Information Retrieval Journal (2019) 22:159–187
https://doi.org/10.1007/s10791-018-9339-9

SOCIAL MEDIA FOR PERSONALIZATION AND SEARCH

Those were the days: learning to rank social media posts for reminiscence

Kaweh Djafari Naini[1] · Ricardo Kawase[1] · Nattiya Kanhabua[2] · Claudia Niederée[1] · Ismail Sengor Altingovde[3]

Received: 13 November 2017 / Accepted: 24 July 2018 / Published online: 11 August 2018

Abstract
Social media posts are a great source for life summaries aggregating activities, events, interactions and thoughts of the last months or years. They can be used for personal reminiscence as well as for keeping track with developments in the lives of not-so-close friends. One of the core challenges of automatically creating such summaries is to decide which posts are memorable, i.e., should be considered for retention and which ones to forget. To address this challenge, we design and conduct user evaluation studies and construct a corpus that captures human expectations towards content retention. We analyze this corpus to identify a small set of seed features that are most likely to characterize memorable posts. Next, we compile a broader set of features that are leveraged to build general and personalized machine-learning models to rank posts for retention. By applying feature selection, we identify a compact yet effective subset of these features. The models trained with the presented feature sets outperform the baseline models exploiting an intuitive set of temporal and social features.

Keywords Learning to rank · Letor · Social media · Personalization · Personalized ranking · Content retention · Social features · Feature selection · Facebook

1 Introduction

Human memory is very effective in keeping us focused on relevant things by forgetting irrelevant information. However, we also quickly forget the details of events or do not completely and/or correctly remember them. This is especially true for episodic memory (Tulving 2002), which is, roughly speaking, responsible for remembering the details of individual events. In episodic memory, the memories of new events interfere with older memories

This paper is an extension of an early version published as a short paper (Naini et al. 2014) and extends it in various ways, as discussed in Sect. 2. More recently, a preliminary/summary version of part of this work (based only on the smaller dataset and only with general ranking models) appeared in a *non peer-reviewed* venue, i.e., as a subsection of an invited book chapter (Niederée et al. 2018).

✉ Ismail Sengor Altingovde
altingovde@ceng.metu.edu.tr

Extended author information available on the last page of the article

as an effect of proactive interference (Underwood 1957). Furthermore, the memories of similar experiences blur into each other very easily, making it difficult to distinguish between the details of individual events (as an effect of retroactive interference McGeoch and McDonald 1931). Thus, the information collected over time in social media applications, such as, Facebook[1] can play an important role for complementing human memory: In the first place, it is created in near real-time and mainly for interaction, sharing and presenting oneself. However, if processed and presented in the right way, it can also be used to revive event memory and support reminiscence. As a foundation, this requires a selective approach for *retention*, which—similar to the focussing and selective role of forgetting in the human brain—helps to decide, which resources are expected to be important for future remembering and reminiscencing.

We are in an unprecedented situation where traces of everyday life and personal history is documented as a side effect of interacting with peers, no longer restricting life logging to major personal events or holidays. By documenting personal life, this information clearly constitutes an asset. Especially, the large volume of photos and videos created and shared by individuals today are considered a valuable part of personal remembrance (Kirk and Sellen 2010). In addition, recent work has shown the interest of users in using social media content for reminiscence and self-reflection as well as the potentials of social media content for this task. In (Zhao et al. 2013) for example, a study with Facebook users has discovered a considerable interest in managing a *personal region* for personal reminiscence and reflection about oneself. Facebook's own investment into its applications *Year in Review*,[2] which aggregates selected content from the past year into a video, and *On this Day*,[3] which presents a user her memories from that day in her Facebook history, highlights at least the expected user interest in this topic (as well as economic opportunities resulting from it).[4]

In the light of the above discussions, we believe that harvesting a personal history from the vast amount of data in social media applications arise as an important and interesting research question. Such summaries are not only useful for personal remembering: they also provide an important source for catching up with what happened in the lives of not-so-close friends (e.g., former class mates), whose activities we do not have time or interests to follow on a day-to-day basis.

Automatically creating social media summaries, which meet human expectations on what to remember and what to forget is, however, a challenging task (Kanhabua et al. 2013; Zhao et al. 2013). As the data involved in typical social media applications are in the form of posts (including text, video and/or audio) and interactions over such posts (such as likes, comments, shares, etc.), the key to create personal summaries automatically is deciding on the posts that need to be included in the summary, i.e., the *memorable* posts. Note, that in this paper we use the term "memorable" in the sense of "worth to be remembered or kept", not in the sense of "easy to remember".

Identifying or rather ranking such memorable posts (as a first step towards creating summaries) is exactly the research question we are adressing in this paper. Similar to the notion of "relevance" in information retrieval, it is not possible to exactly model the "memorable" as this is a partly subjective perception (and hence, a binary classification

[1] https://www.facebook.com.

[2] https://www.facebook.com/yearinreview/.

[3] https://www.facebook.com/onthisday/.

[4] see also https://research.fb.com/facebook-memories-the-research-behind-the-products-that-connect-you-with-your-past/.

model is not likely to be useful); yet one can approximate the notion starting from a broad set of features to rank a user's posts (just like in document retrieval), with the goal of having the most memorable ones at the top positions. Such a ranked list would not only allow browsing of a user's past posts starting from the most memorable ones and scrolling infinitely, if the user has the time and will, but also creating a personal summary from the top-ranked posts. Therefore, in this paper, we introduce learning to rank for retention as a novel research problem and investigate the following questions:

- What are the features that may characterize the memorable posts?
- Can we build general and personalized models for ranking users' posts?
- Can we identify a subset of features that allow building compact ranking models that are as effective as the ones employing all the available features?

Our contributions in this paper to address these questions are as follows:

- To investigate the first question, we designed user evaluation studies involving two complementing sets of participants: a small, yet known set of colleagues/friends (with 41 subjects) and a larger set of workers from a popular crowd-sourcing platform (with 470 subjects). In these studies, participants graded a subset of their own posts using a 5-point likert scale in terms of whether these posts are worth keeping for future needs or not. Using this unique data collection, we first conduct a primary data analysis to investigate to what extent a small set of intuitively chosen features can characterize the memorable posts. We find that the post type and interactions on the post (i.e., number of likes and comments), together with the age of the content, seem to be the best ad hoc evidence to identify the post that may be a candidate for retention.
- While our manual data analysis allowed us to detect a small set of seed features, machine-learning based approaches for similar tasks (say, ranking models for search engines) typically employ all potential features (e.g., up to hundreds or even thousands Macdonald et al. 2012; Yin et al. 2016) that can be extracted from the data, as a feature that is found to be less useful on its own can improve the overall performance of the model when combined with other features. Therefore, we also compiled a broad set of 111 features from our data collection to capture the factors that might influence the multi-faceted retention decisions of users. By leveraging these features, we build general and personalized machine-learning models for ranking memorable posts. Since there does not exist a baseline set of features in the literature for the novel task of ranking for retention, we use the most promising features from our data analysis to train a competitive baseline and compare our models against the latter.

  Our experiments reveal that general models outperform the models with the baseline features and provide relative improvements of up to 16.8 and 20.3%, in terms of the nDCG@5 and nDCG@10 metrics, respectively. Furthermore, the range of the effectiveness scores for these models (i.e., an nDCG score of up to 0.64) is reasonable in comparison to state-of-the-art performance in typical learning-to-rank settings (optimized for relevance); e.g., nDCG@10 is reported to be 0.49 and 0.78 for the Microsoft and Yahoo learning to rank datasets, respectively, in Gigli et al. (2016), and it is less than 0.60 for ranking tweets in Duan et al. (2010). This indicates that our approach in this paper, i.e., training models to rank social media posts for retention, is appropriate and achievable.

To train personalized models, we used the k-nearest neighbors of a user (as in Geng et al. 2008), and obtained moderate yet promising additional gains (i.e., up to another 2% relative improvement in nDCG@5) in certain cases.

- As our last contribution, we focus on feature selection in order to identify a compact yet effective set of the features that are most effective in ranking posts for retention. To this end, we apply a greedy feature selection method that is shown to perform well in learning-to-rank settings (Chelaru et al. 2014; Geng et al. 2007). We show that especially for the higher rank cut-offs, i.e., generating top-15 and -20 rankings of posts, the general models can be trained with a considerably smaller number of features (i.e., between 30 and 72 features instead of all 111) without any adverse effect on the effectiveness, i.e., nDCG scores, but even with occasional positive improvements.

The remainder of this paper is structured as follows. In Sect. 2, we discuss the related work. In Sect. 3, we describe the user evaluation studies and present our data analysis. Section 4 describes the candidate features for retention and Sect. 5 presents the ranking experiments and our main findings. Finally, in Sect. 6, we present our conclusions and their implications for future work.

2 Related work

2.1 Usage of Facebook

Nowadays, social media applications offer a wide variety of functionalities and are also used for very diverse purposes depending upon individual preferences (Joinson 2008; Spiliotopoulos and Oakley 2013). In Spiliotopoulos and Oakley (2013), for example, the authors investigate the motives for social media usage for the case of Facebook. The study focuses on a survey in which participants are asked about their motives for using Facebook based on a number of scenarios, which range from building social connection over photo viewing to inspecting other (also unknown) persons' profiles. In addition, the authors also investigate features for predicting the respective usage scenarios. In our work, we adopt part of the features suggested in this work for capturing usage behavior. Beyond these more expected ways of using social media applications, there are also studies such as Zhao et al. (2013), which show further, less obvious ways of using social media. In their work, Zhao et al. (2013) identify three regions of Facebook functionality, where the *personal region* is used for the management of personal data as a type of *personal locker*. However, the authors point out that due to the focus of Facebook on recent activities the management of data from the past and the transition of data into a personal region imposes several challenges for the user. These findings motivate our work on supporting reminiscence in social media taking a mid- to long-term perspective on content management in social media. To the best of our knowledge, there is no published research work on supporting individual reminiscence with automatically selected Facebook posts. There are, however, the aforementioned Facebook applications *Year in Review* and *On this Day*, which can be considered as first steps in this direction and also show the relevance of the topic.

