Information Sciences 385–386 (2017) 125–137

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Exploiting search history of users for news personalization ☆

Xiao Bai [a], B. Barla Cambazoglu [b], Francesco Gullo [c,*], Amin Mantrach [a], Fabrizio Silvestri [d]

[a] Yahoo Research, Sannyvale, CA, USA
[b] Independent Researcher, Barcelona, Spain
[c] UniCredit, R&D Dept., Rome, Italy
[d] Facebook, London, UK

ARTICLE INFO

Article history:
Received 22 February 2016
Revised 23 October 2016
Accepted 30 December 2016
Available online 30 December 2016

Keywords:
Online news
Personalization
User profiling
Query logs

ABSTRACT

Content personalization is a long-standing problem for online news services. In most personalization approaches users of a news service are represented by topical interest profiles that are matched with news articles in order to properly decide which articles are to be recommended. When constructing user profiles, existing personalization methods exploit the user activity observed within the news service itself without incorporating additional information that can be obtained from other sources.

In this paper we study the problem of news personalization by leveraging usage information that is external to the news service. We propose a novel approach that relies on the concept of "search profiles", which are user profiles that are built based on the past interactions of the user with a web search engine. We extensively test our proposal on real-world datasets obtained from Yahoo. We explore various dimensions and granularities at which search profiles can be built. Experimental results show that, compared to a basic strategy that does not exploit the search activity of users, our approach is able to boost the clicks on news articles shown at the top positions of a ranked result list.

1. Introduction

Online news services have dramatically changed the way people access information. Nowadays, the Web has plenty of news sites. While this large amount of resources provides a fruitful source of information for journalists or other professionals, it may create a problem for normal end users who typically want to reach the desired pieces of information as quickly as possible.

A number of today's online new services, such as Google News and Yahoo News, aim at aggregating different news sources and presenting them to their end users in an organic way. During a session on these news aggregators, users expect to be provided with content that they consider relevant, useful, or interesting. Since every single user has her own set of interests, *personalization* of presented news results becomes an important requirement.

Personalization of a news service is a long-standing challenge. Traditional approaches consist of ranking news articles based on how well they match the user's interests [4,7,17–19,22,26,27,32,35]. Inferring the interests of a specific user (i.e.,

☆ Most of the work was done while all the authors were affiliated with Yahoo Labs, Barcelona, Spain.

* Corresponding author.
E-mail addresses: xbai@yahoo-inc.com (X. Bai), barla@berkantbarlacambazoglu.com (B. Barla Cambazoglu), gullof@acm.org (F. Gullo), amantrac@yahoo-inc.com (A. Mantrach), fsilvestri@fb.com (F. Silvestri).

http://dx.doi.org/10.1016/j.ins.2016.12.038
0020-0255/

building a *user profile*) is a critical aspect that heavily affects the quality of a news personalization system. While earlier systems explicitly asked users to specify their profiles [8,42], it is common today to develop automated user-profiling strategies that do not require any manual effort on the user side [4,19,22].

One of the most valuable information sources used to automatically build user profiles is the online behavior exhibited by users during their interaction with online services. In general, the online behavior can be obtained from endogenous or exogenous sources. In the context of news personalization, endogenous information refers to the interaction of users with the news service itself (e.g., news articles they have read in the past), while exogenous information consists of the user activity that is performed on services other than the news service.

In most existing news personalization systems user profiles are built using endogenous information [13,30,31,34,38]. The rationale is that a news article read by a user represents a clear evidence of her interests. While endogenous information is undoubtedly the most reliable source in automatically discovering user interests, it may not tell us the whole story about the user. Indeed, most users interact with several online services, each serving a different purpose. Hence, it is not uncommon that the interaction with a service reflects user interests that are related to that specific service only and, as such, cannot be unveiled by other services. This means that user interests arising only from endogenous information may correspond to a limited portion of the overall user interests. In this context, exogenous information constitutes a precious source of additional knowledge to complete user profiles and, as such, improve the quality of a news personalization system.

As an example, consider a user from Europe who is used to access an online news service mainly for football news. Suppose that this user is planning a trip to the US and starts interacting with a web search engine to look for flights and accommodation, thus leaving a clear trace in web search logs about her current interest in the US country. Now assume that, while she is still planning her trip, a news about significant changes in the rules for European citizens to enter the US becomes public. This news is clearly interesting for the user, as it might even preclude her access to the country she is planning to visit. In this example a news personalization system relying only on endogenous information would not be able to recognize such news as relevant or useful, as the news content does not match the user's interest about football (the only interest manifested during the user's past interactions with the news service). On the contrary, this news would be recognized as interesting and probably recommended to the user if the system relied on exogenous information derived from web search logs.

Contributions

In this paper we study the novel problem of news personalization by leveraging web search query logs. To the best of our knowledge, the problem of studying the impact of such an exogenous information source on news personalization has never been considered before.

Our claim is that the endogenous information provided by the interaction of users with the news portal can be enriched by exogenous information extracted from web search query logs in order to improve the overall news personalization experience. Specifically, our goal is to understand what kind of information in query logs should be considered to build more complete and higher quality user profiles. This is orthogonal to the specific methods used for constructing user profiles and combining profiles from different sources. In this work we show that very basic methods already suffice to significantly improve the quality of news recommendation, thus attesting that a clear signal on the impact of the web-search source on news personalization exists regardless of the complexity of the employed models. More sophisticated models are clearly expected to be even more effective. For instance, running a topic model on top of search and news profiles together would lead to simultaneously finding latent relationships between the two types of profiles, with consequent benefit with respect to considering each type of profile in isolation. Devising the best ways of building profiles from query logs and combining them with endogenous profiles is however an interesting open problem that we defer to future work.

