Cold-start News Recommendation with Domain-dependent Browse Graph

Michele Trevisiol[1,2] Luca Maria Aiello[1] Rossano Schifanella[3,*] Alejandro Jaimes[1]

[1]Yahoo Labs, Barcelona, Spain {laiello,ajaimes}@yahoo-inc.com
[2]Universitat Pompeu Fabra, Barcelona, Spain {trevisiol}@acm.org
[3]Università degli Studi di Torino, Torino, Italy {schifane}@di.unito.it

ABSTRACT

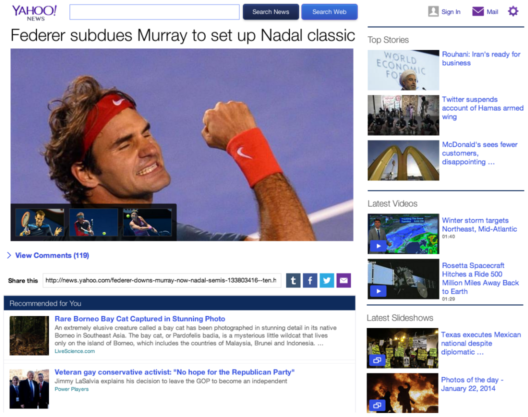
Online social networks and mash-up services create opportunities to connect different web services otherwise isolated. Specifically in the case of news, users are very much exposed to news articles while performing other activities, such as social networking or web searching. Browsing behavior aimed at the consumption of news, especially in relation to the visits coming from other domains, has been mainly overlooked in previous work. To address that, we build a *BrowseGraph* out of the collective browsing traces extracted from a large viewlog of Yahoo News (0.5B entries), and we define the *ReferrerGraph* as its subgraph induced by the sessions with the same referrer domain. The structural and temporal properties of the graph show that browsing behavior in news is highly dependent on the referrer URL of the session, in terms of type of content consumed and time of consumption. We build on this observation and propose a news recommender that addresses the *cold-start* problem: given a user landing on a page of the site for the first time, we aim to predict the page she will visit next. We compare 24 flavors of recommenders belonging to the families of content-based, popularity-based, and browsing-based models. We show that the browsing-based recommender that takes into account the referrer URL is the best performing, achieving a prediction accuracy of 48% in conditions of heavy data sparsity.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

Keywords

BrowseGraph, cold-start, news recommendation, browsing behavior, browsing sessions, recommender systems



Federer subdues Murray to set up Nadal classic

http://news.yahoo.com/federer-downs-murray-now-nadal-semis-133803416--ten.h

Rare Borneo Bay Cat Captured in Stunning Photo
An extremely elusive creature called a bay cat has been photographed in stunning detail in its native Borneo in Southeast Asia. The bay cat, or Pardofelis badia, is a mysterious little wildcat that lives only on the island of Borneo, which includes the countries of Malaysia, Brunei and Indonesia. ...
LiveScience.com

Veteran gay conservative activist: "No hope for the Republican Party"
Jimmy LaSalvia explains his decision to leave the GOP to become an independent
Power Players

Rouhani: Iran's ready for business

Twitter suspends account of Hamas armed wing

McDonald's sees fewer customers, disappointing ...

Winter storm targets Northeast, Mid-Atlantic
01:40

Rosetta Spacecraft Hitches a Ride 500 Million Miles Away Back to Earth
01:29

Texas executes Mexican national despite diplomatic ...

Photos of the day - January 22, 2014

Figure 1: An article page from Yahoo News (compacted layout). Right rail boxes and the infinite-scroll section at the bottom allow the user to browse to other articles.

1. INTRODUCTION

In recent years the consumption of online news has increased rapidly, in contrast with the decline of traditional newspapers[1]. Between 2009 and 2012 the percentage of users visiting news portals raised steadily up to the point to represent the major portion of overall Web traffic[2], comparable to the volume of visits to top domains like Google search[3]. For its importance, richness of content, and abundant user participation, the field of online news has become a crowded arena for research in several areas including retrieval, ranking, recommendation, and personalization [6, 2, 30]. Despite the vast amount of work in the field, there are two aspects of news consumption that are still largely unexplored. First, modern online news providers have turned into more globally connected systems able to attract a wider audience than their core of regular users. News articles are very often shared on different external websites and social media platforms, thus providing a growing number of browsing shortcuts to news portals. To mention two examples, modern search engines serve queries relevant to news stories by directly featuring news articles from major providers, and social media are increasingly used as daily tools for journal-

[*] This work has been performed when the author was a Visiting Scientist at Yahoo Labs, within the framework of the FREP grant.

RecSys'14, October 6–10, 2014, Foster City, Silicon Valley, CA, USA.
Copyright 2014 ACM 978-1-4503-2668-1/14/10 $15.00.
http://dx.doi.org/10.1145/2645710.2645726.

[1] http://stateofthemedia.org/2012/overview-4/key-findings/

[2] http://www.people-press.org/2012/09/27/section-2-online-and-digital-news-2/

[3] http://www.theguardian.com/news/datablog/2012/jun/22/website-visitor-statistics-nielsen-may-2012-google

ists and casual news readers[4] to spread and consume news provided by external parties [2, 17, 28]. Despite such increasing level of integration and mashup, news portals have been studied mostly in isolation. The latter aspect that has drawn very little attention is user *browsing behavior*. Although recent literature is rich in studies about browsing patterns in several online platforms [18, 9, 25], little has been done with respect to the news domain. One reason is that the browsing sessions in online news outlets tend to be short, aimed in most cases to quick catch ups on news [9].

