What can be Found on the Web and How: A Characterization of Web Browsing Patterns

Alexey Tikhonov
Yandex
Moscow, Russia
altsoph@yandex-team.ru

Liudmila Ostroumova Prokhorenkova
Yandex
Moscow, Russia
ostroumova-la@yandex-team.ru

Arseniy Chelnokov
Yandex
Moscow, Russia
achelnokov@yandex-team.ru

Ivan Bogatyy[*]
Google
Mountain View, CA
bogatyi@gmail.com

Gleb Gusev
Yandex
Moscow, Russia
gleb57@yandex-team.ru

ABSTRACT

In this paper, we suggest a novel approach to studying user browsing behavior, i.e., the ways users get to different pages on the Web. Namely, we classified all user browsing paths leading to web pages into several *types* or *browsing patterns*. In order to define browsing patterns, we consider several important points of the browsing path: its origin, the last page before the user gets to the domain of the target page, and the target page referrer. Each point can be of several types, which leads to 56 possible patterns. The distribution of the browsing paths over these patterns forms the navigational profile of a web page.

We conducted a comprehensive large-scale study of navigational profiles of different web pages. First, we demonstrated that the navigational profile of a web page carry crucial information about the properties of this page (e.g., its popularity and age). Second, we found that the Web consists of several typical non-overlapping clusters formed by pages of similar ranges of incoming traffic. These clusters can be characterized by the functionality of their pages.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

Keywords

user browsing behavior, browsing patterns, clustering of web pages

1. INTRODUCTION

In the last years, commercial search engines play an increasingly dominating role in navigating users to the web pages of their interest. Modern search systems not only provide relevant results to user queries, but also serve users with the desired content presented in the most attractive and informative way possible. Given the ambitions of any major search engine to be a "one-stop service center" for all user needs, it is definitely important to understand the ways users find content on the Web.

Motivated by this, we came up with a novel approach to studying traffic generated by user navigation on the Web. In accordance with the previous research in this field [9], our method of user traffic analysis is based on the referrers of web page visits. Each web page visit has one or no referrers: the referrer can be either the previously visited page whose outgoing link to the target page was followed by the user or the referrer can be missing (if the user followed a bookmark, directly typed the target URL into the address bar, etc.). As in previous studies, we consider several types of referrers: *internal* (the referrer is a page located at the same second-level domain as the target page), *external* (the referrer is a page located at another second-level domain), *social* (the referrer is a page which belongs to a social network), *search* (the referrer is a search engine), and the special type of *empty referrer*.

Although, we show that studying referrers of visits to a given page is a useful way to analyze the role of this page on the Web, we also prove that the whole browsing paths that led users to those visits carry even more important information (see Section 5). For example, a search engine or a social network can affect user browsing behavior in several ways: it can directly send a user to the target page of her interest, or it can send a user to the main page of the target domain, or a user may start from a search page and then arrive at the target page only after browsing a few other pages on other web sites. In order to fully analyze user browsing paths, we focus on the following details: where those browsing paths start, how the users starting those paths get to the domain of the target page, how those users get to the target page itself (via which referrer type). So, we define browsing patterns by considering the following points of the browsing path: its

[*]This research was performed while the author worked at Yandex, Moscow, Russia.

WebSci '15, June 28 - July 01, 2015, Oxford, United Kingdom
© 2015 ACM. ISBN 978-1-4503-3672-7/15/06 ...$15.00.
DOI: http://dx.doi.org/10.1145/2786451.2786468

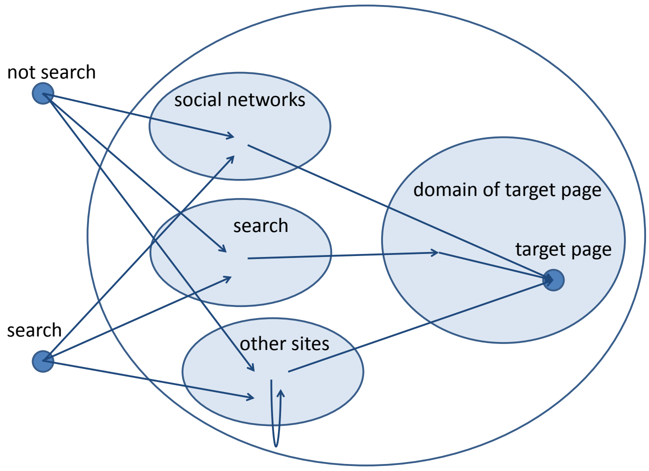

Figure 1: Browsing paths

origin (first visit), the last page before the user gets to the domain of the target page, and the target page referrer.

Finally, we discover several browsing patterns (see Figure 1 for an illustration of some of them). We conducted a large-scale study of these patterns and found that the Web consists of several typical clusters formed by pages of similar types of incoming traffic; for the sake of simplicity we call them *unsearchable, search engine services, entertainment, news, social networks, file hostings*. To sum up, the contributions of this paper are the following.

- We suggest a new method for analyzing user browsing behavior. This method takes into account the following steps of browsing: the starting point, the way of getting to the target domain, and the final transition to the target page.
- We demonstrate that the browsing patterns carry more important information than the types of referrers only. Namely, we train two models which predict several simplest yet important properties of web pages based either on the types of referrers or on the browsing patterns. The model which uses browsing patterns shows much better results.
- Based on the distribution of browsing patterns, all pages can be grouped into several clusters. It turns out that the pages of one cluster are similar to each other in terms of their functionalities.

