The Adressa Dataset for News Recommendation

Jon Atle Gulla
Dep. of Computer Science, NTNU
Trondheim, Norway
jag@ntnu.no

Lemei Zhang
Dep. of Computer Science, NTNU
Trondheim, Norway
lemei.zhang@ntnu.no

Peng Liu
Dep. of Computer Science, NTNU
Trondheim, Norway
peng.liu@ntnu.no

Özlem Özgöbek
Dep. of Computer Science, NTNU
Trondheim, Norway
ozlem.ozgobek@ntnu.no

Xiaomeng Su
Dep. of Computer Science, NTNU
Trondheim, Norway
xiaomeng.su@ntnu.no

ABSTRACT

Datasets for recommender systems are few and often inadequate for the contextualized nature of news recommendation. News recommender systems are both time- and location-dependent, make use of implicit signals, and often include both collaborative and content-based components. In this paper we introduce the Adressa compact news dataset, which supports all these aspects of news recommendation. The dataset comes in two versions, the large 20M dataset of 10 weeks' traffic on Adresseavisen's news portal, and the small 2M dataset of only one week's traffic. We explain the structure of the dataset and discuss how it can be used in advanced news recommender systems.

CCS CONCEPTS

• **Information systems** → **World Wide Web**; *Web searching and information discovery*; Web mining • **Computing methodologies** → Machine learning

KEYWORDS

Datasets, recommender systems, machine learning

ACM Reference format:

J. A. Gulla, L. Zhang, P. Liu, Ö. Özgöbek, X. Su. 2017. The Adressa Dataset for News Recommendation. In *Proceedings of WI '17, Leipzig, Germany, August 23-26*, 2017, 7 pages.
http://dx.doi.org/10.1145/3106426.3109436

ACM ISBN 978-1-4503-4951-2/17/08…$15.00
http://dx.doi.org/10.1145/3106426.3109436

1 INTRODUCTION

Over the last 10 years news recommender systems have gained in popularity and are employed by many online media houses. Today's readers tend to prefer online news rather than traditional paper-based newspapers, and they tend to be less faithful and less patient to news outlets that are not able to provide relevant and timely news [6]. In spite of controversies with filter bubbles and echo chambers, most news readers seem to embrace personalization features on news sites. Google News have for many years used advanced recommendation strategies on their web site with encouraging results [10], and we now see more traditional newspapers reporting interesting results from their personalized news services. For example, when the Washington Post started sending out personalized newsletters, they experienced that click-through rates for the personalized newsletters were three times the average and the overall open rate double of the average for the newspaper's newsletters [2]. After Polaris Media, the third largest media house in Norway, deployed their recommendation solution, the recommended content has been driving 5-10% of all article traffic, and maybe more importantly, the readers spend on the average 21% more time reading a recommended article than a non-recommended one [9].

However, news recommendation is more complex than many other personalization systems for a number of reasons [16]:

- The systems deal with unstructured data that may be hard to interpret and often incomplete, inconsistent, or overlapping in content.
- News stories have short life spans and are often of interest to just a modest geographical area.
- Users' intention varies unsystematically. Sometimes they prefer a certain topic or category, other times they want to see what their friends are reading or just follow the top stories at the moment.
- Users are usually not willing to rate the relevance of online news, and we are left with interpreting various implicit signals to second-guess their interests and satisfaction.

As a result, modern news recommender systems tend to be hybrid solutions, in which collaborative filtering and content-based recommendation are combined. In collaborative filtering we look for similar users and recommend news articles that these similar users have already showed some interests in. Content-based recommendation does not involve other users at all. The idea is rather to build a user profile based on historical behavior and recommend news stories that are similar in topic to what he has read in the past. Technologically this approach makes use of many techniques from information retrieval, like linguistic preprocessing of content data [5] and search queries expanded with profile models [3, 13]. Experiments indicate that users appreciate also additional strategies that boost fresh news, popular news and news that take place in their own neighborhood [7]. The architecture of advanced news recommender systems is illustrated in Figure 1. Some systems choose to merge all these strategies into a single list of recommended news stories, while others let the user activate the strategies they prefer before the recommendations are made.