In Lampe et al. (2008), the authors investigate the changes in usage behavior over time. They find out that changes in user behavior are rarely drastically, except if there are major changes in the functionalities of the applications, as it can, for example, be observed for the introduction of news feed in Facebook. Other works on Facebook focus on the relations

and social capital in Facebook (Ellison et al. 2011; Ellison et al. 2013), representing and measuring social interactions among Facebook users (Gomes and da Graça Campos Pimentel 2014), and modeling discussion threads in social media platforms including Facebook (e.g., see Aragón et al. 2017). In Bauer et al. (2013) the change of desired audience and emphasis of posts over time is investigated in two studies, also stressing the role of older Facebook posts for reminiscence. None of these earlier works address the characterization and ranking of Facebook posts for retention.

2.2 Recency in social media

Similar to traditional online websites (e.g news), social networks have to keep the users engaged on their platform in order to increase their revenues (Chakraborty et al. 2017). Therefore, many social- and professional networks such as Facebook, Twitter, and LinkedIn frequently provide the user with new information to keep them entertained and interested in their website. However, in Chakraborty et al. (2015) the authors show that it is important to also consider the users' information habits—in addition to popularity and recency—in order to avoid bias in coverage. Chakraborty et al. (2017) present an approach of recommending news stories based on trade-offs between recency and relevance by identifying the future impact of news using measures like number of likes and shares. Their results show that considering (estimated) future impact can achieve a good trade-off between recency and relevance for recommending news. Different from these works, we utilize temporal and social features, besides others, to rank user posts for retention.

2.3 Information value assessment

Our problem of identifying memorable posts can also be considered as a special information value assessment problem. Several valuation methods have been proposed, employing a rich variety of criteria. Many approaches take observed usage in the past as the main indication for information value, i.e., for the probability of future use (Chen 2005). A second set of information valuation method is based on time decay models, heavily used in the field of data streams (Cohen and Strauss 2006; Palpanas et al. 2004). An information value assessment approach for photo selection is discussed in Ceroni et al. (2015, 2017). Ceroni et al. present an expectation-orientated approach to support the user in automatically selecting important photos for long-term storage (preservation), reminiscence, and revisiting. In their work, the authors use a variety of item-level and collection-level features for ranking photos according to expected importance or future benefit (e.g., reminiscence) In our work, we also use expectation-oriented approach based on users annotation of their social network profile and use a learning-to-rank approach for filtering content.

2.4 Technology and its "memory" effects

A recent study (Sparrow et al. 2011) has shown that search technology, such as Google, affects human memory. Similarly, shared retrieval-induced forgetting in a social network can reshape the memories of speakers and listeners involved in a conversation, so-called collective memories (Coman and Hirst 2012). Typically, such studies shed light on how human remembering (and forgetting) interacts with technology use. This understanding can benefit the development of methods that aim at complementing the human ability to

remember and forget. This is also the case for our approach, which aims to support the remembering of events. A slightly different, but related approach is taken in two studies (Bowen and Petrelli 2011; Kalnikaité and Whittaker 2011), where possible "digital mementos" (as a digital counterpart of physical mementos) are investigated from a Human–Computer Interaction (HCI) perspective as a way of supporting or triggering the remembering of past events. Technology support for better remembering is also analysed in Crete-Nishihata et al. (2012): The authors show that what they call personal memory technologies can also help cognitively impaired persons to better remember or reconstruct the past. In this paper, by investigating features that characterize the memorable social media posts and building automatic rankers based on these features, our ultimate goal is to assist users on finding out which of their posts deserve to be retained for harvesting a personal history, and which others don't. This is aimed to support user's memory in past events.

2.5 Personal information management (PIM)

The growing amount of personal data brings social media applications closer to typical personal information management problems (Zhao et al. 2013). This also applies to our approach for considering the mid- to long-term perspective of social media usage. PIM tries to understand best practice of users in storing, retrieving, and (re-)using information and to develop new methods and tools for this purpose (Jones 2008). Originally, PIM mainly focused on information on a user's desktop (and on non-digital information) and was subsequently extended to also incorporate activities in the Web (e.g., for search Dumais et al. 2003). A promising direction is the Semantic Desktop (Sauermann et al. 2006) which introduces a personalized semantic layer on top of desktop objects. The requirements for long-term management of personal content, of which social media content is part of, are also considered in Marshall (2011), where the need for selecting content as well as the difficulty of this task is emphasized. Furthermore, the work on temporal organization of personal information in Knoll et al. (2009) is relevant for our final goal of creating life summaries, since it investigates time driven organization and visualization of personal information such as *Personal Narratives*.

2.6 Learning to rank (LETOR)

Modern search engines employ several features to obtain a ranking of web pages for a given query. In the last decade, this led to a new family of algorithms in the field of so-called learning to rank (LETOR), where automatic models are trained to effectively combine these large number of features. The three common categories of LETOR algorithms, namely, pointwise, pairwise and listwise, as well as a variety of methods under each category are discussed in detail by exhaustive surveys of Li (2011) and Liu (2011). Most recently, a detailed elaboration on the internals of ranking solutions in Yahoo! search engine demonstrates that such LETOR algorithms are actually the state of the art for generating rankings in large-scale commercial systems (Yin et al. 2016). Beyond ranking web documents, LETOR approaches are also applied to ranking problems in various domains, such as microblogs (e.g., see Berendsen et al. 2013; Duan et al. 2010), news documents (e.g., see Kanhabua and Nørvåg 2012), videos (e.g., see Chelaru et al. 2014). We refer the reader to the work of Liu (2011) for further application domains of LETOR approaches including (but not limited to) question answering, multimedia retrieval, text summarization and online advertising. In contrary to these previous works, we apply LETOR to rank

Facebook posts for retention using a unique dataset curated for capturing users preferences for memorable posts.

2.7 Our prior work

This paper is an extension of a previously published short paper (Naini et al. 2014) and extends it in various ways. Our previous work focused on a preliminary study with just 20 participants for gaining first insights regarding the features that are relevant for content retention in social media. In contrast, the work presented here is based on a much larger data basis, which enables a systematic feature analysis as well as automatic ranking of posts for retention. Finally note that, a preliminary/summary version of part of this work (based only on the smaller dataset and only with general ranking models) appeared in *a non peer-reviewed* venue, i.e., as a subsection of an invited book chapter (Niederée et al. 2018).

3 User evaluation study

We have performed a series of two evaluation studies based on Facebook data and a Facebook App. Those studies had a twofold purpose: a) we wanted to better understand, what are user's expectations towards the retention of their own content in Facebook from different time periods and b) we wanted to collect a groundtruth of memorable posts, which we can use in our later experiments. The first evaluation is an extension of a preliminary study that has been described in Naini et al. (2014). For a deeper understanding of user expectations we conducted a second evaluation including a larger number of users recruited via crowdsourcing. In this section, we first describe these evaluation studies and then provide an analysis of the collected data.

3.1 Setup and methodology

For encouraging and facilitating participation, we prepared an intuitive evaluation system in the form of a Facebook app. In order to participate, users have to log in with their Facebook credentials and grant the app the permissions to access some of their Facebook information, such a the profile, timeline, and friendship connections. After that, participants were presented with a running list of their posts.

Participants had to assess their posts using a 5-point Likert scale answering the following question: *Which posts do you think are relevant to and worth keeping for the future needs?* Note that, together with the latter question, we also provide the following context: "To facilitate your decision, imagine that you are 5 years in the future and you are looking back to your best moments on Facebook. What would you like to see?". By doing so, we ensure that we have correctly guided participants to annotate posts based on the aforementioned interpretations of *retention* and *memorable* (cf. Sect. 1), but not in the sense of *easy*

to remember or just *useful*; and hence, our annotation studies and resulting datasets are in line with the goal of this work.[5] Once a post is evaluated (with a rating from 0 (irrelevant) to 4 (extremely relevant)), it fades out providing space for further posts to scroll up. The evaluation interface of a single post contains information about its author, creation date, description, image, etc.

Using the above framework, we conducted two evaluation studies that essentially differ in the number of participants and the way they are selected, as described in the following.

3.1.1 Evaluation study-I

The first study was performed between the second week of November 2013 and the third week of February 2014. We had 41 participants, 24 males and 17 females, with age ranging from 23 to 39 years old. Participants were recruited through research communities, including colleagues from the authors' institutions, students, and their friends (and hence, we refer to collected data as the Lab dataset hereafter.) In this evaluation the participants were asked to judge about 100–200 of their posts (yet there were a few users who annotated less or more posts, which are all kept in the dataset). It is important to note that we are not judging participant's memory skills, but instead we are collecting their personal opinions regarding the retention preferences. Due to that, we presented participants' posts in a chronological order starting from the latest.

In total, the dataset includes 8494 evaluated posts, essentially covering the period from 2014 back to 2009 (detailed statistics will be presented later). Additionally, once the users provided us authorization to access their data on the Facebook platform, we were able to collect general statistics that help us to depict their use of Facebook social network. We believe that this first evaluation study, despite a relatively small number of participants, is still interesting and worthwhile since it is ensured to be based on real users with real profiles, i.e., does not include untrustworthy participants, as can happen in the more uncontrolled setup described next.

3.1.2 Evaluation study-II

In November 2014, we conducted a second evaluation with a larger number of participants from a popular crowdsourcing platform, CrowdFlower. The task for the workers and online evaluation system was the same as in the first evaluation. To begin the evaluation study, the workers had to follow a link to our system in the Human Intelligence Task (HIT) page at the crowdsourcing platform, and login with their Facebook account. Only those who had a Facebook account of at least 4 years old were allowed to participate (so that posts from a time span that is comparable to that of the first evaluation could be evaluated) and each worker had to evaluate at least 100 posts to complete the evaluation task. Each participant got 25 posts randomly selected from each year, from 2014 back to 2010. In cases where the users evaluated more than 100 posts or the Facebook profile of the user had less than 25 post for each year, they got older posts to evaluate. Overall, we ended up with the so-called Crowd dataset including 57,281 annotations from 470 users.

[5] While these annotation studies aim to capture user preferences for memorable posts, investigating the underlying reasons triggering such preferences is a very exciting question that needs inter-disciplinary research and is not in the scope of this paper.