Our approach focuses on users who have used both the online news service and the search service. For each user, we record the terms contained in the queries that the user issued to the search engine and, for every query of the user, we record the terms contained in the titles and abstracts of the top 10 results returned by the search engine as answers to the query. These terms altogether constitute what we call the *search profile* of the user. For the personalization task, we consider the search profile of a user coupled with her *news profile*, which is the basic profile built based only on the past interactions of the user with the news service. More precisely, for a given user, both her search profile and her news profile are used to score the news articles, by computing: (1) the cosine similarity between the vector representing the search profile and the vector representing the news content, and (2) the cosine similarity between the news profile vector and the news content vector. We then produce a unified ranking that takes into account both the search profile score and the news profile score by resorting to two alternative methods traditionally used in the literature: (*i*) *score aggregation*, where the two initial scores are combined into a new single score that is eventually used for producing the ultimate ranking, and (*ii*) *rank aggregation*, where the two initial rankings are aggregated into a single ranking through a voting strategy.

We conduct a thorough experimental evaluation to verify whether and when such a combination of search profiles and news profiles can improve the quality of the news personalization task compared to using news profiles in isolation. The main findings arising from our experimental evaluation are as follows:

- The combination of search profiles with news profiles considerably improves upon using news profiles only, and the score aggregation method outperforms the rank aggregation method.

- Using search profiles consisting of query terms and the terms contained in the titles of the top 10 search results leads to a significant improvement, while including the terms contained in the top 10 abstracts does not increase the personalization quality further.
- Employing search profiles leads to improvement for both active users (expected) and inactive users (positively surprising).
- The quality of search profiles depends on the number of queries used to build the profiles. In our experiments we observe an improvement upon the strategy that relies only on news profiles when a user issues no less than 300 queries in a period of 3 months, i.e., when a user issues around 3 queries per day, on average.
- Building search profiles using three months of search history consistently improves the quality of recommendation upon the case where the search history spans a shorter period. On the other hand, extending the time period further (e.g., from four months up to six months) does not bring additional improvement upon the three-month case.
- As user interests evolve with time, more recent search profiles should reflect user interests better and thus ensure higher quality recommendations. In our experiments we show that using search profiles that are one-month old improves the quality of recommendation by up to 5.7% with respect to using profiles that are two to six month old.

Roadmap

The rest of the paper is organized as follows. Section 2 introduces how we build search profiles and combine them with news profiles. Section 3 reports on our experimental evaluation. Section 4 discusses related work. Section 5 concludes the paper.

2. Search-Enhanced news personalization

In this section we report the details of how we leverage the web search history of users to improve their news personalization experience. We restate that the goal of this work is to assess the impact of search query logs on news personalization in terms of what piece of information should be used to have better user profiles. The strategies we leverage for both profile construction and news ranking are the basic ones. As we will show in Section 3, these basic approaches are already enough to achieve considerable improvement in the quality of the news-personalization task, thus confirming that news personalization can benefit from web-search information regardless of the complexity of the employed models. More sophisticated solutions exist, such as statistical models, collaborative-filtering techniques, or click-through/session-based approaches for constructing user profiles from query logs (e.g., [10,43]), as well as machine learning approaches to combine multiple user profiles. Employing these solutions can further improve the quality of our approach and studying their impact constitutes an interesting open problem that we defer to future work.

2.1. Constructing search profiles

We construct the search profile of a user by using the information extracted from the query logs of a web search engine. Query logs record all actions that users perform on the search service. Specifically, they keep track of the time a query was issued, by whom, and the top-k result web pages returned by the search engine as answers to the query. For each result web page, we have access to its URL, title, and an abstract summarizing the content of the page.

Previous work has shown that queries are a good proxy for representing user interests, especially in a personalization task [20]. In general, however, queries on their own contain very few terms and, as a consequence, search profiles built by considering only query terms may easily suffer from a sparsity issue. A possible solution is to exploit the additional information contained in the top results of a query. The fact that such web pages are returned as an answer to the query by the underlying search engine is an implicit evidence that their content is likely to be relevant to the query and they can thus be safely exploited to expand the query-term-only search profiles. In particular, we enrich the search profiles by considering titles and abstracts of the top result pages. We hereinafter refer to search profiles built using only query terms, query terms plus title, and query terms plus title and abstract as, *query-based, title-enriched*, and *abstract-enriched* search profiles, respectively.

More formally, we construct a user profile as follows. Given a topic space $\mathcal{T}$ of dimensionality N_f, a user profile is represented as an N_f-dimensional numerical vector, where each element i denotes the degree of user interest in the topic i in $\mathcal{T}$. In this work we resort to the basic bag-of-words model to define the topic space, therefore N_f corresponds to the number of distinct terms (i.e., 1-grams) that form the vocabulary. The degree of user interest in the topic (term) i is computed by employing a standard TF-IDF strategy, whose details are provided next.

Let N_u be the total number of users and N_q be the total number of queries issued to the search engine by all users in a selected time period. The terms of the complete set of queries can be represented as an $(N_q \times N_f)$-dimensional integer matrix **Qw**, where each entry Qw_{ij} stores the number of times term j appears in query i. The title terms and the abstract terms of the top results of each query can be represented in an analogous way by $(N_q \times N_f)$-dimensional matrices **Tw** and **Aw**, respectively. Matrices **Qw, Tw**, and **Aw** basically keep track of the TF part. The information about the queries issued by the various users is instead stored in a binary matrix **Qu** of size $N_q \times N_u$, where $Qu_{ij} = 1$ if and only if user j issued query i.

