A survey on predicting the popularity of web content

Tatar *et al. Journal of Internet Services and Applications* 2014, **5**:8
http://www.jisajournal.com/content/5/1/8

Journal of Internet Services and Applications
a SpringerOpen Journal

RESEARCH ARTICLE Open Access

A survey on predicting the popularity of web content

Alexandru Tatar[1*], Marcelo Dias de Amorim[1], Serge Fdida[1] and Panayotis Antoniadis[2]

Abstract

Social media platforms have democratized the process of web content creation allowing mere consumers to become creators and distributors of content. But this has also contributed to an explosive growth of information and has intensified the online competition for users attention, since only a small number of items become popular while the rest remain unknown. Understanding what makes one item more popular than another, observing its popularity dynamics, and being able to predict its popularity has thus attracted a lot of interest in the past few years. Predicting the popularity of web content is useful in many areas such as network dimensioning (e.g., caching and replication), online marketing (e.g., recommendation systems and media advertising), or real-world outcome prediction (e.g., economical trends). In this survey, we review the current findings on web content popularity prediction. We describe the different popularity prediction models, present the features that have shown good predictive capabilities, and reveal factors known to influence web content popularity.

Keywords: Web content; Social media; Popularity; Prediction

1 Introduction

In the digital world, web content has become the main attraction. Whether it is useful information and entertainment to Internet users or a business opportunity for marketing companies and content providers, web content is a valuable asset on the Internet. At the same time, the growth in social media innovation, the ease of content creation and low publishing costs has created a world saturated with information. For example, every minute, users around the world send more than 300,000 tweets [1], share more than 680,000 pieces of content on Facebook [2], and upload 100 hours of video on YouTube [3]. Yet the online ecosystem adheres to a "winner-take-all" society: the attention is concentrated on only a few items. In this context, identifying the web content that will become popular becomes of utmost importance. Online users, flooded by information, can reduce the clutter and focus their attention – the most valuable resource in the online world – on the most relevant information for them. In a world where companies spend up to 30% of their budget on online marketing [4], early detection of the next rising star of the Internet can maximize their revenues through better ad placement. Moreover, given the ever-growing consumer Internet traffic, content-distribution networks can rely on popularity prediction methods to proactively allocate resources according to the future users' demand.

But predicting the popularity of web content is a challenging task. First, different factors known to influence content popularity, such as the quality of the content or its relevance to users, are difficult to measure. Then, other factors, such as the relationship between events in the physical world and the content itself are hard to capture and included in a prediction model. Moreover, at a microscopic level, the evolution of content popularity may be described by complex online interactions and information cascades that are difficult to predict [5-7].

Predicting the popularity of web content has become an active area of research and, while still in an incipient phase, a large number of prediction methods for different types of web content have been proposed in the latest years. In this article we review the current state of research in this field, identify trends, and suggest domains that can benefit from these studies. To the best of our knowledge there has been no prior attempt to summarize this research area. The closest to our work is the

*Correspondence: tatar@npa.lip6.fr
[1]LIP6/CNRS – UPMC Sorbonne Universités, 4 Place Jussieu, 75005 Paris, France
Full list of author information is available at the end of the article

survey proposed by Yu and Kak, which describes the different real-life outcomes that can be predicted using social media (e.g., election results, box-office revenues, marketing impact) [8]. In our work we focus on a different prediction objective related to social media: predicting the amount of attention that web content will generate on the Internet.

The remainder of the paper is organized as follows. We narrow down the scope of this survey in Section 2 and briefly review the evolution of this research area in Section 3. We continue with a presentation of the most popular types of web content analyzed so far (Section 4) and describe the measures used to evaluate the prediction performance (Section 5). In order to structure the prediction methods, we propose a classification in Section 6 and describe the prediction methods based on this classification in Section 7. We present the factors known to influence content popularity (Section 8) and review the predictive features that have already been used in a prediction model in Section 9. Finally, we conclude with a presentation of some representative domains that could benefit from web content popularity prediction (Section 10) and look at potential future directions in Section 11.

2 Scope of the survey

Let us now define the scope of this survey. The term *web content* is effectively generic as it broadly defines any type of information on a web site. It can refer both to the subject of the information and the individual item used to deliver the information. In this survey we define web content as any individual item (in the form of text, image, audio, or video), publicly available on a web site, which contains a measure that reflects a certain level of interest showed by an online community.

On the Internet, the popularity of web content can have different connotations. If by content we refer to the subject of the content, such as a person or an organization, then popularity could be expressed by a greater web presence or activity. From a different perspective, one may see web content as an individual web link and define popularity as the popularity of the link (the quantity and quality of inbound links). For the scope of this survey, we consider popularity from the standpoint of the relationship between an individual item and the online users who consume it.

Seen from this perspective, there are different metrics used to quantitatively evaluate web content popularity. The classical way of doing this is to measure the number of views. However, this information is often hidden from the online users and crawling engines. For example, social networking sites, for various reasons, usually do not disclose this information to the online users [9]. But nowadays, with the growing prevalence of Web 2.0 platforms, there are new indicators – publicly available – that reflect users' interest. In response to the publication of a web content, users can now provide a direct feedback, through comments and ratings, or further share it in their online social circles (using, for example, Facebook, Twitter, or Digg). These metrics capture different levels of user engagement and provide valuable information, complementary to view counts: rating improves the quality of publications, comments increase the time spent on a web page, and sharing gives content a greater notoriety. In general, it has been observed that there is a moderate correlation between the different popularity metrics [9-13], as they probably capture different types of habits on the Internet (to observe, comment, rate, or share). In this context, studying these metrics individually or how they relate to each other [14,15] provides a wider and better perspective of what the popularity of a web content actually means.

3 A brief history of the evolution of popularity prediction methods

The beginning of this research area can be found in the early studies on users' web access patterns [16-19]. An important observation of these initial studies was that the distribution of users' requests for web pages is highly skewed and could be described by a Zipf's law [18]. Online videos, accounting for a significant amount of Internet traffic, have been one of the main attraction of these early measurements [20-27]. During this initial phase, researchers have looked at the degree of skewness in the popularity of videos [20,21,24,25] (to determine potential benefit of caching videos) and analyzed which probability distribution best describes the video access patterns (to understand the mechanism that explains users' consumption patterns [28]). These studies revealed that the interest generated by a web content is transient, heterogeneous, and often unpredictable [24,28].

After the properties of Web access patterns have been sufficiently well understood the challenge became to actually predict content popularity [29,30]. The first prediction methods were built on the observation that there is a strong positive correlation between the popularity of a web content at different stages during its lifetime [28,30]. As a result the first prediction methods consisted in linear regression functions that use the amount of attention that a web content generates early after publication to predict its popularity afterwards. The prevalence of Web 2.0 platforms, rich in metadata about how users interact with the web content and with peer online readers have further contributed to a fast evolution of this research area. Prediction methods based on the online social connections created between the users have been proposed, content

published on various web sites has been analyzed, and different measures about web content popularity have been considered.

These initial prediction methods were simple but often inaccurate for web content that remains attractive for longer periods of time [30]. An important step forward has been made with the finding that the evolution of web content popularity over time can be described by a only small number of temporal patterns [31,32]. Thus, more accurate prediction methods that include information about the evolution patterns of content popularity, have been proposed [32-34]. But the content published on a web site is part of a global information ecosystem as it can spread on several web sites and reach consumers through different communication mediums. So, a further breakthrough in the design of more accurate prediction methods has been made with the development of algorithms that can extract and cross-correlate information from different web domains [14,35,36].

4 Types of web content

Users attention is spread across multiple web sites and various types of web content. Some of the most popular types of web content studied so far include: user-generated videos that account for a great percent of Internet traffic [37]; news articles, massively diffused through social networking sites [38] and heavily consumed on mobile devices [39]; stories published on social news aggregators that provide an even greater exposure to the most popular content on the Internet; and items (comments, photos, or videos) published on social networking sites, the most popular platforms to share information and encourage users' participation on a global scale.

Examples of the variety of web content, gathered from different web sites and used in the context of popularity prediction, are illustrated in Figure 1, together with information about the number of items and the time period covered by each data set.

Online videos. YouTube, the world's largest video sharing platform with 100 hours of upload per minute [3] and more than 1 trillion worldwide views per year [40], has been the main focus of the existing studies. The site's content, with more than 200 million unique videos, covers a broad range of topics and is sustained by a big and active online community [41]. Studying the popularity of YouTube content is challenging given the ever-growing number of videos, the many features that the platform provides (e.g., video recommendations, internal search, online social networking), and the limitations associated with the retrieval of a representative sample of videos [42].

The popularity of YouTube videos, commonly expressed by the number of views in research studies, follows a heavy-tailed distribution that, depending on the data set and the method used to fit the distribution, can be described by power-law with exponential cut-off [28], Weibull [41], log-normal [42], or Gamma distributions [43]. But the popularity of videos over time is highly non-stationary. From a high-level point of view, the popularity growth of videos over time can be represented by power-law or exponential distributions [44]. A more fine-grained analysis exhibits even more complex and diverse patterns. For instance, Crane and Sornette found that, while the activity around most YouTube videos can be described by a Poisson process, many videos reveal similar activity around the peak period that can be accurately described by three popularity evolution patterns [31]. Similar temporal evolution patterns have been observed by Figueiredo [45] and even more diverse shapes have been discovered by Gorsun et al. [32].

In addition to YouTube, the popularity of videos published on other platforms has been studied (e.g., Daum [28], Dutch TV [44], DailyMotion [46,47], Yahoo! video [47], Veoh [47], Metacafe [47], Vimeo [34]), but on a smaller scale, and no significant differences have been signaled in terms of popularity distributions.

Online news. The primary source of information in the digital world, news, are created in large numbers and massively diffused through online social networks [38]. Compared to videos that catch users' attention for a longer period of time, the interest in news articles fades quickly, within days after publication [14,48]. The popularity of online news, frequently expressed by the number of comments (the number of views are rarely disclosed by news platforms), is also highly skewed, and can be described by power-law [49,50] or log-normal [51] distributions.

Social bookmarking sites. The third major type of web content analyzed so far is represented by stories posted on social bookmarking sites such as Digg [30,52,53], Slashdot [29], or Reddit [53,54]. Content published on these sites experiences an even greater rate of change with stories reaching their attention peak in the first six hours after publication and being completely saturated within one day [30]. Prediction becomes even more difficult in this setting given the complex interactions between users [55,56] and the promotion algorithm based on the collective opinion of users [57,58]. The popularity of the content published on these sites is described by a heavy-tailed nature that is best represented by Weibull [53] or log-normal distributions [30,52,59].