Table 1 Basic statistics for the Lab and Crowd datasets

	Lab dataset	Crowd dataset
No. of users	41	470
No. of annotated posts	8494	57,281
Avg no. of annotated posts (per user)	207.170	121.874
Min no. of annotated posts	12	100
Max no. of annotated posts	1128	326
Female participants	17 (41%)	136 (29%)
Male participants	24 (59%)	334 (71%)
Age range of participants	23–39	18–65
Year of evaluation (duration)	2013 and 2014 (2 days)	2014 (5 days)

At the end of the evaluation task the participants were asked a few questions to collect personal information about their age, level of education, and country, in case that not all this information is available in their Facebook profile. After answering the questionnaire, the participant could complete the task by entering a code provided from our external evaluation website. On the average, the task was completed in 102 s. Note that, the pay per task was 5 cents, a reasonable amount for a simple task that does not require any background knowledge or skills and that took in average less than 2 min to complete. Further, it is worthy to mention that previous work (Mason and Watts 2009) has demonstrated that higher monetary incentives does not necessarily improve quality in crowdsourced tasks.

As untrustworthy workers are not unlikely in crowdsourcing platforms (e.g., Gadiraju et al. 2015) and assessing inter-rater agreement is not possible (as each participant should annotate only her own posts), we applied other measures to improve the quality of the collected data. In addition to enforcing the condition that each Facebook profile page has to be at least 4 years old, we also cross-checked information provided in the questionnaire against that in the participant's Facebook profile to identify untrustworthy profiles, i.e., those with contradictory information. Furthermore, we tracked the IP address of the users accessing our website in case one is accessing it with more than one Facebook account. After filtering data from such (potentially) untrustworthy workers (in total 54), we ended up with 470 participants. As before, the participants have allowed our application to access their profile information, timeline and their friendship graph on the Facebook platform.

Dealing with privacy issues: In both user evaluation studies, we took extra care regarding the participants' privacy and to comply with Facebook's Platform Policies.[6] It is declared and guaranteed that collected data will not be disclosed to third parties. Furthermore, the data cached represent the minimal amount of required information for the experiments. We emphasize that none of the experimental analysis and results presented in this paper include findings that may cause identifying a particular participant or her data. Especially, the IP address information is not used for tracking users for any kind of their personal activities; but it is used only and exactly once to detect malicious users that connected to our system from multiple accounts, and after filtering such users from the dataset, IP addresses are not used for any other purpose during our analysis. We also used adequate

[6] https://developers.facebook.com/policy/.

Table 2 Top-5 countries of the participants in the Crowd dataset

Country	Percentage
India	12.2
Philippines	8.1
Bulgaria	5.5
Venezuela	5.5
Italy	3.9
Others	64.8

Table 3 Educational level of the participants in the Crowd dataset

Education level	Percentage
Some high school (no diploma)	7.2
High school (diploma)	15.7
Some college (no degree)	18.2
BSc/MSc	44.1
Associate/professional/vocational/tec. degree	14.3
Others	13.5

Table 4 Number and percentage of the evaluated posts per year

Lab dataset	Year	No. of posts	Percentage
	≤ 2009	1140	13.42
	2010	1367	16.09
	2011	724	8.52
	2012	1657	19.51
	2013	3303	38.89
	2014	303	3.57
Crowd dataset	Year	No. of posts	Percentage
	≤ 2009	3514	6.13
	2010	7571	13.22
	2011	10,840	18.92
	2012	10,425	18.20
	2013	13,635	23.80
	2014	11,296	19.72

data protection mechanisms (including encrypted communication and password protection), while storing the data used in the analysis.

3.2 Evaluation results and data analysis

3.2.1 Basic statistics

In Table 1, we summarize the details of the datasets obtained from the first and second evaluation studies, namely, Lab and Crowd datasets, respectively. As expected, the Crowd dataset is not only larger but also much more diverse with respect to age and gender. We

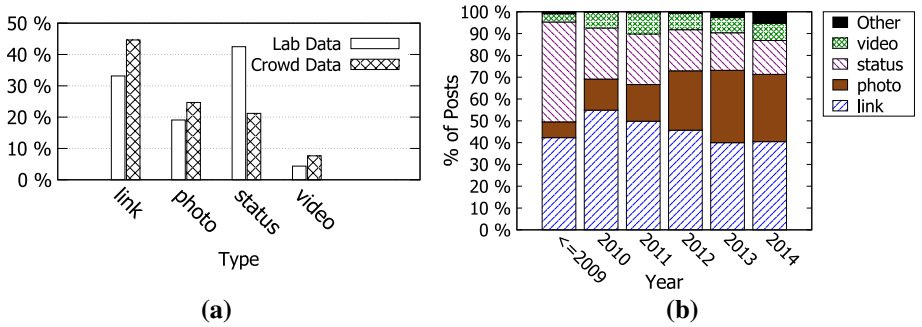

Fig. 1 **a** Percentage of post types for each dataset, **b** percentage of post types per year for the Crowd dataset

also observed considerable diversity for the country and education level of the participants in the latter dataset, as shown in Tables 2 and 3, respectively.

In Table 4, we provide the number of the annotated posts for each year from 2009 to 2014. For earlier years of 2007 and 2008 we don't have a large enough number of posts and hence, we aggregate them with those from 2009. We can observe that in the Lab dataset, there is an imbalance in the distribution of data annotated from each year, as the percentage of posts annotated per year varies from about 3–38% between 2009 and 2014. In contrast, our Crowd dataset seems to be more stable in this sense, especially between the years 2011–2014 (as the percentages are in the range of 18–23% for all years in this period).

At the time of our evaluation, Facebook had seven types of posts (namely, link, checkin, offer, photo, question, swf and video) that basically describes the type of content attached to a post. In Fig. 1a, we present the distribution of these types among the evaluated posts in our studies. In the Lab dataset, the most popular post type is status update (42.5%) followed by shared links (33.1%), photos (19%) and videos (4%). The second dataset, Crowd, has a slightly different distribution where posts of type shared link (44.4%) is the most popular and followed by photos (24.7%), status updates (21.1%) and videos (7%). Note that, in both datasets we disregard the other post types that are infrequent (i.e., less than 1%).

We also investigated the distribution of different post types over years, presented in Fig. 1b for the Crowd dataset. Our observation is that there is a clear increase in the use of photos and videos over time. The number of photos increased from 7% in 2009 to about 30% in 2014. For video we have an increase from 3 to 7% in 2014. These numbers are taken from our larger Crowd dataset, but we can observe a similar trend in the Lab dataset. Several factors help us to explain this change in behavior. First, the catch up of broadband connection allowed users to quickly upload large amounts of data (photos and videos). Second, the dissemination of smart phones with embedded cameras played an important role. Nowadays, anyone can quickly take a snapshot and upload it on the Web. Statistics from photo sharing website Flickr[7] show that the most used cameras are, by far, embedded smart phone cameras.[8] The rate of links and status information changes over years, however, there is no clear trend seen.

[7] http://www.flickr.com.

[8] http://www.flickr.com/cameras.

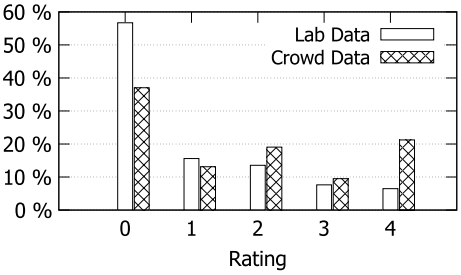

Fig. 2 Distribution of the user ratings for each dataset. The average of ratings is 0.92 for the Lab dataset and 1.65 for the Crowd dataset

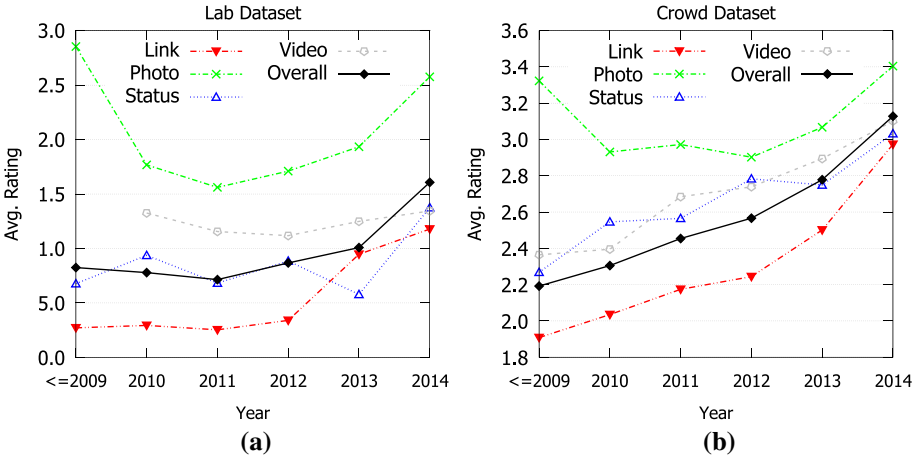

Fig. 3 Average rating of all posts per creation year (the solid black line) and average rating of posts for each content type per creation year (dashed lines)

3.2.2 Analysis of evaluation results

In this section, we present an analysis of the evaluation results and focus on a set of promising features (from the categories of social, temporal and network features, as will be described in the next section) that are most likely to be useful for identifying memorable posts.

We first take a look at the overall distribution of ratings in our datasets, shown in Fig. 2. We observe that in both datasets the portion of posts with rating 0 dominates with 57% for the Lab dataset and 37% for the Crowd dataset. In contrast, the fraction of posts that are given the highest rating is only 6 and 21% in the Lab and Crowd datasets, respectively. This indicates that participants consider a significant fraction of their posts worthless to retain for future, and justifies our work that aims to characterize this relatively small portion of posts, which are memorable, and generate rankings to present such posts at top.