To accommodate this mixture of recommendation strategies, we need datasets for training and evaluation that go beyond hat is found in traditional recommender system datasets. In the RecTech project we are now releasing a new dataset, the Adressa compact dataset, that includes features for a wide range of recommendation strategies as well as access to the full text of all referred news articles. This paper introduces the Adressa compact dataset and explains how it can be used for training advanced news recommender systems.

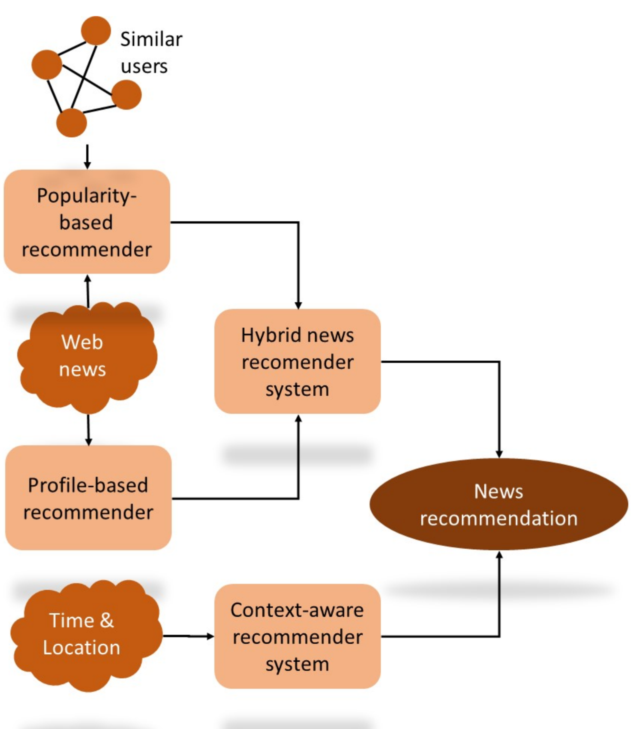


Figure 1: Architecture of advanced news recommender systems

The structure of the rest of the paper is as follows. In Section 2 we briefly discuss relevant datasets for recommender systems in general and news recommender systems in particular. After introducing the RecTech project in Section 3, we present the structure of the Adressa compact dataset in Section 4. In Section 5 we explain some important aspects of the dataset and show how the fields should be interpreted and used. Section 5 discusses the status of our datasets in RecTech and is followed by the conclusions in Section 6.

2 RELATED WORK

Evaluating the quality of recommender systems is an intricate issue. Since the ultimate goal is conversion, the systems should be evaluated on the basis of conversion rates, i.e. the share of all visits to a news site that turn into "purchases" in terms of new subscribers or additional ad revenues. Online studies are however often difficult to carry out, and it may also be difficult to generalize results from systems that are only used in a particular online setting. The alternative is to establish an offline environment that simulates an operative system working on a controllable set of data.

In an offline setting, training and evaluating recommender systems requires datasets that support the learning methods adopted by the systems. Many well-known datasets are set up for collaborative filtering systems and typically contain a number of events on the form

<user, item, rating>.

Each event tells us that a user has given an item a particular rating, e.g. user *Özlem* has given the movie *Alien: Covenant* a score of 4 out of 5. The rating scales vary, but as you see from Figure 1, many datasets has opted for a 1-5 scale.

