

Studying Page Life Patterns in Dynamical Web

Alexey Tikhonov, Ivan Bogatyy, Pavel Burangulov, Liudmila Ostroumova,
Vitaliy Koshelev, Gleb Gusev

Yandex

16 Leo Tolstoy St., Moscow, 119021 Russia

{altsoph, loken17, burangulov, ostroumova-la, vakoshelev, gleb57}@yandex-team.ru

ABSTRACT

With the ever-increasing speed of content turnover on the web, it is particularly important to understand the patterns that pages' popularity follows. This paper focuses on the dynamical part of the web, i.e. pages that have a limited lifespan and experience a short popularity outburst within it. We classify these pages into five patterns based on how quickly they gain popularity and how quickly they lose it. We study the properties of pages that belong to each pattern and determine content topics that contain disproportionately high fractions of particular patterns. These developments are utilized to create an algorithm that approximates with reasonable accuracy the expected popularity pattern of a web page based on its URL and, if available, prior knowledge about its domain's topics.

Categories and Subject Descriptions: H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval

General Terms: Theory, Experimentation

Keywords: temporal profiles; web page popularity; time series; classification.

1. INTRODUCTION

A lot of effort has been put recently into understanding the dynamical aspects of collective attention on the web. A number of papers were devoted to the analysis and prediction of information diffusion, evaluating it in terms of user influence [1] and popularity of different pieces of content in social media, such as topics [9], blog posts [4], links [1, 7], and news [10].

From a search engine perspective, it would be useful to study the global activity dynamics of web pages, that is, the patterns that the total traffic of a single page can follow. First, some state-of-the-art crawling methods employ user browsing behavior while optimizing their policies [6]. We believe the performance of such crawling policies may be improved by incorporating the prediction of new pages' future

popularity. Second, the role of the web as a primary source of up-to-date information becomes increasingly important, which leads to a significant portion of pages becoming useless several days after their creation. A proper understanding of conditions that cause traffic decay of a page may enable us to appropriately detect and remove lifeless content from a search engine index, thus reducing computational resources needed to process a query.

In this paper, we provide a large-scale temporal analysis of web traffic by examining the stream of visits to web pages from users in a European country, employing the browsing log from a popular toolbar. In this study we focus on dynamical activity profiles of web pages whose lifespan is limited to a short period of activity spike. We define *activity profile* of a page as the time series of its daily visits count.

A possible approach to the analysis of the temporal activity is splitting the whole variety of observed time series into a limited number of patterns. We build a multiclass classifier trained on a manually annotated dataset of activity time series. Each time series is judged as belonging to one of the following five categories: the four spike patterns corresponding to the presence or absence of significant activity before and after the peak moment and the case of a temporal profile which does not look like a spike. We apply the obtained classifier to a large-scale dataset of pages and use their predicted labels as estimates of their popularity patterns. These estimates enable us to study the dependence of web page activity profiles on different factors, such as the second-level domain of a page and the topic of the domain.

We observe a noticeable connection between the popularity pattern of a page and the topic of its domain. For each of the four spike patterns, we extract topics with the highest fraction of the considered spike pattern (measured against frequency in the whole dataset). The obtained results can mostly be well explained and look intuitive given the common knowledge about the user browsing behavior.

Encouraged by these observations, we train several multiclass classifiers that predict the (estimated) popularity pattern of a page by (1) the topic of its domain, (2) its URL's tokens, and (3) prior distribution over popularity patterns for the pages with the same domain. The obtained results argue that the temporal popularity profile of a newly discovered web page may be somewhat accurately predicted by its URL. To the best of our knowledge, this is the first study devoted to the temporal popularity patterns of Internet pages.

The remainder of the paper is organized as follows. The next section is a review of related works in social media domain. We describe the data we use for designing activity

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

SIGIR '13, July 28–August 1, 2013, Dublin, Ireland.

ACM 978-1-4503-2034-4/13/07 ...\$15.00.