Using the above notation, the query-based search profiles of the selected users are represented as an $(N_u \times N_f)$-dimensional matrix **Uq** defined as $\mathbf{Uq} = \mathbf{Qu}^T\mathbf{Qw}$. Similarly, the title-enriched search profiles are given by the matrix

$\mathbf{Ut} = \mathbf{Qu}^T(\mathbf{Qw} + \mathbf{Tw})$, while the matrix $\mathbf{Ua} = \mathbf{Qu}^T(\mathbf{Qw} + \mathbf{Tw} + \mathbf{Aw})$ corresponds to the abstract-enriched search profiles. To properly account for term importance, the entries of the three matrices **Uq, Ut**, and **Ua** are scaled using an IDF function computed on the corresponding user profiles. Specifically, each count in **Uq, Ut**, and **Ua** is multiplied by a scaling term computed as the logarithm of the ratio between the total number of queries in the log and the number of queries where the corresponding term appears. IDF is just one among many possible functions that can be used to alleviate the shortcomings of excessively frequent terms.

Note that matrices **Uq, Ut**, and **Ua** contain the search profiles of all users in the selected set: the profile of a single user i can be obtained by simply selecting the ith row of the matrix of interest.

2.2. Combining search profiles with news profiles

In a real news recommender system every time a user j accesses the system, she is provided with a ranked list of n news articles. Each news article a_l is assigned a relevance score se_{jl} that expresses how relevant a_l is for user j. Specifically, the score se_{jl} reflects how well news a_l matches the news profile of user j. A common approach to compute this relevance score is to set it equal to the cosine similarity between the news profile vector and the news vector. The scores $\{se_{jl}\}_{l=1}^{n}$ determine the ranking positions $\{pe_{jl}\}_{l=1}^{n}$ ($pe_{jl} \in [1..n]$) associated with the articles in the list: higher scores correspond to lower ranking positions.

To leverage search profiles, we associate each news article a_l with a further relevance score ss_{jl}, which is computed as the cosine similarity between the search profile of user j and news a_l. The relevance scores $\{ss_{jl}\}_{l=1}^{n}$ in turn yield a further ranking $\{ps_{jl}\}_{l=1}^{n}$.

In order to combine relevance scores and/or ranking positions given by search profiles and news profiles, we rely on two basic strategies, namely *score aggregation* (denoted SP_Score, where SP stands for search profiles) and *rank aggregation* (denoted SP_Rank). The difference between the two approaches is that SP_Score aims at combining the two relevance scores and using this combined score to infer a news ranking, whereas SP_Rank directly combines the two rankings in order to derive the final ranking. Specifically, the combined score Ss_{jl} provided by SP_Score is computed as a linear combination of the min-max-normalized se_{jl} and ss_{jl} scores (normalization performed to project the two rankings onto a common [0, 1] range). We experiment with various values of the parameter used to control the combination. More details on this are in Section 3. The final ranking produced by the SP_Rank method is computed by applying the well-known Borda-count election method to the two rankings $\{pe_{jl}\}_{l=1}^{n}$ and $\{ps_{jl}\}_{l=1}^{n}$.

3. Experiments

In this section we report our experimental evaluation that aims at assessing the validity of the proposed search-profile-based methods SP_Score and SP_Rank. We first describe the experimental setting in Section 3.1, while in Section 3.2 we discuss the results.

3.1. Setting

Dataset

We use the click logs of Yahoo News and the query logs from Yahoo Web Search.[1] We rely on the news click logs of a random day and build search profiles by using the queries that were issued at most six months before that day. We restrict our evaluation to a sample of the users who clicked on at least one news article on the test day and issued at least 1000 queries during the three-month period before the test day. This results in a set of about 70K users, for whom a total number of 140K independent news recommendations have been produced during the test day. Fig. 1 shows the distribution of the number of recommendations that are provided for the users in our dataset.

Methods

We implement the proposed SP_Score and SP_Rank as discussed in Section 2. As far as the SP_Score method, we set the parameter that controls the linear combination between the search profile score and the news profile score to 0.5, as we empirically observed that this value gives good results in most cases.

The main goal of the evaluation is to compare SP_Score and SP_Rank to a baseline method that relies on news profiles only, where the news profiles shared by the proposed methods and the baseline are built by keeping track of the content of the past news read by a user. In particular, the baseline method is a hybrid news-personalization system that exploits only news profiles.[2] It combines (i) content-based information given by the cosine similarity between the news profile of a user and the vector representing the content of a news article, and (ii) collaborative-filtering-like information taking into account how relevant a news article is for other users most similar (in terms of news profile) to the user at hand. More precisely, for each user u and term t in the vocabulary, a weight w_{ut} is computed as the number of times user u has clicked on a news

[1] Publicly available at https://webscope.sandbox.yahoo.com/catalog.php?datatype=r&did=75 .

[2] The baseline method is part of the news personalization module currently used in Yahoo.

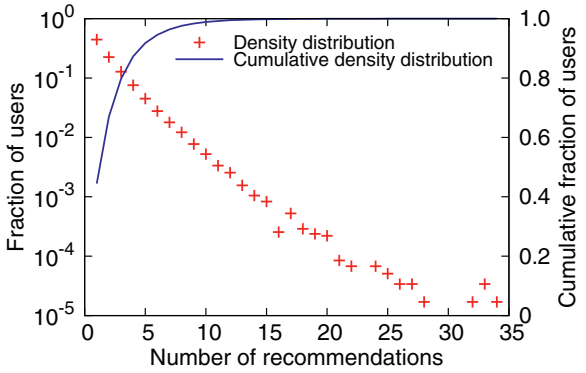


Fig. 1. Distribution of recommendations.

article containing term t. The ultimate news profile vector of user u corresponds to an N-dimensional real-valued vector $\vec{v}_u$ (where N is the vocabulary size), whose entries $\vec{v}_u(t)$, for each term t, are computed as the logarithm of the ratio between w_{ut} and the number of clicks on the same term t of other users who have clicked on news articles similar to those clicked by u. This way news profiles rely on both content-based information (weights w_{ut}) and collaborative filtering (scaling given by the weights of other similar users). Recommendations are made by ranking news articles by a combination of cosine similarity between news profiles and news vectors and popularity (in terms of absolute number of clicks) of the article.

