Kille, B., Hopfgartner, F., Brodt, T., & Heintz, T.

The plista dataset

Kille, B., Hopfgartner, F., Brodt, T., & Heintz, T. (2013). The plista dataset. In Proceedings of the 2013 International News Recommender Systems Workshop and Challenge on - NRS '13. ACM Press. https://doi.org/10.1145/2516641.2516643.

The plista Dataset

Benjamin Kille, Frank Hopfgartner
Technische Universität Berlin
Ernst-Reuter-Platz 7, 10587 Berlin, Germany
{firstname.lastname}@tu-berlin.de

Torben Brodt, Tobias Heintz
plista GmbH
Torstraße 33–35, 10119 Berlin, Germany
{firstname.lastname}@plista.com

ABSTRACT

Releasing datasets has fostered research in fields such as information retrieval and recommender systems. Datasets are typically tailored for specific scenarios. In this work, we present the *plista* dataset. The dataset contains a collection of news articles published on 13 news portals. Additionally, the dataset comprises user interactions with those articles. We inctroduce the dataset's main characteristics. Further, we illustrate possible applications of the dataset.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Information Filtering; H.3.5 [**Online Information Services**]: Web-based Services

General Terms

Documentation, Experimentation

Keywords

News, Recommender Systems, Dataset

1. INTRODUCTION

Nowadays, more and more people read their news online rather than on traditional print media. With the increasing importance of online news portals, we can also observe an increasing need for personalised news services that recommend those articles that are relevant to the users' information need. Although research on adaptive news retrieval and recommendation has been performed for many years (e.g., [12]), most research has been focused on rather small datasets or on datasets which have been designed for a different purpose, casting doubt on the scalability of these approaches. Notable exceptions include [8, 13]. Although, given the restricted access nature of their dataset, further research remains infeasible for most researchers.

In this paper, we present a novel dataset, referred to as *plista* dataset, which has been made publicly available by the plista GmbH to foster research on news recommendation. The dataset has been created in the context of a research cooperation between plista GmbH and TU Berlin. plista is a company that runs a content and advertisement recommendation service on thousands of premium websites (e.g., news portals, entertainment portals). Whenever a user reads an article on one of their customers' web portals, their service provides a list of related articles. According to Brodt [3], plista processes over 5 000 requests per second and recommends millions of articles to even larger numbers of users every day. Given their large market share (mostly in the German speaking world), the dataset allows to track users' clicking behaviour on independent news portals, opening the gate for research on cross-domain news recommendation, user modeling and other related research topics. The dataset has initially been released as part of the ACM RecSys'13 Challenge on News Recommender Systems [19], where researchers were motivated to develop novel recommendation algorithms using this dataset and evaluate them in real-time based on real user feedback. The data have been recorded in a time frame of 4 weeks ranging from June 1 - 30, 2013.

In this paper, we provide a detailed description of this dataset. In Section 2, we first discuss the significance of test corpora for academic research and introduce popular examples. Section 3 represents the main part of this work. Therein, we illustrate the various characteristics of the *plista* dataset. In Section 4, we discuss specific features related to users. We pay particular attention towards the cross-domain aspect. Section 5 outlines the fields whom the dataset can be applied to. We conclude in Section 6.

2. RELATED WORK

In this section, we discuss existing datasets. We first present datasets related to information retrieval. Subsequently, we introduce datasets released for recommender systems research. Since the early days of recommender system research, the provision of standard test corpora has guaranteed a fair comparison of novel and state-of-the-art recommender techniques. The origins of these test corpora lie in the information retrieval (IR) domain where standard test collections have a long tradition, mainly due to the implementation of the Text REtrieval Conference (TREC) initiative [22]. In this section, we first outline the specifics of IR datasets and then discuss differences of test corpora that are used for evaluating recommender algorithms.

2.1 Information Retrieval Datasets

Test collections got first promoted by Cleverdon et al. [5], who introduced a test dataset in a controlled setting for the evaluation of computer-based retrieval engines, often referred to as *Cranfield Paradigm*. In their work, they performed various retrieval experiments on different test databases in a controlled

environment. Constraining the dataset helped them to identify available relevant documents which is helpful in drawing a conclusion on the quality of the output of a retrieval engine. Cleverdon [4] conducted further experiments with alternative indexing languages constituting the performance variables under investigations. These experiments are known as *Cranfield II*. The setting of a classical IR experiment can be divided into three components:

1. A static *test collection* of documents. The purpose of test corpora is to provide common datasets that enable comparison of research approaches.

2. A *set of queries* that are created based on the content of the documents of the test collection. The queries serve, together with the collection, as input for the retrieval engine.

3. A set of documents judged to be relevant or non-relevant to each query (*relevance assessments*). Retrieval results for each query will be compared to these judged documents to pose a statement about the performance of the retrieval engine.

In a typical search task, as many relevant documents as possible have to be retrieved. Two assumptions underlie the methodology: First of all, users only want to retrieve results which are relevant to their query and are not interested in non-relevant results. Furthermore, the relevance of a document to a query is uniform to all. Saracevic et al. [17] distinguishes between five types of relevance: topical, cognitive, motivational, system, and situational relevance. A thorough discussion about these different types is given by Borlund [2]. Within TREC, the commonly used relevance type is topical relevance, which is associated with the "aboutness" of given documents.
Over the years, various data corpora have been released that address different domains such as Patents, Blogs, or Multimedia [11, 14, 18]. With a multitude of test collections available [21], the news domain has been amongst the most prominent content. For further information on the use of IR test collections, the reader is referred to Clough and Sanderson [6] who provide an up-to-date discussion on the subject.