We envision several applications of our research. For instance, it was previously observed that a large portion of Web traffic is generated without any help of search systems. For example, [14] shows that only 6.8% of browsing sessions start from a search engine. In order to adopt more traffic (which directly translates into profits for advertising-funded web services) and to help users to complete the tasks faster, it is very important for search engines to predict or, at least, to be aware of the final search goals of its users when developing new SERP functionalities and expanding the range of supported search tasks. Our methods can also be useful for social networks which aim to hold users inside as long as possible: the understanding of user needs can be helpful for their purposes.

The rest of the paper is organized as follows. In the next section, we describe the previous research on browsing behavior analysis on the Web. Then, in Section 3, we describe our dataset and, in Section 4, we present our classification of browsing paths. The major part of our analysis is presented in Sections 5 and 6. First we present some preliminary experiments and then cluster pages according to their incoming browsing paths. Section 7 concludes the paper and outlines the directions of future research.

2. RELATED WORK

In this section, we give an overview of some papers relevant to our research. First, we describe previous studies on the analysis of user browsing behavior and the influence of search engines and other sites on this behavior. Second, we discuss several papers on various applications of user browsing sessions. These studies serve as a motivation for our research. Finally, we describe some papers, where the methods similar to ours are used.

Note that some of the papers mentioned in this section are written about ten years ago. We expect that the user behavior on the Web could change in the past years, therefore the conclusions made below on the basis of that papers can be partly outdated. This paper closes this gap by presenting up-to-date analysis of user browsing behavior.

2.1 Browsing behavior and search engines

User browsing behavior on the Web and the influence of search engines on that behavior is analyzed in [14]. It is shown that search engines influence about 13.6% of the users' Web traffic. The influenced traffic includes visits to search engine home pages, SERPs, and all their descendants, i.e., all pages which users visit during browsing sessions started at SERPs. For an average user, one fifth of all the sites visited by the user are visited only starting from search engines. Different characteristics of browsing sessions are also presented: the distribution of session length (i.e., the time spent), the average time spent on a page, the number of pages per session, the number of unique sites per session, etc.

The authors of [4] show that the wide-spread use of search engines biases the overall traffic toward popular sites. They estimate the impact of search engines on the popularity evolution of web pages. The conclusion is that when search engines rank pages based on their popularity, it takes much more time for a new page to become popular even if the page is of high quality. Based on these results, we assume that search engines are able to also affect user habits by developing new ranking methods and providing new functionalities on their SERPs that increase the variety of tasks that can be solved with a search engine. We believe that a search engine is able to fight for more user traffic that is concerned with some needs that users currently satisfy in different ways without search engines.

On the contrary, [1] argues that the use of search engines actually has an egalitarian effect. New web sites have a greater chance of being discovered as long as they are about specific topics that match the interests of users as expressed through their search queries.

In [6], browsing paths are studied from the individual's point of view. The authors analyze and compare the behavior of different demographic groups. On the contrary, in the current study, we are focused on the properties of pages and the browsing paths leading to them. User browsing habits were also studied in [8]. It was shown that users tend to prefer some domains and may click on search results representing these domains even if some other result on SERP are more relevant.

2.2 Browsing sessions and their applications

Search browsing trails and their importance for search engines is also studied [2, 3, 18, 20]. It is shown in [20] that, according to various metrics, the pages appeared in search trails are useful for search engines. In [3] search browsing trails leading to web pages are used in order to improve ranking of the documents on SERP.

User browsing behavior is successfully used for constructing new models of relevance and authority [12]. Therefore, studying the relation between user browsing patterns and visited web pages is important for extracting more precise and complete information about the utility of these web pages. For example, it is demonstrated in [21] that the information about where a web page is located, on average, in user browsing sessions (i.e., at the beginning, at the end, or in the middle) allows to improve the state-of-the-art BrowseRank.

It is shown in [11, 13] that the popularity of a web page as well as the dynamics of this popularity essentially depend on and can be predicted on the basis of the location (URL) of this page. Based on these results, we assume that the types of browsing patterns leading to a page are also substantially determined by its location on the Web. Therefore, studying user browsing patterns can help to understand the role of sites and web pages on the Web and, vice versa, the patterns of the browsing paths leading to new web pages can be predicted based on their location. These considerations were a partial motivation for this study.

2.3 Studies with analysis/methods similar to ours

In [9], all pages are divided into Content (news, portals, games, verticals, multimedia), Communication (email, social networking, forums, blogs, chat), and Search (Web search, item search, multimedia search). The authors also consider different referrer types in order to analyze the way users move from page to page, within and across second-level domains, and within and across page types, and analyzed how search interacts with other types of navigation. In the current paper, we broaden the analysis performed in [9] in several directions. First, our aim is to study navigational profiles of individual web pages, while [9] study the statistics aggregated over types of referrers. Second, in addition to the types of referrers, we consider whole browsing paths and their types. We demonstrate that the types of browsing patterns carry more important information than the types of referrers (see Section 5). At last, we identified several typical clusters formed by pages of similar ranges of incoming traffic.

The authors of [19] analyze the correlation between *who* searches, *what* she searches, and *how* she searches. They cluster all users of a search engine based on *what* they search and analyzed the obtained clusters. The analysis we perform in this paper is similar in spirit, but instead of analyzing and clustering users we analyze and cluster web pages.

3. DATA AND TYPES OF USER VISITS

We consider all URLs stored in the log of the browser toolbar of the most popular search engine in Russia *yandex.ru* (60% market share[1]) during the period of three months: from August 1, 2013 to October 31, 2013. From all the obtained URLs we remove ones visited by less than 20 different users, because we want to avoid analyzing URLs that are designed for individual users such as private pages (e.g., personal e-mail pages, which can be visited from several devices with different IPs). The remaining 175M pages, which form our data set under study, altogether attract 65% of the overall traffic (i.e., 65% of web page visits are on these pages).