As a further baseline, we consider a recency-based approach that is quite popular in the context of news personalization. Specifically, according to this method the news articles in each pageview are re-ranked in descending order of their publishing time. For details about the notion of pageview please see blow. We refer to this recency-based approach as TimeB.

Performance assessment

The interaction between a user and the news site is as follows. Every time a user accesses the system, she is provided with a list of 20 news articles, which are primarily ranked by the baseline method exploiting news profiles only. We refer to a pair ⟨user, news list⟩ as a *pageview*. Our goal is to re-rank the 20 news articles in each pageview by employing the proposed SP_Score and SP_Rank methods.

We evaluate the quality of the news rankings produced by our methods by resorting to the *Normalized Discounted Cumulative Gain* (NDCG) metric [5,36]. NDCG measures the quality of a ranked list of items/documents by giving more importance to the items ranked at the top positions of the list. If the user is not satisfied with what is immediately proposed to her, she will need to scroll down with the risk of losing attention. The NDCG aims at measuring this phenomenon, by discounting the recommendations at lower positions of the ranking. This perfectly conforms with the news-personalization context, where, regardless of the device, only a few slots are available to display recommendations.

In our use case, for each pageview p, we define the *relevance* $rel(a)$ of an article a on p as equal to 1 if the user clicked on a, 0 otherwise. Let ρ denote a ranking of the news articles present in the list of a pageview and let $\rho(i)$, for all $i \in [1..20]$, denote the article at position i of the ordered list defined by ρ. The *Discounted Cumulative Gain* (DCG) of a ranking ρ is defined as:

$$DCG(\rho) = rel(\rho(1)) + \sum_{i=2}^{20} \frac{rel(\rho(i))}{\log_2(i+1)}.$$

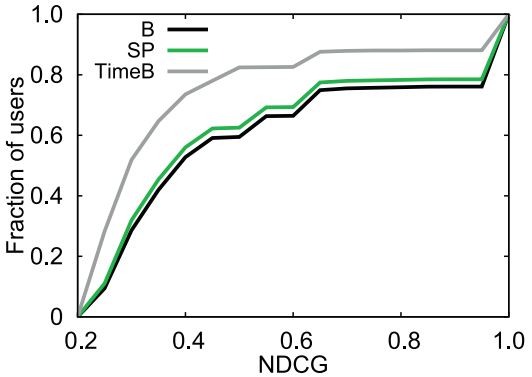
For each pageview p, let also ρ^* denote the *ideal ranking* of the articles in the news list of p, that is an ordered list where the articles having relevance equal to 1 are all put in the front of the list, while the articles with relevance 0 follow them (ties broken arbitrarily). The *NDCG* of a ranking ρ is finally defined as the ratio between the *DCG* of ρ and the *DCG* of the ideal ranking ρ^* [23]:

$$NDCG(\rho) = \frac{DCG(\rho)}{DCG(\rho^*)}.$$

The main goal of our evaluation is to assess whether the proposed search-profile-based methods yield higher *NDCG* values than the baseline. Specifically, in each set of experiments, we focus on the average *NDCG* value (i.e., averaged over all pageviews), on the cumulative distribution of *NDCG* values, as well as on assessing whether the difference between two overall sets of *NDCG* values (i.e., for all pageviews) is statistically significant. In particular, we assess statistical significance by employing the Wilcoxon signed rank test [15]. This choice is motivated since (*i*) the Wilcoxon test does not require for the statistics to be tested to follow any specific distribution, and (*ii*) it is a paired test, which is needed in our context as, for any set of experiments, we compare pairs of observations coming from two competing methods (i.e., NDCG values obtained for a specific pageview).

	avg. $NDCG$
B	0.522
SP	0.501
TimeB	0.402

(a) Average $NDCG$.



(b) $NDCG$ distribution.

Fig. 2. *NDCG* results of the news-profile-only baseline (B), the recency-based baseline (TimeB) and a strategy based on search profiles only (title-enriched search profiles, 3-month training period).

3.2. Results

In the following we report and discuss the main experimental findings observed with our empirical evaluation. Particularly, we are interested in evaluating six critical aspects:

(1) Usefulness of search profiles both in isolation and in combination with news profiles,
(2) Important features at the base of search profiles,
(3) Benefits of search profiles for active and inactive users,
(4) Volume of search queries needed for building satisfactory search profiles,
(5) Time horizon to be considered for constructing search profiles,
(6) Impact of recency on the quality of search profiles.

In the following we provide detailed discussions on each of these aspects.

1. Do search profiles improve the quality of news personalization? First of all, even though our proposal considers search profiles in combination with news profiles, we believe it is anyway worth taking a look at the performance while using search profiles in isolation. We report this experiment in Fig. 2 and we observe that the results confirm what is suggested by common sense: the search-profile-only strategy is not enough to outperform the news-profile-only strategy (denoted as B in the figure). This was expected, as past interactions with the news service is the primary source of information to discover user interests in news. What is more interesting is that the difference between the two strategies is tangible but not particularly evident. This suggests that there is a good chance of observing consistent improvements when combining search profiles with news profiles. The experiments below confirm this claim. Before moving to that, we point out that Fig. 2 also reports on the results of the recency-based baseline TimeB, which recommends news based on their recency. Results show that such a baseline performs evidently worse than the news-profile-only baseline B, and even worse than the search-profile-only strategy. Thus, we avoid reporting its results in the rest of the experiments. For easiness of presentation, we hereinafter use "news-profile-only baseline" and "baseline" interchangeably to refer to the news-profile-only baseline.