Social networking services. Designed with the idea of facilitating interactions among people on the Internet,

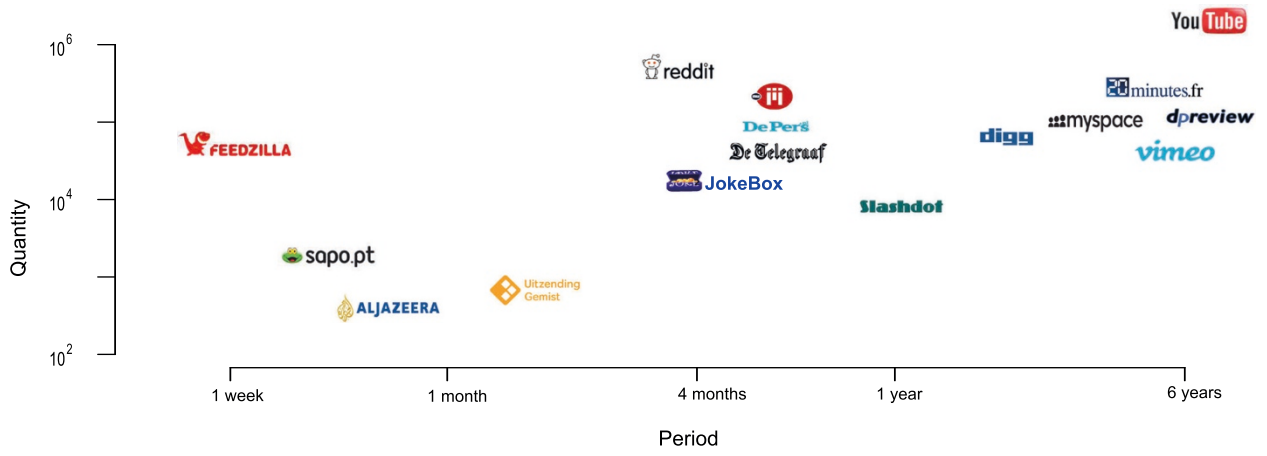

Figure 1 Data sets used as case studies to evaluate the performance of prediction methods. On a log-log scale we depict the total number of items and the cumulative time period covered by each data set.

these sites allow users to build and maintain online social relationships with people that share common interest, background, or real-life relationships. While there are different types of social networking services the most popular are the ones built on the idea of content sharing. Microblogs, such as Twitter and Weibo, are a specific type of social networking services that have been extensively studied. These platforms are probably the most dynamic representation of social media. Users create and share information in the form of short messages, known as tweets, containing up to 140 characters. When a user posts a (re)tweet it becomes visible to all its followers (i.e., members of the social group). Content can easily spread through the social connectivity graphs as followers can further share the content to their own list of followers. Two metrics have been used to measure the popularity of a tweet: the number of users that receive a message in their tweet feed [60], or most commonly, the number of retweets. The popularity of tweets is also highly skewed and can be described by a power-law distribution [61-63].

Tweets are probably one the most ephemeral type of web content as they become popular very fast and they quickly die out. For example, studies conducted on Tencent Weibo found out that an insignificant number of tweets get retweeted after one day [63]. Similarly, a study on Twitter revealed that most tweets receive half of their retweets within the first hour after publication [64]. Useful predictions thus need to be done in the order of minutes after the post of a tweet.

In addition to these main categories, content published on other web sites have been used for popularity prediction tasks such as threads published on discussion forums (DPReview, MySpace [65]) and movie ratings on IMDb [36]. Due to the relevancy of the results we also include in our analysis the prediction results for the content published on two applications: an interactive video sharing application (Zync) [66] and a joke sharing application (JokeBox) [67].

5 Evaluating the prediction models

To provide a more explicit description of the prediction algorithms, let us introduce the terminology and the measures used to evaluate the efficiency of the prediction methods.

Terminology. Let $c \in C$ be an individual item from a set C observed during a period T. We use $t \in T$ to describe the age of an item (i.e., duration since the time it was published) and mark two important moments: indication time t_i, representing the time we perform the prediction and reference time t_r, the moment of time when we want to predict content popularity. Let $N_c(t_i)$ be the popularity of c from the time it was published until t_i and let $N_c(t_r)$ be the value that we want to predict, i.e., the popularity at a later time t_r. We define $\widehat{N}_c(t_i, t_r)$ the prediction outcome: the predicted popularity of c at t_r using the information available until t_i. Thus, the better the prediction, the closer $\widehat{N}_c(t_i, t_r)$ is to $N_c(t_r)$.

Evaluation. We distinguish two prediction goals: (i) Numerical prediction – predict the exact value of the popularity, (ii) Classification – predict the popularity range that an item is most likely to fall in.

5.1 Numerical prediction

There are different ways to assess the efficiency of a numerical prediction [68]. Mean Squared Error (MSE – Equation 1) is used to report the average of the squared

errors. By taking the square root of MSE, one can express the error in the same dimension as the estimated value (RMSE – Equation 2). One important limitation of squared errors is that they put too much weight on the effect of outliers, and in this case reporting the absolute errors is a good alternative (MAE – Equation 3).

Absolute errors can be meaningfully interpreted if one knows the range of the actual popularity values. Otherwise, a good way of expressing the prediction performance is through relative errors such as the Mean Relative Error (MRE – Equation 4) and Mean Relative Squared Error (MRSE – Equation 5). Relative measures are also useful to compare the efficiency of prediction algorithm across studies, as in most cases the popularity values have widely different ranges (e.g., the number of views on YouTube is several orders of magnitude greater than the number of comments on a news web site). Special attention should be paid when using these error measures for zero-inflated variables as the relative error is undefined when the actual value is zero.

Another way of expressing the prediction error is through the Relative Squared Error (RSE – Equation 6), Root Relative Squared Error (RRSE – Equation 7), and Relative Absolute Error (RAE – Equation 8). The error in this case is expressed relative to the performance of a simple predictor, the average of the actual values (computed on the training data set).

The quality of a numerical prediction can also be reported using the correlation coefficient or the coefficient of determination (R^2). Compared to the previous measures, which show how the estimated values diverge from the actual ones, these evaluation criteria can only express the degree of linear association between the two variables (predicted and actual values).

$$\text{MSE} = \frac{1}{|C|}\sum_{c\in C}\left(\widehat{N}_c(t_i,t_r) - N_c(t_r)\right)^2. \quad (1)$$

$$\text{RMSE} = \sqrt{\frac{1}{|C|}\sum_{c\in C}\left(\widehat{N}_c(t_i,t_r) - N_c(t_r)\right)^2}. \quad (2)$$

$$\text{MAE} = \frac{1}{|C|}\sum_{c\in C}\left|\widehat{N}_c(t_i,t_r) - N_c(t_r)\right|. \quad (3)$$

$$\text{MRE} = \frac{1}{|C|}\sum_{c\in C}\left|\frac{\widehat{N}_c(t_i,t_r) - N_c(t_r)}{N_c(t_r)}\right|. \quad (4)$$

$$\text{MRSE} = \frac{1}{|C|}\sum_{c\in C}\left(\frac{\widehat{N}_c(t_i,t_r) - N_c(t_r)}{N_c(t_r)}\right)^2. \quad (5)$$

$$\text{RSE} = \frac{\sum_{c\in C}\left(\widehat{N}_c(t_i,t_r) - N_c(t_r)\right)^2}{\sum_{c\in C}\left(\widehat{N}_c(t_r - \overline{N}(t_r)\right)^2}. \quad (6)$$

$$\text{RRSE} = \sqrt{\frac{\sum_{c\in C}\left(\widehat{N}_c(t_i,t_r) - N_c(t_r)\right)^2}{\sum_{c\in C}\left(\widehat{N}_c(t_r - \overline{N}(t_r)\right)^2}}. \quad (7)$$

$$\text{RAE} = \frac{\sum_{c\in C}\left|\widehat{N}_c(t_i,t_r) - N_c(t_r)\right|}{\sum_{c\in C}\left|(\widehat{N}_c(t_r - \overline{N}(t_r)\right|}. \quad (8)$$

5.2 Classification

This prediction problem can also be addressed as a classification task, where, assuming that the popularity range is known, one can split this interval in k non-overlapping popularity ranges. Thus, given the k possible outcomes the prediction goal is to correctly predict the popularity class of a web content.

Various metrics are available to evaluate the quality of a classification method [68,69]. *Accuracy*, one of the most reported metric, is used to express the proportion of correctly classified instances. This measure is nevertheless inappropriate when dealing with highly imbalanced classes, which can often be the case when referring to web content popularity, characterized by a heavy-tail nature. For example, a possible experiment could be to learn a classifier that predicts which videos will get more that 10^6 views on YouTube - a "small" class (1%) according to a recent study [70]. A simple rule, that decides that all videos receive less than 10^6 views, will correctly predict 99% of the cases. Thus, a good level of accuracy is obtained without even learning any prediction rule on how to detect the popular items.

To measure the performance of the classifier on a "small" class, a good alternative is to use *precision*, *recall*, or *F-score* (the harmonic mean between *precision* and *recall*). But *F-score* measures the performance of a classifier for only one class. To report the aggregate performance over multiple classes, a good solution is to use the *macro-average* measure (average *F-score* over all k classes).

6 A classification of web content popularity prediction methods

To structure the prediction methods, we propose a classification that groups the methods according to the type

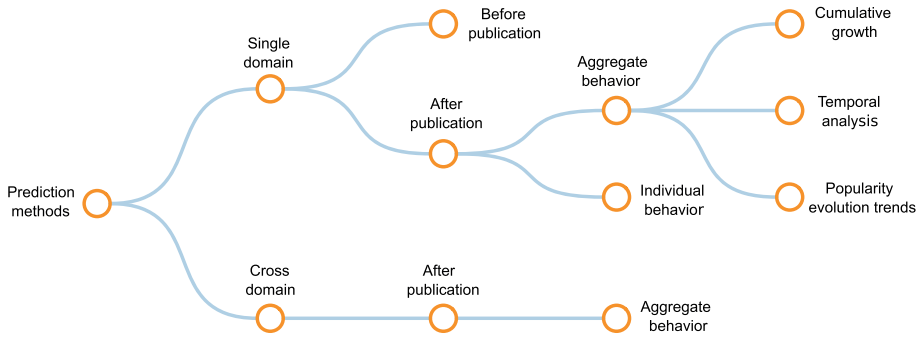

Figure 2 A classification of content popularity prediction methods.

and the granularity of information used in the prediction process (Figure 2).

6.1 Single domain

We define a domain as the web site where an individual item resides, regardless if it has been created or shared from an external source (e.g., news article shared on social bookmarking sites). Methods under this category are used to predict content popularity using only the information available on the web site.

6.1.1 Before publication

One of the most challenging objectives is to predict the popularity before the publication of a web content, relying only on content metadata or the online social connections of the publisher.

6.1.2 After publication

The alternative is to include in the prediction model data about the attention that one item receives after its publication.

Aggregate behavior. A common approach is to deduce future content popularity using the aggregate users' attention after the publication of a web content. This solution can further be separated in three main categories:

- Study the *cumulative growth* of attention, i.e., the amount of attention that one item receives from the moment it was published until the prediction moment.
- Perform a *temporal analysis* of how content popularity evolves over time until the prediction moment.
- Use clustering methods to find web items with similar *popularity evolution trends*.