In this work we address these two aspects in combination and exploit them in a task of news recommendation in a *cold-start* scenario. We study the way users browse news content in relation to the *type* of online domain they were browsing *before* landing on the news page, also known as *referrer* domain. Our contribution begins with finding that browsing is a meaningful phenomenon to study also for online news, as the browsing graphs have a coherent and well-formed structure also in this domain. We find that the referrer partly determines the type of news consumed and the time when it is consumed. In the wake of previous studies about the impact of the referrer domains on user browsing behavior [4], we use the browsing graph induced by the referrer of the browsing sessions to predict the next article a newcomer will visit right after she lands on a page of the site. Using a very large sample of viewlogs from Yahoo News (∼500M entries), we compare 24 flavors of recommenders for next-article consumption, including popularity-, item-, and browsing-based models. We find that browsing-based recommendations achieves the best overall precision@1 among all the methods: up to 48% in conditions of heavy volatility of news articles and of high data sparsity arising from the large amount of candidate articles to recommend.

We summarize our main contributions as follows:

- We introduce the *BrowseGraph* in the context of news and we define the notion of domain-dependent *BrowseGraph* (*ReferrerGraph*), namely a graph composed by the browsing sessions of users coming from the same referrer domain, *e.g.*, *facebook.com*. (§3).
- We study the *BrowseGraph* built from a large sample of browsing activity from Yahoo News. We explore it with respect to time and topic, providing insights on relations between the domain of origin, the type of news consumed and the temporal patterns (§4).
- We provide a method to recommend the next article to read in a cold-start scenario using the information from the *ReferrerGraphs*. Our recommender outperforms a number of item-, popularity-, and browsing-based baselines (§5).

We are not aware of any other work that use the *BrowseGraph* for news recommendation. Moreover, we exploit the referrer domain introducing and studying for the first time the *ReferrerGraph*, and show how its information helps to improve the accuracy of the recommendation.

2. RELATED WORK

Cold-Start Recommendation. Cold-start problem recommendation refers either to new items or to new users, and we consider the latter case. Some solutions to the problem require an initial, even though small, set of preferences about the newcomer user [22, 21, 1](*warm-start* scenario). The user profile can be populated by asking to the users to rate a set of items first or importing their preferences from external systems [19]. When a minimal user profile is available, a common approach is to apply association rules in order to expand the user profile. Sobhanam *et al.* [22] used clustering algorithms to find the most similar known items and then they extended the profile of the users with item related to their tastes. Shaw *et al.* [21] used also non-redundant rule sets showing how they can improve the results. Yang *et al.* [31] presented a Bayesian-inference based recommendation system that exploit the social network structure of the user in order to perform personalized recommendation. In this paper, the only information we have is the external referrer URL from which the user is coming from, and the news article the user is currently watching. While no one has considered the first information in the field of recommender system, the latter one could be integrated in a content-based approach, that is one of the baseline we implemented (§5.2).

News Recommendation. Despite the progress of recommender systems in general, news recommendation is still a very active area [5]. The majority of news recommender systems are based on user information collected over time [12, 5, 10, 11], relying on logged-in users only and combining collaborative filtering and content-based approaches. However, when reading news website the user is most often not logged-in, making impossible (or at least very difficult) the creation of a reliable user profile. There are approaches based on models of similarity among news articles [16, 10, 20] considering different key factors like textual similarity, recency, coherence, novelty and popularity. Tsagkias and Blanco [26] proposed a language intent model that extracts a query from the user browsing session. In this way they model the user interest using the queries of the users. There are also approaches looking for the most authoritative news sources as a proxy for quality [6, 24]. However, most of previous work relies on the user historical profiles. Our work is different since it is based on the *BrowseGraph*, using previous collected behavior of users grouped based on the external domain where they come from.

BrowseGraph and Referrer URL. In recent years there have been numerous studies on user traffic in term of navigation/browsing of web pages. Browsing sessions have been leveraged for different tasks such as recommending photos and photostreams [3], predicting the users demographic [8], or optimizing the web crawler [13]. Liu *et al.* [14, 15] introduced the *BrowseGraph*, a graph built by the users navigation patterns where the nodes are webpages and the edges are transitions made by users. They compared different centrality metrics computed on the standard hyperlink graph model, on the *BrowseGraph*, and on their combination, finding that the *BrowseGraph* rank has higher quality. The *BrowseGraph* has been used to rank items like photos in Flickr [25], where the external referrer domain (*i.e.* the last domain visited by the user before entering in Flickr), was considered to attribute more importance to the images with highest external visibility. The referrer domain has been proved to be useful for studying the popularity of media items (such as YouTube videos [7]), and to characterize the

[4] http://bit.ly/1bQG1uL

[5] http://recsys.acm.org/recsys13/nrs/

type of session the user is likely to perform [4]. However, we are not aware about studies that exploit this important information in the context of news or that use the *ReferrerGraphs* for new recommendation purposes as we do.

3. BROWSEGRAPH IN THE NEWS DOMAIN

To investigate the activity of news consumption and browsing we analyze a sample extracted from 2 months of the Yahoo News viewlog, also considering the browsing activity of users coming from different (*families* of) domains. In this section, we describe the raw data we use and its preprocessing (§3.1) and how we leverage it to generate the *BrowseGraphs* (§3.2).

3.1 News site viewlogs

Each article page in Yahoo News contains a nearly inexhaustible variety of options to transition to other article pages. As shown in Figure 1, the typical news page contains a right rail with several boxes of recommended news (*e.g.*, recent articles) and an infinite-scroll list of personalized news at the bottom. To capture the user's browsing activity we consider a sample of the site's *viewlog*[6], where each pageview is recorded with different fields. First, *BCookie* that is an anonymized user identifier computed from the browser cookie. *CurrentURL* and *ReferrerURL* represent, respectively, the URL of the page the user is currently visiting and the URL of the page from which she transitioned to the current one, possibly including external domains. Last, the *User-Agent* identifies the browser in use, and the *Time* indicates when the page was visited. By preserving only the traffic from well-known browser identifiers (to remove spider-generated requests), we are left with approximately 0.5 billion pageviews. To model browsing transitions we group pageviews in *sessions*, consisting of a sequence of pageviews made by the same user in a *continuous* time frame. As is standard practice [23], we split sessions by timeout (inactivity between two pageviews is longer than 25 minutes) but also when the user leaves the site for another domain, disregarding the timeout. In addition to the browsing transitions, we gather also the information about the *topic* of each article page (editorially assigned), the article headline and its body. There are 22 article topics in total, some of which are listed in Table 3.