As the goal of the paper is discovery and analysis of different patterns of browsing traffic, we start with a description of basic properties of pages and their visits. First of all, we identify two special types of pages, which play a special role in browsing behavior: *search* type and *social* type. A page is regarded as a *social* page, if it belongs to one of the 22 most popular social networks in Russia. Social networks play a very important role in the Web navigation nowadays, since users exhibit a specific browsing behavior there [10]. Search engines are self-contained and they tend to hold users inside as long as possible, this may lead, for instance, to many internal transitions. We also define pages of *search* type, since we are especially interested in studying the influence of search engines on user browsing. We fix three most popular search engines in the country: *yandex.ru, google.ru*, and *mail.ru* (these search engines cover 97% of Russian search traffic). We classify SERPs of these search engines as pages of *search* type.

For each URL, we analyze user browsing paths passing through it as we describe in the next section. In order to do that, we first extract referral trees as it is described in [9]. First, we define browsing sessions as sequences of web page visits for a particular user, where the time passed between any consecutive web page visits is less than 30 minutes — a widely used threshold defining search or browsing sessions [17]. Then, for each user and for each browsing session of this user, we define *a referral forest* (i.e., a group of *referral trees*) in the following way. Each entry in the browsing logs describing a particular user visit is a triple ⟨*timestamp, referrer, target page*⟩. We draw a directed edge from an entry $\langle t_1, ref_1, targ_1\rangle$ to another entry $\langle t_2, ref_2, targ_2\rangle$ of the same session iff the following conditions hold:

1. $targ_1 = ref_2$,

[1] *http://company.yandex.com/*

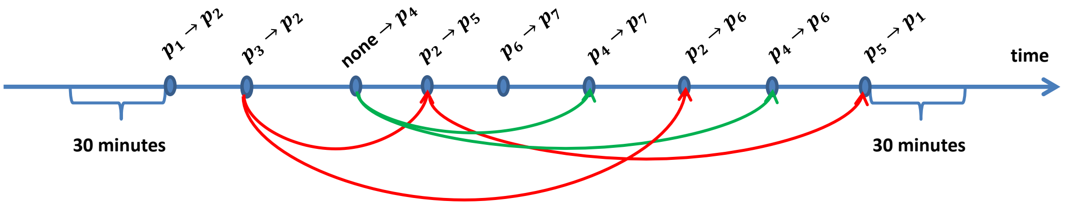

Figure 2: Example of browsing session and its referral forest. Arrows connect consecutive transitions.

Algorithm 1: Determining the type of the referrer

```
if $ref = none$ then
    type = RefNone;
else
    if $ref \in$ a social network then
        if $ref, targ \in$ the same domain then
            type = RefSocialInt;
        else
            type = RefSocialExt
    else
        if $ref \in search$ then
            type = RefSearch;
        else
            if $ref, targ \in$ the same domain then
                if $ref =$ main page then
                    type = RefMainPage;
                else
                    type = RefInt;
            else
                type = RefExt;
```

2. $t_1 < t_2$,

3. there is no entry $\langle t, ref, targ \rangle$ such that $targ = ref_2$ and $t_1 < t < t_2$.

After this procedure, all visits of a user within one browsing session are divided into several referral trees. Figure 2 presents an example of one browsing session consisting of 4 referral trees: the first one consists of green arrows, the second one is formed by red arrows, and the remaining two consist of only one web page visit each ($p_1 \to p_2$ and $p_6 \to p_7$). Note that we have $ref = none$ in a triple $\langle t, ref, targ \rangle$, if the referral page is not defined. In particular, we have no referrer, if a user used bookmarks, directly typed the URL of *targ* into the address bar, etc.

The first triple of a referral tree is called its *origin*. The origin has *none* type, if the triple describes a user visit made via a direct request of the target URL to the browser: URL is typed in the address bar, clicked from a bookmark, clicked from another page visited in the previous sessions, etc. Alternatively, a user may not request the target URL, if the session starts with a search query usually typed in the browser address bar. The second scenario reflects the search intent of a user. In order to fully analyze this intent, we say that the origin of a referral tree is of *search* type, if either the referrer of the origin is SERP, or its target is SERP.

For each user visit to a URL (i.e., for each triple in the browsing logs, where the target page is equal to this URL), we define the *browsing path* as the sequence of web page visits that form the branch of the corresponding referral tree leading from its origin to the URL. For example, on Figure 2, we have one path for p_1: $p_2 \to p_5 \to p_1$ and one path for p_5: $p_2 \to p_5$.

For every URL visit, we collect several *features* of its browsing path defined above. We are interested in where the path started, how the target domain was found, and how the target page itself was reached. Therefore, we define the following features of a path:

- type of the target page's referrer,
- type of the target domain referrer, i.e., the last page before the user finally entered the domain containing the target page,
- type of the path's origin.

We have already discussed the types of origins in this section. In the next section, we formally define the types of referrers and the types of domain referrers which we use to describe browsing paths.

4. DESCRIPTION OF BROWSING PATHS

We analyze referrers of URL visits in a way similar to [9]. Specifically, we use 7 types of URL referrers: RefNone, RefSocialInt, RefSocialExt, RefSearch, RefMainPage, RefInt, and RefExt. The description of these types can be found in Table 1 (also see Algorithm 1 for determining the type of referrer). As we describe in the previous section, we distinguish the pages of social networks and SERPs. We are also interested in whether the visit is from the page on the same second-level domain (internal) or not (external). And, if the visit is from the same second-level domain, then we check, whether the visit is directly from the main page of that domain or not.