Individual behavior. Instead of treating each user's action equally, one may further refine the prediction model by taking into account individual user behavior.

6.2 Cross domain

Explaining popularity from the perspective of single domain is limited due to the complex users' interactions across different platforms. Methods under this class draw conclusions by extracting and transferring information across web sites.

7 A survey on popularity prediction methods

Several popularity prediction methods have been proposed in the last decade, from simple linear regression functions to complex frameworks that cross-correlate information from different web sites. We describe these methods following the classification proposed in Section 6 and present their performance on predicting the popularity of different types of web content. A summary of these methods is also presented in Table 1.

7.1 Single domain

In the vast majority of cases, prediction methods rely entirely on the information available on the web site where content has been published.

7.1.1 Before publication

Predicting the popularity of an item before its publication is particularly useful for web content characterized by a short lifespan. News articles, which are time-sensitive by nature, fall under this category and have been analyzed in two studies [71,72].

Tsakias et al. address this prediction task as a two-steps classification problem: predict if news articles will receive comments and if they do, if the number of comments will be high or low [71]. The proposed prediction method is a random forest classifier trained on a large number of features (textual, semantic, and real-world). Using several Dutch online news sources the authors show that one can accurately predict which articles will receive comments and observe that the performance degrades significantly

Table 1 Summary of the popularity prediction methods presented in the survey

Class	Methods	Data sets	Benchmark model	Performance/Remarks
Before publication	SVM, Naive Bayes, Bagging, Decision Trees, Regression [72]	Feedzilla		Shows an accuracy of 84% in predicting the popularity range of a news article.
Before publication	Random Forests [71]	AD, De Pers, FD, NUjiji, Spits, Telegraaf, Trpuw, WMR		Good performance in identifying which articles will receive at least one comment.
Cumulative growth	Constant growth [29]	Slashdot		Good performance in predicting the number of comments one day after the publication of an article (MSE = 36%).
Cumulative growth	Constant scaling [30]	Digg, YouTube	Constant growth, Log-linear	Outperforms the constant growth and the log-linear models in terms of MRSE.
Cumulative growth	Log-linear [30]	Digg, YouTube	Constant growth, Constant scaling	Outperforms the constant growth and the constant scaling models in terms of MSE.
Cumulative growth	Survival analysis [65]	DPreview, MySpace		Using the information received in the first day after the publication it can detect with 80% accuracy which threads will receive more than 100 comments.
Cumulative growth	Logistic regression [61]	Twitter		The model can successfully identify which messages will not be retweeted (99% accuracy) and those that will be retweeted more than 10,000 times (98% accuracy).
Temporal analysis	Multivariate linear regression [33]	YouTube	Constant scaling	An average improvement of 15% in terms of MRSE compared to the constant scaling model.
Temporal analysis	Reservoir computing [77]	YouTube	Constant scaling	Minor improvement compared to the constant scaling model.
Temporal analysis	Time series prediction [32]	YouTube		Designed for frequently-accessed videos. Good performance in predicting the daily number of views.
Temporal analysis	kSAIT [63]	Twitter	Regression-based methods	Predict the number of tweets using information from the first hour after content publication. An improvement of up to 10% compared to regression-based methods.
Popularity evolution patterns	Hierarchical clustering [32]	YouTube		Designed for rarely-accessed videos. The model shows good performance for short-term predictions but significantly larger ones for long-term predictions.
Popularity evolution patterns	MRBF [33]	YouTube	Constant scaling, Multivariate linear regression	An average improvement of 5% in terms of MRSE compared to multivariate linear regression and 21% compared to constant scaling model.
Popularity evolution patterns	Temporal-evolution prediction [34]	YouTube, Vimeo, Digg	Log-linear	Significant improvement compared to the log-linear method. The model can be used to predict the temporal evolution of popularity.
Individual behavior	Social dynamics [81]	Digg	Log-linear	It incorporates information about the design of the web site. Shows an accuracy of 95% in identifying which articles will get on Digg's front page.
Individual behavior	Conformer Maverick [67]	JokeBox	Collaborative filtering solutions	Adequate for platforms that rank content based on user votes. Better performances than collaborative filtering solutions.
Individual behavior	Bayessian networks [64]	Twitter		MRE of 40% when predicting the total number of tweets using the information received in the first five minutes after publication.
Cross-domain	Linear regression [36]	IMDb, Twitter, YouTube		Designed to predict movie ratings using social media signals. The best performance was achieved when using textual features from Twitter and the fraction of likes over dislikes from YouTube.

Table 1 Summary of the popularity prediction methods presented in the survey *(Continued)*

Cross-domain	Linear regression [14]	Al Jazeera		Results show that a model based on social media reactions in the first ten minutes has the same performance as one based on the number of views received in the first three hours.
Cross-domain	Social transfer [35]	YouTube, Twitter	SVM basic	Shows a 70% accuracy in identifying which videos will receive sudden bursts of popularity (60% improvement over a model that uses only the information available on YouTube).

when trying to predict if the volume of comments will be high or low.

Bandari et al., using the number of tweets as an indicator of news popularity, formulate the prediction task both as a numerical and a classification problem [72]. The authors show that predicting the exact popularity of news articles is prone to large errors ($R^2 = 0.34$), but that predicting ranges of popularity is more effective, with an accuracy of 84% when identifying articles that would receive a small, medium, or large number of tweets.

7.1.2 After publication

Aggregate behavior The methods under this category have been used to predict web content popularity based on the aggregate users' attention received early after content publication.

Cumulative growth. One of the first solutions, used to predict the popularity of Slashdot stories, is proposed by Kaltenbrunner et al. [29]. The model, which we will refer to as `growth profile` (we adopt the terminology used in [30]), assumes that, depending on the time of the publication, news stories follow a constant growth that can be described by the following function:

$$\widehat{N}_c(t_i, t_r) = \frac{N_c(t_i)}{P(t_i, t_r)}, \tag{9}$$

where $P(t_i, t_r)$ is a rescaling parameter and represents the average growth of a story from t_i to t_r

$$P(t_i, t_r) = \frac{1}{|C|} \sum_{c \in C} \frac{N_c(t_i)}{N_c(t_r)}. \tag{10}$$

The effectiveness of this method was tested on a large corpus of Slashdot stories and shows a reasonable performance in predicting the popularity of stories using the aggregate users' reactions in the first day after news publication (average MRE of 36%).

Describing future popularity as a linear relationship of the popularity at earlier stages is also proposed by Szabo and Huberman under the `constant scaling` model [30]:

$$\widehat{N}_c(t_i, t_r) = \alpha_2(t_i, t_r) N_c(t_i). \tag{11}$$

Parameter α is computed in such a way that the model minimizes MRSE and is described by the following expression:

$$\alpha(t_i, t_r) = \frac{\sum_{c \in C} \frac{N_c(t_i)}{N_c(t_r)}}{\sum_{c \in C} \left[\frac{N_c(t_i)}{N_c(t_r)}\right]^2}. \tag{12}$$

Szabo and Huberman also observe a positive correlation between the popularity of an item early after its publication and its popularity at a later stage and propose a logarithmically transformed linear regression model (`log-linear`) expressed as

$$\widehat{N}_c(t_i, t_r) = \exp\left(\ln N_c(t_i) + \beta_0(t_i, t_r) + \frac{\sigma_0^2(t_i, t_r)}{2}\right). \tag{13}$$

For the coefficients of Equation 13, β_0 is computed on the training set using maximum likelihood parameter estimation on the regression function $\ln N_c(t_r) = \beta_0(t_i, t_r) + \ln N_c(t_i)$ and σ_0^2 is the estimate of the variance of the residuals on a logarithmic scale.

This method shows good predictive performance on several data sets: Digg stories [30], YouTube videos [30], articles published on a French news platform [73], and Dutch online news articles [51]. For example, Tsagkias et al. observe that, by using the number of comments received in the first ten hours after the publication of news articles, one can attain good performances in predicting the final number of comments (average MRSE of 20%) [51].

A different approach is proposed by Lee et al. [65]. Instead of predicting the exact amount of attention the authors study the possibility of predicting if a web content will continue to receive attention from online readers after a certain period of time. The prediction model proposed for this problem (Cox proportional-hazards regression) is a widely used method in `survival analysis` that allows one to model the time until an event occurs (a typical event is "death", from which the term survival analysis is derived). While the main utilization of this method could be to predict the lifetime of a web content, by changing the definition of an event, the method

can also be used for popularity prediction tasks. The solution proposed by Lee et al. is to consider as event the time when a web content will reach a popularity value above a certain threshold. The performance of this method was tested on threads from two online discussion forums, DPreview and MySpace, with popularity expressed as the number of comments per thread. Using different statistics related to the users' comment arrival rate the authors show that, by observing user activity in the first day after publication, the method can detect with 80% accuracy the threads that will receive more than 100 comments.

Regression-based methods have been frequently used for this prediction task. Tatar et al. use a simple linear regression based on the early number of comments to predict the final number of comments for news articles [74]. The authors observe that there is no significant improvement when using specialized prediction models as a function of the category and the publication hour of an article. Marujo et al. study the problem of predicting the number of clicks that news stories will receive during one hour. Various prediction methods have been tested (multiple linear regression, regression-based trees, bagging, and additive regression) using different features extracted from the news web platform. The authors show that by combining different regression algorithms one can obtain fairly good results (MRE = 12%) in predicting the number of clicks received by news articles during one hour. Cho et al. use a linear model on a logarithmic scale to predict popularity ranges for political blog posts [75]. The authors show that, by looking at the number of page views in the first 30 minutes, one can classify articles in three classes of popularity with 86% accuracy. A different approach is proposed by Tatar et al. who study the performance of three popularity prediction methods (simple linear regression, linear-log, and constant scaling) to order news articles based on their future number of comments [50]. Using a data set of news articles and comments, the authors show that, out of the three methods, a simple linear regression is the most adequate for this prediction task, suggesting that a smaller least squares error does not imply a smaller ranking error.

Predicting the popularity of web content, based on the aggregate user behavior, has also been addressed as a classification problem. Jamali and Rangwala use the number of comments that Digg stories receive in the first ten hours to predict the final Digg score [56]. By training different classification methods the results indicate that it is possible to predict the popularity class of a Digg story with an accuracy of 80%, 64%, and 45% when separating stories in 2, 6, and 14 ranges of popularity. Hong et al. study the problem of predicting the number of retweets for Twitter posts [61]. The authors address this problem as a multiclass classification task, where, for a given tweet the goal is to predict the range of popularity and not the exact retweet count. Using a logistic regression classification function and various content, topological, and temporal features the authors show that they can successfully predict which messages will not be retweeted (99% accuracy) and those that will be retweeted more than 10,000 times (98% accuracy).