2.2 Recommender System Datasets

Although IR datasets have been used for evaluating different recommendation algorithms (e.g., [20]), focusing on topical relevance (i.e., the "aboutness") of given documents is not always feasible in a recommender system context. This applies especially for collaborative recommender algorithms where the content of items is not as important as the item's relation to other items in the dataset.
The main aim of recommender algorithms is to identify items in a dataset that might be interesting for a specific user. A more specific task is to estimate the user's interest by estimating the rating that this user would give to an item. Such information is usually not provided by standard IR test collections, raising demand for novel test collections that allow for addressing mentioned research challenges.
In recent years, various test corpora from different domains such as books, music and jokes [9, 10, 23] have been released. These datasets consist of user ratings in the form < User, Item, Rating >. Differing from IR research, the most research activities have been on recommending movies rather than news, mainly due to the release of a large dataset as part of the Netflix Prize, an open competition for predicting user ratings for movies [1].
With the release of the *plista* dataset, we intend to foster research in the field of item-centric news recommendation. In contrast to user-centric recommendation, item-centric recommendation focuses on evaluating system performance. We continue by illustrating the dataset's characteristics in the subsequent section.

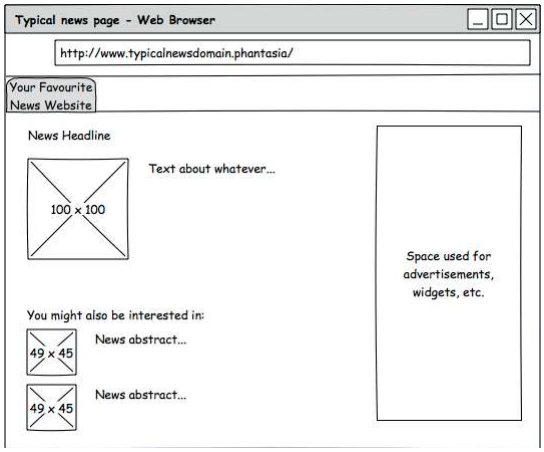

Figure 1: Exemplary news article website.

3. DATASET DESCRIPTION

In this section, we introduce the *plista* dataset. In the context of this dataset, we refer to news recommendations as those news articles that are promoted to users when they read articles on online news portals. The recommended articles are usually displayed at the end of a page in a small widget box labelled "You might also be interested in", "Recommended articles" or similar. Figure 1 illustrates the position of the recommendations on a typical news portal page. While some publishers provide their own recommendations, more and more providers rely on the expertise of external companies such as plista who do provide such recommmendation services. Having a large customer base, plista processes millions of user visits in real time on a daily basis. Publishers benefit not only from their own user interactions, but are additionally able to leverage user interactions occurring on other portals by obtaining recommendations from service providers such as plista.
The *plista* dataset is a logfile dump of the activity records that plista processed in June 2013 for recommending news articles in real time. While plista provides this service for thousands of online portals, this dataset contains records for a limited number of news portals, covering different spectra of the news world such as general, sports-related, or information technology related news. Since plista's domestic market is Central Europe, all news providers publish articles in German.
The corpus consists of four types of activities that have been performed by two types of actors on selected online domains: Adding and updating articles (done by the online editors of the respective news portal) as well as reading an article and clicking on a recommendation (the latter two activities being performed by the online customer, i.e., the readers of the online portals). We refer to adding articles as CREATE, updating articles as UPDATE, reading articles as IMPRESSION, and following recommendation links as CLICK.
The data in their raw form are represented as JSON encoded objects grouped by their type and the day they had been recorded. This grouping results in a total of 120 files accounting for 65GB. We distinguish two basic types of JSON formats. One such type is used for CREATES and UPDATES. Those are represented as flat

Table 1: Dataset statistics: Category, Alexa rank, plista's CTR, number and proportion of *impressions*, *clicks*, *creates* and *updates* for each publisher. The bottom part summarises the data with the sum (Σ), mean value (μ), and standard deviation (σ) for all four characteristics. Note that there were three impressions and two clicks without publisher assigned. We disregard those in the table. Further note that `www.wohnen-und-garten.de` did not receive any clicks. This publisher does not use the recommendation service. The same holds for `www.cnet.de`, `www.zdnet.de`, and `www.silicon.de`.

				Impressions		Clicks		Creates		Updates	
Publisher	Category	Rank	CTR	Count	Rate	Count	Rate	Count	Rate	Count	Rate
www.ksta.de	general	616	0.0126	19 102 953	0.2268	240 404	0.2195	18 435	0.2620	945 997	0.1828
www.sport1.de	sports	93	0.0233	20 595 846	0.2489	489 288	0.4467	3 544	0.0472	73 759	0.0143
www.gulli.com	IT	489	0.0035	4 727 153	0.0561	16 731	0.0153	299	0.0504	1 701	0.0003
www.tagesspiegel.de	general	460	0.0166	10 767 860	0.1279	178 554	0.1630	10 155	0.1443	35 959	0.0069
www.computerwoche.de	IT	1,486	0.0065	1 477 734	0.0175	9 622	0.0088	3 318	0.0472	7 510	0.0015
www.cio.de	IT	4,669	0.0144	498 943	0.0059	7 182	0.0066	3 311	0.0471	4 716	0.0009
www.cfoworld.de	finance	39,593	0.0133	29 112	0.0003	387	0.0004	72	0.0010	156	0.0000
www.tecchannel.de	IT	n/a	0.0106	1 313 698	0.0156	13 976	0.0128	650	0.0092	4 282	0.0008
cnet.de	IT	4,685	0	82	0.0000	0	0.0000	0	0.0000	0	0.0000
www.zdnet.de	IT	1,693	0	186	0.0000	0	0.0000	0	0.0000	0	0.0000
www.silicon.de	IT	1,250	0	12	0.0000	0	0.0000	0	0.0000	0	0.0000
www.wohnen-und-garten.de	gardening	7,956	0	397 883	0.0046	0	0.0000	39	0.0042	941	0.0002
www.motor-talk.de	automotive	186	0.0056	24 945 336	0.2962	139 179	0.1271	30 530	0.4340	4 099 095	0.7922
Σ				**84 210 795**		**1 095 323**		**70 353**		**5 174 116**	
μ			**0.0131**	**6 477 754**		**84 256**		**5 412**		**398 009**	
σ				**9 235 861**		**146 282**		**9 262**		**1 141 818**	