In Fig. 3 we compare the *NDCG* results achieved by the proposed SP_Score and SP_Rank methods to the baseline. The results of our methods reported here refer to search profiles built considering a 3-month training period and exploiting terms from each query issued along with the title of its top-10 result web pages (i.e., title-enhanced search profiles given by the matrix **Ut** defined in Section 2.1). The figure shows that our methods clearly outperform the news-profile-only baseline in terms of both average *NDCG* and overall distribution of *NDCG* values. Importantly, as reported in Fig. 3(a), the differences between the proposed methods and the baseline are statistically significant. Among the two proposed methods, SP_Score exhibits in general better accuracy: this is motivated by the fact that its profile-combining strategy is more fine-grained than SP_Rank (see Section 2.2).

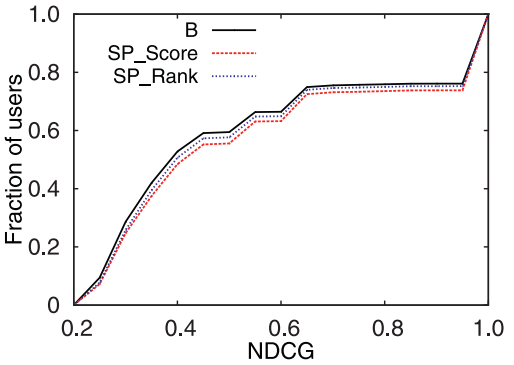
Therefore, based on the findings above, we can state that it is possible to improve the quality of news personalization by exploiting the web search history of a user.

2. What are the important features to be considered in a search profile? To answer this question, we study the impact of building search profiles at different granularities, i.e., by considering query terms only (i.e., query-based search profiles given by the matrix **Uq** defined in Section 2), or including information from titles (i.e., title-enhanced search profiles given by the matrix **Ut** in Section 2) or titles plus abstracts (i.e., abstract-enhanced search profiles given by the matrix **Ua** in Section 2) of the top-10 web pages returned as results to the query by the underlying search engine.

The results of this experiment are reported in Fig. 4. The first finding is that query terms alone are too sparse to allow any method to obtain a clear improvement upon the news-profile-only baseline. In fact, using query terms only, our SP_Score

	avg. $NDCG$	p-value<0.05 (vs. B)
B	0.522	—
SP_Score	0.545	yes
SP_Rank	0.533	yes

(a) Average $NDCG$ and statistical significance.

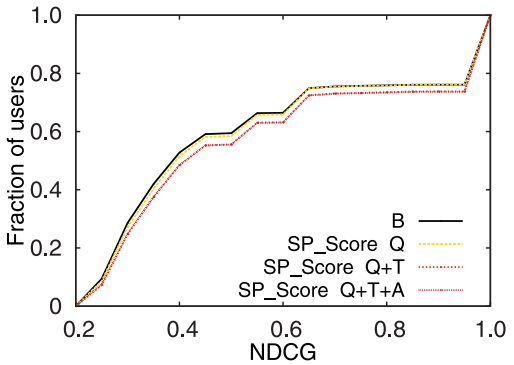


(b) $NDCG$ distribution.

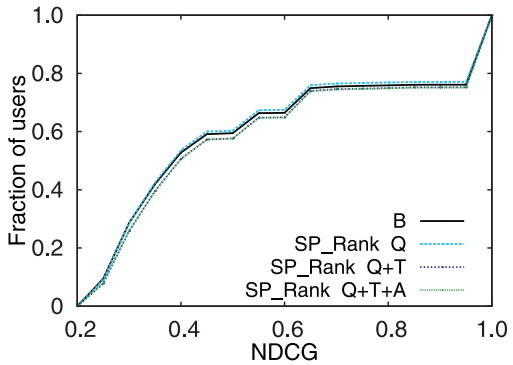
Fig. 3. *NDCG* results of the baseline (B) and the proposed SP_Score and SP_Rank methods (title-enriched search profiles, 3-month training period).

		avg. $NDCG$	p-value<0.05 (vs. B)
B		0.5217	—
SP_Score	Q	0.5259	no
	Q+T	0.5449	yes
	Q+T+A	0.5453	yes
SP_Rank	Q	0.5155	no
	Q+T	0.5328	yes
	Q+T+A	0.5334	yes

(a) Average $NDCG$ and statistical significance.



(b) $NDCG$ distribution (SP_Score).



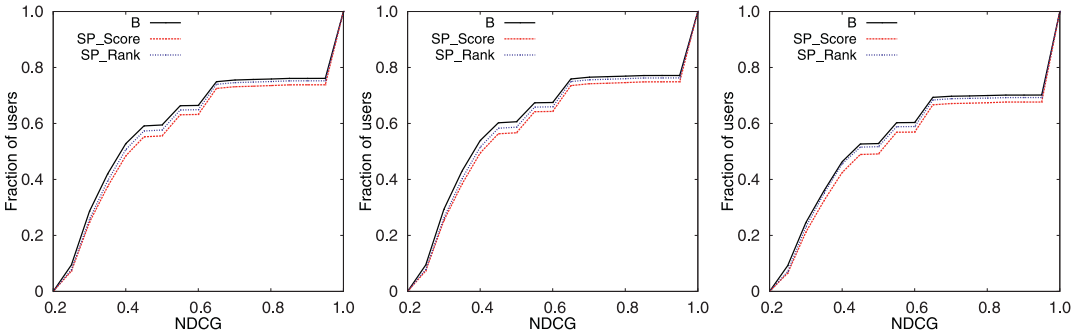
(c) $NDCG$ distribution (SP_Rank).