Temporal analysis. For web content that captures users' attention for longer periods of time (e.g., certain videos that are viewed during several months or even years) it has been observed that the aggregate-based prediction models are prone to large errors [30]. To improve the prediction effectiveness, one solution is to design models that can weight users' attention differently based on the recency of the information relative to the prediction moment. For this type of evaluation, the aggregate user behavior is sampled in equal-size intervals of duration δ where $x_c(i)$ is the popularity of an item c during the ith interval, and $X_c(t_i)$ is the vector of popularities for all intervals up to t_i: $X_c(t_i) = [x_c(1), x_c(2), x_c(3) \ldots, x_c(i)]^T$ ($N_c(t_i) = \sum_{j=1}^{i} x_c(t_j)$).

Pinto et al. rely on this approach to predict the popularity of YouTube videos [33]. Using a sampling rate of one day the authors use a `multivariate linear regression` expressed as

$$\widehat{N}_c(t_i, t_r) = \Theta(t_i, t_r) X_c(t_i). \tag{14}$$

The parameters of the model, $\Theta(t_i, t_r) = [\theta_1, \theta_2, .., \theta_i]$ are computed to minimize MRSE under the new definition of estimated popularity. Using a collection of YouTube videos this model shows a significant improvement compared to the `constant scaling` model. For instance, predicting the popularity of a video one-month after its publication using data from the first week shows an average improvement of 14% over the `constant scaling` model. The main drawback of this algorithm, as mentioned by the authors, is that in order for the prediction methods to be effective, additional exploration is needed to decide on the optimal history length and the sampling rate.

`Reservoir computing` [76], a novel paradigm in recurrent neural networks, is proposed as a model that could consider more complex interactions between early and late popularity values (between $X_c(t_i)$ and $N_c(t_r)$). More specifically, this technique is used to build a large recurrent neural network that allows one to create and evaluate nonlinear relationships between $X_c(t_i)$ and $N_c(t_r)$ [77]. On a small sample of YouTube videos this model shows a minor improvement over the `constant scaling` model in predicting the daily number of views based on the observations received in the previous ten days.

For videos that are popular over long periods of time (those that receive views during at least half a year), Gursun et al. [32] observe that the daily number of views can be modeled through a `time series prediction` model using Autoregressive Moving Average (ARMA). Thus, the popularity of a video at a given day n, $x_c(n)$, can be predicted using the following formula:

$$x_c(n) = \sum_{i=1}^{p} \alpha_i x_c(n-i) + \epsilon_n + \sum_{j=1}^{q} \theta_j \epsilon_{n-j}, \tag{15}$$

where $\alpha_1, \ldots, \alpha_p$ are the parameters of the autoregressive model, $\theta_1, \ldots, \theta_q$ are the parameters of the moving average, and $\epsilon_n, \epsilon_{n-1}, \ldots$ are the white noise error terms.

The model shows good performance in predicting the number of daily views based on the viewership received in the previous week ($p = q = 7$), with an average MRE error of 15%. The main limitation of this method is that it has a very high computational cost as it requires one ARMA model for each video. To improve the scalability of the model the authors use principal component analysis (PCA) as follows: 1) use PCA to find the main principal components that can approximate the time series for the entire collection of videos and 2) apply ARMA modeling to the principal components instead of the individual time series. This solution significantly improves the scalability of the model (e.g., it requires 20 ARMA models to make predictions for the entire collection of videos) and shows a minor decrease in the prediction accuracy (MRE = 0.12 when using individual ARMA models compared to MRE = 0.14 when using principal component analysis).

Kong et al. propose `kSAIT` (top-k Similar Author-Identical historic Tweets), an algorithm that can predict the popularity of tweets one, two, or three days after publication based on the retweet information received in the first hour [63]. The underlying assumption of this algorithm is that, tweets are retweeted in a similar manner depending on the author of the tweet. The prediction algorithm is thus user-specific (there is one prediction function for each user) and uses as predictive features only users' retweeting behavior as it does not include any information about content itself or about users' centrality in the graph of social interactions. Each tweet is described by a set of features (e.g., retweet acceleration, retweet depth) derived from the time-series of the retweets published in the first hour after publication by the direct and n-level followers, the publication time of the tweet, and information about the users who retweeted the original tweet. When a new tweet is posted, the algorithm computes the similarity of the tweet and all other tweets published by the same user, selects the top-k most similar tweets, and estimates the popularity of the target tweet as an average of the popularity of the top-k most similar tweets.

The performance of the algorithm was evaluated on a data set from Tencent Weibo and compared to several regression-based methods. The algorithm shows good prediction performance (an improvement of up to 10% in terms of MAE compared to regression-based methods), but training a personalized function for each user makes it difficult to be implemented in large-scale social networks.

Popularity evolution trends. Several studies reveal that the evolution of content popularity over time can accurately be described by a small number of temporal patterns [31,32,45,78]. Crane and Sornette provide one of the first evidence of this fact while analyzing the popularity evolution of YouTube videos [31]. The authors observe that a Poisson process describes the attention around the majority of videos (90% of the videos) and the remaining ones follow three popularity evolution trends (illustrated in Figure 3). These trends are characterized by a single popularity peak but different patterns in which the popularity grows and declines. For more accurate predictions it is important to know that content exhibits well-known temporal dynamics, as the prediction function can be adapted to the specific shape of popularity evolution.

One of the first models that exploits the temporal evolution patterns is proposed by Gursun et al. [32]. While analyzing the viewership around YouTube videos the authors observe two overall categories of videos: those that are consistently popular over time and those that are viewed during a small period of time. The second category is characterized by short-time popularity bursts and can be described by a small number of temporal patterns. To reveal these patterns the authors use `hierarchical clustering` based on the time-series of videos popularity during 64 days centered on the peak. This strategy reveals that, for videos that are viewed during short periods of time, there are ten common shapes that describe the temporal evolution for most of the videos. Once these shapes are detected the prediction task consists in mapping videos to the clusters that best describe their evolution until the prediction moment (t_i) and in using the temporal evolution trends of the clusters to deduce future video popularity. On a sample of YouTube videos, this method shows good performance in making short-term predictions (predict the number of views in the next day) but significantly larger ones in making long-term predictions.

Pinto et al. put forward an improvement to the `multivariate linear regression` model by proposing a solution that captures the similarity between videos in terms of their temporal evolution patterns [33]. The model assumes that the temporal popularity evolution of a subset of videos is representative for the

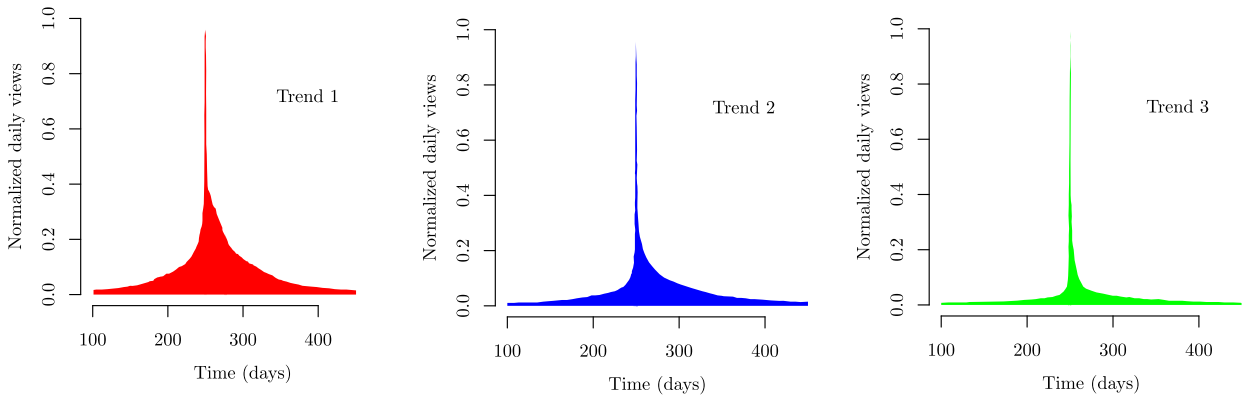

Figure 3 Example of three popularity evolution trends discovered by Crane and Sornette [31] (and similar with some of the trends presented by Figueiredo [45] and Gursun et al. [32]). The figure shows the average number of views over time normalized by the number of views during the peak day.

entire population and could be used to improve the prediction accuracy. More specifically, the prediction model, called `multivariate radial basis function` (MRBF), is described by the following relationship:

$$\widehat{N}_c(t_i, t_r) = \Theta(t_i, t_r) X_c(t_i) + \sum_{c_1 \in C_1} w_{c_1} \mathrm{RBF}_{c_1}(c), \qquad (16)$$

where $C_1 \in C$ is the representative subset of videos and w_{c_1} is the weight associated with each item. RBF stands for the Radial Basis Function with Gaussian kernel (chapter 6 [79]) and measures the similarity between the target video and each video in C_1. Training MRBF model involves finding the optimal parameters Θ and w_{c_1} to minimize MRSE, setting the optimal values of RBF kernel, and finding a representative set of videos. The performance of this model shows an average improvement of 5% over `multivariate linear regression` and 20% compared to the `constant scaling` model.

Ahmed et al. propose a model that uses a more granular description of the temporal evolution of content popularity [34]. Instead of using a set of representative items to describe the entire evolution of content popularity, this model selects representative members during regular intervals of duration δ and defines rules to model the transitions among subsequent intervals.

The representative members for each interval are computed using Affinity Propagation clustering algorithm [80]. To calculate the similarity between items, the authors derive two features from $X_c(t_i)$: one that compares if two items receive the same proportion of users' attention and another one that measures if the two items experience a similar popularity growth. Once the clusters of popularity are identified, they are grouped into a probabilistic framework used to describe the evolution of content popularity between clusters over time. Thus, by knowing to which cluster an individual item is most likely to belong at time t_i, one can predict its popularity at a future moment of time t_r.

The performance of the model was tested on three data sets (YouTube, Vimeo, and Digg) and shows a significant improvement over the `log-linear` model. For example, when using the observations received in the first 24 hours to predict the popularity four hours ahead, this model shows a MRSE error of 1% for Digg and 3.5% for Vimeo and YouTube; a significant improvement compared to the `log-linear` model that shows a performance of 17% for Digg, 24.2% for Vimeo, and 29.7% for YouTube.

Individual behavior Instead of treating each user's reaction equally in the prediction process, models under this category draw conclusions based on individual user behavior.

`Social dynamics`, the model proposed by Lerman and Hogg, describes the temporal evolution of web content popularity as a stochastic process of user behavior during a browsing session on a social media site [81]. In its original form, the model is designed according to the characteristics of the social bookmarking site Digg: stories can be found in three sections of the site (front, upcoming, and friend list pages), users can express their opinions through votes, and stories are arranged in pages or promoted to different sections of the site based on the dynamics of votes.

User behavior is modeled through a set of states that describe the possible actions that one can take on a site: browse through the different sections, read news stories, and cast votes to further recommend them to the Digg community. Browsing sessions are dynamic as

stories circulate through the site (i.e., they may appear on different sections of the site or change position on the page) depending on the voting results. Individual user behavior is thus linked to the collective behavior, which in the end explains how stories receive votes over time. More specifically, the number of votes a story receives depends on its visibility and general interest. Visibility is expressed as the probability of finding a story in different sections of the site and interest is linked to the quality of the story estimated by the voting dynamics.