objects comprising relatively few properties (cf. Table 2). In contrast, the second type of JSON format dedicates to Clicks and Impressions. This type exhibits a more complex structure. The higher level of complexity arises form a larger quantity and more complex structure of data to be encoded. The JSON object contains basically three elements. First, a type definition classifies the objects as either an instance of Click or Impression. Second, an element describes the context of the object. Hereby, context refers to the active user. The context element includes information such as geographic location and browser environment. Additionally, the context element contains publisher-related data including estimated socio-economic user characteristics and item categorisation. Third, a recommendation element illustrates what news articles had been recommended. If the objects is of type Click, the recommendation element also mentions which recommendation had been clicked.

The dataset contains a total of 84 210 795 impressions, 1 095 323 clicks, 70 353 creates, along with 5 174 116 updates. An overview of the domains, their traffic rank in Germany (as determined by Alexa[1]), plista's own click-through rate (CTR)[2] and recorded activities is shown in Table 1. As evident from the traffic rank, some of the selected portals are among the most popular websites in the German speaking world, while others have a far smaller customer base. This is also reflected by the numbers of user interactions that have been recorded. Table 1 also displays the mean value and standard deviation of all four activities. The comparably large standard deviation values reveal the high variation between individual news portals.

In the remainder of this section, we provide a detailed description of the data corpus at hand, focusing on impressions (Section 3.1), clicks (Section 3.2), creates (Section 3.3) and updates (Section 3.4).

[1] `http://www.alexa.com/topsites/countries/DE`

[2] The CTR assesses the effectiveness of the news recommender systems. It is defined as the ratio of clicks over impressions.

3.1 Impressions

An impression record is created whenever a user reads an article on one of the available news portals. As shown in Table 1, the dataset contains more than 84 million impressions distributed over 13 news portals. When dealing with preference data, we typically observe two phenomena: (i) sparsity, and (ii) popularity bias. The former refers to the observation that users facing large item collections will interact with a comparably small fraction of items. The latter refers to the observation that a small fraction of items comprises a large fraction of preferences. In contrast, a large fraction of items comprises very few preferences. Both phenomena have been observed in preference data for movies [15], books [23], and music [9].

Figure 2 summarises the impressions for publisher `www.ksta.de`. The first histogram shows the distribution of impressions by users. We observe that most users interact rarely with the news portal resulting in a large fraction of users with few impressions. This confirms the presence of a markable sparsity phenomenon. This implies that for the majority of users we observe few interactions with the news portals. This represents an immense challenge to the system. The system has predict preferences on the basis of few data points. The second histogram shows the distribution of impressions by items. Again, we observe that most items attain few impressions. This observation confirms that existence of a noticeable popularity bias. Users focus their attention towards a relatively small subset of items. These items appear to be generally suited recommendations. The less popular articles represent a much more challenging recommendation tasks. The first bar chart displays the distribution of impression frequencies over the time of day. We observe a large amount of Impressions during day time. In constrast, we observe few Impressions occurring during night time. This confirms our intuition in that users consume articles promtly as the emerge. In the night, when less articles are published, the consumption decreases. Furthermore, we observe that the distribution peaks at the lunch time. This indicates that a large subset of users reads news articles during