Fig. 4. *NDCG* results of the baseline (B) and the proposed SP_Score and SP_Rank with different information considered to build the search profiles: query-based search profiles (Q), title-enriched search profiles (Q + T), abstract-enriched search profiles (Q + T + A).

method slightly outperforms the baseline, but the difference is not statistically significant. Instead, augmenting the search profiles with both queries and titles (Q + T) gives much better results: the differences with respect to the baseline are statistically significant for both SP_Score and SP_Rank. Further enriching the search profiles with terms in the abstracts clearly keeps the difference from the baseline statistically significant and leads to a slight ulterior improvement with respect to using Q + T terms. The improvement is however not that evident: the average *NDCG* only slightly increases (0.07% for

	all users		active users		inactive users	
	avg. $NDCG$	p-value<0.05 (vs. B)	avg. $NDCG$	p-value<0.05 (vs. B)	avg. $NDCG$	p-value<0.05 (vs. B)
B	0.522	—	0.522	—	0.522	—
SP_Score	0.545	yes	0.588	yes	0.538	yes
SP_Rank	0.533	yes	0.573	yes	0.526	yes

(a) Average $NDCG$ and statistical significance



(b) $NDCG$ distribution:
all users (left), active users (middle), inactive users (right)

Fig. 5. *NDCG* results of the baseline (B) and the proposed SP_Score and SP_Rank methods for different users (title-enriched search profiles, 3-month training period).

SP_Score and 0.11% for SP_Rank), and the difference between the Q + T *NDCG* values and the Q + T + A *NDCG* values is not statistically significant. A possible explanation is that the terms contained in the abstract but not in the title are usually contextual terms that add only little information to what is already provided by the query + title terms themselves.

Considering the increased dimensionality of the resulting search profiles when using abstracts, we can thus conclude that building search profiles using query + title terms is perhaps the best choice in terms of trade-off among accuracy, computational effort, and space needed to store the profiles.

3. Is there any difference between active and inactive users? The next aspect we focus on is to which extent the improvement exhibited by our search-profile-based methods distinctly affect users who are active/inactive in the news site. We define active and inactive users as those who clicked on at least 100 and less than 100 news articles during a 3-month training period, respectively. The ultimate goal is to understand whether our strategy is valid also for users who have a weaker interaction with the news site, i.e., users who have less than 100 clicks on the news website during the 3-month training period. The results are reported in Fig. 5. According to the figure, for either active or inactive users, both SP_Score and SP_Rank methods achieve better *NDCG* results than the news-profile-only baseline, and the differences are statistically significant. The impact of this finding is noteworthy, as it clearly assesses that the proposed methods leveraging exogenous information improve the quality of news recommendation, even for those users who exhibit weak interaction with the news site.

4. How many search queries are needed when building a search profile in order to observe quality improvements? We now shift the attention to the problem of assessing how much web search history is actually needed for observing an improvement in the quality of news personalization. In Fig. 6 we report the results achieved by aggregating title-enriched search profiles at different granularities (in terms of number of queries): from 200 to 1000 queries. The queries we consider in the various samples are randomly selected from the ones issued by each user during a three-month period. The figure shows that the improvement of SP_Score upon the baseline starts right after 200 queries and gets progressively larger. The improvement of SP_Rank happens later: at around 600 queries. The difference from the baseline becomes statistically significant at around 300 queries (SP_Score) and 700 queries (SP_Rank), respectively. In summary, we can state that the quality of the news personalization system can evidently benefit from the use of web search history for a number of 300 queries issued in a period of 3 months (i.e., an average of around 3 queries per day).

5. How much time should the historical information span in order to produce high-quality recommendations? How does the quality vary with the increase in time span? The objective here is to analyze the behavior of the proposed search-profile-based methods when the training period varies. We aim at discovering the impact of the amount of historical

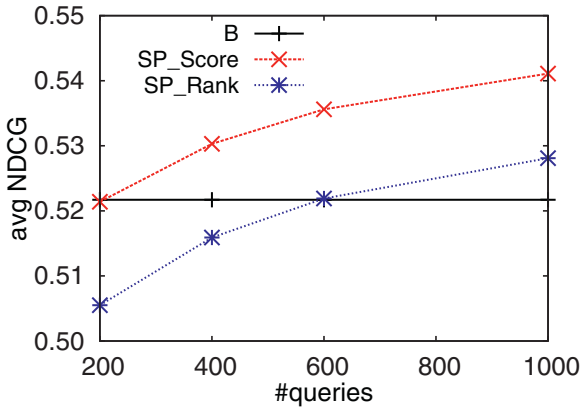


Fig. 6. *NDCG* results of the baseline (B) and the two proposed methods (SP_Score and SP_Rank) with varying the number of queries issued (title-enriched search profiles).

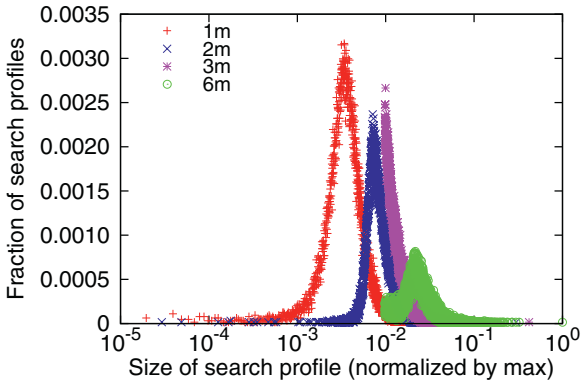


Fig. 7. Distribution of profile size.

information collected for each user on the performance of the search profile for that user. In particular, we consider title-enriched search profiles based on queries issued on a time period spanning one month, two months, ..., up to six months before the test day.