3.2 Domain-dependent BrowseGraph

We aggregate the session data over all users to generate a *BrowseGraph* [14], namely a graph where the nodes are pages and the directed links connect pages that have been visited in sequence during a browsing session. The edges of the graph therefore have the following structure:

$$\langle Pageview_{source}, Pageview_{destination}, weight\rangle ,$$

where *weight* is the total volume of transitions, aggregated over all sessions. Because the *BrowseGraph* captures patterns of user flow, it has been leveraged in several applications, ranging from estimation of page importance [15, 14] to multimedia ranking [3, 25]. In the news domain, the *BrowseGraph* is particularly valuable as the links between articles may vary depending on time (*e.g.*, turnover of the "top stories" during the day) or even on the user, as the set of links shown may vary in a personalized fashion. In addition, we go beyond the study of the general browsing patterns by comparing the browsing behavior of users who land on the news website from *different domains*. We aim to verify empirically the idea that users coming from different types of external web services are interested (or exposed) to different types of content and therefore behave differently.

To decompose the overall *BrowseGraph* G into subgraphs G_d that account for the sessions originated from a specific domain d, we use only the sessions whose first referrer URL matches the domain d. For instance, a user who accesses a news page from a tweet will start a new session that will be part of the Twitter *ReferrerGraph*. We refer to the subgraph G_d as the *ReferrerGraph* for the domain d. In particular, we consider 9 source domains. Three search engines: *Bing*, *Google*, and *Yahoo*; three social networks: *Facebook*, *Reddit*, and *Twitter*; and the *homepage* of the news portal, a special case of referrer URL that represents a significant entry point for Yahoo News users. In addition, we also consider two aggregated *ReferrerGraphs* created by the union of the graphs in the *search engines* and *social networks* respectively (we call them *Search* and *Social*), as some of the characteristics of the *ReferrerGraphs* of domains belonging to the same family are quite similar (see §4). Last, since the consumption of news items is strictly dependent on time, we define a *temporal BrowseGraph* G^t as the *BrowseGraph* originated by the browsing sessions occurring in an hourly time slot t. We partition the *BrowseGraphs* on hourly intervals, ending up with $1,440$ temporal graphs for each domain, for a total of $12,960$ graphs.

4. ANALYSIS

In this section we report the structural properties of the full BrowseGraph and of the *ReferrerGraphs* induced by the different domains of origin (§4.1) and a study of their evolution in time (§4.2).

4.1 Domain-dependent news consumption

We find the distribution of the number of hops per session to be broadly distributed (not shown) but with very low average values (Table 1), in agreement with previous work that found the user interaction with news portals being short and time-constrained [9]. Despite that, the BrowseGraph built from the full set of user sessions is connected, with a greatest weakly connected component that spans up to 95% of all the pages and whose nodes are ∼5 hops away on average (statistics are summarized in Table 2). This means that although the individual browsing interactions with the news portal are short, the collective browsing behavior weaves an implicit network of associations between articles whose points are on average 5 hops away. The connectivity of the BrowseGraph appears to be scale-invariant, as very similar connectivity values are found for the *ReferrerGraphs*, the most disconnected one being Twitter, with 87% of nodes in its giant component. The average in-degree can be considerably high due to the large number of possible connections that an article page has with others (as illustrated in §3), and it is by far the largest in the *homepage* graph, which dominates in terms of traffic volume. Although the degree distributions vary considerably (Figure 2), the edge weight distributions are closer to each other, the vast majority of edges having very low weights.

A natural question is whether the graphs are different just in terms of structural properties or also with respect to the

[6] The processing was performed in aggregate on anonymized data

Table 1: Average number of hops during browsing sessions with different referrer domains.

Full	**Homepage**	**Google**	**Yahoo**
1.94	3.11	1.81	1.97
Bing	**Facebook**	**Twitter**	**Reddit**
1.79	1.34	1.24	1.12

Table 2: Structural statistics of the *ReferrerGraphs* ($\langle d \rangle$ indicates the average shortest path length and GCC indicates the Giant Connected Component).

Graph	*#Nodes*	*#Edges*	*Density*	*%GCC*	$\langle k_{in} \rangle$	$\langle d \rangle$
full	745,720	10,017,826	$1.8 \cdot 10^{-5}$	0.95	2551	5.14
homepage	257,465	3,516,661	$5.3 \cdot 10^{-5}$	0.99	1830	4.15
google	163,411	928,364	$3.5 \cdot 10^{-5}$	0.93	400	3.98
yahoo	116,403	490,239	$3.6 \cdot 10^{-5}$	0.95	229	2.91
bing	70,665	308,824	$6.2 \cdot 10^{-5}$	0.91	224	3.34
facebook	24,058	84,837	$1.6 \cdot 10^{-4}$	0.95	141	3.31
twitter	5,065	8,922	$3.5 \cdot 10^{-4}$	0.87	39	3.17
reddit	2,840	5,851	$7.3 \cdot 10^{-4}$	0.95	81	3.67

type of their nodes. To measure the overlap between graphs we compute the Jaccard similarity between the set of their nodes (Figure 3a). As one might expect, similarity is lower between the two major families: search and social. Surprisingly though, there are conspicuous differences also within each group. For instance, Twitter and Reddit have only ~20% of the overall amount of their nodes being covered by both. This means that the users coming from Twitter are visiting only a small portion of the news articles visited by users coming from Reddit. In other words, the users interest is strongly dependent by the type of website they are coming from. We spot also significant differences in the type of news content consumed in the different networks. To measure that, we count the frequency of articles belonging to each of the news topics (see §3.1), and we rank topics by their frequency in each network (Table 3). Naturally, the most popular type of news content in all the cases is related to general-type news. But except for the top position, the rankings show substantial differences, with celebrity-related news being the main interest for users coming from search engines, while blogs, sports, and entertainment are the most popular topics in Facebook, Twitter and Reddit respectively.