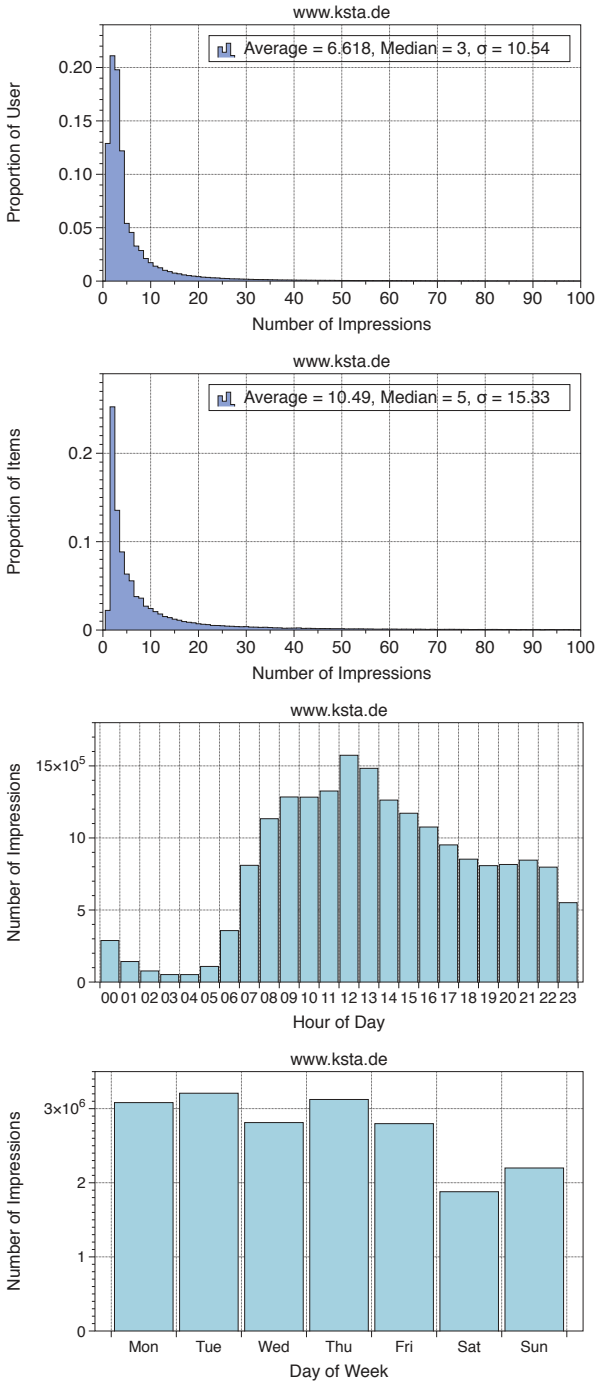

Figure 2: IMPRESSIONS **of** `www.ksta.de` **summarised from top to bottom as histogram of impressions by user, histogram of impressions by items, distribution of impressions over time by hour of day, and weekday.**

their lunch break. The other bar chart depicts the distribution of impression frequencies over the day of week. We observe that `www.ksta.de` receives a noticeably lower amount of impressions during the weekend. Again, this confirms our expectation in that users tend to read articles that relate to general news rather in the regular working week. Note that other publishers' histograms for impressions by users/items exhibit similar distributions. The bar charts for impressions by time of day and day of week show variations. Due to space limitations we confine ourselves to present figures referring to `www.ksta.de`.

3.2 Clicks

CLICKS represent users following links to articles that have been recommended while reading a news article. The dataset comprises approximately 1 million such click events. This gives an overall ratio of about 1 click per 80 impressions. The distribution of clicks on the different news portals can be seen in Table 1. The ratio of clicks to impressions (CTR) reflects the quality of the recommender systems. Figure 3 shows how this ratio develops over time for the news portals `www.ksta.de` and `www.tagesspiegel.de`. We selected both portals since they offer general news. Thus, their contents will largely overlap. We compare the CTR on a daily basis. `www.ksta.de` values are represented as blue circles while orange triangles represent the CTR values of `www.tagesspiegel.de`. We observe that the values cover an approximate range of $[0, 0.02]$. This confirms our observation that on average 1 out of 80 impressions coincides with a click event. The CTR of `www.ksta.de` surpasses the CTR of `www.tagesspiegel.de` with few exception. Click objects exhibit the same attributes as impressions (see Table 3). Additionally, clicks include a reference to the initial impression along with the recommendations provided. Thus, we can see what recommendations the users have disregarded. Figure 4 illustrates the timely development of CLICKS, IMPRESSIONS, and the relative CTR. Hereby, we sum CLICKS and IMPRESSIONS for all those publishers whose recommendations have been clicked at least once. This excludes publishers `www.cnet.de`, `www.zdnet.de`, `www.silicone.de`, and `www.wohnen-und-garten.de` (cf. Table 1). The upper chart summarises the daily number of IMPRESSIONS and CLICKS on a logarithmic scale. The lower plot depicts the difference between the daily CTR and the average CTR over the whole time frame ($\mu(CTR) = 0.0129$). We observe a low CTR for the first day. Additionally, we observe a comparably high CTR for the first 3 weeks. In contrast, the last week's CTR remains below average. The differences between the daily CTR and the average CTR cover the approximate spectrum of $[-0.007, 0.007]$. This indicates that the maximum CTR in the observed period is ≈ 0.0199. Conversely, the minimum CTR is at ≈ 0.0059.

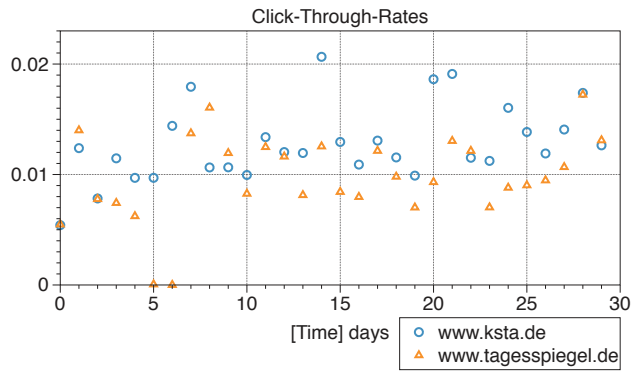

Figure 3: Click-Through-Rate (CTR) for `www.ksta.de` **and** `www.tagesspiegel.de`**. Both portals offer general news allowing us to compare them. CTR is defined as the ratio of clicks over impressions.**