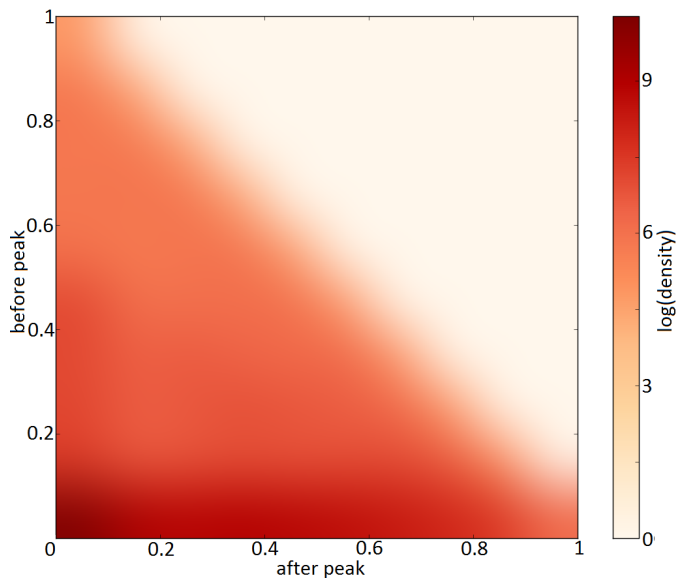


Figure 1: Heat map captures distribution of activity profiles in the space of two factors: portions before and after peak. The stripe on the right indicates density per unit cell. There are no pronounced clusters in this space.

profiles and provide analysis of their patterns in Sections 3 and 4 respectively. In Section 5 we describe the effect of domain topics on their page popularity patterns. We describe results of prediction of popularity patterns in Section 6. Section 7 concludes the paper.

2. RELATED WORK

There is little work on the analysis of the popularity dynamics of web pages. In [8] Radinsky et al. propose a physical model that captures popularity variation of a query, a document, and a query-document pair. However, their study deals with traffic from a particular search engine only, which may be biased. Furthermore, the authors focus on modeling and predicting future popularity, while our task focuses on analyzing the distribution of popularity patterns.

In [11] Yang et al. performed a clustering of two sets of time series, first one quantifying mentions of popular phrases in news sources and the second one measuring appearance of hashtags in Twitter, both using hour granularity. In [5] authors cluster time series of Twitter hashtags usage at day granularity relying on two factors: the portion of activity before the peak day and the portion of activity after it. The results obtained in the two papers are rather similar and capture the four types of spikes corresponding to the presence of absence of significant activity before and after the peak moment. Relying on these results, we propose a classification algorithm to determine the types of patterns that the temporal series under consideration follow.

3. DATA PREPARATION

All the experiments in this paper were performed on a fully anonymized web page visits log from a search engine¹

¹yandex.com

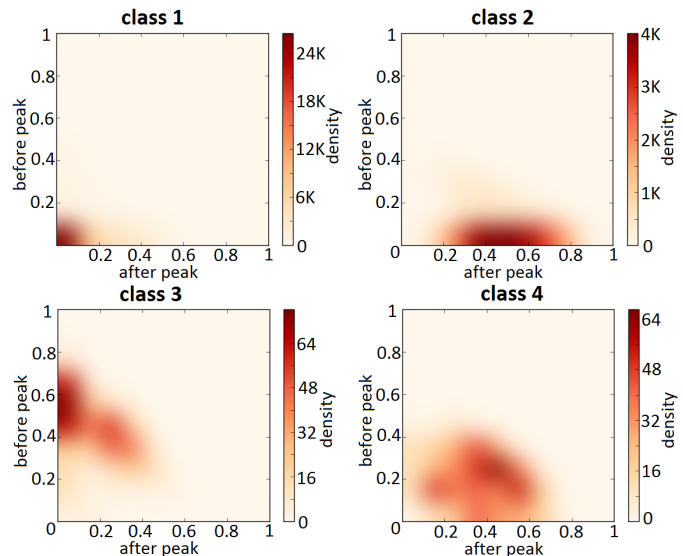






Figure 2: Heat maps capture distributions of activity portions before and after peak for different temporal patterns of activity spikes. The stripes indicate density per unit cell.

browser toolbar, used by millions of people across different countries.