We argue that, in order to fully analyze user browsing behavior, it is not enough to consider only URL referrers. It is important to understand how user started the path and which pages this path passed through. Therefore, we broaden the study from [9] by considering other features of a browsing path. To this end, we consider 4 types of the domain referrer, i.e., the last page before the user reaches the domain of the target page: DomNone, DomSocial, DomSearch, DomOther (see Table 1 for the description). As we discussed

Type of target ref	**description**
RefNone	no referrer
RefSocialInt	referrer and target belong to one domain which is a social network
RefSocialExt	referrer belongs to another domain which is a social network
RefSearch	referrer is a search engine (its main page or its SERP)
RefMainPage	referrer is a main page of the same domain, which is not a search engine, nor a social network
RefInt	referrer belongs to the same domain, but not *social*, *search*, or *main page*
RefExt	referrer belongs to another domain, but not *social* or *search*

Type of domain ref	**description**
DomNone	no domain referrer
DomSocial	domain referrer is a social network
DomSearch	domain referrer is a search engine
DomOther	domain referrer is some external page, but not of the type *search* or *social*

Type of origin	**description**
OriginSearch	referrer of the first visit of the path is *search*, or it is none and the target is *search*
OriginNone	not search

Table 1: Description of browsing path features.

in the previous section, we also consider two types of origin: OriginNone and OriginSearch (see Table 1). Finally, we have 56 types of paths (*browsing patterns*), which correspond to $7 \times 4 \times 2 = 56$ possible value combinations of three features: type of URL referrer, type of domain referrer, and type of origin. Note that some combinations of features are impossible, others can be rare. In our dataset 36 browsing patterns are presented. We argue that these types of paths leading to a page and their frequencies provide the opportunity to characterize the role of this page on the Web.

It turns out that the most frequent browsing patterns in our dataset are

- (OriginNone,DomNone,RefInt), i.e., the path starts after more than a half an hour of inactivity or with the empty referrer and a user reaches the target page through several transitions between different pages of the same domain. This browsing pattern covers a half of all visits.
- (OriginNone,DomOther,RefExt), i.e., a user reaches the target page through several transitions between different pages located on some other domains. This pattern covers 15% of all visits.
- (OriginNone,DomNone,RefSocialInt), i.e., a user get to the target page through several transitions between different pages of the same domain which is a social network. This browsing pattern reflects a typical use of a social network. This pattern covers 11% of all visits.

We noticed that the top 18 browsing patterns cover about 99% of all user visits. Therefore, we further merge the remaining patterns into one pattern *others* and get 19 types of browsing paths. These types are used in Sections 5 and 6 in order to form a navigational profile of a web page: each page is described by 19 numbers, each number is the fraction of incoming paths of the certain type.

5. PRELIMINARY EXPERIMENTS

The goal of this section is to demonstrate that 1) the types of browsing paths leading to a page carry crucial information about its properties related to the ways users utilize this page, the possible aims and context; 2) browsing patterns carry more valuable information than the types of referrers. To this end, we consider the following simplest yet important properties of a URL:

- **Popularity.** Popularity of a page measured as the total number of user visits during the considered time interval.
- **Number of "/".** Site-level distance from the main page measured using the number of "/" in the URL, zero for the main page. This characteristic shows how deeply the page is placed on its second-level domain and obviously correlates with the role of this page on the Web: main pages are usually hubs, they contain links to the content pages, while the deeper the page, the greater the probability that it is a content page, an image, a file, etc. Our assumption is that the ways users get to a page depend on how far this page is from the main page.
- **Age.** Current age of a page, i.e., the difference between the last day of the considered period (October 31, 2013) and the day the URL was discovered by the search engine (which intensively crawls and re-crawls hundreds of millions of web-sites on a daily basis). It is interesting to examine whether users use different ways to get to older and younger pages.
- **Size of domain**, i.e., the number of documents (in the search index) which belong to the domain of the URL. This characteristic shows whether the page belongs to a large domain or not.

Note that all the above properties are quantitative. So, we can try to predict them using the distribution of user visits over the fixed set of browsing patterns for each page. We train a proprietary implementation of Friedman's gradient boosted decision tree-based machine learning algorithm [5] with the features described later in this section. From the set of all URLs in our dataset (see Section 3) we sample a random subset of 500K URLs. Half of the pages forms the training set, the others form the test set. As it is usually

	F_{ref}		$F_{pattern}$		F_{all}	
Log popularity	RefNone	36%	(OriginNone,DomNone,RefNone)	18%	RefNone	20%
	RefMainPage	19%	(OriginNone,DomSearch,RefInt)	14%	(OriginNone,DomSearch,RefInt)	12%
	RefExt	16%	Other	13%	Other	12%
Number of "/"	RefInt	41%	(OriginNone,DomNone,RefInt)	29%	RefInt	32%
	RefMainPage	18%	(OriginNone,DomNone,RefMainPage)	14%	RefMainPage	10%
	RefSocial	11%	(OriginNone,DomNone,RefSocialInt)	12%	(OriginSearch,DomSearch,RefInt)	7%
Age	RefSearch	25%	(OriginNone,DomNone,RefNone)	14%	(OriginNone,DomSearch,RefSearch)	11%
	RefNone	25%	(OriginSearch,DomSearch,RefSearch)	13%	(OriginSearch,DomSearch,RefInt)	10%
	RefExt	15%	(OriginNone,DomSearch,RefSearch)	12%	(OriginNone,DomNone,RefNone)	9%
Log size of domain	RefMainPage	28%	(OriginSearch,DomSearch,RefInt)	14%	(OriginSearch,DomSearch,RefInt)	13%
	RefNone	27%	(OriginNone, DomNone, RefInt)	14%	(OriginNone,DomNone,RefInt)	11%
	RefInt	13%	(OriginNone,DomOther,RefInt)	10%	(OriginNone,DomOther,RefInt)	10%

Table 2: The most important features and their weights for each property and each set of features.