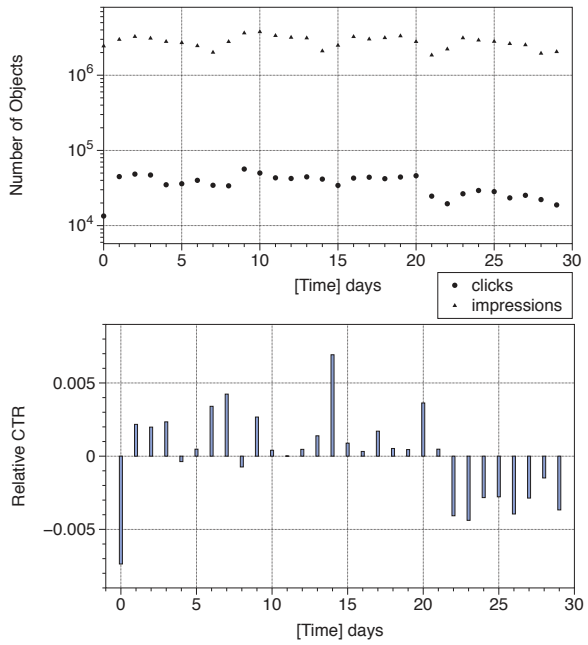

Figure 4: Timely development of CLICKS (represented as points), IMPRESSIONS (represented as triangles), and the relative CTR over time. The time is scaled in days. We measure relative CTR as the difference between the day-specific CTR and the average CTR (μ(CTR) = 0.0129). All values refer to the total objects.

3.3 Creates

CREATES represent new articles added on the news portals. A total of 70 353 news articles had been added to the different article collection in the observed time frame (see Table 1). Table 2 lists attributes assigned to CREATES as well as UPDATES. Each object is assigned an identifier, a reference to one of the news portal, an URL, along with a timestamp refering to the creation time. Additionally, a title, a text snippet as well as an image may be provided. Occasionally, those attribues were missing at first. This might be an indication for breaking news were additional texts or alternative headlines are added as more details become available. Figure 3.3 illustrates the timely development of the article collections of `www.ksta.de` (left-hand side) and `www.tagesspiegel.de` (right-hand side). Blue circles represent CREATES. The abscissa is scaled by days. The ordinate is logarithmically scaled. For `www.tagesspiegel.de` the number of created articles slightly fluctuates during the first week. Later, the number of created articles stagnates at arround a constant level. In contrast, the `www.ksta.de` exhibits a distinctive pattern. Notably more items are created during the working week compared to the weekend.

3.4 Updates

Updates represent an altered version of an existing news item being entered to the news portal replacing the initial article. An interesting observation from Table 1 is the high proportion of news portal `www.motor-talk.de` regarding UPDATES with 79.22 %. An investigation of those events reveals that the news portal offers users the chance to comment on displayed news articles. Unfortunately, every comment results in an update of the article causing the large amount of recorded updates. The actual news content remains rather unchanged. Figure 3.3 shows the amount of updates over time for `www.ksta.de` (left-hand side) and `www.tagesspiegel.de` (right-hand side). We observe that articles are much more frequently updated than entered into the system. This confirmls the intuition that as more and more details become available, news editors revise the initially published articles. The `www.tagesspiegel.de` exhibits two particularities. First, on the days 6 and 7 the number of updates appears strikingly low. The number of updates on those two days even deceeds the number of articles created. Second, the number of updates on day 22 exceeds all other days by far. The `www.ksta.de` does not show comparable trends. Conversely, the progress exhibits the same pattern as the clicks with major difference between working days and weekend days.

Table 2: Attributes of CREATES and UPDATES. Besides the type of the attribute, the table lists their availability at time of creation.

Attribute	Type	Availability
ID	Integer	100.00 %
publisher	Integer	100.00 %
title	Text	90.36 %
URL	URL	100.00 %
image	URL	65.23 %
text	Text	81.94 %
creation time	timestamp	100.00 %

4. USER DETAILS

In this section, we discuss aspects related to the users whose interactions have been recorded. We mainly focus on the availability of cross-domain appearances. Cremonesi et al. [7] distinguish cross-domain recommendation scenarios by the overlap of the sets of users or items on two distinct domains. We refer to an individual news portal as a *domain*. Alternatively, one could describe the setting as cross-site appearances. We consider the domains distinct since each news article is linked to one news portal at most. Thus, we deal with a scenario where the sets of items do not overlap while the sets of users do. Although the dataset contains records from millions of users, the dataset does not reveal any personal information that can be used to identify individual users. Users are identified through session IDs. These IDs allow us to track the users' news consumption behaviour. The observation is limited to the news portals included in the data set. In total, 14 897 978 unique session IDs have been recorded. Note that the dataset contains $\approx 84 \cdot 10^6$ IMPRESSIONS. This gives us an average news consumption of ≈ 5.6383 per session ID.

Table 3 provides an overview of the user-related attributes that can be found in the data set. All attributes have been submitted by the users' browsers. The table lists the number of distinct values as well as the entropy H, the maximum entropy value $\max H$, and their ratio for all attributes. Entropy values close to 0 signal a distribution with a dominating value. In contrast, values close to the maximum indicate a uniform distribution. We observe that most attributes' entropy ranges in the center between dominating values and uniform. There are two exceptions. Both *Language* and *Time To Action* appear to have dominating value. In the case of *Language* this is likely due to the fact that the articles are written and consumed in the German language. As mentioned above, users' privacy is an important issue. Therefore, information have