Interstingly, the average of ratings differs considerably between the Lab Data Set (0,92) and the Crowd data set (1,65). Furthermore, Fig. 3 shows that this behavior is consistent over the years and content type—so the difference in ratings cannot be attributed to these parameters. Therefore, we think that this difference might be due to differences in participants' behaviour between the studies: First, as also reflected by the statistics in Table 1, in the Lab dataset there are a few users who made a large number of annotations (up to 5

times larger than the average number of annotations) and their possible bias (towards lower ratings) might have influenced the averages. Secondly, some of the users in the Lab dataset are more familiar annotation studies and, thus, might have interpreted the annotation goal (i.e., scoring posts based on their value to retain/remember for future) in a stricter sense and might have given lower ratings. From a reverse perspective, some workers involved in the Crowd dataset might have been more generous in assigning high ratings. We are aware that without additional after-study questionnaires with participants (which was not anymore possible for our study and still hard to attain for the crowd workers even at the time of collecting data), these discussions will not be conclusive; yet they can at least shed light to the possible causes of such observed differences.

We also analyzed the distribution of ratings wrt. the post types. We find that posts of type *photo* have the highest average rating, namely 1.93 and 3.10 for the Lab and Crowd dataset, respectively. In both datasets, *video* is the type with the second highest average rating (i.e., 1.27 for the Lab and 2.78 for the Crowd dataset). The average ratings of types *status update* and *link* are found to be considerably lower (especially for the Lab dataset), suggesting that posts with type *photo* or *video* are more likely to be memorable.

Content age for retention. Next, we focus on the role of time in deciding on content retention, i.e., whether older content on the average is rated lower than more recent content. For this purpose, we investigate the relationship between the post ratings and age of post. In Fig. 3, the solid line shows the average rating for the different years of content creation.

The figure reveals a clear trend where participants in the evaluation assigned higher ratings to more recent posts. This is in line with the idea of a decay function (as widely used in the field of data streams Cohen and Strauss 2006; Palpanas et al. 2004) underlying the content retention model. A strong decrease in ratings with growing age can especially be observed for the early years of the study (2013 and 2014). Surprisingly, we also see an increase in the rating values for the year 2009 in the figure, which we attempt to explain using a fine-grain analysis of ratings, i.e., per post type, in the following paragraph.

In Fig. 3, we see the trend for the average ratings for individual post types denoted with the dashed lines.[9] Once more, we observe an increase of ratings for the most recent posts. However, we also see very high average ratings for the oldest photos (older than 5 years). Thus, we conjecture that seeing these older (already forgotten) photos again caused some positive surprise for the users, which resulted in higher ratings. Indeed, this perception would also support the idea of creating Facebook summaries for reminiscence, yet we leave its verification (maybe via face-to-face participant interviews) as a future work. Note that the same trend (of rating more recent content as more worthy to retain) holds for both datasets, yet as shown in the figure, the Crowd dataset is exhibiting a smoother behaviour for different post types.

Finally, Fig. 4 demonstrates the same trend from a different perspective (given for only Crowd dataset for brevity). In this figure, the black line indicates the total percentage of posts considered as memorable (i.e., those with a rating greater than 0), which increase consistently over the years, while the red line shows the the percentage of posts rated with 0, exhibiting an opposite trend. Overall, these findings suggest that content age may serve as an important feature to identify memorable posts.

[9] For the Lab dataset, videos are shown for year 2010 and afterwards, as the number of videos before 2010 is very small in this dataset.

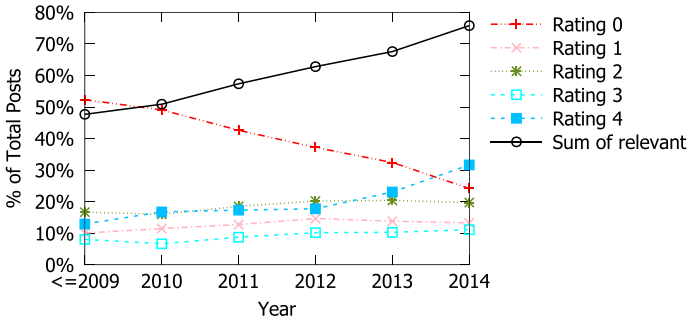

Fig. 4 Percentage of posts per rating for each year. The black line denotes the percentage of all the posts with a rating greater than 0 for each year (for the Crowd Dataset)

Number of likes and comments for retention. On Facebook, it is possible to *comment* for or *like* a particular post, as common forms of expressing community feedback. In our larger Crowd dataset, 70% (80%) of the posts lack any likes (comments), while 26% (18%) of the posts have between 1 and 10 likes (comments), respectively. Figure 5 reveals that for the posts with higher ratings, the average number of likes (comments), is also higher. This trend holds for both datasets, and also confirms our preliminary findings in Naini et al. (2014) that involved a smaller number of participants than those of the studies reported here. This indicates the robustness of this observation. Thus, the number of likes and comments seem to be among the crucial features to characterize the memorable posts.

Network features for retention. To better understand the importance of connections of the users involved (e.g., liked, commented, or tagged) in a post, we analyse for each post a set of network measures capturing two main effects. First, the relationship between the users' social graph and the users involved in a post. Here, our assumption is that posts involving more people from the user's friendship graph may have a higher probability of being relevant for retention than other posts with a few friends in their social graphs. To this end, we compute the feature *overlap of friends*, which is the ratio of the friends of a user to all people who are involved in a post. Secondly, we are interested in the relationships within the social graph of each post to identify differences in their users' connections. In this case, our assumption is that a high connectivity within the users involved in a post can lead to a higher chance that a post is considered relevant for future needs. To this end, we capture the graph connectivity by standard network measures, such as

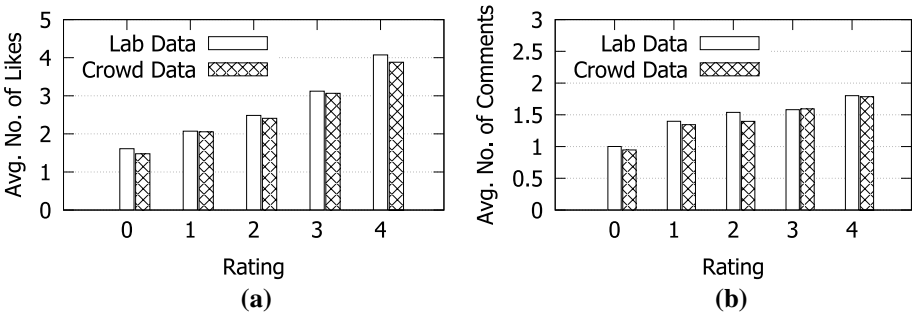

Fig. 5 Average number of **a** likes and **b** comments for the posts per rating

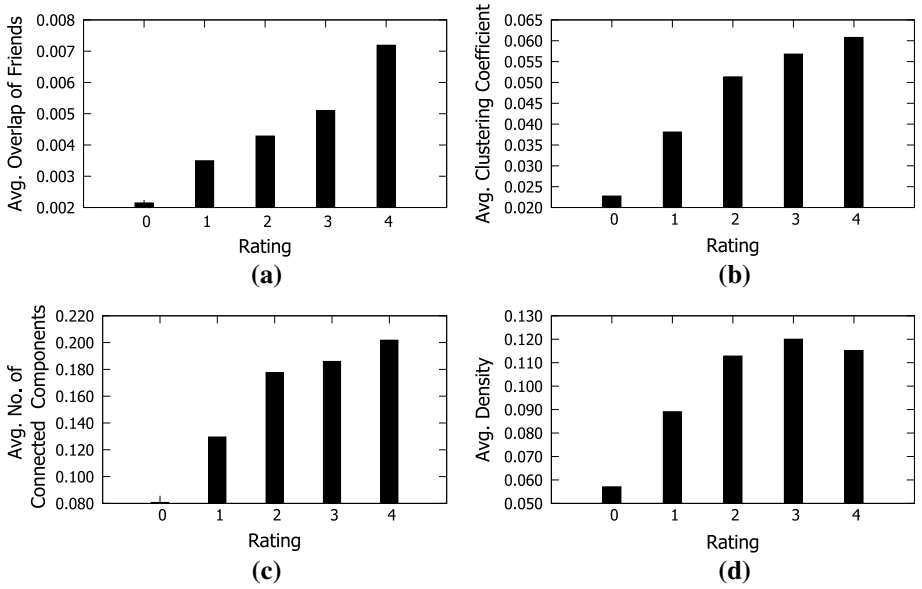

Fig. 6 Average feature score computed over the network of users who liked a post versus post rating, for the network features **a** overlap of friends, **b** clustering coefficient, **c** no. of connected components and **d** density

the *clustering coefficient* (Watts and Strogatz 1998), *number of connected components* (Tarjan 1972), and *density* (Coleman and Moré 1983).

In Fig. 6, we present the average values for these four network features over the posts for each rating (for the Crowd dataset). While computing these features for a given post, we only considered the users who *liked* the post, i.e., the *likes-network*. The figure shows that posts with higher ratings exhibit higher scores for these features, implying that such features may also be useful in identifying memorable posts.

Summary. From the previous statistics and analysis, we can deduce first ideas for determining features that have a high impact in the identification of memorable posts. Roughly speaking, recent photos and videos with high number of likes/comments and high overlap/connectivity within their social graphs of likes seem to be the best candidates for retention.

4 Candidate features for retention

Our data analysis presented in the previous section allows us to detect a small set of seed features. In this section, we will investigate a more comprehensive feature set, which is inspired by the seed feature set and features employed in related work. Machine-learning based approaches for tasks similar to ours typically employ all potentially useful features (e.g., thousands of features are used for training ranking models for search engines (Yin et al. 2016)) that can be extracted from the data, as a feature that is found to be less useful on its own can improve the overall performance of the model when combined with other features. Therefore, for capturing factors that might influence retention decisions of users, we compiled a broad set of 111 features.

Firstly, we adopted well known categories from document retrieval into our choice of feature classes such as network features, content-based, popularity-related (social) features and temporal features (see e.g. Chapelle and Chang 2011). Furthermore, we took inspiration from feature sets used in other works of social media analysis and information retrieval, e.g. Badache and Boughanem (2014, 2015b) using basic social media features, Chelaru et al. (2014) using features related to comments and likes for ranking Youtube videos, and Pantel et al. (2012) addresing the impact of *number of shares* and *likes* in search. The use of temporal features is inspired by the use of features of this class (e.g., Age) in other domains, such as IMDB movies (Badache and Boughanem 2015a, 2017), news (Kanhabua and Nørvåg 2012) and tweets (e.g., see Berendsen et al. 2013). Our use of privacy-related features is motivated by related work such as Krishnamurthy and Wills (2008), Liu et al. (2011).