The differences in terms of graph structure, their size and type of nodes impact directly the type of articles that are consumed the most or that are perceived as most interesting by the users. To gauge that, we consider two metrics of news importance, namely the *pageview count* and the *PageRank centrality* computed on the weighted graphs. We apply each metric separately to the *ReferrerGraphs* and we compute pairwise the Kendall's τ similarity between the ranks. Figure 3b displays the values for PageRank, similar results are found for the viewrank (not shown for brevity). To discount for the different dimensionality, τ is measured only on the elements contained in the intersection of the two sets. To account for the noise that can be potentially introduced by the permutations on the latest positions on the rank (i.e., articles with very similar scores in the long-tail of the popularity curve), we repeat the same measure on the top 100 and 1000 articles, obtaining very similar results. Rankings tend to be more similar within domain families ($\tau \in [0.50, 0.68]$ for search and $\tau \in [0.20, 0.27]$ for social) rather across families ($\tau \in [0.14, 0.4]$).

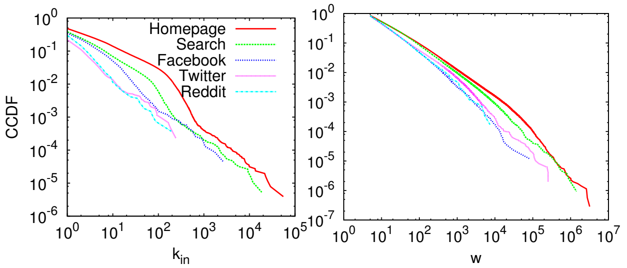
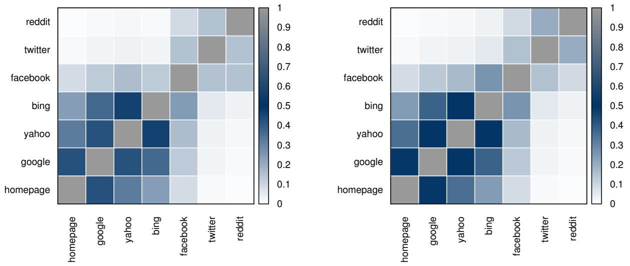


Figure 2: Complementary Cumulative Distribution Function (CCDF) of indegree (k_{in}) and edge weight (w) in some *ReferrerGraphs*. Search graphs are collapsed in one curve due to their similar distributions.



(a) *Jaccard similarity of node sets*

(b) *Kendall* τ *between news pageranks*

Figure 3: Node overlap between graphs and article ranking comparison.

4.2 News consumption in time

We study the evolution in time of the news consumption among the different *ReferrerGraphs*, with respect to the referrer domains and the topics.

4.2.1 Time and domain

Time plays a central role in news content consumption as news articles tend by their nature to become rapidly stale. In Figure 4 (left) we plot the distribution of the relative volume of views that articles receive in time (hours). In Figure 4 (right) we show the same measure but on a normalized time axis that starts from the article publishing time ($t = 0$) up to the last visit received ($t = 1$). We refer to this normalized timeline as the article *lifespan*. Consistently with previous work, we find that 80% of visits are received within the first 30 hours after the article publication and before the first 20% of the overall article lifespan.

One question, however, is whether the temporal aspect of news consumption depends on the type of network the user is coming from. To investigate this aspect, we repeat the previous measures separately on the three *ReferrerGraphs homepage*, *search*, and *social* in separation. For each of them, we measure the distribution of the total volume of visits at each point of the article normalized lifespan (Figure 5). A phenomenon of rapid decay emerges in all the three cases, however the curves exhibit major differences in skew. While news consumption through the homepage tends to happen in earlier stages of the lifespan, accesses through social networks and search engines are shifted towards later stages. In particular, for the *social* domain we observe evident peaks of accesses during late stages of the article life. Even if these

Table 3: Most popular news topics for different browse graphs. We grouped *Google*, *Bing*, and *Yahoo* into the same network (*search*) since their ranks are nearly identical.

Full	Homepage	Search	Facebook	Twitter	Reddit
News	News	News	News	News	News
Celeb.	Video	Celeb.	Entertain.	Sports	Blogs
Finance	Celeb.	Finance	Celeb.	Finance	Politics
Video	Finance	Video	Video	Video	Sports
Sports	Sports	Sports	Finance	Entertain.	Technology
Politics	Politics	Movie	Blogs	Lifestyle	Finance
Movie	Movie	Politics	Sports	Movie	Movie
Lifestyle	Lifestyle	Blogs	Photos	Photos	Video
Blogs	Blogs	Music	Lifestyle	Celeb.	Lifestyle
Music	Music	Entertain.	Movie	Music	Celeb.
Entertain.	Entertain.	Lifestyle	Politics	Politics	Health

peaks account for rather small percentage of the whole traffic (up to 2%), they still represent a non-negligible number of accesses. Examples of news that largely contribute to the visit volume inside the peaks belong to the "trivia" type of stories that are most commonly seen on social networks[7].