Fig. 7 shows the distribution of the size of search profiles, computed as the number of queries issued by a user divided by the maximum number of queries among all users. The distributions for search profiles based on queries issued during four and five months are not reported for the sake of readability of the figure. Results of our methods are instead reported in Fig. 8.

For all time periods considered, the figure shows that both SP_Score and SP_Rank are significantly better than the news-profile-only baseline, and increasing the training period always leads to better accuracy, although the improvement tends to decrease with increasing time period. Indeed, in Fig. 8(b), where we report whether the difference between the results of two consecutive time periods is statistically significant, we can see that this observation only holds for the time periods of up to three months, while for the remaining time periods the differences are not statistically significant. Based on this finding, we can therefore conclude that the richer the search profile is, the more useful the search signal is in the news personalization task, at least up to a three-month time period. Considering time periods larger than three months does not lead to any consistent performance improvement.

6. How does the recency of constructed user profiles affect the quality of news personalization? Herein, we focus on the problem of understanding how recent the search profiles should be in order to guarantee good performance. To this end, we perform the following experiment. We build search profiles considering different time windows before the test day. Particularly, we set the size of the time window equal to one month and we let such a window slide back from the test day month by month, up to six months ago. To be more clear, assuming for example January 1st 2014 as the test day, we consider search profiles built based on the queries issued during December 2013 (1 month back), November 2013 (2 months back), ..., and July 2013 (6 months back).[3]

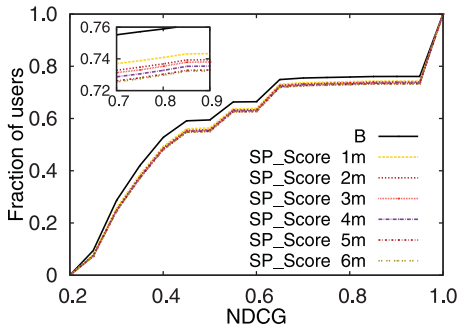
[3] Note that this experiment is different from the experiment we conducted for experiment 5, where, when we talk about a time period of k months, we refer to search queries issued during the *whole* period between the test day and the k months before the test day.

	1 month		2 months		3 months		4 months		5 months		6 months	
	avg. $NDCG$	p-value <0.05 (vs. B)	avg. $NDCG$	p-value <0.05 (vs. B)	avg. $NDCG$	p-value <0.05 (vs. B)	avg. $NDCG$	p-value <0.05 (vs. B)	avg. $NDCG$	p-value <0.05 (vs. B)	avg. $NDCG$	p-value <0.05 (vs. B)
B	0.522	—	0.522	—	0.522	—	0.522	—	0.522	—	0.522	—
SP_Score	0.540	yes	0.543	yes	0.545	yes	0.546	yes	0.548	yes	0.549	yes
SP_Rank	0.524	yes	0.530	yes	0.533	yes	0.534	yes	0.536	yes	0.537	yes

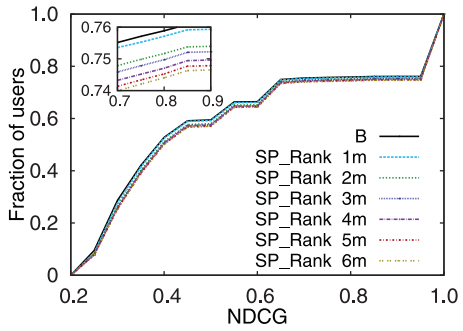
(a) Average $NDCG$ and statistical significance vs. the baseline.

	p-value<0.05				
	2M vs. 1M	3M vs. 2M	4M vs. 3M	5M vs. 4M	6M vs. 5M
SP_Score	yes	yes	no	no	no
SP_Rank	yes	yes	no	no	no

(b) Statistical significance among the various time periods.



(c) $NDCG$ distribution (SP_Score).



(d) $NDCG$ distribution (SP_Rank).

Fig. 8. *NDCG* results of the baseline (B) and the proposed SP_Score and SP_Rank with different training periods to build the search profiles (title-enriched search profiles).

The results of this experiment are shown in Fig. 9. Such results show that the recency of the search profiles clearly matters. Indeed, the further the time period considered to build the search profiles from the test day, the lower the accuracy of the recommendation. Particularly, the average *NDCG* when considering a 1-month-back period is 5.7% and 5.6% larger than the average *NDCG* resulting from a 6-month-back period for the SP_Score and SP_Rank methods respectively.

4. Related work

News personalization has become an extremely active research area in the last years [28]. Existing approaches are usually broadly classified into *collaborative filtering* [41], *content-based* [33], and *hybrid* [9].

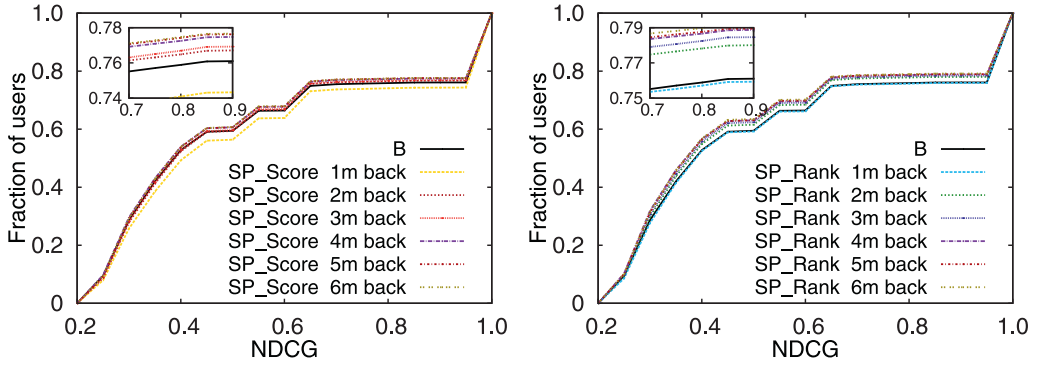
Collaborative-filtering-based news-personalization systems [13,40] recommend news to any specific user based on the ratings of other users who share similar interests with her. A well-known limitation of such approaches is the so-called *item cold-start problem*, which concerns the hardness of recommending items that have very few ratings or no ratings at all. This weakness is particularly problematic in the context of news personalization, given the inherent highly-dynamic nature of the items (i.e., news) to be recommended. For this reason, content-based systems are more common [4,6,7,18,19,22,27,44]. The general idea behind such systems is to build a user profile based on the user's past activity on the news website and recommend news based on how well they match that profile. Collaborative filtering and content-based systems are also combined together into the so-called hybrid news-personalization systems [12,21,29–31,47].