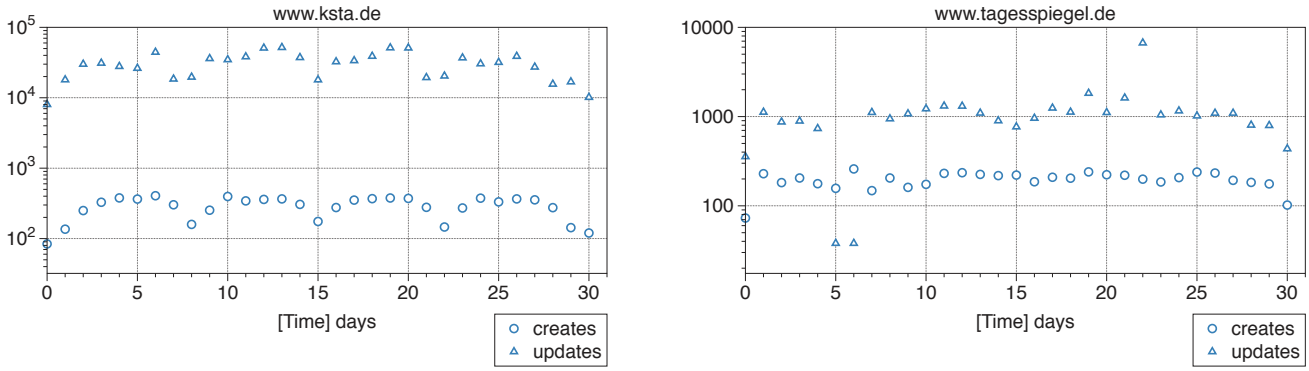

Figure 5: Evolution of articles over time for `www.ksta.de` (left-hand side) and `www.tagesspiegel.de` (right-hand side). Circles represent articles added to the news portals. Triangles symbolise articles being updated. We observe that articles are much more likely to be updated than added. The ordinate is scaled logarithmically.

Table 3: A selection of attributes assigned to impressions.

Attribute	Distinct Values	H	max H	H/ max H
Browser	172	3.5732	7.4263	0.4812
Device	6	0.8485	2.5850	0.3282
Location	1346	4.1937	10.3945	0.4035
Do Not Track	3	0.6754	1.5850	0.4261
ISP	17501	5.2359	14.0952	0.3715
Language	200	0.5628	7.6439	0.0736
Operating System	91	2.9708	6.5078	0.4565
Time To Action	47	0.0601	5.5546	0.0108
Widget	95	3.3918	6.5700	0.5163

been pseudonymised, i.e., details such as ISP, language and operation system are provided as generic ID rather than the actual value. The location attribute has been created by mapping the users' IP addresses with the registered metropolitan region. Given the publishers' focus on German speaking customers, it is not surprising that most users live in Germany, Austria or Switzerland and speak German. The relatively low entropy value of the location attribute indicates a biased distribution. This comes with no surprise since the two general news providers are the online representation of two regional newspapers from the metropolitan areas of Berlin and Cologne, respectively. The other attributes have not been normalised further. This explains, for example, the large number of different languages that have been reported by the browser, i.e., variations of the same language have not been merged. The dataset has a comprehensive description of all attributes attached. The description illustrates the value ranges and data structure.

Since users are identified based on their session ID, they can be tracked over several independent news providers (assuming that all providers rely on the plista recommendation service). This ability opens opportunities to apply cross-domain recommendation techniques. Hereby, we consider each news portal as a separate domain. Table 4 outlines the overlap of user identifiers on ten of the publishers. We observe a substantial overlap for some combinations of publishers. `www.motor-talk.de` and `www.gulli.com` share the largest number of users in the timeframe where the data have been recorded. Additionally, Table 4 lists the percentage of users who visit other news portals included in the dataset for each publisher. We observe that the proporation of such users ranges in

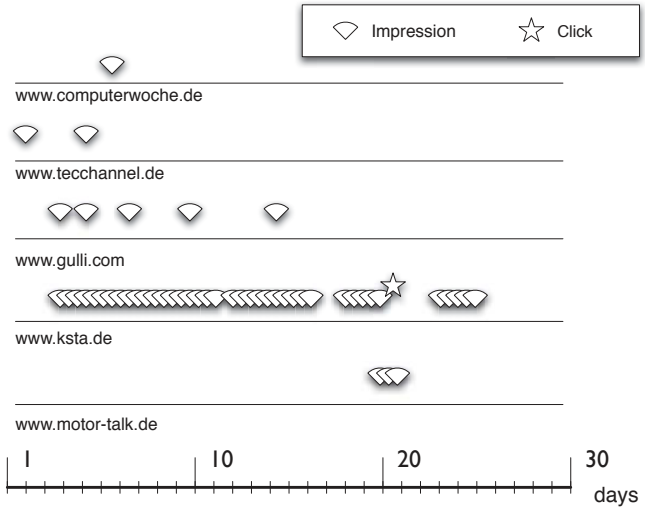

Figure 6: IMPRESSIONS and CLICKS observed for user with session identifier 1047533856. We observe that the user is mainly interested in articles of `www.ksta.de`. Besides the user visits a total of four additional news portals. All those portals publish predominantly information technology related news. The user clicked only once on a recommendation. The time scale is set to days in June 2013.

[0.1039, 0.3654]. These proportions should suffice to evaluate recommendation methods utilising cross-domain preference data. Figure 6 displays interactions of an exemplary user (identified with session ID 1047533856). The user interacts with five different news portals within the four week period. Most interactions occur with `www.ksta.de`. In addition, the user reads mostly IT-related news. The user has utilised the recommendation service only once. We observe that targeting this very user with knowledge derived from other domains yields potential to improve their experience. For instance, as an IT-related news article is added to `www.ksta.de`, we might recommend this very article since the user exhibits a noticeable interest in information technology.

Table 4: Number of users with IMPRESSIONS on the two intersecting news portals. We blanked the diagonal elements since the publishers coincide. On the bottom we list the number of users whose impressions span across several news portals including the specific publisher in the respective column. Note that we consider each user only once leading to less hits than summing up the individual values for all publishers listed in the table. Finally, we show the total number of users per publisher and the ratio over users visiting several news portals and the total number of users.