Dataset	Items	Users	Ratings	Density (%)	Ratings scale
MovieLens 1M	3,883 movies	6,040	1,000,209	4.26	[1-5]
MovieLens 10M	10,682 movies	71,567	10,000,054	1.31	[1-5]
MovieLens 20M	27,278 movies	138,493	20,000,263	0.53	[1-5]
Netflix	17,770 movies	480,189	100,480,507	1.18	[1-5]
MoviePilot	25,058 movies	105,137	4,544,409	0.17	[1-5]
Last.fm 360K	294,015 artists	359,347	17,559,530	0.017	[1, 5]
Yahoo Music	624,961	1,000,990	262,810,175	0.042	[1, 5]
Jester	150 jokes	124,113	5,865,235	31.5	[-10, 10]
Book-crossing	271,379 books	92,107	1,031,175	0.004	[1, 10] + implicit
YOW	5,921 articles	28	10010	6.0	[1, 5] + implicit
Plista	70,353 articles	14,897,978	84,210,795	0.008	Click counts
Adressa 2M compact	923 articles	15,514	2,717,915	0.19	Click counts, reading times

Figure 2: Comparison of some well-known datasets

The MovieLens datasets are from the University of Minnesota and come in different sizes and different time periods [11]. MovieLens 20M contains more than 20 million 5-star ratings of movies and was logged from January 1996 to March 2015. Only users that rated at least 20 movies were included, and only movies that had at least one rating was included. There is some text data included in the form of movie titles and genres, but otherwise little that can support content-based or context-aware recommendation strategies.

Netflix is another well-known large dataset with explicit ratings of movies [1]. The dataset contains more than 100 million movie ratings that were recorded by Netflix subscribers between December 1999 and December 2005. Only movie IDs are used to identify the movies, leaving out both movie titles and genres from the dataset. Similar limitations are found in the Yahoo Music dataset, which includes more than 260 million ratings [4].

Movielens, Yahoo Music and Netflix are all very large datasets that have proven useful in developing collaborative filtering solutions. The explicit ratings are used to construct patterns of similar users, but the lack of movie descriptions makes it hard to test content-based strategies with these datasets. For news recommender systems these datasets are of limited use.

YOW from Carnegie Mellon University is a small dataset from the news domain that contains both explicit ratings and implicit signals like mouse and keyboard scrolling [15]. It also contains some textual data about classes (news categories) that can be used to construct simple user profiles. There is unfortunately little contextual information in the dataset, and the textual part is too small to allow full-fledged user profiles to be extracted.

A more interesting news dataset is the Plista dataset from plista GmbH and TU Berlin [8]. Logs from 13 German news portals were collected for four weeks in June 2013 as part of the ACM RecSys'13 challenge. The dataset contains more than 84 million article views (called *impressions* in the dataset), though no explicit ratings or reading times are included. User's interest is solely based on whether he reads or does not read a particular article. The dataset also contains some textual data and has been used to train both collaborative and content-based recommender systems. Time-related data like publication date are included, but there is no information about geographical locations.

An overview of some popular datasets is found in Figure 2 (adapted and expanded from [12]). The numbers for the Adressa dataset are added at the end of the table and explained in the next section.

3 THE RECTECH PROJECT

The Adressa dataset was prepared as part of the RecTech project on recommendation technology. This is a 40 million NOK (4.2 million Euro) industry-led research project that was started in 2016 and is partly funded by the Research Council of Norway. The project owner is Adresseavisen, while the other partners are Cxense in Oslo, NTNU in Trondheim and VTT in Finland. Adresseavisen (with short form "Adressa") is the largest newspaper in the Polaris Media news house, with around 140,000 daily readers.

RecTech's objective is to develop the next generation recommender systems for online news recommendation. The recommendation platform itself is provided by Cxense, while NTNU is developing new methods for fine-grained user profiling and deep content analysis. Computational linguistics, machine learning and Big Data architectures are central in this work. The projects builds on the experiences from an earlier prototype called the SmartMedia mobile news recommender systems [14]. Adresseavisen serves as a Living Lab for the project, as new components are developed and deployed on real traffic.