The authors validate the model on a small sample of Digg stories by studying user' reactions to the publication of stories and by taking into account the online connections created between Digg users. By using this algorithm, the authors reveal that they can predict in 95% of the cases which stories will become popular enough to reach Digg's front page.

For platforms that allow users to cast positive and negative votes on the content, Yin et al. propose `Conformer Maverick`, a model used to predict content popularity based on users' voting profiles [67]. The underlying assumption of the model is that, in the voting process, users can have two behaviors: obey the general users' opinion (the "conformers") or be against them (the "mavericks"). The profile of a user is in-between these two extremes but in general one trait prevails.

The first step is to build user profiles based on the voting history by comparing individual votes with the overall appreciation of the content, i.e., if the majority of votes is positive or negative. These profiles are later used to decide if an item will become popular by analyzing early user votes. Receiving positive votes from conformers and negative ones from mavericks is then considered as a good indication that an item will be appreciated by the majority. Using data from a joke sharing application the algorithm shows a better performance than a collaborative filtering solution.

Zaman et al. propose a probabilistic model based on Bayesian inference to predict the popularity of Twitter messages [64]. The predictive features are content-agnostic and based on retweets time-series and the social connectivity graph of the Twitter users. The model is based on the assumption that Twitter users have similar actions with regard to the post of a tweet that creates a pattern in the evolution of tweets popularity. In particular, the probability of a (re)tweet to be retweeted depends on the number of followers and the distance from the user that originally generated the tweet. Using a small data set of 52 tweets, the method shows a good performance (given the difficulty of the task), with an average MRE error of 40% using the retweeting information received in the first five minutes after the publication.

7.2 Cross domain

The second major category of methods is used to predict web content popularity using information from multiple web domains: extract data from one domain (e.g., social media) and transform it into knowledge to predict web content popularity in another domain (e.g., the site where content was published). Currently, only methods that predict content popularity *after publication* based on the *aggregate behavior* have been proposed.

Oghina et al. use data from Twitter and YouTube to predict movie ratings on IMDb [36]. By training a linear regression model on several textual features extracted from Twitter and various statistics from YouTube (likes, dislikes, and comments) the authors show that they can accurately predict movie ratings on IMDb. The authors indicate that the best performance is obtained by combining the ratio of likes over dislikes from YouTube activity with the subjective terms (positive and negative unigrams about the movies) extracted from Twitter.

The algorithm proposed by Roy et al., `Social Transfer`, extracts information from Twitter to detect videos that will experience sudden bursts of popularity on YouTube [35]. The model consists of the following steps: extract popular topics from Twitter, associate these topics to YouTube videos, and compare the popularity of videos on Twitter with their popularity on YouTube. A disproportionate share of attention on Twitter compared to YouTube is then used as strong evidence that a video will experience a sudden burst of popularity.

Topics are learned by analyzing Twitter stream, extracting topical words, and finding topics from words with semantic similarity. Each topic has a certain popularity on Twitter based on its prevalence in the Twitter stream and the time it first appeared. The algorithm uses the Social Transfer framework [82] to map videos – using only the textual information from the title and video description – to topics extracted from Twitter. The popularity of a video on Twitter, expressed by the popularity of its topic, is then compared to its popularity on YouTube (represented by number of views) and, if the difference is significant, the video is considered susceptible to receive a sudden burst of attention.

Using data from YouTube and Twitter, and by training a support vector machine classifier, the algorithm shows that it can predict with 70% accuracy which videos will experience a significant increase in popularity on a daily basis. This strategy shows an improvement of almost 60% compared to a model that uses only the information available on YouTube.

Castillo et al. propose a prediction method that collects information about the early attention that news articles receive on social networks to predict the total number of

page views on a news site [14]. The statistical method used for this task is a `multiple linear regression` that uses as input the following variables: number of Facebook shares, number of tweets and retweets, entropy of tweet vocabulary, and the mean number of followers sharing the articles on Twitter. Using a collection of Al Jazeera news stories, the authors show that a model based on the social media signals received in the first ten minutes after publication achieves the same performance as one based on the number of page views received in the first three hours.

The effectiveness of cross-domain prediction methods indicate that, when information related to a web content is spread across multiple web sites, aggregating information from multiple sources can significantly improve the prediction accuracy. In particular, the information extracted from Twitter proved very useful in learning more accurate prediction models. The benefit of using social streams as an additional source of information can be explained by the fact that sharing is one of the most popular methods to reach information on the Internet. And, as sharing rarely happens inside the originating web domain, this information provides an additional – and more reactive – perspective about the actual popularity of a web content.

8 What makes web content popular?

The magic formula of what makes a web content popular is still unknown but some of the ingredients have been discovered. The content of a web item (e.g., the topic, message, or quality) plays a major role in its future success [83], but there are other elements (e.g., dissemination factors, promotion, or social influence) that have a significant contribution. Identifying the factors that impact content popularity is important in building more accurate prediction models by understanding which are the significant variables (i.e., variables that show a causal relationship) that should be used in a model or in finding alternative proxy variables when the original variable is difficult to measure. In this section we present the factors known to have a strong impact on web content popularity and we indicate in Section 9 which variables have already been used in a prediction model.

Content matters in the amount of attention that a web item will receive. Emotion is one of the most important drivers for online audience. Videos, evoking strong and mostly positive emotions, are more likely to be shared within online communities [84]. Similarly, content that generates high-arousal emotions (e.g., awe, anxiety) disseminates faster on the Internet and captures a larger amount of users' interest [85,86]. The quality of the content [87,88] and its geographic relevance [89,90] are also positively correlated with content popularity. The topic of the content is also important as content popularity is susceptible to bursts of attention in response to real-world events [91]. On the other hand, there are elements that have a negative impact on content popularity. One of them is the presence of multiple versions of the same content that tends to limit the popularity of each individual copy [28].

There are also several content-agnostic factors that have a strong impact on the popularity growth [92]. Popular Internet services, such as search tools, recommendation systems, and social sharing applications can extent web content visibility and increase its popularity. Taking the example of YouTube, the internal search engine accounts for most of the views, followed by the recommendation systems and the social sharing tools [12,92]. But the outcome of these services also play an important role in how popular a web content will become. For example, it has been observed that videos acquire a greater number of views if they are recommended in the related list of other popular videos [15,93] and the higher the position of a video in the list the greater the number of views [94]. The recommendation system thus creates a strong linked structure between similar videos, which influence each other in terms of popularity [95]. This information can be extremely valuable to newborn videos that can have a greater chance of becoming popular if they manage to create links – by choosing a relevant title, description, or keyword set – with similar popular videos.

Social sharing acts as an additional catalyst of user attention. Diffusing videos through social networks, blogs, or e-mail services generates peaks of attention during short periods of time [70]. Similarly, the social connections created within a site play an important role in how popular a web content will become. For example, it has been observed that in the early stages after the publication of a web content the greater the social network of the publisher the greater the increase in content popularity. Finally, social influence can have a non-negligible consequence on the popularity growth. A study conducted by Salganik et al. reveals that, when users are informed about the collective decisions of other individuals, the popularity of songs are driven by a "rich-get-richer" effect [87].

9 Predictive features

Accurate predictions depend on the predictive characteristics of the variables used in the model. While most prediction models proposed so far use the popularity at early moments as the only predictive variable there have been several attempts to include other features in a prediction model. We provide a brief summary of the various

features used in the prediction models and report their predictive performance.

Characteristics of content creators. The online media ecosystem is populated by content creators (independent producers, professional bloggers, mainstream mass media, or news agencies) with different but relatively stable – and maybe predictable – audience. Including the identity of the content creator in a prediction model is exploited by Bandari et al. who notice that the publisher of a news article is one of the strongest predictor of the number of tweets that a news article will generate [72].

Textual features. Certain words or key phrases that probably refer to hot or controversial topics often produce a significant amount of attention. There have been two efforts to include textual features in a prediction model. Tsagkias et al. extract the top-100 most discriminative terms from various news sources and observe that these terms have a strong performance in predicting which articles will be highly commented [71]. Similarly, Marujo et al. show that popular key-phrases have a strong predictive power in predicting the number of views for news articles [96].

Content category. Designing specialized prediction models depending on the category of the content showed little benefit in predicting the popularity of videos [33] and new articles [72,96]. The only notable exceptions have been signaled for YouTube *Music* videos [33] and news articles related to *Technology* section [72]. The low predictive performance of using this information in a prediction model can be explained by the overlapping scope of categories, with content often belonging to multiple categories at once [28,72].

Named entity identification. Popular entities in the real world (people, locations, or organizations) can often be a catalyst of user attention in the online sphere. Tsagkias et al. observe a strong impact in including popular entities from Netherlands in a prediction model designed to spot news articles that will receive a high number of comments [71].

Sentiment analysis. The specific emotion triggered by a web content is highly correlated with its online popularity [86] but extracting the correct sentiment and learning how to use this information for popularity prediction is a difficult task. The subjectivity of the language has shown little predictive power in predicting the volume of tweets for online news stories [72]. However, it has been observed that articles that are written in a more positive or negative voice, associated with strong emotions (e.g., admiration or anger), are good indicators of how viral articles will become [85]. In addition, Oghina et al. observe that subjective terms from the discussions about movies on Twitter can successfully be used as a predictive variable in predicting movie ratings on IMDb [36].

Social media signals. As we saw in Section 7.2, social media conveys valuable information about web content popularity. Castillo et al. show that the attention that news articles generate across social networks (number of Facebook shares, number of tweets and retweets, the language of the Twitter messages) is effective in predicting the popularity of articles on a news site [14]. Oghina et al. successfully use information from Twitter and YouTube to predict movie ratings on IMDb. Another example of the predictive power of social media has been reported by Roy et al. who show that the popularity of a topic on Twitter provides a good indication that a YouTube video will experience a sudden burst in popularity [35].

Social sharing viewing behavior. Yahoo! Zync is an application that allows users to share and jointly manipulate video content in real time. Shamma et al. study how users' actions during a sharing session can be used to predict the popularity of YouTube videos and observe that these interactions are strong indicators of videos with a high number of views [66].

Real-world features. Content published in online media is strongly related to real-world events but transferring information from the physical to the online world is very challenging. An attempt to employ real-world information in the predictions process has been done by Tsagkias et al. who show that there is an insignificant benefit in using the weather conditions (average temperature in Netherlands) to predict the number of comments for news articles [71].

10 Shaping the future: Applications of web content popularity prediction

In the modern information age accurate popularity predictions can prove valuable to different actors: online users can filter more easily the huge amount of information; content producers and content providers can better organize their information and build more effective delivery platforms; and advertising networks can design more sophisticated and profitable advertising strategies. However, predicting the popularity of web content, as useful as it seems, has been employed in few real-world applications. We review the current practical uses of these methods and propose new applications that could benefit from this research area.