Property	F_{ref}	$F_{pattern}$	F_{all}
Log popularity	0.522	0.737	0.752
Number of "/"	0.264	0.295	0.297
Age	0.072	0.111	0.111
Log size of domain	0.173	0.297	0.305

Table 3: Coefficient of determination R^2 for three groups of features.

done for heavy-tailed distributions, we predict the logarithm of popularity and domain size instead of their initial values by minimizing mean squared error on the training set. For other properties we minimize mean squared error for non-transformed values.

We compare three groups of features.

1. As we described in Section 4, we have 19 browsing patterns. So, each page has 19 features $p_1, \ldots, p_{19}$, where p_i is the fraction of incoming paths of type i for this page. We denote the set of these features $F_{pattern}$.
2. We have 7 types of URL referrers, so the second group of features F_{ref} consists of features $r_1, \ldots, r_7$, where r_i is the fraction of URL referrers of type i among all visits to this page.
3. Finally, we join the above two sets and obtain the set of all (26) features F_{all}.

Table 3 reports the quality of predictions in terms of the coefficient of determination [16]:

$$R^2 = 1 - \frac{\sum_i (\hat{p}_i - p_i)^2}{\sum_i (p_i - \bar{p})^2},$$

where p_i is the value of a property, $\hat{p}_i$ is its estimation, and $\bar{p}$ is the average value of this property. The closer R^2 to 1, the better the model fits our data. Note that values $R^2 > 0$ mean that the model provides a better prediction than the mean value of the data.

As it can be seen from Table 3, all the properties can be better predicted by $F_{pattern}$ than by F_{ref}. Moreover, the prediction by $F_{pattern}$ is almost as good as by F_{all}, since the information about web page referrers is implicitly included into the information about the browsing patterns.

Note that the quality of prediction differs significantly for different properties. For example, the logarithm of popularity can be predicted much better than the other characteristics of a web page. It is encouraging, since the popularity is probably the most important characteristic of a web page from both users' and search engines' points of view. For instance, the popularity prediction is successfully used in web crawling [11, 13]. The fact that the popularity is correlated with the user browsing behavior is not surprising. For example, the popular pages are more likely to be placed at higher positions on SERPs, they are more likely to be typed directly in the browser address bar, etc. Also, when we predict the logarithm of popularity, then $F_{pattern}$ outperforms F_{ref} by a large margin.

On the contrary, if follows from Table 3 that the age of a page is the most difficult property to predict. The possible reason for this is that the age is not strongly connected with some specific user browsing habits. Note that, in this case, the set of features F_{all} provides the same quality as $F_{pattern}$ do solely.

We also analyzed the importance of different features. In Table 2, we present the top three features (for each property and for each set of features) according to the weighted contribution into our prediction model (see Section 10.13 in [7] for the description of those weights). The contribution measures weighted improvement of the loss function over all employments of a feature during the learning process.

We observe some expected results: e.g., for the prediction of the site-level distance from the main page (number of "/"), the most useful browsing paths are the ones including internal pages or main pages. Indeed, a page is close to the main page, if it is referred by the main page. On the

	pages	visits	avg. visits per page	avg. numb of "/"	avg. age of pages in days	avg size of domain	avg/median length of path
Cluster 1	44%	37%	135	2.5	278	9.0 M	10.0 / 6
Cluster 2	8%	6%	117	1.2	165	4.2 M	11.5 / 3
Cluster 3	10%	17%	268	1.8	322	3.8 M	5.6 / 3
Cluster 4	21%	24%	176	2.1	393	2.5 M	5.3 / 3
Cluster 5	12%	9%	125	1.1	372	0.4 M	16.9 / 3
Cluster 6	5%	7%	210	1.7	300	4.7 M	5.8 / 2

Table 4: Quantitative characteristics of clusters.

	44%	8%	10%	21%	12%	5%
44%	**0.01**	1.17	0.42	0.46	1.21	0.86
7%	1.14	**0.08**	0.95	0.90	1.19	0.88
11%	0.67	0.98	**0.25**	0.46	1.05	0.66
19%	0.43	0.96	0.30	**0.04**	1.01	0.53
13%	1.22	1.22	1.07	1.00	**0.00**	0.91
5%	0.88	0.93	0.68	0.52	0.91	**0.02**

Table 5: Sizes of clusters and the Euclidean distance between their centroids for clusterings of two random samples of 500K pages.

contrary, a page is probably far from the main page, if it is reached after some internal browsing.

Another interesting observation is that, among the set F_{all}, the most important features often belong to the set $F_{pattern}$ (see, e.g., the rows *Age* and *Log size of domain* in Table 2).

Our main conclusion is that the set of features $F_{pattern}$ not only gives additional information when we combine it with the set F_{ref}, but $F_{pattern}$ can often be more useful than F_{ref}. In some cases, the information contained in navigational profiles almost fully covers the information contained in the distribution of referrers (see Table 3, the rows *Number of "/"* and *Age*).