We extract records made in the 80-day period from August 27, 2012 to November 14, 2012. For each visited URL in the log, we calculate the number of hits the URL received on each of those 80 days. The obtained time series represent the temporal activity profiles we study.

We filter the dataset by imposing two conditions. First, we remove all pages that have non-zero visits on the first day of the considered time frame. Those pages are likely to have been created before the time period we examine, making us unable to analyze the beginning of their lifespan. In particular, this filters out the already popular web pages, allowing us to focus on the short-lived pages that form dynamical part of the Web. Second, we remove pages whose maximal activity does not reach 100 hits per day. This allows us to filter out non-popular pages, leaving only the most important content for the study. In total, the collected dataset describes activity profiles of approximately 3.1M web pages.

4. TEMPORAL PATTERNS OF PAGE LIFE

This section describes the particular classes of temporal activity patterns that we distinguish in our study. Both the results in the previous works and our own observations confirm the validity of the following five-types classification model. The first four types represent different types of activity spike with a pronounced single peak. The slope of a spike before the peak can be either sharp (a quick rise of popularity in one or two days) or *soft* (the activity rises steadily throughout a longer period of time). Similarly, the slope of a spike after the peak can be *light* or *substantial*. Different combinations of slope types before and after the peak form the four types of spikes: (1) sharp rise, light tail , (2) sharp rise, substantial tail , (3) soft rise, light tail , (4) soft rise, substantial tail . The fifth type

represents profiles that (5) do not look like a spike. This includes several pronounced peaks, or even a period of stable user attention strictly limited to a certain period of time.

In [5], Lehman et al. proposed to cluster the temporal profiles of user attention by projecting them onto the space of the following two factors: the portion of activity before the peak day and the portion of activity after that. Unfortunately, this approach does not suit well the case of web pages, which are not necessarily very popular within our study and thus their activity profiles are noisier and allow for more variety within each of the spike types (see the distribution of activity profiles in the space of those two factors on Fig. 1).

Instead, we perform a multiclass classification as follows. We manually examine a sample of 1377 URLs, annotating their activity profiles with five labels (1)-(5) described above. We then train a Friedman’s gradient boosting decision tree model [2] on the obtained judged dataset, extracting the significant properties of the daily visits graphs to use as features. We used 12 features, with the most significant three being: the fraction of visits that happen 3 days after the peak and later, the time difference between the first visit to the page and its peak day, and the fraction of visits before the peak day.

The classes 3 and 4 (soft rise case) are rather rare, constituting approximately 1.5% and 0.5% respectively in a random sample. As a result, the trained algorithm is biased to detect them even less often than they occur. We overcome this problem as follows. After the first round of annotating a random sample and training the algorithm, we use the obtained algorithm to filter the whole database for URLs that are likely to fall into classes 3 and 4. After that, we annotate this particular sample in order to obtain a better representation of the rare classes. In total, we annotated 408, 149, 170, 65, and 585 profiles falling into classes 1 to 5 respectively.

Tab. 2, A) represents the confusion matrix of the classification model averaged over test for 10-fold cross validation on the extended judged dataset.

A)

		predicted				
		1	2	3	4	5
annotated	1	354	23	12	1	18
	2	19	110	0	5	15
	3	3	1	126	12	28
	4	1	2	18	32	12
	5	18	21	48	6	492

B)

		1	2	3	4	5
1	1	-0,54	-0,07	-0,20	-0,65	
2	-0,54	1	-0,14	-0,01	-0,28	
3	-0,07	-0,14	1	0,06	0,11	
4	-0,20	-0,014	0,06	1	0,12	
5	-0,65	-0,28	0,11	0,12	1	

Table 2: A) Confusion matrix of the classification model that predicts the temporal pattern of a web page activity profile; B) Correlation matrix of frequencies of popularity patterns over a set of second-level domains.