The capacity to predict the viewership or the engagement around web content can be used as a tool for content optimization. For example, news web sites select and organize articles from a highly dynamic content pool. Instead of relying on human editors (a practice that is still common nowadays), web sites can refine their decisions through automatic solutions using online content optimization methods [97]. Agarwal et al. show that they can

significantly increase the number of clicks for Yahoo! news articles if, instead of using human editors to select and arrange the articles, one uses automatic selection algorithms that measure users' interest in news articles [97]. In this context, accurate popularity predictions, used to highlight and recommend articles, can improve user experience and boost the site's traffic. Moreover, by monitoring and reacting to social media signals, editors can increase the traffic through social media optimization solutions [11]. Currently, to the best of our knowledge, there have been no online evaluations of how the prediction methods described in this survey could be used to increase the traffic on a web site. In an offline setting Marujo et al. show good performance in predicting the number of clicks that news articles will receive during one hour [96]. Still in an offline evaluation mode, Tatar et al. explore the efficiency of two popularity prediction methods to rank news articles based on the future number of comments and show that the `log-linear` model could be an effective method for this ranking problem [73]. It is also important to understand how users react to information about the predicted popularity. Even if web content shows a mild resilience to self-fulfilling prophecies [98] the prediction outcome can become a strong form of social influence that inflates or dampens the success of a web item. One solution to this problem is to create a feedback loop to listen to users' reactions and adjust the decisions depending on how the audience is responding to the prediction outcome.

Information about the future amount of interest that web content will generate can also be valuable in online advertising as an alternative to existing contextual ad placement models [99]. The possibility to quickly spot the future popular items on the Internet creates the opportunity of additional profits for advertising agencies. Popularity prediction methods can also be used in the context of online marketing. For example, suppose that a company initiates a marketing campaign on social media with the goal of reaching a certain number of online users. To measure the success of a campaign one possible strategy is to wait until the interest in web content fades away. A more useful solution would be to monitor and predict in real-time the amount of interest that a certain post will generate and decide more quickly if additional publications would be needed [60].

Faced with an ever increasing traffic demand, content providers and content delivery networks set large-scale caching infrastructures to distribute copies of the web content across multiple locations. Optimal placement of replicas [100] – in terms of location and number of copies under bandwidth and storage constraints – depend on how accurate one can predict the future users' demand: which content will be popular [70], its geographic locality of interest [89], and the amount of attention that it will generate.

Cache replacement policies (i.e., decide on which item to evict from a cache when there is no available space) remain an important issue for the performance of a proxy cache. Traditional cache replacement algorithms use the historical information about content requests to decide which item to keep in the cache to maximize a certain performance metric (e.g., hit rate, the amount of saved bandwidth) [101]. Two of the most used replacement policies, even nowadays, are LFU (prioritize the most requested item) and LRU (prioritize the most recently requested item). One way to improve the efficiency of cache replacement algorithms is to actually integrate popularity prediction methods in the cache replacement decision. Famaey et al. propose P-LFU, an adaptation of LFU that determines which content to evict from the cache based on the predicted future demand [102]. Four generic functions (linear, power-law, exponential, and Gaussian) have been used to predict future content demand with the exponential distribution showing the most accurate results. Using a workload trace from a Video-on-Demand service the authors show that popularity prediction methods can increase the cache hit-rate with up to 10% compared with LFU cache replacement strategy.

Another domain where popularity prediction can prove valuable is mobile data offloading. Under the increasing consumption of mobile data traffic, telecom operators look for new solutions to reduce the traffic from cellular networks. Opportunistic networks have recently been proposed as an appealing solution to offload content with non-real time constrains, where, instead of using the cellular network infrastructure, mobile users can retrieve content from collocated peers [103].

In this context, one possible strategy is to benefit from the spatio-temporal mobile users requests, proactively replicate (prefetch) popular web items into mobile users' cache according to the predicted future demand, and rely on device-to-device communications to treat future content requests. To increase the performance of this strategy, the decision of what content to replicate should reflect content popularity dynamics, i.e., to replicate web content according to the predicted demand. If popularity predictions are accurate enough, future content requests could be handled by collocated mobile users and thus bypass the communication with the infrastructure and reduce network traffic and battery consumption. If, however, predictions are inaccurate, this will lead to an inefficient use of mobile and network resources. In this scenario, the benefit of predictions could be even greater if, in addition to predicting the number of users interested in a web content, one could also predict which users will trigger the requests. Social networks have become an important

mechanism for information spreading and by learning the social structures created by users one can understand the patterns of information diffusion [7,104] which further gives one the ability to predict when a user will be interested in a certain content. This type of approach has been proposed by Malandrino et al. that show that, by predicting information cascades, mobile users' requests can be treated in advance which can lead to a reduction of up to 50% of the cellular data traffic during periods of high data traffic loads [105].

Major Web search engines such as Google, Yahoo!, and Bing are always looking for new ranking factors to improve the relevancy of their search results. Over the years, search algorithms have become extremely sophisticated including hundreds of ranking factors based on the content of a web page or its importance in the Web graph. But even these complex algorithms may sometimes fail to retrieve the relevant information. For example, when searching for relevant information in the Blogosphere, Gonçalves et al. observe that commercial search engines (UOL, Yahoo!, and Google) failed to correctly retrieve an important percent of the relevant blogs on the first page of the results [106]. In the same study the authors show that the results of the query can significantly improve by including the popularity of the blog in the search algorithm. A search engine that includes the collective users' opinion about web content in its algorithm – if one has access to the different popularity statistics – is probably the future evolution of search engines [107,108]. Predicting web content popularity could fit well in this context as the outcome of a search query could prioritize future popular web content over expired one.

11 Summary and outlook

In this article we reviewed the current state-of-the-art on web content popularity prediction methods. We presented the different prediction methods, reported their performance, and suggested several applications that can benefit from these findings.

Even if research on predicting the popularity of web content has been an active area in the latest years there are many avenues that wait to be explored. We suggest some possible directions for the future work.

Predicting long-term popularity evolution. Most of the previous studies address the problem of predicting the exact amount of attention that a web content will receive up to a future moment in time. While this is useful for timely detections of popular items, a greater impact would come from a long-term evolution forecast [34,45] (i.e., to predict how content popularity evolves over time). Knowing this can reveal how content progresses through the different stages of popularity: initial growth, peak period, decline, and even popularity rebounds. This information can help online advertisers or content delivery networks in making more profitable decisions, focusing on a content during its popularity peak and wasting fewer resources on expired items.

Building richer models. In addition to early popularity measures, several studies analyzed the predictive power of various features. We believe that this direction has not been fully explored and further work is needed in finding more powerful predictive features. For example, except for Bandari et al. that use the identity of the publisher in the prediction model [72], to our knowledge no other work has studied the predictive power of a content publisher. Yet news columnists and video publishers attract a significant and maybe predictable audience on their own.

The topic of the web content plays an important role in its future popularity. The daily agenda of discussions on the Internet and mainstream media is centered on major topics with limited and different life cycles. Thus, capturing trending topics and learning how to include them in prediction models can lead to a major breakthrough in the prediction accuracy. Research in this field has made important advances in the recent years. Leskovec et al. show that the attention that online users pay to certain topics can accurately be described by six time-series shapes [109]. Similarly, Nikolov et al. propose an algorithm that can accurately predict the trending topics on Twitter earlier (with an average of 1.43 hours) than the internal algorithm used by Twitter [110].

For web content characterized by a very short lifecycle it has been observed that timely predictions present a real challenge. For example, news articles quickly become popular and "die-out" within hours. One way to improve the predictability of news would be to extract recurrent events over time, observe the level of interest that they generate, and predict when these future events will take place. Predicting global events in various fields (e.g., economy, seismology, society), as challenging as it may seem, is nevertheless plausible. Radinsky et al. propose two algorithms for this prediction task: PROFET, an algorithm that predicts the terms used in the future news stories based on the historical web query patterns [111]; and Pandit, a system that can predict future events given an existing news event [112].

Understanding and merging user activity stemming from different web channels is an important direction to follow. Up to now, Twitter feed has been used as the main source of information. But there are other potential directions to explore. For example, analyzing Web users' query behavior can unveil important insights about the popularity of certain topics and the ability to predict search queries, as showed by Radinsky

et al. [113], could be incorporated in a popularity prediction model. Wikipedia is also a valuable source of information. Important real-life events are quickly recorded on Wikipedia and real-time monitoring of this channel can be transformed into valuable information in a prediction model. Wikipedia Live Monitor is a good example of automatic monitoring tool that detects breaking news events by studying simultaneous user activity for certain topics edited in different languages [114,115].

Beyond popularity predictions. Studying online content popularity should be used not only to better understand the dynamics of content consumption but also to improve various web services. For instance, by understanding which factors influence content popularity, content producers can design the genome of popular content. Although there are many factors that are difficult to control, creating content that is original (remember that multiple copies of the same content has a negative impact on popularity [28]), fresh (the advantage of the first-comer [92]), emotional (stronger emotions are correlated to content viralness [85]), and by tagging it with popular keywords (to appear in more popular recommendation lists [93]) can increase the likelihood of web content becoming popular. Then, online advertisers should try to figure out how to seize the opportunity of finding popular content in advance and design novel monetization strategies. Finally, there are few reports on how content popularity prediction can be used to design more effective networking solutions. Yet predicting web content popularity dynamics can be used to design more scalable content delivery solutions that proactively replicate content according to the future users' demand.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
AT, MDdA, SF, and PA summarized the papers covered in this article, classified the existing literature, and proposed future perspectives for this research area. All authors read and approved the final manuscript.

Acknowledgements
We are grateful to the anonymous reviewers for their valuable comments and suggestions. The work presented in this paper has been carried out at LINCS (www.lincs.fr) and was partially supported by EINS, the Network of Excellence in Internet Science, FP7 grant 28802.

Author details
[1]LIP6/CNRS – UPMC Sorbonne Universités, 4 Place Jussieu, 75005 Paris, France. [2]Communication System Group – ETH Zurich, 35 Gloriastrasse, 8092 Zurich, Switzerland.