The fact that the navigational profiles carry valuable information about various properties of web pages motivated us to perform deeper analysis of these profiles. Therefore, in the next section, we try to distinguish different types of pages in the aspect of the ways users navigate to them. We do this by clustering the pages according to their navigational profiles.

6. CLUSTERING

In the previous section, we demonstrated that the types of browsing paths leading to a page can be used to predict important properties of this page, which characterize its role on the Web. In this section, we aim to distinguish different types of pages in the aspect of the ways users navigate to them. To this end, we cluster all pages according to the distribution of browsing patterns.

As we discussed before, each page is represented by a 19-dimensional vector $v = (p_1, \dots, p_{19})$, where p_i is the fraction of incoming paths of type i leading to this page. We sampled 500K pages uniformly at random from our dataset and clustered the set of vectors representing them using expectation-maximization algorithm[2]. The number of clusters was defined by cross-validation[3]: the data was partitioned into 10 parts uniformly at random, each of the parts was then left aside in turn as a test set, a clustering model was trained on the other 9 parts, and the value of the log-likelihood was calculated for the test set. The average of the obtained 10 log-likelihood values was used as the clustering performance to optimize the number of clusters.

The described procedure gave 6 clusters. In order to verify that our clustering procedure is stable, we performed the same procedure for another random sample of 500K pages. Again, we obtained 6 clusters with similar sizes, centroids, and properties (see Table 5). We also compared the expectation-maximization clustering algorithm with k-means algorithm (with 6 clusters) on one sample of 500K pages. With k-means we also obtained very similar clusters. The Rand index [15], which measures how similar are the obtained clusters, is equal to 0.81 in our case.

Table 4 presents the quantitative characteristics of the obtained clusters: their sizes (percentage of unique URLs and user visits to them), average properties of URLs in a cluster (popularity, age, site-level distance from the main page, size of the corresponding domain), and average lengths of the paths leading to pages in each cluster. In the next section, we describe the obtained clusters and discuss the results from Table 4 in detail.

6.1 Description of clusters

In this section, we describe the obtained clusters. We first look at several characteristics of the clusters and, based on them, we give an interpretation of the obtained clusters. The characteristics we consider are the following:

- Typical domains of URLs in the clusters;
- Typical features of the browsing paths leading to the pages from the cluster;
- Typical browsing patterns.

Let us first explain what we mean by *typical*. For example, if we want to find the typical browsing patterns for a cluster

[2] *weka.sourceforge.net/doc.dev/weka/clusterers/EM.html*

[3] *http://www.statsoft.com/Textbook/Cluster-Analysis#vfold*

	typical domains	typical features	typical browsing patterns
Cluster 1	mobile applications sites about cars online dating	RefInt – 2.5	(OriginNone,DomNone,RefInt) – 2.9 (OriginNone,DomSearch,RefInt) – 2.3 (OriginSearch,DomSearch,RefInt) – 2.2
Cluster 2	subdomains of search domains (translate.google.com, realty.yandex.ru, etc.)	RefSearch – 17.5 OriginSearch – 6.8	(OriginSearch,DomSearch,RefSearch) – 246 (OriginNone,DomSearch,RefSearch) – 67.6 (OriginSearch,DomSearch,RefInt) – 28.2
Cluster 3	forums entertainment adult content	RefMainPage – 10.6	(OriginSearch,DomSearch,RefMainPage) – 11.6 (OriginNone,DomNone,RefMainPage) – 11.0 (OriginNone,DomSearch,RefMainPage) – 10.2
Cluster 4	news sites entertainment	RefNone – 4.7 DomSearch – 2.3 DomExt – 2.1	(OriginSearch,DomSearch,RefSearch) – 14.6 (OriginNone,DomSearch,RefSearch) – 5.4 (OriginSearch,DomExt,RefInt) – 4.5
Cluster 5	social networks	RefSocialInt – 44.2	(OriginSearch,DomSearch,RefSocialInt) – 87.8 (OriginNone,DomNone,RefSocialInt) – 47.0
Cluster 6	file hostings social networks	RefSocialExt – 35.2 RefExt – 28.7 DomSocial – 10.2	(OriginNone,DomSocial,RefInt) – 38.0 (OriginNone,DomSocial,RefSocialExt) – 35.3 (OriginNone,DomExt,RefExt) – 31.7

Table 6: Typical domains, features, and patterns of clusters and their relative frequencies.

C, then, for each browsing pattern P, we compute the value

$$F_C(P) = \frac{N_{P,C}}{N_C} \cdot \frac{N - N_C}{N_P - N_{P,C}} \,,$$

where N is the total number of user visits to the 500K pages in the dataset, N_P is the number of user visits with the browsing pattern P, N_C is the size of C, $N_{P,C}$ is the the number of user visits to pages in C with the browsing pattern P. In other words, $F_C(P)$ reflects how frequent P is in the cluster C in comparison with its frequency outside C (further we call this value the *relative frequency*). The domains/patterns/features of the largest relative frequencies are called *typical* for the cluster C.

The most typical domains, features, and browsing patterns, as well as their relative frequencies, can be found in Table 6. Note that we do not present the names of the typical domains themselves, we present only their description. The reason is that these domains are mostly in Russian, therefore, we manually analyzed the most typical 25 domains and summarized our observations in Table 6.

Now let us present the description of the obtained clusters. We give a name to each cluster which corresponds to the most typical pages in the cluster, although this does not mean that all the pages correspond to this name.

Cluster 1: Unsearchable.
This is the largest cluster in our dataset (44% of URLs). As we show in Section 6.2, users do not use search engines (RefSearch) to find these pages.