The selected features can be categorized into five groups described as follows, while each feature is individually described in Table 5:

- *Temporal features* The inclusion of temporal features is inspired by the idea that retention preferences are influenced by a decay function as it was also confirmed by the data analysis in the previous section. For temporal features, we consider the temporal aspect of the post in terms of creation date, age, and lifetime. While *age* is the time between the evaluation and creation date, i.e., the time the post was created, *lifetime* is measuring the active time of a post starting at the time it was created to the last update. We also use variants of the age feature, which use the time of the last update and the time of the last commenting, respectively, instead of the creation time. Note that, an interesting direction that we have not explored in this work and left as future work is using fine-grain details, such as the *hour of the day* or *day of the week* (e.g., maybe posts from the weekends can be considered as more valuable to remember by the users).
- *Social features* The social features capture core signals of social interaction in a Social Web application, covering the features that are typically used in Facebook analysis: *number of likes*, *number of comments*, and *number of shares*.
- *Content-based features* We use the *type* of posts as well as some specific features extracted from the metadata of the post (as provided by Facebook) such as the *status type*, *hasLink*, *hasIcon*, and *app type*. To respect user privacy, the only text-based feature in our set is the length of text included in posts and comments. In other words, we do not utilize the textual content of the posts.
- *Privacy features* These are based on the privacy settings for a post that are specified by its owner to restrict the access of this post to a particular set of user.
- *Network features* Based on our analysis in the previous section, for each post we extract seven different network features as presented in Table 5. We compute these features from three different graphs for each post, namely, the graph of users who liked the post, graph of users who commented on the post, and graph of all users who liked, commented or tagged in the post. We employ the implementations of these features as provided by the Gephi project.[10,11]

[10] https://github.com/gephi.
[11] https://gephi.org/.

Table 5 List of candidate features for retention

Feature	Category	Description
No. of likes	Social	No. of people who like this post
No. of comments	Social	No. of comments on the post
No. of shares	Social	No. of shares of the post
No. of tagged users	Social	No. of users mentioned in a post
No. of likes on comments	Social	Total no. of likes included in the comments
Created-time	Temporal	The creation time of the post
Lifetime	Temporal	The time between creation and last update of the post
Age (creation time)	Temporal	The age of the post between the day of the evaluation and the time the post is created, updated or commented last time
Age (update time)	Temporal	
Age (last comment)	Temporal	
Privacy settings	Privacy	A privacy class setting for access to the post, e.g. *everyone*, *friends_of_friends*, *all_friends*, *custom*, *self*, *null*
No. of users (allowed)	Privacy	No. of the specific users or friends (in lists) who can see the post
No. of users (denied)	Privacy	No. of the specific users or friends (in lists) who are not allowed to see the post
Privacy friends	Privacy	No.of users (in a customized category) who can see the post, e.g. *some_friends*
Has description	Privacy	Post has a description of the privacy settings
Type of Post	Content-based	Type of a post with values such as *link*, *status*, *photo*, *video*, etc
Status Type of Post	Content-based	Description of the type of a status update. Values are *added_photos*, *added_video*, *created_group*, etc
Has message	Content-based	The post contain a status message
Has story	Content-based	Text from stories that are not intentionally generated by users, e.g., when someone else posts on the person's profile
Has description	Content-based	A description to a particular content, e.g., a website
Has link	Content-based	A link is attached to the content
Has caption	Content-based	The caption of a link is in the post
Has icon	Content-based	The post has a link to an icon representing the type of this post
Length of message	Content-based	The length of the corresponding textual section in the post

Table 5 (continued)

Feature	Category	Description
Length of story	Content-based	
Length of description	Content-based	
Length of comments	Content-based	Total length of the comments for the post
Send by mobile-APP	Content-based	Post includes information about the app that was used to publish it
Type of APP	Content-based	The type of the app the post was published by. Values are *mobile*, *others*, and *null*
Self posted	Content-based	Content is posted by the user herself
Self liked	Content-based	The post is liked by the user herself
Self commented	Content-based	The user commented on her post
Overlap of friends	Network	The ratio of the user's friends involved in the social graph created for a post
Clustering coefficient	Network	The clustering coefficient of the social graph created for a post
Degree	Network	The degree of the social graph created for a post
Connected components	Network	The no. of connected components in the social graph created for a post
Density	Network	The density of the social graph created for a post. A complete graph has all possible edges and density equal to 1
Diameter	Network	The diameter is the maximal distance in the social graph created for a post
Modularity	Network	Modularity of a post measures how each graph in its social graph is decomposed into modular communities

We also apply a personalized normalization to the social and network features to capture the individual characteristics and behavior of users more accurately. Furthermore, each categorical feature (like *type*) is mapped to multiple binary features (e.g., *type_IsLink*, *type_IsPhoto*, *type_IsStatus*, etc.). After these normalization and binarization steps, we end up with 111 features including 5 temporal, 8 social, 39 content-based, 13 privacy, and 46 network features. In Table 5, we provide a brief description for each feature.

5 Ranking posts for retention

In the previous section we have presented a number of candidate features that provide the foundation for developing a method for identifying memorable posts. We will use ranking for this purpose, thereby adopting strategies from the web search domain. This domain makes heavy use of machine learned rankers and they are also employed in commercial search engines (e.g., see Yin et al. 2016). Translating our setting into a search problem, we link a user to a query and a user's posts to documents retrieved by a query. For training, an m–dimensional feature vector F is constructed for each post p of a given user. This feature vector is enriched with the rating r, which the user had assigned to post p in the evaluation study. In the testing phase, pairs (u, F) of users u and post Feature vectors F are presented to the learned model and the model returns a ranked list of posts for u. For evaluating the performance of the model, we use Normalized Discounted Cumulative Gain (nDCG) (Järvelin and Kekäläinen 2002), as a typical metric from literature. This is a rank-sensitive metric, which considers graded labels. We report nDCG scores at the cut-off values of {5, 10, 15, 20}. In addition to general ranking models, we also investigate personalized ranking models for better understanding commonalities as well as personal differences in rentention preferences.

5.1 General ranking models with feature selection

In majority of our experiments, we employ a well-known algorithm for learning-to-rank, namely RankSVM (Joachims 2002) (as our experiments with other approaches—reported later in this section—did not yield any improvements over RankSVM). While building a model, instead of single data instances, RankSVM considers the pairs of instances (posts of a user, in our case). We apply *leave-one-out cross validation* for both of our datasets (i.e., in our case, each user serves once as the test instance for whom we evaluate the ranking, while all other users' annotations are used for training the model; and then evaluation scores are averaged over all users). Our choice of leave-one-out cross validation is based on the fact that the datasets used in our study are very hard to obtain (i.e., the participants do not only allow access to their posts, but they should also annotate them for retention, as the latter task is subjective and cannot be done by another person) and hence, they are not as large as the datasets used in other ranking scenarios.

For the Lab dataset, we use all 8,494 (posts) from 41 users, as described before. For the larger Crowd dataset, we randomly took 100 posts per user to avoid class imbalance (as there were some users who evaluated much more than 100 posts), which resulted in 47,000 posts for 470 users.

To the best of our knowledge, we are the first to propose ranking social media posts for retention, hence, in the literature, there does not exist a baseline set of features that is specified for our task. Therefore, we train two baseline models taking into account our findings

on features for retention from our data analysis presented in Sect. 3, and considering features that are found useful in other ranking scenarios.

In the first baseline, `Social`, we use basic social features, namely, the *number of likes*, *number of comments* and *number of shares* (and their versions normalized per user). We choose the latter features as they are the most intuitive popularity signals in social web and hence, likely to be involved in practical applications, such as the Facebooks apps discussed before.[12] Our data analysis has also yielded evidence that number of likes and comments can be useful for identifying memorable posts. Furthermore, the merit of these features are shown in other ranking scenarios: In particular, all three basic social features are employed in ranking of IMDB movies in Badache and Boughanem (2014, 2015b) while *number of comments* and *number of likes* are employed in Chelaru et al. (2014) for ranking Youtube videos, and the utility of the features *number of shares* and *likes* (beyonds others) for search are investigated in Pantel et al. (2012).

For the second baseline, `Social+Age`, in addition to basic social features we use a temporal feature, *age* (wrt. the creation time), to build our models, as this feature is again found very promising in Sect. 3. As in the previous case, temporal features, such as `Age`, are also shown to be useful for ranking in other domains, such as IMDB movies (Badache and Boughanem 2015a, 2017), news (Kanhabua and Nørvåg 2012) and tweets (e.g., Berendsen et al. 2013).

Figure 7 reveals the performance of RankSVM for ranking posts using all the proposed features for the Lab and Crowd datasets. As a first observation, we see that the baseline models differ in performance for the two datasets. The `Social` baseline performs better for the Lab while `Social+Age` baseline performs better for the Crowd dataset. It is a bit surprising that the temporal feature does not improve the results for the Lab dataset. However, as Fig. 3 shows, the relationship between post age (more specifically, creation year) and rating is much stronger for the Crowd set than the Lab dataset. In particular, for the Crowd set, there is a consistent increase of average ratings for more recent years, while the curve for the Lab dataset (especially for the years except the last one) is almost horizontal. This could explain the observed behaviour. Nevertheless, in the following, all the expressions claiming an improvement over a baseline refers to the baseline that performs better for the dataset in question.

Our results presented in Fig. 7 further show that the candidate features presented in Sect. 5 are actually very useful, and using all these features (denoted as `All`) for training a ranker yields relative effectiveness improvements of up to 9.21% (from an nDCG@5 score of 0.58–0.63) and 16.8% (from 0.52 to 0.61) over the baselines, for the Lab and the Crowd dataset, respectively. For the latter set, relative improvements in nDCG scores are even larger for the higher cut-off values of 10, 15 and 20; being 20.4, 22.9 and 26.2%, respectively.