The Cxense platform for news recommendation and monitoring was used to extract the dataset, which covers one week of web traffic from February 2017 on the *www.adresseavisen.no* web site. All the 923 articles in the dataset are in Norwegian, and the average article length is 518.6 words. Of the 15,514 readers in the dataset, 672 are registered subscribers. A subscriber has access also to articles behind the pay wall, and we can follow him from one session to another, and form one device to another. We would expect that the user logs of subscribers to be fairly accurate with respect to their interests and preferences. For non-subscribers there are two aspects that need to be taken into account:

- There might be articles behind the pay wall that they would like to read and are consistent with their interests, but are not available to them. Hence, they are not clicking on or reading all the articles of interest to them.
- Every session constructs a new user ID. A particular reader will be associated with a new user ID every time he initiates a new session. He will also get different user IDs for different devices.

This means that the user profiles of non-subscribers are less complete in the current form of the dataset. Methods from cross-device tracking may alleviate some of these problems by linking sessions together that belong to the same user.

With a total of 2,717,915 article views, the density of the dataset is about 0.19%. A calculation of density values per day shows that the density is around 0.21% most days, though the data for day 1 is very sparse with a density of only 0.11%.

4 STRUCTURE OF DATASET

The raw data set extracted from the Cxense platform is split into three separate folders and contains a wide range of attributes that are not all very useful in recommender systems. From the raw data we construct a compact dataset that is comprised of two parts:

- A table of reading events. Each row includes 18 attributes that describe the event itself, summarize the article viewed, and identify the user viewing the article.

- Additional data about the articles.

The attributes of the event table are listed in Figure 3. Each reading event is given an internal unique ID. There is time stamp for the event, and there are two Boolean attributes that reveal if the event is at the beginning or end of a user session. The *activeTime* attribute records the time spent reading the particular article. The table then identifies the web page for the news article and also the web page that the user came from (referrerURL). Each article has a publication date, is given an internal document ID, has a title, is associated with exactly one news category, and contains a specific number of words. The last part of the table describes important properties of the reader. The reader has a unique user ID, is located in a particular city, region and country, and uses a particular device with a particular operating system.

There is no explicit rating of news stories, but the are implicit signals of interests in terms of click counts and time spent reading the articles may be used to calculate scores.

Attribute	**Description**	**Example**
eventID	Id of reading event	1082287123 (integer)
Time	The time of the event	1487572383 (Unix time)
sessionStart	Indicates if the event is the first event in the session	False
sessionStop	Indicates if the event is the last event in the session	False
activeTime	The active time spent on a page	23 (seconds)
canonicalURL	URL of the visited page	"http://adressa.no"
referrerURL	The URL of the referrer page	"www.facebook.com"
documentId	Internal ID of page	"9757814edc2d346dfcf6f54e349f404c4e9775cf"
Title	Title of the article	"Test av 19 grovbrød"
Category	News category	"sport"
wordCount	Number of words on page	545 (words)
publishTime	Date of publication	"2017-02-20T09:45:47.000Z"
userID	The cross-site user identifier	"cx:i8i85z793m9j4yy0:cv8ghy3v45j8"
City	City inferred from user's IP address	"Verdal"
Region	Region inferred from user's IP address	"Nord-Trøndelag"
Country	Country inferred from user's IP address	"no"
os	Operating system	"Windows"
deviceType	The type of device	"Desktop"

Figure 3: Fields of the Adressa compact dataset

In addition to the reading event table each news article is associated with some more detailed information about its content:

- *Author*: the author of the news articles
- *Keywords*: any keywords or tags annotated with the article.
- *Entities*: extracted entities and prominent terms from the news article. Each entity or term is described by a label (e.g. "Lionel Messi"), a type (e.g. "person"), a count (e.g. 2 occurrences in the article) and a weight (e.g. 0.78).
- *Body*: The entire text – in Norwegian – of the news article.