4.2.2 *Time and topic*

The referrer domain is just one variable that might impact the temporal patterns of content consumption for news, and the type (or topic) of the article can also have a role on that. In Figure 6 we show the aggregated volume of views in time for news articles belonging to six different categories. The relative positions of the accesses from homepage, social sites, and search engines change depending on the topic. The baseline consumption behavior is given by the general type of news, for which the view volume from the homepage is consistently higher than the volume from search engines, which is in turn higher than the one from social networks. For blogs, the view volume is more similar across the three macro-networks and the curves intersect more often, with two clear phases. First, the number of visits from social media goes above search for a short time, likely because blog posts in important news sites are usually written by bloggers who are heavily involved in online social networking. Last, the accesses from homepage and search become comparable in volume. Similar observations hold for sports, movies, and celebrities, that get a higher volume of accesses in later stages. In the case of celebrities in particular, we observe that accesses from social exceed even the ones from the homepage, after a certain point. This may happen when news about specific events (*e.g.*, academy awards) cause an outburst in the social media discourse. Last, an interesting case emerges from visual news such as photo-galleries related to news events. In this case, the accesses from social and search are comparable in the early life of the news, while the volume from search and homepage are comparable in the later stages. This delineates a scenario in which images lend themselves to spread easily in social media.

4.2.3 *Rank variation*

Besides studying the attention received in time by the overall set of articles, we check how the attention received by an article changes in relation to others. Ranking the articles by their viewcount, we explore how the rank changes in time and across *ReferrerGraphs*. We consider *Homepage*, *Search*, and *Social ReferrerGraphs* separately and for each of them we compute an hourly view rank R_t for all the articles they contain. Then, for each set of articles published in the hourly time slot t_i, we compute the Kendall's τ between their view rank at t_i and the view ranks in subsequent hours, more formally $\tau(R_{ti}, R_{ti+j}), \forall j \geq 1$. Then, we shift each measurement back in time by i hours, so that all sets of articles start from time 0, and we average all the measurements, with resulting curves in Figure 7. The lower the value of τ, the farther the ranking at time t is from the ranking of at the original publication time. In all the cases we observe the values decrease rapidly in the first 5 hours and a steady-state occurs within the first 24 hours. This finding is consistent with the volume of views dropping of several orders of magnitude in few hours. The τ value after 2 days is, on average, not higher than 0.55, meaning that the final ranking changes considerably from the initial one. Although the trend of the three curves is analogous, they have different offsets. Articles accessed via search change their relative position less and the ranking stabilize slightly quicker, while on the other extreme accesses from social networks impact the view rank more.

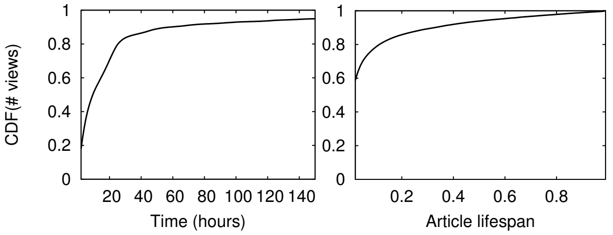


Figure 4: Cumulative number of page views in time

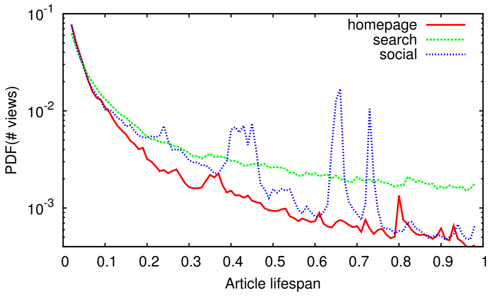


Figure 5: PDF of the number of views received in each of three *ReferrerGraphs* over the normalized lifespan of the news, from the publication ($x = 0$) to the last visit ($x = 1$).

5. COLD-START PREDICTION OF NEXT VIEW

Item recommendation is a crucial task in news sites as they have to deal with a rapidly changing pool of thousands of fresh articles and millions of users, each one with a specific range of interests. In such a scenario, profiling users with their explicit (*e.g.*, comments, article saved, printed, shared) and implicit (*e.g.*, views, time spent) activities on-site is an effective way to recommend new content that matches the user interest. However, personalization is not possible in cases of *cold-start*, when a user who is a newcomer or is not logged-in lands on the site. In this context, the information of the *BrowseGraph* can help, as the activity of previous users provide a collective trace of previous browsing patterns that can be recommended also to the new user. In particular, we show that the *ReferrerGraphs* are particu-

[7] *e.g.*, http://abcn.ws/1fPc0zu and http://abcn.ws/1iX4nHD

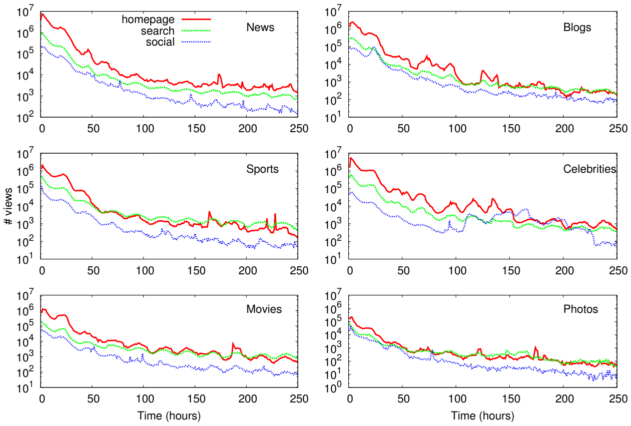


Figure 6: Number of views in each *ReferrerGraph* in time, breakdown by news topic.

larly effective to this end. Next, we formally define the recommendation problem (§5.1), describe a number of methods to address it (§5.2), compare them on a large scale dataset (§5.3), and discuss the results.

5.1 Problem definition

A newcomer user $u \in U$ is given, who begins a new session at time t on page $p_{start} \in P$, with referrer domain $d \in D$. The task consists in predicting a page p_{next} that u will visit right after p_{start}. We restrict the problem to users whose sessions will include at least an additional pageview after p_{start}. We consider, for simplicity, a time line quantized in discrete 1-hour slots and we assume to know the information about the browsing sessions generated by other users in the previous time slot t-1. To be able to draw a comparison also with text- or popularity-based recommenders, we consider an additional set of metadata for every page $p \in P$, specifically, v_p^{t-1}: cumulative number of pageviews at time $t-1$; cat_p: the page's topical category; h_p: the page textual headline; b_p: the textual body of the page.