Apart from the relative improvements over the intuitive baselines, we believe that the range of the effectiveness scores for our general models (i.e., an nDCG score of up to 0.64) is reasonable in comparison to state-of-the-art performance in typical learning-to-rank settings optimized for relevance. For instance, a recent work reports that over Microsoft and Yahoo challenge datasets (each with around 30K queries), a state-of-the-art ranker yields nDCG@10 scores of 0.49 and 0.78, respectively (Gigli et al. 2016). For ranking tweets, an

[12] For instance, while Facebook's "Year in Review" does not disclose how the content for each user is tailored, it is stated that the *number of mentions* in the posts is used to determine the top-10 topic list for the platform itself.

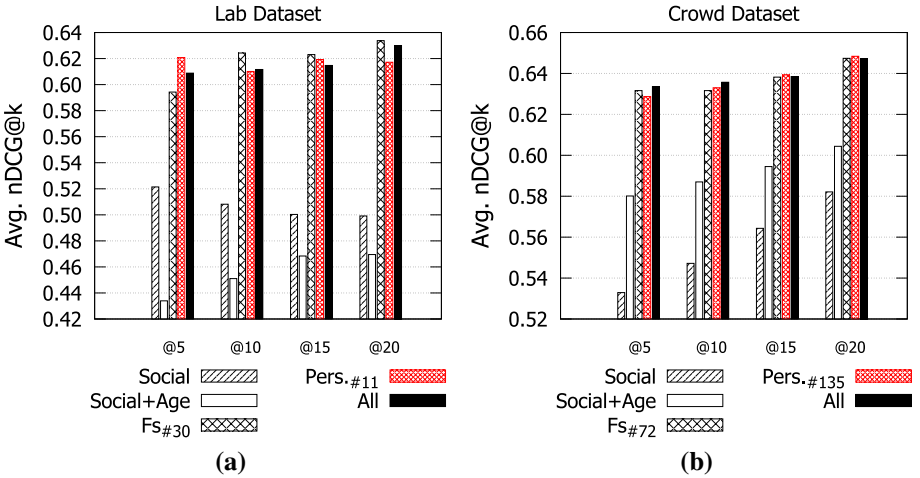

Fig. 7 Effectiveness of the ranking models for **a** Lab and **b** Crowd dataset. `Social` and `Social+Age` denote the baselines, `All` denotes the general ranking model with all features, FS_X denotes the general model with X features (after feature selection) and $Pers._K$ denotes the personalized model using K nearest neighbors of each user

approach again with RankSVM is shown to yield nDCG@10 scores less than 0.60 (Duan et al. 2010). This indicates that our approach in this paper, i.e., training models to rank social media posts for retention, is appropriate and effective.

Before proceeding with additional experiemnts to investigate the impact of features on the ranking effectiveness, at this point, we also experiment with other LETOR approaches that are employed in the RankLib[13] package to evaluate their performance for our task. In Fig. 8, we present nDCG@20 scores for four popular algorithms from the latter package, namely, RankNet (Burges et al. 2005), ListNet (Cao et al. 2007), AdaRank (Xu and Li 2007) and LambdaMART (Wu et al. 2010), versus RankSVM. For all algorithms, models are trained using all the available features for the Crowd dataset. For each method, we tuned their paramters to report the best-case performance.

Figure 8 reveals that RankNet and ListNet methods, which are based on neural networks,[14] and AdaRank are inferior to LambdaMART and RankSVM. Comparing the latter two, we observed that the quality of rankings are almost the same, and we decided to continue with our previous choice of providing results only for RankSVM in the rest of the paper.

After analyzing the ranking performance using all features, we investigate whether we can obtain the same performance using only some of these features, an experiment that also helps us to determine the most impactful features for the trained models. Features selection methods are often used in the context of classification tasks (e.g. Chang and Lin 2008), while there are a few works on applying them in the context of learning-to-rank (Dang and

[13] https://sourceforge.net/p/lemur/wiki/RankLib.

[14] While it may seem tempting to also apply deep learning approaches for this problem, our datasets are not large enough to allow such experiments, i.e., with multi-layer neural networks, and hence, this is left as a future work.

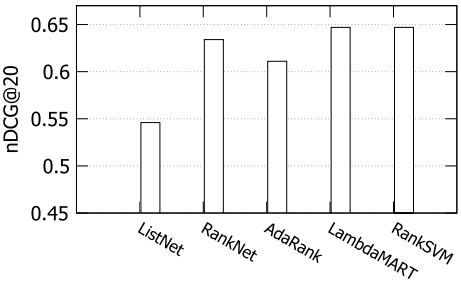

Fig. 8 Effectiveness of different LETOR approaches trained with all available features (for the Crowd dataset)

Croft 2010; Geng et al. 2007; Gigli et al. 2016; Naini and Altingovde 2014). In our work, we employ GAS (Greedy search Algorithm of Features Selection) by Geng et al. (2007). GAS not only computes the effectiveness of individual features, but it also considers their pairwise similarity. In more detail, we compute to which extent the top-20 ranking generated by two different features correlate. In this context the similarity of two ranked list is computed using Kendall's Tau metric. The feature selection is a greedy process: In each iteration, after selecting the feature with the highest score, the scores of all other features is discounted based on their similarity to the selected feature. The algorithm stops, when the desired number of features N is reached. In our case, we experiment with different values of N, considering all possible values from 1 to the total number of features (111).

Figure 7 also shows the results for the feature selection strategy with the best-performing value of N, which is found to be 30 (27% of all features) and 72 (64.9%) for the Lab and the Crowd datasets, respectively. Remarkably, although they are trained with a subset of all features, these smaller models still yield comparable (and sometimes, slightly better) effectiveness wrt. the models using all features, especially for the Crowd dataset.

For this latter experiment, we analyze the features selected by GAS in each fold (recall that we have leave-one-out cross validation) to identify the most promising features for the task of ranking posts. In particular, we obtained the rank of features in each fold and averaged these values to have a score for each feature. Then, we determined top-25 features with the highest scores for our Lab and Crowd datasets, separately.

In Table 6, we present the common features that appear among the top-25 features of both datasets (the rank column denotes the position of the feature in top-25 list for a given dataset). We observe that these 16 features fall into four of the categories described before, while no features from the privacy category could get into the list. It turns out that temporal features (along with their variants) and basic social features (no. of likes and comments) are among the most effective for the ranking models. In contrary to the results discussed for the `Social+Age` Baseline, temporal featrures are also in the top list for the Lab dataset in this analysis. We think that this might be due to the interaction of the temporal features with the other features: In particular, `Social+Age` baseline trains a ranker with four main types of features (*number of likes*, *number of comments*, *number of shares* and *age*) as well as some of their normalized variants (per user, etc.), which at the end still add up to less than 10 features. In contrast, for the top-25 features, we used the highest ranked features by the selection algorithm at each fold (as we use leave-one-out cross validation) and averaged the features scores. Thus, in each fold, the best-performing model could have used several features, average being 30 for the Lab dataset. That is, the performance of features in Table 6 is obtained when they are used together with 30 other features on the average, and we think that the interrelationships between these features may have increased the success of temporal features in comparison to the much smaller baseline model.

Table 6 The common features in the top-25 features computed for Lab and Crowd datasets (along with the feature's rank in each list)

Category	Feature	Rank in Lab dataset	Rank in Crowd dataset
Temporal	Age (created time)	1	7
Temporal	Created time	2	2
Temporal	Age (last updated time)	3	16
Temporal	Lifetime	4	12
Temporal	Age (last comment)	5	15
Social	No. of likes	9	17
Social	No. of comments	12	20
Social	Pers. No. of likes	24	22
Network	Overlap. No. of friends (all)	16	9
Network	Density (all)	17	18
Network	Pers. density (all)	7	14
Content-base	Type	6	21
Content-base	Pers. length message	8	4
Content-base	Length story	10	5
Content-base	Pers. length story	15	3
Content-base	Length description	22	23

In addition, there are network features computed over all the users involved in the post (i.e., those who liked, commented or tagged), as well as content-based features, namely, type and length of the post in the top-25 features. This list verifies our analysis presented in Sect. 3, and further demonstrates that it is helpful to have various variants of the same feature (e.g., normalized or computed in alternative ways), as a learning algorithm can benefit from all. Finally, some features (like the content length) that may not seem to be promising on its own at a first glance turn out to be useful when used in combination with others.

5.2 Personalized ranking models

So far, we considered a general ranking model learnt for all the users. However, in search domain, recent studies have shown that it is beneficial to build query-dependent ranking models, as queries significantly differ from each other (e.g., Can et al. 2014; Geng et al. 2008; Zhang et al. 2012). In particular, Geng et al. (2008) propose to use k-Nearest Neighbor (kNN) method so that for a given query first its nearest neighbors are found in the training set and then a customized ranker is learnt using only these neighbor instances. Analogously, in our setup, it is natural to hypothesize that similar users may have similar motivations and preferences while deciding on the memorable posts. Hence, we also apply a kNN based strategy to build more personalized ranking models.

We represent each user with a vector of three key features, namely, the *number of posts*, *number of friends*, and *number of connections* among the user's friends, which may reflect the coherence in the user's network. We anticipate that these user centric features best capture the activity level of a user in a social media application, and users with similar activity patterns can exhibit similar behavior while deciding on the memorable posts. To determine the nearest neighbors of a user, we compute the Euclidean distance between the pairs of these feature vectors, and choose the ones (k of them) that yield the smallest distances.

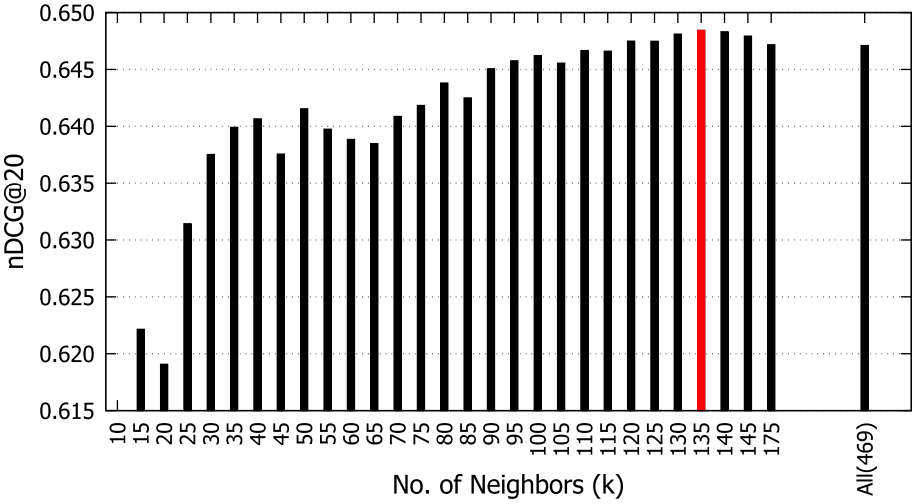

Fig. 9 Effectiveness of the personalized ranking model versus number of neighbors, k, for kNN (for the Crowd dataset)

Then, for each test user, only these k nearest neighbors (and their posts) are used to train the RankSVM algorithm.