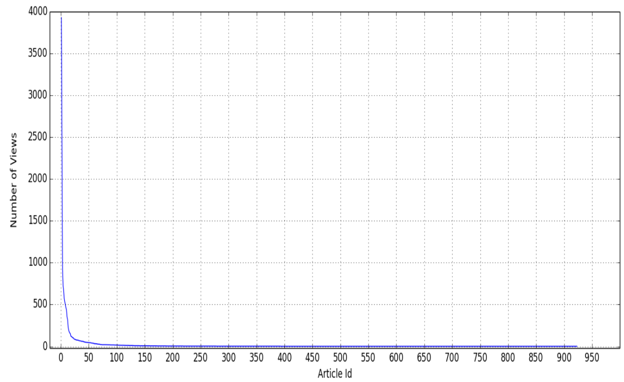
As shown at the end of the table in Figure 2, The Adressa 2M compact dataset contains 923 news articles, 15,514 readers, and about 2.7 million events. Each of these events corresponds to a user reading a particular news article.

As seen from Figure 4, normal news stories make up about 43% of the articles included in the dataset. Slightly less than 6 entities are mentioned per article of this category. There are also many sports and culture articles in the dataset, and they have a higher frequency of entities than normal news stories. Each culture article includes on average 11 entities, which is slightly less than the almost 12 entities in the average travel article. The 60 articles categorized as "Pluss" are articles that are only available to paid subscribers of Adresseavisen.

Category	**No. of articles**	**% of articles**	**Average no. of entities**
"Nyheter" (news)	397	43.0	5.86
"100sport" (100sport)	136	14.7	7.53
"Kultur" (culture)	91	9.9	11.0
"Pluss" (paid content)	60	6.5	9.59
"Meninger" (opinions)	56	6.1	8.84
"Forbruker" (consumer)	41	4.4	5.92
"Familie og oppvekst" (family)	24	2.6	8.05
"Bolig" (housing)	23	2.5	6.3
"Bil" (car)	19	2.1	8.63
"Sport" (sport)	18	2.0	5.88
"Reise" (travel)	18	2.0	11.94
"Været" (weather)	11	1.2	7.11
"Digital" (digital)	8	0.9	10.38
"Folk" (people)	6	0.7	10.83
"Live studio"	4	0.4	1.67
"Student"	2	0.2	3.0
"Video"	2	0.2	1.0
"Tjenester" (services)	1	0.1	2.0
"Kamera" (camera)	1	0.1	9.0
"Nyttig" (useful)	1	0.1	14.0
"Jobb" (job)	1	0.1	3.0
"Tema" (theme)	1	0.1	4.0
"Migration catalog"	1	0.1	2.0
"Incoming"	1	0.1	1.0
Average per category	*38.5*	*4.2*	*6.6*

Figure 4: Number of articles per news category and average number of entities per article per category.

In Figure 5 we see how often the 923 articles have been viewed in the dataset. There are some articles that are very popular, but a very large share of the articles have been viewed only a handful times. Only 392 articles (32.34%) are viewed more than one time, and no more than 22 articles (18.15%) are viewed more than 100 times. The two most popular articles, which both come from the "news" category, are viewed 3.929 and 2.436 times in the course of these 7 days.



THE DATASET

Figure 5: Number of article views per article

Figure 4 and 5 indicates that the dataset is somewhat unbalanced. There are a few articles that are popular, and almost 70% have been viewed only one single time. Similarly, whereas one news category accounts for almost half the articles published, 9 out of 24 categories have only one or two articles.

5 ANALYZING THE DATASET

5.1 Subscribers and Non-Subscribers

The dataset makes a distinction between subscribers and non-subscribers. Subscribers have access to all articles on the news portal, including the news stories that are categorized as "Pluss". Since they are stored in the system, we use the same user ID for subscribers every time they start a session and on all devices. Non-subscribers do not have access to "Pluss" articles, and their ID changes for every user session and for every device they are using.