Received: 21 November 2013 Accepted: 4 July 2014
Published: 13 August 2014

References
1. Telegraph (2013). http://www.telegraph.co.uk/technology/twitter/9945505/Twitter-in-numbers.html
2. Facebook statistics (2013). https://newsroom.fb.com/News
3. YouTube Statistics (2013). http://www.youtube.com/yt/press/statistics.html
4. Internet Advertising Bureau (2013). http://www.iabuk.net/about/press/archive/uk-digital-adspend-up-125-to-almost-55bn
5. Cha M, Mislove A, Adams B, Gummadi KP (2008) Characterizing social cascades in Flickr. In: Proceedings of the First Workshop on Online Social Networks. ACM, Seattle, Washington, USA, pp 13–18
6. Sadikov E, Medina M, Leskovec J, Garcia-Molina H (2011) Correcting for missing data in information cascades. In: Proceedings of the Fourth ACM International Conference on Web Search and Data Mining. ACM, Hong Kong, China, pp 55–64
7. Cheng J, Adamic L, Dow PA, Kleinberg JM, Leskovec J (2014) Can cascades be predicted? In: Proceedings of the 23rd International Conference on World Wide Web (WWW). ACM, Seoul, Republic of, Korea, pp 925–936
8. Yu S, Kak S (2012) A survey of prediction using social media. arXiv preprint arXiv:1203.1647
9. Bernstein MS, Bakshy E, Burke M, Karrer B (2013) Quantifying the invisible audience in social networks. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, Paris, France, pp 21–30
10. Agarwal D, Chen B-C, Wang X (2012) Multi-faceted ranking of news articles using post-read actions. In: Proceedings of the 21st International Conference on Information and Knowledge Management. CIKM '12. ACM, Maui, Hawaii, USA, pp 694–703
11. Lifshits Y (2010) Ediscope: Social analytics for online news. Technical Report YL-2010-008. Yahoo! Labs
12. Figueiredo F, Benevenuto F, Almeida JM (2011) The tube over time: characterizing popularity growth of youtube videos. In: Proceedings of the Fourth ACM International Conference on Web Search and Data Mining. ACM, Hong Kong, China, pp 745–754
13. Yao Y, Sun A (2013) Are most-viewed news articles most-shared? In: AIRS. Lecture Notes in Computer Science, vol. 8281. Springer, Singapore, pp 404–415
14. Castillo C, El-Haddad M, Pfeffer J, Stempeck M (2013) Characterizing the life cycle of online news stories using social media reactions. arXiv preprint arXiv:1304.3010
15. Chatzopoulou G, Sheng C, Faloutsos M (2010) A first step towards understanding popularity in youtube. In: INFOCOM IEEE Conference on Computer Communications Workshops. IEEE, San Diego, CA, pp 1–6
16. Cunha CR, Bestavros A, Crovella ME (1995) Characteristics of WWW client-based traces. Technical report, Computer Science Department, Boston University
17. Almeida V, Bestavros A, Crovella M, de Oliveira A (1996) Characterizing reference locality in the WWW. In: Parallel and Distributed Information Systems (PDIS). IEEE, Miami Beach, FL, pp 92–103
18. Breslau L, Cao P, Fan L, Phillips G, Shenker S (1999) Web caching and Zipf-like distributions: Evidence and implications. In: INFOCOM'99. Eighteenth Annual Joint Conference of the IEEE Computer and Communications Societies, vol. 1. IEEE, New York, NY, pp 126–134
19. Barford P, Bestavros A, Bradley A, Crovella M (1999) Changes in web client access patterns: Characteristics and caching implications. World Wide Web 2(1-2)‡15–28
20. Chesire M, Wolman A, Voelker GM, Levy HM (2001) Measurement and analysis of a streaming media workload. In: Proceedings of the 3rd Conference on USENIX Symposium on Internet Technologies and Systems, vol. 3. USENIX Association, San Francisco, California
21. Almeida JM, Krueger J, Eager DL, Vernon MK (2001) Analysis of educational media server workloads. In: Proceedings of the 11th International Workshop on Network and Operating Systems Support for Digital Audio and Video. ACM, Port Jefferson, New York, USA, pp 21–30
22. Sripanidkulchai K, Maggs B, Zhang H (2004) An analysis of live streaming workloads on the internet. In: Proceedings of the 4th ACM SIGCOMM Conference on Internet Measurement. ACM, Taormina, Sicily, Italy, pp 41–54
23. Yu H, Zheng D, Zhao BY, Zheng W (2006) Understanding user behavior in large-scale video-on-demand systems. In: ACM SIGOPS Operating Systems Review, vol. 40. ACM, Leuven, Belgium, pp 333–344

24. Cherkasova L, Gupta M (2004) Analysis of enterprise media server workloads: access patterns, locality, content evolution, and rates of change. IEEE/ACM Trans Netw 12(5)781–794
25. Gill P, Arlitt M, Li Z, Mahanti A (2007) Youtube traffic characterization: a view from the edge. In: Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement. ACM, San Diego, California, USA, pp 15–28
26. Cha M, Kwak H, Rodriguez P, Ahn Y-Y, Moon S (2007) I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system. In: Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement. ACM, San Diego, California, USA, pp 1–14
27. Zink M, Suh K, Gu Y, Kurose J (2009) Characteristics of youtube network traffic at a campus network - measurements, models, and implications. Comput Netw 53(4)501–514
28. Cha M, Kwak H, Rodriguez P, Ahn Y-Y, Moon S (2009) Analyzing the video popularity characteristics of large-scale user generated content systems. IEEE/ACM Trans Netw. (TON) 17(5)1357–1370
29. Kaltenbrunner A, Gomez V, Lopez V (2007) Description and prediction of slashdot activity In: Web Conference, 2007. LA-WEB 2007. Latin American. IEEE, Santiago, Chile, pp 57–66
30. Szabo G, Huberman BA (2010) Predicting the popularity of online content. Commun ACM 53(8)80–88
31. Crane R, Sornette D (2008) Robust dynamic classes revealed by measuring the response function of a social system. Proc Nat Acad Sci 105(41)15649–15653
32. Gursun G, Crovella M, Matta I (2011) Describing and forecasting video access patterns. In: INFOCOM, 2011 Proceedings IEEE. IEEE, Shanghai, pp 16–20
33. Pinto H, Almeida JM, Gonçalves MA (2013) 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. WSDM '13. ACM, Rome, Italy, pp 365–374
34. Ahmed M, Spagna S, Huici F, Niccolini S (2013) 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. WSDM '13. ACM, Rome, Italy, pp 607–616
35. Roy SD, Mei T, Zeng W, Li S (2013) Towards cross-domain learning for social video popularity prediction. IEEE Trans Multimedia 15(1255-1267)
36. Oghina A, Breuss M, Tsagkias M, de Rijke M (2012) Predicting IMDb movie ratings using social media In: Proceedings of the 34th European Conference on Advances in Information Retrieval. ECIR'12. Springer, Barcelona, Spain, pp 503–507
37. Cisco Visual Networking Index: Forecast and Methodology (2014). http://www.cisco.com/c/en/us/solutions/collateral/service-provider/ip-ngn-ip-next-generationnetwork/white_paper_c11-481360.html
38. Kwak H, Lee C, Park H, Moon S (2010) What is twitter, a social network or a news media? In: Proceedings of the 19th International Conference on World Wide Web. ACM, Raleigh, North Carolina, USA, pp 591–600
39. Reynolds Journalism Institute: News consumption on mobile media (2008). http://www.rjionline.org/research/rji-dpa-mobile-media-project/2013-q1-research-report-1
40. Mashable (2013). http://mashable.com/2011/12/31/youtube-in-2011
41. Cheng X, Dale C, Liu J (2008) Statistics and social network of youtube videos. In: 16th International Workshop on Quality of Service. IWQoS. IEEE, Enschede, pp 229–238
42. Borghol Y, Mitra S, Ardon S, Carlsson N, Eager D, Mahanti A (2011) Characterizing and modelling popularity of user-generated videos. Perform Eval 68(11)1037–1055
43. Cheng X, Dale C, Liu J (2007) Understanding the characteristics of internet short video sharing: Youtube as a case study. arXiv preprint arXiv:0707.3670
44. Avramova Z, Wittevrongel S, Bruneel H, De Vleeschauwer D (2009) Analysis and modeling of video popularity evolution in various online video content systems: power-law versus exponential decay. In: First International Conference on Evolving Internet (INTERNET'09). IEEE, Cannes/La Bocca, pp 95–100
45. Figueiredo F (2013) 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. WSDM '13. ACM, Rome, Italy, pp 741–746
46. Carlinet L, Huynh T, Kauffmann B, Mathieu F, Noirie L, Tixeuil S (2012) Four months in daily motion: Dissecting user video requests. In: Wireless Communications and Mobile Computing Conference (IWCMC). IEEE, Limassol, Cyprus, pp 613–618
47. Mitra S, Agrawal M, Yadav A, Carlsson N, Eager D, Mahanti A (2011) Characterizing web-based video sharing workloads. ACM Trans Web (TWEB) 5(2)8
48. Dezsö Z, Almaas E, Lukács A, Rácz B, Szakadát I, Barabási A-L (2006) Dynamics of information access on the web. Phys Rev E 73(6)066132
49. Mishne G, Glance N (2006) Leave a reply: An analysis of weblog comments. In: Third Annual Workshop on the Weblogging Ecosystem. Edinburgh, UK
50. Tatar A, Antoniadis P, de Amorim MD, Fdida S (2012) Ranking news articles based on popularity prediction. In: Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012). IEEE, Istanbul, pp 106–110
51. Tsagkias M, Weerkamp W, De Rijke M (2010) News comments: Exploring, modeling, and online prediction. In: Proceedings of the 32Nd European Conference on Advances in Information Retrieval. ECIR'2010. Springer, Milton Keynes, UK, pp 191–203
52. Lerman K, Ghosh R (2010) Information contagion: An empirical study of the spread of news on Digg and Twitter social networks. In: ICWSM, vol. 10. The AAAI Press, Washington, DC, USA, pp 90–97
53. Wang C, Ye M, Huberman BA (2012) From user comments to on-line conversations. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '12. ACM, Beijing, China, pp 244–252
54. Wallenta C, Ahmed M, Brown I, Hailes S, Huici F (2008) Analysing and modelling traffic of systems with highly dynamic user generated content. University College London, Tech. Rep. RN/08/10
55. Gómez V, Kaltenbrunner A, López V (2008) Statistical analysis of the social network and discussion threads in Slashdot. In: Proceedings of the 17th International Conference on World Wide Web. ACM, Beijing, China, pp 645–654
56. Jamali S, Rangwala H (2009) Digging digg: Comment mining, popularity prediction, and social network analysis. In: International Conference on Web Information Systems and Mining (WISM). IEEE, Shanghai, China, pp 32–38
57. Lerman K, Galstyan A (2008) Analysis of social voting patterns on digg. In: Proceedings of the First Workshop on Online Social Networks. ACM, Seattle, WA, USA, pp 7–12
58. Tang S, Blenn N, Doerr C, Van Mieghem P (2011) Digging in the digg social news website. IEEE Trans Multimedia 13(5)1163–1175
59. Van Mieghem P, Blenn N, Doerr C (2011) Lognormal distribution in the digg online social network. Eur Phys J B 83(2)251–261
60. Kupavskii A, Umnov A, Gusev G, Serdyukov P (2013) Predicting the audience size of a tweet. In: ICWSM. Cambridge, Massachusetts, USA, The AAAI Press
61. Hong L, Dan O, Davison BD (2011) Predicting popular messages in Twitter. In: Proceedings of the 20th International Conference Companion on World Wide Web. ACM, Hyderabad, India, pp 57–58
62. Ma H, Qian W, Xia F, He X, Xu J, Zhou A (2013) Towards modeling popularity of microblogs. Front Comput Sci 7(2)171–184
63. Kong S, Ye F, Feng L (2014) Predicting future retweet counts in a microblog. J Comput Inform Syst 10(4)1393–1404
64. Zaman T, Fox EB, Bradlow ET (2013) A bayesian approach for predicting the popularity of tweets. arXiv preprint arXiv:1304.6777
65. Lee JG, Moon S, Salamatian K (2012) Modeling and predicting the popularity of online contents with Cox proportional hazard regression model. Neurocomputing 76(1)134–145
66. Shamma DA, Yew J, Kennedy L, Churchill EF (2011) Viral actions: Predicting video view counts using synchronous sharing behaviors. ICWSM, Barcelona, Spain
67. Yin P, Luo P, Wang M, Lee W-C (2012) A straw shows which way the wind blows: ranking potentially popular items from early votes. In: Proceedings of the Fifth ACM International Conference on Web Search and Data Mining. ACM, Seattle, Washington, USA, pp 623–632
68. Witten IH, Frank E (2005) Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, Burlington, Massachusetts, USA
69. Manning CD, Raghavan P, Schütze H (2008) Introduction to Information Retrieval vol. 1. Cambridge University Press, Cambridge, NY
70. Broxton T, Interian Y, Vaver J, Wattenhofer M (2010) Catching a viral video In: International Conference on Data Mining Workshops (ICDMW). IEEE, Sydney, NSW, pp 296–304