In Fig. 9, we present the performance of personalized models versus k, the number of neighbors in kNN, which is in the range [1, 469] for the Crowd dataset (the trend for the Lab dataset is similar and not shown here for brevity). Note that, the figure does not include k values greater than 175, as the effectiveness score does not vary much after this point. The figure shows that training with very small number of neighbors (e.g., less than 25) may cause losses in the model effectiveness, and the best results are obtained for k =135.

In Fig. 7, we also present the performance for personalized ranking of posts.[15] As the best results are obtained when we set the number of neighbors k to 11 for the Lab dataset and to 135 for the Crowd dataset, we report only these cases. Our results are encouraging in that, for both datasets, the personalized approach can provide gains in comparison to using a general ranking model for various cut-off values (cf., compare the fourth and last bars in Fig. 7 for each cut-off value). Most remarkably, for the Lab dataset, while nDCG@5 score is 0.608 for the general model using all features, personalized model achieves a score of 0.620, providing a relative improvement of about 2%. We envision that in a real setup where millions of users exist with different habits of interacting with the social applications, the idea of building ranking models customized for individual users might improve the effectiveness even more.

Finally, we also experimented with feature selection in the personalized setup, where we applied the GAS strategy for the k nearest neighbors of each user. It turns out that, feature selection diminishes the benefits obtained by building personalized rankers and hence, we do not provide results for this experiment. Given that the training is already restricted to a

[15] While we regret to make the reader refer to back to check this figure, we preferred to present the performance of all ranking models in a single figure for the sake of comparability and brevity.

small set of neighbors, we conclude that it may not pay off to apply feature selection when we aim to build specific models per user.

Note that, a final concern for building personalized ranking models could be efficiency. In the case of the web search, training a model for each query can imply prohibitive online processing costs, as the users typically expect search results in less than a second (Geng et al. 2008). However, in our case, this would be less of a concern; as ranking the posts for retention is not an everyday task for a user, but an application that is most likely to be executed periodically, such as de-fragmenting your hard-drive. Hence, the additional processing latency for online model building can be tolerated by the users, for the promise of a better final ranking. Furthermore, it is still possible to improve the efficiency using offline pre-processing techniques, such as clustering, as proposed in an earlier work (Geng et al. 2008). Thus, both from the effectiveness and efficiency perspectives, we conclude that building personalized ranking models for retention arises as a promising direction.

5.2.1 Summary

Our experiments presented in this section show that general models trained with 111 candidate features yield reasonable effectiveness (nDCG scores over 0.61 for all cut-off values and datasets) and outperform intuitive baselines (using social and temporal features) with a large margin (up to 26%) for ranking posts for retention. We also demonstrated that these general models can be made more compact by feature selection, and even after this, the performance is comparable to the models using all the features. Finally, we built personalized ranking models that can provide a relative improvement of about 2% over the general models.

6 Conclusions

In this article, we lay the foundations towards the creation of life summaries from a social media platform, Facebook. This is a non-trivial challenge that requires accurate ranking of memorable posts, i.e. posts worth remembering, in a user's timeline. In order to address this challenge, one first needs to assess users' perception of what is important for retention in a social platform. To this end, we conducted two user evaluation studies: The first study involved 41 participants from the research communities and yielded 8494 annotated posts, while the second study involved 470 participants recruited from a crowdsourcing platform and yielded 57,281 annotated posts.

On this invaluable corpus, we conducted a primary data analysis and identified a small set of seed features that are most likely to characterize memorable posts. Next, leveraging a broader set of candidate features extracted for each annotated post, we trained both general and personalized models to rank the posts. These rankers are effective, as they can outperform a practical baseline that employ the most intuitive features identified during our data analysis, and as they yield effectiveness scores comparable to the recent works that again employ machine-learnt ranking models for a different yet related purpose, namely, traditional document retrieval. A question that still remains open for exploration is whether it is possible to further increase the effectiveness of the rankers by taking into account the textual content of the posts, which lies in a grey area involving hot debates on user privacy issues.

In our experiments, by applying feature selection, we could identify a compact set of features that captures the most discriminative representatives of different feature categories as we define here (namely, content-based, temporal, social, network, and privacy), and yield ranking models that are as effective as those with all the features. This is also valuable, not only for building models more efficiently in large scale systems, but also for figuring out the directions we need to concentrate in future user studies for a more fine-grained understanding of the human retention preferences in social media applications.

In summary, in this paper we essentially show that building models for automatic ranking of posts for retention is an attainable task. In addition, we identify a large number of features that are useful for this task and show that certain subsets of these features can be equally effective. Due to the subjectivity of the task, those subsets are differing for different datasets, which are likely to represent individual groups of users with different preferences for retention. Motivated by this finding, we build personalized ranking models and show that their performance is encouraging. This latter finding also indicates an exciting direction that calls for deeper research, namely, building personalized models that take into account a richer set of user characteristics, which can benefit from interdisciplinary research about human behaviour.

In our future work, we also plan to address grouping of related posts of a user for structuring the information space and develop effective ways of generating concise and diverse summaries over such groups of posts for retention. Another promising direction for future work is to evaluate the learned models for retention "in the wild", i.e., to find participants for evaluating the model on their own posts and to further refine the model based on the evaluation results.

Funding I.S. Altingovde is supported by Turkish Academy of Sciences Distinguished Young Scientist Award (TUBA-GEBIP 2016). This work was partially funded by the DFG Project "Managed Forgetting" (Contract Number NI-1760/1-1).

References

Aragón, P., Gómez, V., García, D., & Kaltenbrunner, A. (2017). Generative models of online discussion threads: State of the art and research challenges. *Journal of Internet Services and Applications*, *8*(1), 15:1–15:17.

Badache, I. & Boughanem, M. (2014). Social priors to estimate relevance of a resource. In *Fifth information interaction in context symposium, IIiX '14*, Regensburg, Germany, August 26–29, 2014 (pp. 106–114).

Badache, I., & Boughanem, M. (2015a). Document priors based on time-sensitive social signals. In *Advances in information retrieval: 37th European conference on IR research, ECIR 2015*, Vienna, Austria, March 29–April 2, 2015 (pp. 617–622).

Badache, I., & Boughanem, M. (2015b). A priori relevance based on quality and diversity of social signals. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, Santiago, Chile, August 9–13, 2015 (pp. 731–734).

Badache, I., & Boughanem, M. (2017). Fresh and diverse social signals: Any impacts on search? In *Proceedings of the 2017 conference on conference human information interaction and retrieval, CHIIR 2017*, Oslo, Norway, March 7–11, 2017 (pp. 155–164).

Bauer, L., Cranor, L. F., Komanduri, S., Mazurek, M. L., Reiter, M. K., Sleeper, M., & Ur, B. (2013). The post anachronism: The temporal dimension of facebook privacy. In *Proceedings of the 12th ACM workshop on privacy in the electronic society, WPES '13* (pp. 1–12). ACM: New York.

Berendsen, R., Tsagkias, M., Weerkamp, W., & de Rijke, M. (2013). Pseudo test collections for training and tuning microblog rankers. In *The 36th international ACM SIGIR conference on research and development in information retrieval, SIGIR '13*, Dublin, Ireland, July 28–August 01, 2013 (pp. 53–62).

Bowen, S., & Petrelli, D. (2011). Remembering today tomorrow: Exploring the human-centred design of digital mementos. *International Journal of Human-Computer Studies*, *69*(5), 324–337.

Burges, C., Shaked, T., Renshaw, E., Lazier, A., Deeds, M., Hamilton, N., & Hullender, G. (2005). Learning to rank using gradient descent. In *Proceedings of the 22nd international conference on machine learning* (pp. 89–96). ACM.

Can, E. F., Croft, W. B., & Manmatha, R. (2014). Incorporating query-specific feedback into learning-to-rank models. In *Proceedings of SIGIR '14* (pp. 1035–1038).

Cao, Z., Qin, T., Liu, T.-Y., Tsai, M.-F., & Li, H. (2007). Learning to rank: From pairwise approach to listwise approach. In *ICML*.

Ceroni, A., Solachidis, V., Niederée, C., Papadopoulou, O., Kanhabua, N., & Mezaris, V. (2015). To keep or not to keep: An expectation-oriented photo selection method for personal photo collections. In *Proceedings of ICMR '15* (pp. 187–194).

Ceroni, A., Solachidis, V., Niederée, C., Papadopoulou, O., & Mezaris, V. (2017). Expo: An expectation-oriented system for selecting important photos from personal collections. In *Proceedings of ICMR '17* (pp. 452–456).

Chakraborty, A., Ghosh, S., Ganguly, N., & Gummadi, K. P. (2015). Can trending news stories create coverage bias? on the impact of high content churn in online news media. In *Computation and journalism symposium*.

Chakraborty, A., Ghosh, S., Ganguly, N., & Gummadi, K. P. (2017). Optimizing the recency-relevancy trade-off in online news recommendations. In *Proceedings of WWW '17* (pp. 837–846).

Chang, Y.-W., & Lin, C.-J. (2008). Feature ranking using linear SVM. In *Proceedings of WCCI causation and prediction challenge* (pp. 53–64).

Chapelle, O., & Chang, Y. (2011). Yahoo! learning to rank challenge overview. In Chapelle, O., Chang, Y., Liu, T.-Y. (Eds.), *Proceedings of the learning to rank challenge, volume 14 of proceedings of machine learning research* (pp. 1–24). Haifa: PMLR.