To identify subscribers in the dataset, we need to inspect the canonicalURL attribute to see whether the user has read any articles of the "Pluss" type. If there is no "Pluss" string in the canonicalURL, it means that both subscribers and non-subscribers may read the article. If a user has read the following article, for example, we cannot know whether this is a subscriber or not:

canonicalURL : http://www.adressa.no/meninger/kronikker/2017/02/06/Nav-og-velferdsordningene-er-livsviktige.-Systemet-har-bare-%c3%a9n-stor-feil-14174815.ece

On the other hand, if the user is accessing the news article below, which is marked with "Pluss" in the canonicalURL attribute, we know that this has to be a subscriber:

canonicalURL : http://www.adressa.no/pluss/magasin/2017/02/06/Fjellseter-1907-og-2017-14156717.ece

5.2 Location and Time

There are three attributes that identify the geographical location of the reader: *city*, *region* and *country*. If an event is marked with

City: "bjugn"
Region: "sor-trondelag"
Country: "no"

the reader is assumed to be sitting in Bjugn, which is a city in the Sør-Trøndelag county in Norway. This information is based on IP addresses associated with the reader.

Now, the recommender system may also need to identify the location of the news story, if it should recommend local news to the reader. This is possible with the additional content data available for each article in the dataset. There is a particular attribute, *Entities*, that lists all identified entities in the article with their entity types and prominence. If there are any locational entities present in the article, these can be used to calculate the geographical proximity of reader and news event. For example, if an article viewed by this person in Bjugn mentions Trondheim, which is marked as a locational entity in the article content data set, we can canculate that there is a 102 km distance between the reader (Bjugn) and what we may assume is the location of the news event (Trondheim). Keep in mind, though, that there may be many locational entities in an article, entities may be ambiguous, and they may not even have anything to do with where the event took place. For example, Heimdal is a location both in Norway and in the US, but it may also refer to an old God that has nothing to do with any location at all.

An important issue in all news recommender systems is recency. The life span of news articles depends on their category, but is in any case much shorter than the life spans of books, movies, and other items you may buy on online shopping sites. Since most conventional news stories last only 1-2 days, it does not make sense to recommend older stories, unless they come from particular categories like "opinions" or "consumer" that have somewhat longer life cycles. The age of an article is computed by comparing its publication date (*publishTime*) with the time the user was reading the article (*Time*).

5.3 Identifying a Session

There is no concept of session in the dataset itself, but we can use the attributes *userID*, *Time*, *sessionStart*, and *sessionStop* to string reading events together into a user session.

Assume that we want to construct the reading session of a user at a particular point of time. The first thing we must do is to find the first event of the session. This is the record, in which the attribute *sessionStart* is true for this particular user. We then follow the user's next record with from the Time and userID attributes and check if the session stops or not by inspecting the *sessionStop* attribute. If *sessionStop* is true, we are at the end of the session and we know that there are two events in this session. If *sessionStop* is false, we need to keep this record and find the next record until we run into the final one with *sessionStop* set to true. The sum of records, including the first one with sessionStart set to true and the last one with sessionStop set to true, forms a user session.

The process is illustrated in Figure 6 below. The first reading event of the session is at time t_0. The *sessionStart* attribute is true to indicate that this is the first event of the session, and the *sessionStop* attribute is false to indicate that there are more events of the session. The next event at time t_1 has both these attributes set to false, which means that we are not at the end of the session yet and need to find a third event for the user at time t_2 ($t_0 < t_1 < t_2$). We see that the event at t_2 terminates the user session, as *sessionStop* is set to true, and we have then identified a user session that involves clicking on three different news articles.

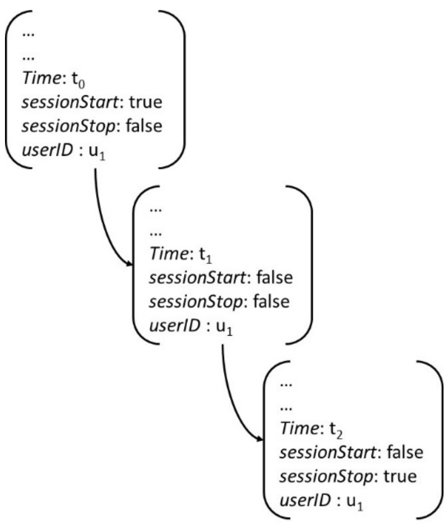


Figure 6: Constructing a session form reading events

5.4 Implicit Signals

The implicit feedback from user is recorded at two levels of granularity:

- *Click counts*: Every row in the reading event table constitutes an act of clicking and possibly reading a news article. At the most basic level, reading an article may be interpreted as an interest in the article's topic, while not reading the article means that the topic is uninteresting.
- *Reading times:* The dataset also records the time spent reading the news articles. This time may be used directly as an indication of interest, but it may be wise to normalize this time with respect to the length of the article (wordCount). Using *activeTime* and *wordCount*, you may estimate how much of the article the user has actually read, which may serve as a more fine-grained indication of interest than just the click rates.

5.5 Content-Based Representations

A central part of the RecTech dataset is the availability of all textual data of the news stories as well as some relevant meta data. We have chosen to include the full body text in case someone wants to do their own analysis of the content. For many applications, though, it will probably be sufficient to use the keywords and entities provided in the dataset and ignore further analysis of the article body.

For a particular article on the finances of the RBK football club, the event dataset and the additional article content data give us the following information:

Title:	"RBK finansieres av internasjonale spillselskaper" (RBK is funded by international betting companies)
Author:	"Morten Klein"
Category:	"meninger" (opinions)
Keywords:	RBK, Byåsen håndball, spillemidler, sponsor
Entities:	football club RBK (Rosenborg), RBK director Tove Moe Dyrhaug as well as a number of other clubs and some betting companies.
wordCount:	545 words
Body:	«Det er med undring at jeg leser uttalelsene til RBKs daglige leder Tove Moe Dyrhaug i Adresseavisen 4. februar. Moe Dyrhaug uttaler seg i forbindelse med støtten på 1 million kroner undertegnede og Norgesspill.com har valgt å gi til Byåsen Håndball Elite. ...»

The category and the keywords are manually tagged by the journalists of Adresseavisen and may be of variable quality. The entities are automatically defined using standard methods for named entity recognition and high-frequency phrase extraction.

6 DISCUSSION

The Adressa dataset is split into two parts, the reading event dataset and the additional article content data. The first part can be downloaded from SmartMedia's web pages at NTNU with no restrictions. As some of the news articles were published behind pay walls, this additional information in the second part is not published directly on the web, but will be freely available on request.

In total, the dataset contains more features and support a wider range of recommendation strategies than the public datasets that are in use today. In particular we have added many content attributes that should make it easier to include content-based recommendation strategies in the systems. Since other large datasets today emphasize the collaborative filtering part, we think our dataset will be a good complement.

In the news domain there are no explicit ratings available. Click counts and reading times reveal users’ intentions to some extent, but the lack of proper ratings is a challenge in all news recommender systems.

The current 2M dataset suffers from being rather unbalanced. There are a few categories that dominate the articles published, and only a small share of the articles have satisfactory click rates. On top of this, we can only follow subscribing readers from one session to another, but these constitute only 4.3% of the total number of readers. The unbalanced nature of the dataset is probably symptomatic for the news domain in general, but is particularly problematic when the total dataset is so small. We are therefore in the process of preparing a new dataset, the Adressa 20M dataset, that is 10 times larger and covers 10 weeks of traffic on www.adressa.no. Also, the raw logs of these compact datasets will be available as a raw data 2M and raw data 20M dataset at some later stage.

7 CONCLUSIONS

The Adressa dataset is one of the most comprehensive open datasets for training and evaluating recommender systems. As opposed to other datasets, it contains features that support not only collaborative filtering and content-based recommendation, but also advanced contextual strategies involving time and location. The full text of all news articles, with identified keywords and weighted entities, is also available on request.

This paper describes the compact 2M dataset. We are in the process of preparing a dataset with the same structure that is 10 times larger and will be made available on the same terms.

We believe that such public datasets are useful to the recommender system community. There are unfortunately very few datasets available for advanced hybrid recommender systems or context-aware recommender systems. Good datasets are needed to evaluate the effects of new and more advanced recommendation strategies. Hopefully, the Adressa dataset will encourage also others to publish datasets that enable us to develop solutions with better personalization experiences.