5.2 Prediction methods

All the prediction algorithms we consider are determined by the combination of three components we describe next.

5.2.1 Selection of candidate pages

Full neighbors set (full). After the initial visit of p_{start}, the target user could transition, in principle, to any other page in P. However, we measure that in the 95% of the cases, p_{next} is included among the set $\Gamma_G^{t-1}(p_{start}) = \{p_i | (p_{start}, p_i) \in E(G^{t-1})\}$, namely the out-neighbors of p_{start} in the *BrowseGraph*created from the browsing sessions occurring during the timeslot t-1. This happens because, even though the cardinality of $\Gamma_G^{t-1}(p_{start})$ can be very big (recall the degree distribution in Figure 2), most of the browsing links between p_{start} and its neighbors are created shortly after the news are published. For this reason, an effective strategy would be to consider only the set $\Gamma_G^{t-1}(p_{start})$ as output range for the prediction. We call this selection strategy **"full"**, after the full *BrowseGraph*we use to perform the selection.

Referrer neighbors set (ref). In §4 we observed that the type of referrer URL determines to a certain extent the type of news consumed. One might argue that adapting the candidate page selection based on the user domain of origin d, could potentially improve the prediction accuracy. We could therefore restrict the output range to the neighbors of p_{start} in the domain-dependent graph G_d^{t-1}. Using G_d^{t-1} instead of the full graph G^{t-1} implies a drop in the chance of finding p_{next} in the set $\Gamma_G^{t-1}(p_{start})$ from 95% to a minimum of 48% for the Yahoo graph and a maximum of 72% for Homepage. However, based on our previous analysis, the subset of remaining pages could have a higher likelihood of being good candidates for prediction. Similarly to the previous case, we name this selection strategy **"ref"**.

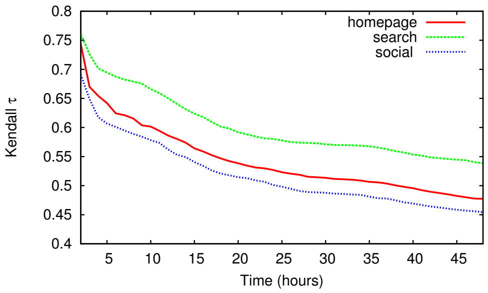


Figure 7: Kendall's τ calculated between the view rank at time t and the view rank at time 0.

Mixed neighbors set (mix). A natural extension is to combine the **ref** and **full** approaches in cascade. By definition, G_d^{t-1} has a subset of the nodes in G^t. Therefore, it may happen that the node p_{start} is not present in the subgraph or does not have any out-neighbor. So, we adopt the following strategy: if $p_{start} \in N(G_d^{t-1}) \wedge k_{out}(p_{start}) > 0$, then use the **ref** strategy, otherwise rollback to **full**. In the following, we refer to this strategy as **"mix"**.

In the following, we call C the set of candidate nodes, disregarding the strategy used to obtain it.

5.2.2 Topical filtering

As we report in Table 4, the probability of transitioning from a page with a topical category cat_p to another page with the same category, computed over all the transitions, varies depending on the domain of origin for that session. In the cases of *Twitter* and *Facebook*, there is a slight tendency to stick to the same topic, whereas for the other domains two consecutive pages in the session tend to belong to different categories. We leverage this information to enrich the initial candidate selection strategy keep only those articles in C that belong to the same topic as p_{start} for *Twitter* and *Facebook*, or to a different topic for the other domains.

5.2.3 Prediction of next page

All the methods use the *BrowseGraph* information to select an initial set of candidate pages C, according to one of the strategies defined above. After that, a criterion for the selection of the predicted next page among the ones in the set is needed. Next, we describe four algorithms, with their shortnames in parenthesis.

Random (rand). A simple baseline that selects at random a node in C.

Content-based (cb). A standard approach to recommend items at cold-start is to select the most similar article to the one the user is currently consuming, according to text-

Table 4: Probability that a user navigates between pageviews of the same category.

Total	Facebook	Twitter	Reddit	Google	Bing	Homepage
0.34	0.59	0.64	0.48	0.44	0.44	0.33

based metrics. When the body of the article is available (35% **of articles in our dataset)** the similarity is computed between the bodies, otherwise their headline (always available) is used. We compute the cosine similarity of the vector representation weighted with TD-IDF of p_{start} with the ones of every $p_i \in C$. Text is preprocessed with stopword removal and stemming [29].

Most Popular (pop). Another typical cold-start recommendation approach is to select the most popular item. We recommend the node in C with the highest view count of the timeslot between $t-2$ and $t-1$, namely the most popular article of the previous hour, in order to keep into account its recency (or freshness).

Edge-based (edge). Consider the weight on the edges that encode the likelihood of transition between nodes according to the browsing traces recorded at time $t-1$. Hence, we predict p_{next} to be the node in $\Gamma_G^{t-1}(p_{start})$ with highest weight on the incoming edge from p_{start}. Depending on the initial candidate selection strategy (§5.2.1), the edges considered (and their weights) will be either the ones in the *BrowseGraph* (for the **full** selection) or the ones in the *ReferrerGraph* (for the **ref** selection).

5.3 Experimental results

We apply our prediction strategies to the sessions of 1,438 hourly timeslots, for an average of $350K$ users per timeslot. We evaluate the goodness of the prediction by measuring its overall **Precision@1**: a true positive occurs when the predicted page is equal to p_{next}, a false positive when that condition does not hold. Additionally, since all the methods we presented lend themselves to produce ranking of pages (based on popularity, similarity, *etc.*), we also measure their Mean Reciprocal Ranking for the top 3 news articles (**MRR@3**). There is always a chance that the correct article cannot be possibly predicted because p_{next} might be not included in the set of candidates $\Gamma_G^{t-1}(p_{start})$ (for example because the article is published at time t and does not exists yet at time $t-1$). We adopt a conservative approach and we count also these cases as false positives. Figure 8 summarizes the prediction results. To have a more detailed picture of the cases in which the different approaches work best, we report separate evaluation results for the sessions with different referrer domains. Twelve bars for each group represent the precision and MRR scores for the combinations of the three selection strategies (**full**, **ref**, **mix**) with the next-node selection methods (**random**, **cb**, **pop**, **edge**). The maximum precision achieved for the different domains partially depends on the dimensionality of the session volume for that domain. This is mainly because the smaller the $\Gamma_G^t(p_{start})$ set, the higher the probability of giving in output a correct prediction just by chance. The most interesting experimental findings lie instead in the offsets between different methods' results.