Chelaru, S., Orellana-Rodriguez, C., & Altingovde, I. S. (2014). How useful is social feedback for learning to rank youtube videos? *World Wide Web*, *17*(5), 997.

Chen, Y. (2005). Information valuation for information lifecycle management. In *Proceedings of international conference on autonomic computing*.

Cohen, E., & Strauss, M. J. (2006). Maintaining time-decaying stream aggregates. *Journal of Algorithms*, *59*(1), 19–36.

Coleman, T. F., & Moré, J. J. (1983). Estimation of sparse jacobian matrices and graph coloring blems. *SIAM Journal on Numerical Analysis*, *20*(1), 187–209.

Coman, A., & Hirst, W. (2012). Cognition through a social network: The propagation of induced forgetting and practice effects. *Journal of Experimental Psychology: General*, *141*(2), 321–36.

Crete-Nishihata, M., Baecker, R. M., Massimi, M., Ptak, D., Campigotto, R., Kaufman, L. D., et al. (2012). Reconstructing the past: Personal memory technologies are not just personal and not just for memory. *Human-Computer Interaction*, *27*(1–2), 92–123.

Dang, V., & Croft, W. B. (2010). Feature selection for document ranking using best first search and coordinate ascent. In *Proceedings of SIGIR'10 workshop on feature generation and selection for information retrieval*.

Duan, Y., Jiang, L., Qin, T., Zhou, M., & Shum, H. (2010). An empirical study on learning to rank of tweets. In *Proceedings of COLING '10* (pp. 295–303).

Dumais, S., Cutrell, E., Cadiz, J., Jancke, G., Sarin, R., & Robbins, D. C. (2003). Stuff i've seen: A system for personal information retrieval and re-use. In: *SIGIR '03* (pp. 72–79).

Ellison, N. B., Gray, R., Vitak, J., Lampe, C., & Fiore, A. T. (2013). Calling all facebook friends: Exploring requests for help on facebook. In *Proceedings of ICWSM '13*.

Ellison, N. B., Steinfield, C., & Lampe, C. (2011). Connection strategies: Social capital implications of facebook-enabled communication practices. *New Media & Society*, *13*(6), 873–892.

Gadiraju, U., Kawase, R., Dietze, S., & Demartini, G. (2015). Understanding malicious behavior in crowdsourcing platforms: The case of online surveys. In *Proceedings of CHI '15* (pp. 1631–1640).

Geng, X., Liu, T.-Y., Qin, T., Arnold, A., Li, H., & Shum, H.-Y. (2008). Query dependent ranking using k-nearest neighbor. In *Proceedings of SIGIR'08* (pp. 115–122).

Geng, X., Liu, T.-Y., Qin, T., & Li, H. (2007). Feature selection for ranking. In *Proceedings of SIGIR'07* (pp. 407–414).

Gigli, A., Lucchese, C., Nardini, F. M., & Perego, R. (2016). Fast feature selection for learning to rank. In *Proceedings of ICTIR '16* (pp. 167–170).

Gomes, A. K., & da Graça Campos Pimentel, M. (2014). Evaluation of media-based social interactions: Linking collective actions to media types, applications, and devices in social networks. In N. Agarwal, M. Lim, & R. Wigand (Eds.), *Online Collective Action* (pp. 75–95). Vienna: Springer.

Järvelin, K., & Kekäläinen, J. (2002). Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, *20*(4), 422–446.

Joachims, T. (2002). Optimizing search engines using clickthrough data. In *Proceedings of KDD'02* (pp. 133–142).

Joinson, A. N. (2008). Looking at, looking up or keeping up with people?: Motives and use of facebook. In *Proceedings of CHI '08*.

Jones, W. (2008). *Keeping found things found: The study and practice of personal information management*. San Francisco: Morgan Kaufmann Publishers Inc.

Kalnikaité, V., & Whittaker, S. (2011). A saunter down memory lane: Digital reflection on personal mementos. *International Journal of Human-Computer Studies*, *69*(5), 298–310.

Kanhabua, N., Niederée, C., & Siberski, W. (2013). Towards concise preservation by managed forgetting: Research issues and case study. In *Proceedings of the 10th international conference on preservation of digital objects, iPres '13*.

Kanhabua, N., & Nørvåg, K. (2012). Learning to rank search results for time-sensitive queries. In *21st ACM international conference on information and knowledge management, CIKM'12*, Maui, HI, USA, October 29–November 02, 2012 (pp. 2463–2466).

Kirk, D. S., & Sellen, A. (2010). On human remains: Values and practice in the home archiving of cherished objects. *ACM Transactions on Computer-Human Interaction*, *17*(3), 10:1–10:43.

Knoll, S., Hoff, A., Fisher, D., Dumais, S., & Cutrell, E. (2009). Viewing personal data over time. In *Proceedings of CHI'2009 workshop on interacting with temporal data*.

Krishnamurthy, B., & Wills, C. E. (2008). Characterizing privacy in online social networks. In *Proceedings of the first workshop on online social networks, WOSN '08* (pp. 37–42). ACM: New York.

Lampe, C., Ellison, N. B., & Steinfield, C. (2008). Changes in use and perception of facebook. In *Proceedings of CSCW '08*.

Li, H. (2011). Learning to rank for information retrieval and natural language processing. In: *Synthesis lectures on human language technologies*. Morgan and Claypool Publishers.

Liu, T. (2011). *Learning to rank for information retrieval*. Berlin: Springer.

Liu, Y., Gummadi, K. P., Krishnamurthy, B., & Mislove, A. (2011). Analyzing facebook privacy settings: User expectations vs. reality. In *Proceedings of the 2011 ACM SIGCOMM conference on internet measurement conference, IMC '11* (pp. 61–70). ACM: New York.

Macdonald, C., Santos, R. L. T., & Ounis, I. (2012). On the usefulness of query features for learning to rank. In *Proceedings of CIKM '12* (pp. 2559–2562).

Marshall, C. C. (2011). Challenges and opportunities for personal digital archiving. In C. A. Lee (Ed.), *I, Digital: Personal Collections in the Digital Era* (pp. 90–114). Chicago: Society of American Archivists.

Mason, W. A., & Watts, D. J. (2009). Financial incentives and the "performance of crowds". *SIGKDD Explorations*, *11*(2), 100–108.

McGeoch, J. A., & McDonald, W. T. (1931). Meaningful relation and retroactive inhibition. *American Journal of Psychology*, *43*(4), 579–588.

Naini, K. D., & Altingovde, I. S. (2014). Exploiting result diversification methods for feature selection in learning to rank. In *Proceedings of ECIR'14* (pp. 455–461).

Naini, K. D., Kawase, R., Kanhabua, N., & Niederée, C. (2014). Characterizing high-impact features for content retention in social web applications. In *Proceedings of WWW (Companion Volume)* (pp. 559–560).

Niederée, C., Kanhabua, N., Tran, T., & Naini, K. D. (2018). Preservation value and managed forgetting. In V. Mezaris, C. Niederée, & R. H. Logie (Eds.), *Personal Multimedia Preservation: Remembering or Forgetting Images and Video* (pp. 101–129). Cham: Springer.

Palpanas, T., Vlachos, M., Keogh, E., Gunopulos, D., & Truppel, W. (2004). Online amnesic approximation of streaming time series. In *Proceedings of ICDE '04*.

Pantel, P., Gamon, M., Alonso, O., & Haas, K. (2012). Social annotations: Utility and prediction modeling. In *The 35th International ACM SIGIR conference on research and development in information retrieval, SIGIR '12*, Portland, OR, USA, August 12–16, 2012 (pp. 285–294).

Sauermann, L., Dengel, A., Elst, L. V., Lauer, A., & Schwarz, M. S. (2006). Personalization in the EPOS project. In *Proceedings of ESWC* (pp. 42–52).

Sparrow, B., Liu, J., & Wegner, D. M. (2011). Google effects on memory: Cognitive consequences of having information at our fingertips. *Science*, *333*, 776–778.

Spiliotopoulos, T., & Oakley, I. (2013). Understanding motivations for facebook use: Usage metrics, network structure, and privacy. In *Proceedings of CHI '13*.

Tarjan, R. (1972). Depth-first search and linear graph algorithms. *SIAM Journal on Computing*, *1*(2), 146–160.

Tulving, E. (2002). Episodic memory: From mind to brain. *Annual Review of Psychology*, *53*(1), 1–25.
Underwood, B. (1957). Interference and forgetting. *Psychological Review*, *64*(1), 49–60.
Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of small-worldnetworks. *Nature*, *393*(6684), 440–442.
Wu, Q., Burges, C. J., Svore, K. M., & Gao, J. (2010). Adapting bboosting for information retrieval measures. *Information Retrieval*, *13*, 254–270.
Xu, J., & Li, H. (2007). Adarank: A boosting algorithm for information retrieval. In *SIGIR*.
Yin, D., Hu, Y., Tang, J., Jr., T. D., Zhou, M., Ouyang, H., Chen, J., Kang, C., Deng, H., Nobata, C., Langlois, J., & Chang, Y. (2016). Ranking relevance in yahoo search. In *Proceedings of KDD '16* (pp. 323–332).
Zhang, X., He, B., Luo, T., & Li, B. (2012). Query-biased learning to rank for real-time twitter search. In *Proceedings of CIKM '12* (pp. 1915–1919).
Zhao, X., Salehi, N., Naranjit, S., Alwaalan, S., Voida, S., & Cosley, D. (2013). The many faces of facebook: Experiencing social media as performance, exhibition, and personal archive. In *Proceedings of CHI '13*.

Affiliations

Kaweh Djafari Naini[1] · Ricardo Kawase[1] · Nattiya Kanhabua[2] · Claudia Niederée[1] · Ismail Sengor Altingovde[3]

Kaweh Djafari Naini
naini@l3s.de

Ricardo Kawase
kawase@l3s.de

Nattiya Kanhabua
nkanhabua@ntent.com

Claudia Niederée
niederee@l3s.de

[1] L3S Research Center, Appelstr. 9a, 30167 Hannover, Germany

[2] NTENT Inc., Barcelona, Spain

[3] Computer Engineering Department, Middle East Technical University, 06580 Ankara, Turkey