ACKNOWLEDGMENTS

This work was carried out as part of the industry-led research project RecTech, project number 245469, supported by the Research Council of Norway’s BIA innovation research program.

REFERENCES

[1] J. Bennet and S. Lanning. The Netflix Prize. In KDDCup’07, 2007.

[2] R. Bilton. The Washington Post tests personalized “pop-up” newsletters to promote its big stories. NiemanLab, 12 May 2016.

[3] T. Brasethvik and J. A. Gulla. A Conceptual Modeling Approach to Semantic Document Retrieval. In Proceedings of the 14th International Conference on Advanced Information Systems Engineering (CAiSE), pp. 167-182, 2002.

[4] G. Dror, N. Koenigstein, Y. Koren, and M. Weimer. The Yahoo! Music Dataset and KDD-Cup’11. In Proceedings of KDD-Cup 2011 Competition, pp. 3-18, 2012.

[5] J. A. Gulla, P. G. Auran, and K. M. Risvik. Linguistic Techniques in Large-Scale Search Engines. In Proceedings of the 6th International Conference on Applications of Natural Language to Information Systems (NLDB’02), pp. 218-222, 2002.

[6] J. A. Gulla, C. Marco, A. D. Fidjestøl, J. E. Ingvaldsen, and Ö. Özgöbek. The Intricacies of Time in News Recommendation. 4th Internationall workshop on News Recommendation and Analytics (INRA’16). UMAP Extended Proceedings, 2016.

[7] J. E. Ingvaldsen, Ö. Özgöbek, and J. A. Gulla. Context-Aware User-Driven News Recommendation. In Proceedings of the 3rd International Workshop on News Recommendation and Analytics (INRA), pp. 33-36. 2015.

[8] B. Kille, F. Hofpgartner, T. Brodt, and T. Heintz. The Plista Dataset. In Proceedings of the International News Recommender Systems Workshop and Challenge (NRS), pp. 16-23, 2013.

[9] H. Kvalheim. iTromsø: Norway’s first fully personalized mobile news site. WAN-IFRA, Vienna, 2016.

[10] J. Liu, P. Dolan, and E. R. Pedersen. Personalized news recommendation based on click behavior. In Proceedings of the 15th International Conference on Intelligent User Interfaces, pp. 31-40. ACM. 2010.

[11] F. Maxwell Harper and Joseph A. Konstan. 2015. The MovieLens Datasets: History and Context. ACM Transactions on Interactive Intelligent Systems (TiiS) 5, 4, Article 19 (December 2015), 19 pages.

[12] M. P. Robillard, W. Maalej, R. J. Walker and T. Zimmermann. Recommendation Systems in Software Engineering. Springer, 2014.

[13] G. Solskinnsbakk, and J. A. Gulla. Combining ontological profiles with context in information retrieval. Data & Knowledge Engineering, 69(3), pp. 251-260. 2010.

[14] M. Tavakolifard, J. A. Gulla, K. C. Almeroth, J. E. Ingvaldsen, G. Nygreen and E. Berg. Tailored news in the palm of your hand: a multi-perspective transparent approach to news recommendation. In Carr, Laender, Loscio, King, Fontoura, Vrandecic, Aroyo, de Oliveira, Lima, and Wilde (Eds.), Proceedings of the 22nd International World Wide Web Conference (WWW’2013) – Companion Volume, pp. 305-308. Rio de Janeiro, May 2013. ACM.

[15] Y. Zhang. Bayesian Gaphical Models for Adaptive Information Filtering. PhD thesis. Carnegie Mellon University. 2005.

[16] Ö. Özgöbek, J. A. Gulla, and R. C. Erdur. A Survey on Challenges and methods in News Recommendation. In Proceedings of the 10th International Conference on Web Information Systems and Technologies (WEBIST), Volume 2, pp. 278-285. 2014.