Fdida, and P. Antoniadis. A survey on predicting the popularity of web content. *Journal of Internet Services and Applications*, 5(1):1, 2014.

[75] Y. Tausczik and J. Pennebaker. Predicting the perceived quality of online mathematics contributions from users' reputations. In *CHI*, pages 1885–1888. ACM, 2011.

[76] Q. Tian, P. Zhang, and B. Li. Towards predicting the best answers in community-based question-answering services. In *ICWSM*, 2013.

[77] M. Tsagkias, W. Weerkamp, and M. De Rijke. Predicting the volume of comments on online news stories. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 1765–1768. ACM, 2009.

[78] B. Uzzi, S. Mukherjee, M. Stringer, and B. Jones. Atypical combinations and scientific impact. *Science*, 342(6157):468–472, 2013.

[79] M. Vasconcelos, J. M. Almeida, and M. A. Gonçalves. Predicting the popularity of micro-reviews: A foursquare case study. *Information Sciences*, 325:355–374, 2015.

[80] C. Wang and D. M. Blei. Collaborative topic modeling for recommending scientific articles. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 448–456. ACM, 2011.

[81] D. Wang, C. Song, and A.-L. Barabási. Quantifying long-term scientific impact. *Science*, 342(6154):127–132, 2013.

[82] X.-J. Wang, X. Tu, D. Feng, and L. Zhang. Ranking community answers by modeling question-answer relationships via analogical reasoning. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 179–186. ACM, 2009.

[83] Y. Wang, Y. Tong, and M. Zeng. Ranking scientific articles by exploiting citations, authors, journals, and time information. In *AAAI*, 2013.

[84] S. Weber and J. Luo. What makes an open source code popular on git hub? In *2014 IEEE International Conference on Data Mining Workshop*, pages 851–855. IEEE, 2014.

[85] X. Wei, H. Huang, C.-Y. Lin, X. Xin, X. Mao, and S. Wang. Re-ranking voting-based answers by discarding user behavior biases. In *AAAI*, pages 2380–2386, 2015.

[86] R. Yan, C. Huang, J. Tang, Y. Zhang, and X. Li. To better stand on the shoulder of giants. In *Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries*, pages 51–60. ACM, 2012.

[87] R. Yan, J. Tang, X. Liu, D. Shan, and X. Li. Citation count prediction: learning to estimate future citations for literature. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 1247–1252. ACM, 2011.

[88] F. Yang, Y. Liu, X. Yu, and M. Yang. Automatic detection of rumor on sina weibo. In *Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics*, page 13. ACM, 2012.

[89] Y. Yao, H. Tong, T. Xie, L. Akoglu, F. Xu, and J. Lu. Joint voting prediction for questions and answers in cqa. In *ASONAM*, 2014.

[90] Y. Yao, H. Tong, T. Xie, L. Akoglu, F. Xu, and J. Lu. Detecting high-quality posts in community question answering sites. *Information Sciences*, 302:70–82, 2015.

[91] Y. Yao, H. Tong, F. Xu, and J. Lu. Predicting long-term impact of cqa posts: a comprehensive viewpoint. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1496–1505. ACM, 2014.

[92] H. Yu, L. Xie, and S. Sanner. The lifecyle of a youtube video: Phases, content and popularity. In *Ninth International AAAI Conference on Web and Social Media*, 2015.

[93] T. Yu, G. Yu, P.-Y. Li, and L. Wang. Citation impact prediction for scientific papers using stepwise regression analysis. *Scientometrics*, 101(2):1233–1252, 2014.

[94] X. Yu, Q. Gu, M. Zhou, and J. Han. Citation prediction in heterogeneous bibliographic networks. In *SDM*, volume 12, pages 1119–1130. SIAM, 2012.

[95] T. Zhou, X. Si, E. Chang, I. King, and M. Lyu. A data-driven approach to question subjectivity identification in community question answering. In *AAAI*, 2012.

[96] H. Zhu, C. Liu, Y. Ge, H. Xiong, and E. Chen. Popularity modeling for mobile apps: A sequential approach. *IEEE transactions on cybernetics*, 45(7):1303–1314, 2015.

[97] Z. Zhuang, E. Elmacioglu, D. Lee, and C. L. Giles. Measuring conference quality by mining program committee characteristics. In *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*, pages 225–234. ACM, 2007.