First, the random baseline achieves always the worst performance, followed by the popularity, content, and edge strategies in order. The popularity approach, **pop**, relies on aggregate information about the amount of page visits, but it disregards where such visits came from, leading to lower performance. The **cb** recommender works only slightly better, our hypothesis is that the selection of next article is not driven by patterns of content similarity. In other words, after having read a news story the user is likely not motivated to keep reading about the same (or similar) stories after. The best method by far is **edge**, meaning that previous transitions from p_{start} and p_{next} constitute the stronger signal for the prediction of future transitions: it reaches 48% and 54% in P@1 and MRR@3 for social referrer domains and 27% and 34% for the search domains.

Regarding the node selection strategies, **ref** outperforms **full** in all the three social domains (except for the **ref-edge** combination in Facebook). The fact that a more specific type of recommendation works better suggests that people coming from social networks tend to retrace the same browsing paths other people from the same referrer domain have already explored, with limited serendipitous discovery. The opposite occurs for the search domains, where **full** beats **ref**. This may happen because query-driven systems provide a wider range of entry points to the news site than the links posted on social networks, thus making the prediction task harder. The same happens for the homepage, where the variability of content displayed is very wide and dynamic. However, for both families, **mix** is the most effective strategy that is able to significantly boost even more the precision for social networks and to fill the performance gap with **full** for the search domain. Homepage, which is the domain originating the highest number of sessions, is the only one in which **full** has top precision. In this case, the behavior of users is so varied that restricting the options to a subgraph turns out to be detrimental for the prediction.

Last, when the topical filtering is applied (§5.2.2), the precision experiences a drop in performance loosing from 10.6% up to 69.5% (not shown in plots). This happens because discarding too many nodes introduces the high risk of ruling out very good candidates (*e.g.*, a node connected to p_{start} with a high-weight edge). In our case, as shown in Table 4, the probability of transitioning to the same topic (or to a different one) is not far from 0.5 in all cases, therefore the topical information is not discriminative enough to filter out nodes without losing the most likely next pages.

The experiments highlight how the referrer URL of a browsing session can help to predict the navigation pattern and improving the next-hop recommendation in news browsing. A recommender that uses the weights of the *BrowseGraph* edges appears to be an effective way to anticipate user needs, especially for people coming from social media. This is particularly important, as it as been shown [2, 17, 28] that social media platforms are playing an increasingly important role on the news propagation[8].

6. CONCLUSION

We present an analysis of the browsing traces extracted from a very large viewlog from Yahoo News, introducing the definition of a special case of the *BrowseGraph* model, namely the *ReferrerGraph*, that consists of a subgraph built from the browsing sessions with homogeneous referrer URL. We find that the browsing graphs of news sites are well-connected despite the tendency to rapid staleness of content and to the typically short user sessions. *ReferrerGraphs*

[8] http://www.journalism.org/2013/10/24/the-role-of-news-on-facebook/

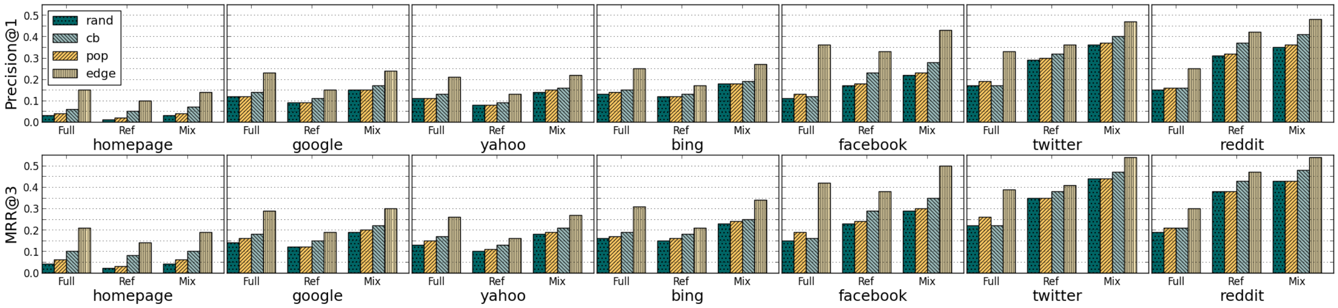


Figure 8: Prediction accuracy for the 9 recommendation strategies, computed for the sessions in each *ReferrerGraph* separately.

built considering 9 major domains appear to be quite non-overlapping, to cover articles of different topics, and to lead to the emergence of different sets of most popular articles. Traffic traces coming from different families of referrer domains have different time consumption patterns: for example, the sessions originating from search engines and social networks tend to consume content slightly after the visits coming from the news site homepage, and with some bursty consumption peaks for social networks caused by occasional spread of viral stories. Last, we build on our analytical findings by showing that the *ReferrerGraph* can be applied to effective article recommendation in a cold-start scenario. In terms of modeling our prediction setup could be extended using continuous-time models, and a comparison between *BrowseGraphs* built at multiple time scales is also a natural extension. Our experiments mainly prove the point that the referrer domain has a big predictive potential that should be considered when building any browsing user model. As our goal is limited to the prediction of the *next page* visited, more general content-based techniques for cold-start [1, 27] are not directly comparable with our approach, although a more extensive comparison would be valuable to gain a broader view on the problem. At any rate, we hope the findings highlighted in this paper lead to a greater consideration of the referrer domain with particular focus on cold-start problems.