Publisher		1	2	3	4	5	6	7	8	9	10
www.ksta.de	1	-	86 555	53 988	134 877	15 863	4 644	640	9 748	1 954	98 679
www.sport1.de	2	86 555	-	69 564	117 068	19 557	4 800	579	13 308	1 262	130 194
www.gulli.com	3	53 983	69 564	-	106 523	50 517	7 254	464	58 079	1 788	222 407
www.tagesspiegel.de	4	134 877	117 068	106 523	-	30 957	7 977	1 051	18 703	2 820	147 177
www.computerwoche.de	5	15 863	19 577	50 517	30 957	-	21 891	1 187	38 604	484	46 178
www.cio.de	6	4 644	4 800	7 254	7 977	21 891	-	1 029	6 436	105	8 288
www.cfoworld.de	7	640	579	464	1 051	1 187	1 029	-	328	8	805
www.tecchannel.de	8	9 748	13 308	58 079	18 703	38 604	6 436	328	-	267	40 942
www.wohnen-und-garten	9	1 954	1 262	1 788	2 820	484	105	8	267	-	4 200
www.motor-talk.de	10	98 679	130 194	222 407	147 177	46 178	8 288	805	40 942	4 200	-
Sum of cross-domain users Σ_{CD}		327 796	365 696	461 657	457 468	165 899	43 573	4 025	137 519	10 677	590 758
Total sum of users Σ_I		1 905 181	3 217 756	1 913 917	2 382 792	536 740	119 257	12 151	426 002	71 913	5 687 083
Σ_{CD} / Σ_I		0.1721	0.1136	0.2412	0.1920	0.3091	0.3654	0.3312	0.3228	0.1485	0.1039

5. DISCUSSION

The *plista* dataset has been created to allow evaluating recommendation algorithms. The spectrum of features enables researchers to assess a variety of recommendation methods. CLICKS and IMPRESSIONS let test collaborative filtering techniques since both yield users' preferences towards news articles. Hereby, the CLICKS represent preferences for news recommendations while the IMPRESSIONS represent preferences for news. In addition, the availability of textual and other content features allows researchers to assess the performance of content-based recommendation techniques. Additionally, those textual features may be enriched by means of semantic data collections. Each object has a timestamp assigned. Thus, recommendation methods utilising the timely context can be evaluated. Additionally, the dataset contains the device attribute. Given which device a users reads news articles on, we can infer their current context. For instance, a user reading news on a desktop PC is likely to be at work. On the other hand, a users who comsumes news on their tablets indicate leisure time activities.

Besides evaluating recommendation algorithms, the *plista* dataset allows us to investigate alternative ways of user modeling. We may take various attributes into consideration to describe groups of users. Language, geographic location, and browser represent examples for such attributes. We may measure the correlations between those attributes and the users' behaviour. Thus, we can derive a clustering and model users by their (fuzzy) mapping onto the set of clusters. The presence of 13 news portals opens a set of new research challenges. As Cremonesi et al. [7] discuss, cross-domain recommender systems yield the potential to transfer consumption patterns thus providing better recommendations. The *plista* dataset comprises user preferences split over 13 news publishers. As shown in Table 4, there is a fraction of users whose preferences include several news portals. Those can be used to evaluate cross-domain recommendation techniques. Hereby, the different news portals cover a wide range of topics including general news, sports, and information technology. This setting enables us to determine whether the content focus represents the most important aspect when transferring consumptions patterns.

Additionally, the dataset includes both popular news portals and publishers with less traffic. We will investigate what impact a news portal's popularity yields when transferring consumption patterns. Another scenario supported by the dataset is popularity prediction. Knowing whether an article will attract a lot of attention is an important factor for news portals. With the features contained in CREATES and UPDATES a machine learning problem can be formulated. The target is to predict how many impressions an newly added article will gather in a fixed time frame. Conversely, an analysis of the popularity of news articles might support providers to phrase their articles in a way maximising the readers' attention. The dataset represents the log of 4 weeks of users interacting with the 13 selected news portals. In the scope of the *News Recommender Systems Workshop and Challenge 2013* participants have the opportunity to evaluate their news recommendation algorithms directly interacting with the *plista* system. The dataset can be used for bootstrapping purposes in this context (for more details see `http://orp.plista.com`, and [16]).

Besides its characteristics discussed above, there are some aspects that must be taken into consideration when using this dataset for research. First, users are identified by their session IDs. This yields two risks. On the one hand, users might have several devices for reading online news such as desktop PCs, tablets, and smartphones. Thus, we end up with several user IDs referring to a specific user. On the other hand, users might share their news-reading devices. For instance, a couple might use the same tablet to read news. In that case, we would observe a user profile mixing up two distinct user profiles. Second, the news portals included in the dataset exhibit noticeable differences both in quantity and quality of their service. Some attract a large amount of users while other have to deal with smaller customer bases. This dictates different objectives for recommendation methods. Third, the content is restricted to German. This may limit the applicability of natural language processing instruments designed for other languages. Fourth, preferences are not expressed on a numeric scale. This limits the applicability of popular evaluation metrics such as root mean squared error (RMSE). Fifth, the texts included in the CREATES and UPDATES represent the text included in the recommendation snippet. The actual news article typically contains more text. Still, the objects include a URL which can be used to

access the full text. Finally, the evaluation of news recommendation typically involves a definition of relevance. Although, clicking on a recomended news article might not reflect what we intend to measure. Users might only like the title of the recommendation snippet. Having read the first few sentences, they might find the article irrelevant.