We evaluate the performance of multi-class classifiers utilizing MAUC (multi-class AUC) as defined in [3]. This measure is simple to implement, does not require weighting of

error types, and is not sensitive to highly unbalanced distribution of prior class probabilities. The value of MAUC averaged over test for 10-fold cross validation is **0.89** for our classification model.

We then apply the classification model to the whole dataset and treat the time series-based predictions as the actual pattern types. As a result, we are able to analyze the pattern types of 3.1M URLs in the context of their domain topic and other properties as explained in the next sections. See Tab. 3 for statistics of pattern distribution over the dataset and Fig. 2 for the distribution of activity portions before and after the peak for different popularity patterns.

class	Σ	1	2	3	4	5
URL count	3.1M	1.6M	0.7M	18K	24K	0.7M
URL portion	100%	52%	24%	0,6%	0,8%	23%
hits portion	100%	28%	20%	0.8%	0.7%	50%

Table 3: Rate statistics including portions of hits gathered by URL’s of different temporal profiles.

5. DOMAIN TOPICS

In this section we study the distribution of pages popularity patterns conditioned by different topics of their domains. For this purpose, we use a public database of domains manually categorized by their topics for a popular search engine. We train a Naive Bayes classifier on this data by employing unigram features. For each page whose activity profile is represented in our dataset, we define its second-level domain and apply the obtained topic classifier to the content pages of that domain. If some topic covers a greater portion of that pages, we attribute the whole domain to that topic. After that, we mark out 154 most frequent topics according to the number of pages of our dataset whose domain is attributed to these topics. Covering about 87% of content, they are employed as domain topic features in the classification tasks of the next section.

For each of the four patterns of spikes, we rank these 154 topics according to the frequency of that pattern within their pages. We present the top topics of the four lists in Tab. 1. As we see, the first profile is more typical for news pages that are meant for a short-term people attention, such as sport events, concerts, etc. The traffic of these pages is attributed to the event moment, it rises and falls quickly. The second pattern is typical for pages, which appear on large entertainment portals devoted to fun or hobby. They gather a high traffic right after their creation and keep popular for some time. The third pattern is determined by rapid fall of popularity after a long rise. It is typical for pages devoted to upcoming events, the most striking example is weather forecast. Another example is a car ads, usually gaining traffic until the car is sold. The third pattern is not well separated from the fourth pattern, which is more frequent for blog posts and other social media content, whose popularity rises relatively slowly till its peak and then falls gradually.

We also approach the following research question: how the frequencies of the five popularity patterns among pages within different domains correlate with each other? For this purpose, we created a dataset of second-level domains (referred to as “small dataset”) that are represented by at least 10 web pages in our dataset of activity profiles. This dataset covers 95.3% of the whole dataset of activity profiles. We

sharp rise, light tail				sharp rise, substantial tail				soft rise, light tail				soft rise, substantial tail			
topic	u	p	s	topic	u	p	s	topic	u	p	s	topic	u	p	s
online betting	65K	89%	1,72	photoart	2,5K	71%	2,96	weather	3K	5,5%	9,5	photo album hosting	27K	5,5%	7,2
news agencies	167K	88%	1,71	amateur photo	2,4K	68%	2,85	car ads	16K	5,1%	8,9	pets	29K	5,2%	6,8
entertainment	117K	87%	1,67	knitting	3,3K	60%	2,50	photo album hosting	27K	2,3%	4,0	blog hosting	30K	5,0%	6,6
biathlon	4.1K	85%	1,66	caricatures	7,5K	59%	2,47	blog hosting	30K	2,2%	3,9	media editors	32K	4,7%	6,2

Table 1: Topics with the highest domination for different patterns of activity profiles. Here u denotes URLs count, p is the URL’s rate of the current popularity pattern, s equals to p normalized by the prior rate of the pattern (given at Tab. 3).

treat the portions of each of the five popularity patterns within a certain domain as its five features. Tab.2, B) represents the Pearson correlation matrix of the pairs of the five domain features over the small dataset. This matrix captures which pairs of patterns are collaborative and which are competitive in the context of their domain. We see, for example, the largest negative correlation of the first and the fifth patterns, which means that short-term and long-term web pages are well-separated at the level of their domains.