At first sight, it is hard to say what ties all the typical domains of the cluster together: we observe a lot of mobile sites, especially games, online dating sites, etc. In order to understand this cluster better, we manually looked at random pages belonging to it and found out that this cluster mostly consists of the pages which are useless for search engines, i.e., they can never be reached directly from SERPs. There are several reasons for this: some pages can be reached only after logging in (e.g., mobile services), others can be reached only by internal search on the web sites, also, for instance, there are a lot of unpopular youtube videos which have no textual description and therefore cannot be indexed by search engines, etc. The main pages and other searchable pages of these domains fell into other clusters.

This cluster consists of the pages which belong to large domains and are located far from the main page. The paths leading to them are usually quite long. The most typical feature in this cluster is RefInt. Typically, users start their browsing path from bookmarks, search, or direct typing of the URL, they immediately get to the target domain and surf there before they reach the target page.

Cluster 2: Search engine services.
Analysis of typical domains shows that this cluster mostly consists of subdomains of search domains, e.g., *disk.yandex.ru, docs.google.com, translate.google.com, realty.yandex.ru,* etc. The most typical features for this cluster are RefSearch and OriginSearch. Table 4 shows that these pages are usually located close to the main page.

Cluster 3: Entertainment.
This cluster mostly consists of entertaining pages, adult content, and forum pages. The main distinctive feature of this cluster is that it gathers almost all traffic from main pages. So, the typical scenario for these pages is the following: a user somehow gets to the main page of the target domain and then clicks on some interesting URLs presented on this page (news, images, posts, etc.). Also, according to Table 4, these pages are usually popular.

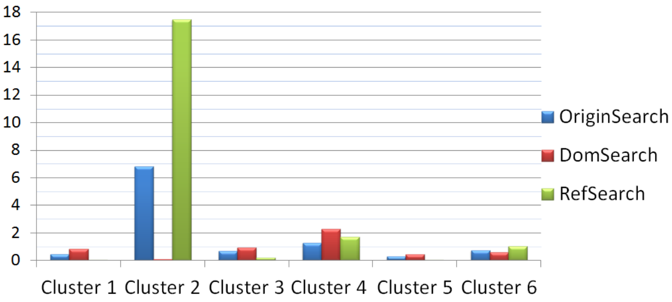

Figure 3: The relative frequencies of search features in a cluster in comparison with the frequencies outside the cluster.

Feature	Conditional entropy
RefSocialInt	0.662
RefInt	0.668
RefSearch	0.676
RefExt	0.677
RefMainPage	0.679
DomOther	0.679
OriginSearch	0.680
RefNone	0.683
OriginNone	0.683
DomNone	0.684
DomSocial	0.684
RefSocialExt	0.684
DomSearch	0.685
Entropy of clusters	0.686

Table 7: Conditional entropy of cluster id given a feature

Cluster 4: News sites.
This cluster mostly consists of the pages belonging to news domains and it also includes entertaining pages. It differs from the previous cluster by the typical ways users reach the pages: they usually start from a search engine, then either get to the target page itself or get to the target domain and surf inside it.

Cluster 5: Social networks.
The typical domains of this cluster are social networks. The most prominent feature of browsing paths is RefSocialInt and browsing patterns are either started from the search or without a referrer. The pages of this cluster are usually close to the main page (for example, users' social network profile pages). On the other hand, some of them have very large paths (16.9 transitions on average), what can be explained by the specificity of social networks, which usually present small pieces of data at different pages and tend to hold users inside as long as possible. This cluster contains almost all social internal traffic.

Cluster 6: File hostings.
This cluster mostly consists of file hostings, although some social networks are also presented. The reason is that music, images, and videos which are stored on file hostings or video/photo sharing platforms are often shared through social networks. Therefore, the paths leading to the pages on file hostings are similar to the paths leading to some social network pages. Let us look closer at the typical browsing patterns of this cluster (see Table 6). There are three typical patterns:

- (OriginNone,DomSocial,RefInt) — users get to the target domain from a social network and then make several internal transitions,
- (OriginNone,DomSocial,RefSocialExt) — a link to the target page is found on a social network page (for example, *instagram.com* is the typical domain for this cluster),
- (OriginNone,DomExt,RefExt) — a link to the target page is found on some external page which is not a social network.

6.2 Analysis of search traffic

In this section, we provide a deeper analysis of how search features are distributed over the clusters. Figure 3 shows the relative frequencies (see Section 6.1) of each search feature (OriginSearch, DomSearch, and RefSearch) for different clusters. For example, in Cluster 2 the feature RefSearch is 17.5 times more frequent than outside this cluster. Note that any value greater than 1 means that the feature is more frequent inside the cluster than outside it. So, typically, users use search to find news sites (Cluster 4), or, obviously, search engine services (Cluster 2). Also, RefSearch very rarely appears in Clusters 1, 3, and 5: Cluster 1 consists of pages useless for search engines, pages of Cluster 3 are usually reached from the main page, Cluster 5 presents the social internal surfing.

6.3 Analysis of features

In order to understand which features of browsing paths play the most important role for clustering pages they lead to, we compute the conditional entropy of clusters given a feature. Namely, for each feature, we consider two variables: the cluster id which takes 6 values and "feature" which takes the values 0 and 1. Then we compute the entropy of the variable cluster id conditioned on the variable "feature". The obtained value quantifies the amount of information needed to describe the cluster id given that the presence or absence of the feature is known. The conditional entropy is always less than or equal to the entropy of the cluster id. The smaller the obtained value for some feature, the more information this feature gives about the cluster.