6. CONCLUSIONS

In this paper, we described the *plista* dataset, a corpus consisting of millions of user interactions with news articles on independent news portals. The dataset has been released in the context of the ACM RecSys'13 Challenge on News Recommender Systems where participants could use it to train models that can be applied for real-time news recommendation. To the best of our knowledge, a comparable dataset has not been made publicly available yet. Therefore, we argue that the datasets provides opportunities to address research challenges in the field of news recommendation such as the role of user context (e.g., based on users' locations), collaborative filtering techniques, cross-domain recommendation or user modelling. Additionally, the dataset contains information about news consumption patterns. Detecting and analysing such pattern yields the potential for news providers to further optimise their systems. We discussed the main characteristics of the dataset including the 4 main data structures (CREATES, UPDATES, IMPRESSIONS, and CLICKS). We highlighted comparabilities to existing datasets as well as special features. In particular, we outlined the dataset's use to foster research on cross-domain recommender systems. You may contact the first author in order to get access to the data.

7. ACKNOWLEDGMENTS

We thank the *plista GmbH* for providing the comprehensive dataset. This work has been supported by the German Federal Ministry of Economics and Technology in the scope of the EPEN project `http://www.dai-labor.de/en/irml/epen/`.

8. REFERENCES

[1] J. Bennett and S. Lanning. The netflix prize. In *KDDCup'07*, 2007.

[2] P. Borlund. The concept of relevance in ir. *JASIST*, 54(10):913–925, 2003.

[3] T. Brodt. The search for the best live recommender system. In *BARS'13: Proceedings of the International SIGIR Workshop on Benchmarking Adaptive Retrieval and Recommender Systems*, page 3, 8 2013.

[4] C. Cleverdon. The effect of variations in relevance assessments in comparative experimental tests of index languages. Technical report, Cranfield Institute of Technology, 10 1970.

[5] C. Cleverdon, J. Mills, and M. Keen. Factors determining the performance of indexing systems. Technical report, ASLIB Cranfield project, Cranfield, 1966.

[6] P. Clough and M. Sanderson. Evaluating the performance of information retrieval systems using test collections. *Information Research*, 18(2), 2013.

[7] P. Cremonesi, A. Tripodi, and R. Turrin. Cross-domain recommender systems. In *IEEE International Conference on Data Mining Workshops*, pages 496–503, 2011.

[8] A. S. Das, M. Datar, A. Garg, and S. Rajaram. Google news personalization: scalable online collaborative filtering. In *Proceedings of the 16th international conference on World Wide Web*, WWW '07, pages 271–280, New York, NY, USA, 2007. ACM.

[9] G. Dror, N. Koenigstein, Y. Koren, and M. Weimer. The Yahoo ! Music Dataset and KDD-Cup Š 11. In *JMLR: Workshop and Conference Proceedings (KDDCup)*, pages 3–18, 2012.

[10] K. Goldberg, T. Roeder, D. Gupta, and C. Perkins. Eigentaste: A constant time collaborative filtering algorithm. *Information Retrieval*, 4(2):133–151, 2001.

[11] E. Graf and L. Azzopardi. A methodology for building a patent test collection for prior art search. In *Second International Workshop on Evaluating Information Access*, 2008.

[12] F. Hopfgartner and J. M. Jose. Semantic user profiling techniques for personalised multimedia recommendation. *Multimedia Syst.*, 16(4-5):255–274, 2010.

[13] J. Liu, P. Dolan, and E. R. Pedersen. Personalized news recommendation based on click behavior. In *IUI*, pages 31–40, 2010.

[14] C. Macdonald, R. L. Santos, I. Ounis, and I. Soboroff. Blog track research at trec. *SIGIR Forum*, 44(1):58–75, 2010.

[15] Y.-J. Park and A. Tuzhilin. The long tail of recommender systems and how to leverage it. In *Proceedings of the 2008 ACM conference on Recommender systems - RecSys '08*, page 11, New York, New York, USA, 2008. ACM Press.

[16] A. Said, J. Lin, A. Bellogín, and A. de Vries. A month in the life of a production news recommender system. In *Proceedings of the CIKM Workshop on Living Labs for Information Retrieval Evaluation*, LivingLab. ACM, 2013. (to appear).

[17] T. Saracevic. Relevance reconsidered. In *CoLIS'96: Proceedings of the Second International Conference on Conceptions in Library and Information Science, Copenhagen, Denmark*, pages 201–218, 1996.

[18] A. F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and trecvid. In *Multimedia Information Retrieval*, pages 321–330, 2006.

[19] M. Tavakolifard, J. A. Gulla, K. C. Almeroth, F. Hopfgartner, B. Kille, T. Plumbaum, A. Lommatzsch, T. Brodt, A. Bucko, and T. Heintz. Workshop and challenge on news recommender systems. In *RecSys'13: Proceedings of the International ACM Conference on Recommender Systems*. ACM, 10 2013.

[20] D. Vallet, F. Hopfgartner, and J. M. Jose. Use of implicit graph for recommending relevant videos: A simulated evaluation. In *ECIR*, pages 199–210, 2008.

[21] E. M. Voorhees and D. Harman. The text retrieval conference (trec): History and plans for trec-9. *SIGIR Forum*, 33(2):12–15, 1999.

[22] E. M. Voorhees and D. K. Harman. *TREC: Experiment and Evaluation in Information Retrieval*. MIT Press, Cambridge, MA, USA, 1 edition, 2005.

[23] C.-N. Ziegler, S. M. McNee, J. A. Konstan, and G. Lausen. Improving recommendation lists through topic diversification. In *Proceedings of the 14th international conference on World Wide Web*, WWW '05, pages 22–32, New York, NY, USA, 2005. ACM.