6. TEMPORAL PATTERN PREDICTION

Encouraged by results of the previous section, we trained several classifiers that predict web page popularity pattern by its URL. We used the following three groups of features: (1) topics of page’s second-level domain (binary ones), (2) weights of Naive Bayesian classifier trained on URL tokens, and (3) the five domain features. The features of (1) and (3) groups are described in the previous section. We trained Friedman’s gradient boosting decision tree models [2] with 10-fold cross validation on (1) the dataset of activity profiles, and (2) the small dataset by employing different combinations of feature groups. Performance of the obtained models is reported in Tab. 4.

dataset	(1)	(2)	(1)+(2)	(3)	(1)+(2)+(3)
whole	63,6	72,0	72,2	68,3	73,0
large dom.	64,1	73,1	73,4	68,5	73,7

Table 4: MAUC (%) of classifiers predicting popularity pattern of a web page by different groups of features.

The prediction performance increases if we restrict the task to large domains that provide richer information on their properties.

7. CONCLUSION

In this paper, we examined different patterns that temporal activity profiles of web pages’ traffic can follow. We proposed an algorithm to automatically classify the activity profiles that have a short-term popularity spike into five different patterns depending on how quickly they gain and lose popularity. The algorithm is subsequently used to conduct a large-scale analysis. The main contribution of this paper is the examination of the degree of dependence that

a page’s activity pattern has upon the factors that can be obtained before the actual pattern is known: (1) the distribution of topics on page’s domain, (2) tokens of page’s URL, and (3) prior distribution over popularity patterns on page’s domain. We show that a significant amount of information is indeed contained in the aforementioned factors, and build a classification algorithm that allows to predict the activity pattern a page will have with a reasonable accuracy.

8. REFERENCES

- [1] E. Bakshy, J. Hofman, W. Mason, and D. Watts. Identifying ‘influencers’ on twitter. In *WSDM*, 2011.
- [2] J. H. Friedman. Stochastic gradient boosting. *CSDA*, 38:367–378, 1999.
- [3] D. J. Hand and R. J. Till. A simple generalisation of the area under the roc curve for multiple class classification problems. *Mach. Learn.*, 45(2):171–186, oct 2001.
- [4] L. Hong, O. Dan, and B. D. Davison. Predicting popular messages in twitter. In *WWW*, pages 57–58, 2011.
- [5] J. Lehmann, B. Gonçalves, J. J. Ramasco, and C. Cattuto. Dynamical classes of collective attention in twitter. In *WWW*, pages 251–260, 2012.
- [6] M. Liu, R. Cai, M. Zhang, and L. Zhang. User browsing behavior-driven web crawling. In *CIKM*, pages 87–92, 2011.
- [7] S. Myers and J. Leskovec. Clash of the contagions: Cooperation and competition in information diffusion. In *ICDM*, pages 539–548, 2012.
- [8] K. Radinsky, K. Svore, S. Dumais, J. Teevan, A. Bocharov, and E. Horvitz. Modeling and predicting behavioral dynamics on the web. In *WWW*, pages 599–608, 2012.
- [9] A. Saha and V. Sindhwani. Learning evolving and emerging topics in social media: a dynamic nmf approach with temporal regularization. In *WSDM*, pages 693–702, 2012.
- [10] G. Szabo and B. A. Huberman. Predicting the popularity of online content. In *WWW*, pages 80–88, 2010.
- [11] J. Yang and J. Leskovec. Patterns of temporal variation in online media. In *WSDM*, pages 177–186, 2011.