The conditional entropies can be found in Table 7. Several observations can be made based on these results. First, as one might expect, the features corresponding to the referrer type are the most important for the clustering procedure. Indeed, the immediate referrers of a page better characterize the role of this page on the Web than some previous steps of the corresponding browsing sessions. On the other hand, for any individual feature, the conditional entropy is quite close to the entropy of the cluster id. This means that the presence of an individual feature gives only a small amount of information about the cluster id, and the whole set of features is needed.

7. CONCLUSION AND FUTURE WORK

In this paper, we suggested a new method of analyzing user browsing behavior. In particular, we suggested a classification of user browsing paths leading to a page, which takes into account the following steps of the browsing: the starting point, the way of getting to the target domain, and the final transition to the target page. We demonstrated that the browsing patterns carry more valuable information than the types of referrers only. In particular, we found that the distribution of the incoming browsing patterns characterizes basic properties of the page such as popularity and age. In addition, based on the distribution of the incoming browsing patterns, the pages can be clustered into six groups which we informally call *unsearchable, search engine services, entertainment, news, social networks, file hostings.* We believe that the new approach can be useful for search engines: for instance, it is probably useless for search engines to adopt more traffic in Cluster 1 and they should not probably care about the pages in this cluster, while it can be very reasonable to fight for traffic in Cluster 3.

We see the following promising direction for future research. In this paper, we analyzed the patterns of paths which lead to web pages, i.e., we analyzed the pages visited *before* the user reached the target page. In order to deeper analyze the role of a page on the Web, it is also interesting to study which pages where visited *after* this page.

Acknowledgements. Special thanks to Pavel Serdyukov for his careful reading and useful comments.

8. REFERENCES

[1] R. Baeza-Yates, A. P. Jr, and N. Ziviani. The evolution of web content and search engines. In *Proceedings of the 8th ACM Workshop on Web Mining and Web Usage Analysis*, 2008.

[2] P. Bailey, R. W. White, H. Liu, and G. Kumaran. Mining historic query trails to label long and rare search engine queries. In *ACM Transactions on the Web*, volume 4 (4), 2010.

[3] M. Bilenko and R. W. White. Mining the search trails of surfing crowds: identifying relevant websites from user activity. In *Proceedings of the 17th international conference on World Wide Web*, pages 51–60, 2008.

[4] J. Cho and S. Roy. Impact of search engines on page popularity. In *Proceedings of the 13th international conference on World Wide Web*, pages 20–29, 2004.

[5] J. H. Friedman. Stochastic gradient boosting. In *Comput. Stat. Data Anal.*, volume 38(4), pages 367–378, 2002.

[6] S. Goel, J. M. Hofman, and M. I. Sirer. Who does what on the web: A large-scale study of browsing behavior. In *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media*, 2012.

[7] T. Hastie, R. Tibshirani, and J. H. Friedman. The elements of statistical learning: data mining, inference, and prediction: with 200 full-color illustrations. *New York: Springer-Verlag*, 2001.

[8] S. Ieong, N. Mishra, E. Sadikov, and L. Zhang. Domain bias in web search. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 413–422, 2012.

[9] R. Kumar and A. Tomkins. A characterization of online browsing behavior. In *Proceedings of the 19th international conference on World wide web*, pages 561–570, 2010.

[10] J. Leskovec, L. Backstrom, R. Kumar, and A. Tomkins. Microscopic evolution of social networks. *Proceedings of the 14th ACM SIGKDD international conference on Knowledge Discovery and Data mining*, pages 462–470, 2008.

[11] M. Liu, R. Cai, M. Zhang, and L. Zhang. User browsing behavior-driven web crawling. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 87–92, 2011.

[12] Y. Liu, B. Gao, T.-Y. Liu, Y. Zhang, Z. Ma, S. He, and H. Li. Browserank: letting web users vote for page importance. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 451–458, 2008.

[13] L. Ostroumova, I. Bogatyy, A. Chelnokov, A. Tikhonov, and G. Gusev. Crawling policies based on web page popularity prediction. In *Advances in Information Retrieval, Lecture Notes in Computer Science, vol. 8416*, pages 100–111, 2014.

[14] F. Qiu, Z. Liu, and J. Cho. Analysis of user web traffic with a focus on search activities. In *WebDB*, pages 103–108, 2005.

[15] W. M. Rand. Objective criteria for the evaluation of clustering methods. In *Journal of the American Statistical Association*, volume 66(336), pages 846–850, 1971.

[16] C. R. Rao. Linear statistical inference and its applications. *Wiley, New York*, 1973.

[17] A. Spink, M. Park, B. J. Jansen, , and J. Pedersen. Multitasking during web search sessions. In *Information Processing and Management*, volume 42(1), pages 264–475, 2006.

[18] A. Tolstikov, M. Shakhray, G. Gusev, and P. Serdyukov. Through-the-looking glass: utilizing rich post-search trail statistics for web search. In *Proceedings of the 22nd ACM International Conference on Information and Knowledge Management*, pages 1897–1900, 2013.

[19] I. Weber and A. Jaimes. Who uses web search for what: and how. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 15–24, 2011.

[20] R. W. White and J. Huang. Assessing the scenic route: measuring the value of search trails in web logs. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 587–594, 2010.

[21] M. Zhukovskiy, A. Khropov, G. Gusev, and P. Serdyukov. Introducing search behavior into browsing based models of page's importance. In *Proceedings of the 22nd international conference on World Wide Web companion*, pages 129–130, 2013.