71. Tsagkias M, Weerkamp W, De Rijke M (2009) Predicting the volume of comments on online news stories. In: Proceedings of the 18th ACM Conference on Information and Knowledge Management. ACM, Hong Kong, China, pp 1765–1768
72. Bandari R, Asur S, Huberman BA (2012) The pulse of news in social media: Forecasting popularity. In: ICWSM. The AAAI Press, Dublin, Ireland
73. Tatar A, Antoniadis P, de Amorim MD, Fdida S (2014) From popularity prediction to ranking online news. Soc Netw Anal Min 4(174)
74. Tatar A, Leguay J, Antoniadis P, Limbourg A, de Amorim MD, Fdida S (2011) Predicting the popularity of online articles based on user comments. In: Proceedings of the International Conference on Web Intelligence, Mining and Semantics. ACM, Sogndal, Norway
75. Kim S-D, Kim S-H, Cho H-G (2011) Predicting the virtual temperature of web-blog articles as a measurement tool for online popularity. In: 11th International Conference on Computer and Information Technology (CIT). IEEE, Pafos, Cyprus, pp 449–454
76. Maass W, Natschläger T, Markram H (2002) Real-time computing without stable states: A new framework for neural computation based on perturbations. Neural Comput 14(11)2531–2560
77. Wu T, Timmers M, Vleeschauwer DD, Leekwijck WV (2010) On the use of reservoir computing in popularity prediction. In: Proceedings of the 2010 2nd International Conference on Evolving Internet. INTERNET '10. IEEE, Valencia, Spain, pp 19–24
78. Yang J, Leskovec J (2011) Patterns of temporal variation in online media. In: Proceedings of the Fourth ACM International Conference on Web Search and Data Mining. ACM, Hong Kong, China, pp 177–186
79. Hastie T, Tibshirani R, Friedman J, Franklin J (2009) The Elements of Statistical Learning: Data Mining, Inference and Prediction. 2nd edn. Springer, NY
80. Frey BJ, Dueck D (2007) Clustering by passing messages between data points. Science 315(5814)972–976
81. Lerman K, Hogg T (2010) Using a model of social dynamics to predict popularity of news. In: Proceedings of the 19th International Conference on World Wide Web. ACM, Raleigh, North Carolina, USA, pp 621–630
82. Roy SD, Mei T, Zeng W, Li S (2012) Socialtransfer: cross-domain transfer learning from social streams for media applications. In: Proceedings of the 20th ACM International Conference on Multimedia. ACM, Nara, Japan, pp 649–658
83. Figueiredo F, Almeida JM, Benevenuto F, Gummadi KP (2014) Does content determine information popularity in social media?: a case study of youtube videos' content and their popularity. In: Proceedings of the 32Nd Annual ACM Conference on Human Factors in Computing Systems. CHI '14. ACM, Toronto, ON, Canada, pp 979–982
84. Guadagno RE, Rempala DM, Murphy S, Okdie BM (2013) What makes a video go viral? An analysis of emotional contagion and internet memes. Comput Hum Behav 29(6)2312–2319
85. Berger JA, Milkman KL (2012) What makes online content viral? J Market Res 49(2)192–205
86. Berger J (2011) Arousal increases social transmission of information. Psychol Sci 22(7)891–893
87. Salganik MJ, Dodds PS, Watts DJ (2006) Experimental study of inequality and unpredictability in an artificial cultural market. Science 311(5762):854–856
88. Hogg T, Szabo G (2009) Diversity of user activity and content quality in online communities. In: Proceedings of 3rd International Conference on Weblogs and Social Media (ICWSM). The AAAI Press, San Jose, California, USA
89. Brodersen A, Scellato S, Wattenhofer M (2012) Youtube around the world: geographic popularity of videos. In: Proceedings of the 21st International Conference on World Wide Web. ACM, Lyon, France, pp 241–250
90. Huguenin K, Kermarrec A-M, Kloudas K, Taïani F (2012) Content and geographical locality in user-generated content sharing systems. In: Proceedings of the 22nd International Workshop on Network and Operating System Support for Digital Audio and Video. ACM, Toronto, Ontario, Canada, pp 77–82
91. Ratkiewicz J, Flammini A, Menczer F (2010) Traffic in social media I: paths through information networks. In: IEEE Second International Conference on Social Computing (SocialCom). IEEE, Minneapolis, MN, USA, pp 452–458
92. Borghol Y, Ardon S, Carlsson N, Eager D, Mahanti A (2012) 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. ACM, Beijing, China, pp 1186–1194
93. Davidson J, Liebald B, Liu J, Nandy P, Van Vleet T, Gargi U, Gupta S, He Y, Lambert M, Livingston B, Sampath D (2010) The YouTube video recommendation system. In: Proceedings of the Fourth ACM Conference on Recommender Systems. ACM, Barcelona, Spain, pp 293–296
94. Zhou R, Khemmarat S, Gao L (2010) The impact of YouTube recommendation system on video views. In: Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement. IMC '10. ACM, Melbourne, Australia, pp 404–410
95. Zhou R, Khemmarat S, Gao L, Wang H (2011) Boosting video popularity through recommendation systems. In: Databases and Social Networks. ACM, Athens, Greece, pp 13–18
96. Marujo L, Bugalho M, Neto JPDS, Gershman A, Carbonell J (2013) Hourly traffic prediction of news stories. arXiv preprint arXiv:1306.4608
97. Agarwal D, Chen B-C, Elango P, Motgi N, Park S-T, Ramakrishnan R, Roy S, Zachariah J (2008) Online models for content optimization. In: Advances in Neural Information Processing Systems. MIT Press, Cambridge, MA, pp 17–24
98. Salganik MJ, Watts DJ (2008) Leading the herd astray: An experimental study of self-fulfilling prophecies in an artificial cultural market. Soc Psychol Q 71(4)338–355
99. Ghose A, Yang S (2009) An empirical analysis of search engine advertising: Sponsored search in electronic markets. Manag Sci 55(10)1605–1622
100. Applegate D, Archer A, Gopalakrishnan V, Lee S, Ramakrishnan KK (2010) Optimal content placement for a large-scale VoD system. In: CoNEXT. ACM, Philadelphia, Pennsylvania, USA
101. Podlipnig S, Böszörmenyi L (2003) A survey of web cache replacement strategies. ACM Computing Surveys (CSUR) 35(4)374–398
102. Famaey J, Iterbeke F, Wauters T, DeTurck F (2013) Towards a predictive cache replacement strategy for multimedia content. J Netw Comput Appl 36(1)219–227
103. Han B, Hui P, Kumar V, Marathe M, Shao J, Srinivasan A (2012) Mobile Data Offloading Through Opportunistic Communications and Social Participation. IEEE Transactions on Mobile Computing 11(5)821–834
104. Galuba W, Aberer K, Chakraborty D, Despotovic Z, Kellerer W (2010) Outtweeting the twitterers-predicting information cascades in microblogs. In: Proceedings of the 3rd Conference on Online Social Networks. USENIX Association, Boston, MA, pp 3–3
105. Malandrino F, Kurant M, Markopoulou A, Westphal C, Kozat UC (2012) Proactive seeding for information cascades in cellular networks. In: INFOCOM IEEE Conference on Computer Communications Workshops. IEEE, Orlando, FL, USA, pp 1719–1727
106. Gonçalves MA, Almeida JM, dos Santos LG, Laender AH, Almeida V (2010) On popularity in the blogosphere. Internet Comput IEEE 14(3)42–49
107. Bao S, Xue G, Wu X, Yu Y, Fei B, Su Z (2007) Optimizing web search using social annotations. In: Proceedings of the 16th International Conference on World Wide Web. ACM, Banff, Alberta, Canada, pp 501–510
108. Facebook search engine (2012). http://www.searchenginejournal.com/facebook-seo-beastrank-ranking-factors-facebook-search-engine/49696/
109. Leskovec J, Backstrom L, Kleinberg J (2009) Meme-tracking and the dynamics of the news cycle. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '09. ACM, Paris, France, pp 497–506
110. Nikolov S (2012) Trend or no trend: a novel nonparametric method for classifying time series. Master's thesis, Massachusetts Institute of Technology, Cambridge, MA, USA
111. Radinsky K, Davidovich S, Markovitch S (2008) Predicting the news of tomorrow using patterns in web search queries. In: Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology-Volume 01. IEEE Computer Society, Washington, DC,USA, pp 363–367
112. Radinsky K, Davidovich S, Markovitch S (2012) Learning causality for news events prediction. In: Proceedings of the 21st International Conference on World Wide Web. ACM, Lyon, France, pp 909–918
113. Radinsky K, Svore K, Dumais S, Teevan J, Bocharov A, Horvitz E (2012) Modeling and predicting behavioral dynamics on the web. In: Proceedings of the 21st International Conference on World Wide Web. ACM, Lyon, France, pp 599–608

114. Steiner T, van Hooland S, Summers E (2013) MJ no more: using concurrent Wikipedia edit spikes with social network plausibility checks for breaking news detection. In: Proceedings of the 22nd International Conference on World Wide Web Companion. International World Wide Web Conferences Steering Committee, Rio de Janeiro, Brazil, pp 791–794
115. Steiner T (2014) Telling breaking news stories from Wikipedia with social multimedia: A case study of the 2014 winter olympics. arXiv preprint arXiv:1403.4289

doi:10.1186/s13174-014-0008-y
Cite this article as: Tatar *et al.*: A survey on predicting the popularity of web content. *Journal of Internet Services and Applications* 2014 **5**:8.