Acknowledgments

This work was partially funded by Grant TIN2012-38741 (Understanding Social Media: An Integrated Data Mining Approach) of the Ministry of Economy and Competitiveness of Spain and by the SocialSensor FP7 project, funded by the EC under contract no. 287975.

7. REFERENCES

[1] D. Agarwal and B.-C. Chen. flda: Matrix factorization through latent dirichlet allocation. In *WSDM*. ACM, 2010.

[2] C. Castillo, M. El-Haddad, J. Pfeffer, and M. Stempeck. Characterizing the life cycle of online news stories using social media reactions. In *CSCW*. ACM, 2014.

[3] L. Chiarandini, P. Grabowicz, M. Trevisiol, and A. Jaimes. Leveraging browsing patterns for topic discovery and photostream recommendation. In *ICWSM*. AAAI, 2013.

[4] L. Chiarandini, M. Trevisiol, and A. Jaimes. Discovering Social Photo Navigation Patterns. In *ICME*. ACM, 2012.

[5] A. S. Das, M. Datar, A. Garg, and S. Rajaram. Google news personalization. In *WWW*. ACM, 2007.

[6] G. M. Del Corso, A. Gullí, and F. Romani. Ranking a stream of news. In *WWW*. ACM, 2005.

[7] F. Figueiredo, F. Benevenuto, and J. Almeida. The Tube over Time : Characterizing Popularity Growth of YouTube Videos. In *WSDM*, 2011.

[8] J. Hu, H.-J. Zeng, H. Li, C. Niu, and Z. Chen. Demographic prediction based on user's browsing behavior. In *WWW*, 2007.

[9] R. Kumar and A. Tomkins. A characterization of online browsing behavior. In *WWW*. ACM, 2010.

[10] L. Li, D. Wang, T. Li, D. Knox, and B. Padmanabhan. SCENE: a scalable two-stage personalized news recommendation system. In *SIGIR*. ACM, 2011.

[11] C. Lin, R. Xie, L. Li, Z. Huang, and T. Li. Premise: personalized news recommendation via implicit social experts. In X. wen Chen, G. Lebanon, H. Wang, and M. J. Zaki, editors, *CIKM*, pages 1607–1611. ACM, 2012.

[12] J. Liu, P. Dolan, and E. R. Pedersen. Personalized news recommendation based on click behavior. In *IUI*. ACM, 2010.

[13] M. Liu, R. Cai, M. Zhang, and L. Zhang. User browsing behavior-driven web crawling. In *CIKM*. ACM, 2011.

[14] Y. Liu, B. Gao, T.-Y. Liu, Y. Zhang, Z. Ma, S. He, and H. Li. BrowseRank: letting web users vote for page importance. *SIGIR*, 2008.

[15] Y. Liu, M. Zhang, S. Ma, and L. Ru. User Browsing Graph : Structure , Evolution and Application. In *WSDM*, 2009.

[16] Y. Lv, T. Moon, P. Kolari, Z. Zheng, X. Wang, and Y. Chang. Learning to model relatedness for news recommendation. In *WWW*. ACM, 2011.

[17] R. M. C. McCreadie, C. Macdonald, and I. Ounis. News article ranking: leveraging the wisdom of bloggers. In *RIAO*, 2010.

[18] A. L. Montgomery, S. Li, K. Srinivasan, and J. C. Liechty. Modeling online browsing and path analysis using clickstream data. *Marketing Science*, 23(4):579–595, 2004.

[19] A. M. Rashid, G. Karypis, and J. Riedl. Learning Preferences of New Users in Recommender Systems : An Information Theoretic Approach. In *ACM SIGKDD Explorations Newsletter*, volume 10, page 90, Dec. 2008.

[20] A. Said and A. Bellogín. News Recommendation in the Wild : Recommendation Algorithms in the NRS Challenge. 2013.

[21] G. Shaw, Y. Xu, and S. Geva. Using Association Rules to Solve the Cold-Start Problem in Recommender Systems. In *Advances in Knowledge Discovery and Data ...*, pages 21–24, 2010.

[22] H. Sobhanam and a. K. Mariappan. Addressing cold start problem in recommender systems using association rules and clustering technique. In *ICCCI*, pages 1–5. Ieee, Jan. 2013.

[23] A. Spink, M. Park, B. J. Jansen, and J. Pedersen. Multitasking during web search sessions. *Inf. Process. Manage.*, 42(1):264–275, 2006.

[24] I. Trajkovski. Pagerank-Like Algorithm for Ranking News Stories and News Portals. *ICT Innovations*, 231:87–96, 2013.

[25] M. Trevisiol, L. Chiarandini, L. M. Aiello, and A. Jaimes. In *SIGIR*. ACM, 2012.

[26] M. Tsagkias and R. Blanco. Language intent models for inferring user browsing behavior. In *SIGIR*, page 335, New York, New York, USA, 2012. ACM Press.

[27] M. Tsagkias and R. Blanco. Language intent models for inferring user browsing behavior. In *SIGIR*. ACM, 2012.

[28] M. Tsagkias, M. de Rijke, and W. Weerkamp. Linking online news and social media. In *WSDM*. ACM, 2011.

[29] W. Wagner. Steven bird, ewan klein and edward loper: Natural language processing with python, analyzing text with natural language toolkit. *Lang.Resour.Eval.*, 44(4):421–424, 2010.

[30] C. Wang, M. Zhang, L. Ru, and S. Ma. Automatic online news topic ranking using media focus and user attention based on aging theory. In *CIKM*. ACM, 2008.

[31] X. Yang, Y. Guo, and Y. Liu. Bayesian-inference based recommendation in online social networks. In *IEEE INFOCOM*, pages 551–555. Ieee, Apr. 2011.