Regardless of their specific category (i.e., collaborative filtering, content-based, or hybrid), existing news-personalization systems do not rely on information external to the news website. As a result, they all suffer from the so-called *user cold-start problem*, i.e., the problem of providing effective recommendations to users who exhibit poor interaction with the news website. In this work we aim at overcoming this issue by leveraging external information coming from search query logs.

A related body of research looks at microblogging services like Twitter to deliver personalized news [1,2,14,24,25,39]. Our work departs from this existing literature first of all because we exploit another source of external information, i.e., web search query logs. Also, and more importantly, our work is noticeably different in spirit. In all those works, in fact, user profiles are built based on the microblogging service only, which makes the resulting microblogging-based news-personalization module *alternative* to the default module present on the news website [14]. We instead use external information in order to

	avg. $NDCG$					
	1 month back	2 months back	3 months back	4 months back	5 months back	6 month back
B	0.522	0.522	0.522	0.522	0.522	0.522
SP_Score	0.540	0.521	0.518	0.512	0.512	0.511
SP_Rank	0.524	0.508	0.503	0.498	0.498	0.496

(a) Average $NDCG$.



(b) $NDCG$ distribution (SP_Score). (c) $NDCG$ distribution (SP_Rank).

Fig. 9. *NDCG* results of the baseline (B) and the proposed SP_Score and SP_Rank methods with varying the recency of the search profiles (title-enriched search profiles).

complete user profiles deriving from the interaction with the news website so as to achieve a direct impact on the default news-personalization module itself.

A number of works presented in the literature deal with the problem of incorporating externally-provided OLAP-based aggregate ratings into recommender systems [3,45,46]. This problem is only marginally related to the problem we tackle in this paper, as it follows the general direction of content recommendation based on external data, but it however remains different from several perspectives. First of all, it does not explicitly deal with news personalization, but with recommending content in general, especially movies, for which aggregate-rating providers are more easily available (e.g., IMDB) than in the case of news articles. Second, aggregate ratings cannot be considered as a real external information source as they are still derived from the interaction between users and items that are at the basis of the recommendation. Our goal is instead to exploit an external source such as search query logs that do not directly express any interaction between users and items to be recommended but can anyway unveil useful additional interests.

Finally, the problem of exploiting web-search information, such as query logs, click-through data, or session data, for personalization of online services has been extensively studied in the literature. However, this body of research has focused on personalization of services that are inherent to web search itself, such as type resolution of entities in a web-search query [37], enrichment of web-search queries by query expansion [11] or web-search results [16]. Our work instead exploits information from web-search queries to improve personalization of an external service, i.e., a news portal.

5. Conclusions

We addressed the problem of news personalization by leveraging information extracted from web search query logs. We devised a method that represents the interests of a users based on the web search queries she issued, the titles of the pages returned as a result to the queries, as well as the displayed snippets. We evaluated two strategies for combining personalized news rankings obtained by exploiting web search history with news rankings obtained through common user interactions with the news site. Our experiments indicate that exploiting search profiles leads to considerable improvements upon using traditional news-interaction-based profiles only.

In the future we plan to dig into the methods used for constructing search profiles and to combine search profiles and news profiles. In particular, as a first attempt, we will study the impact of using topic model on top of search and news profiles, so as to better capture the latent relationships between the two types of profile. We also plan to apply the same idea to other services that may provide user-interaction data (e.g., social networks). In general, in fact, the entire web history of a user can potentially be used for personalization.

References

[1] F. Abel, Q. Gao, G.-J. Houben, K. Tao, Analyzing user modeling on twitter for personalized news recommendations, in: Proc. of Int. Conf. on User Modeling, Adaptation, and Personalization (UMAP), 2011, pp. 1–12.
[2] F. Abel, Q. Gao, G.-J. Houben, K. Tao, Twitter-based user modeling for news recommendations, in: Proc. of Int. Joint Conf. on Artificial Intelligence (IJCAI), 2013.
[3] G. Adomavicius, R. Sankaranarayanan, S. Sen, A. Tuzhilin, Incorporating contextual information in recommender systems using a multidimensional approach, ACM Trans. Inf. Syst. (TOIS) 23 (1) (2005) 103–145.
[4] J.-w. Ahn, P. Brusilovsky, J. Grady, D. He, S.Y. Syn, Open user profiles for adaptive news systems: Help or harm? in: Proc. of Int. Conf. on World Wide Web (WWW), 2007, pp. 11–20.
[5] R. Baeza-Yates, B. Ribeiro-Neto, et al., Modern Information Retrieval, 463, ACM press New York, 1999.
[6] T. Bansal, M. Das, C. Bhattacharyya, Content driven user profiling for comment-worthy recommendations of news and blog articles, in: Proc of Int. ACM Conf. on Recommender Systems (RecSys), 2015, pp. 195–202.
[7] D. Billsus, M.J. Pazzani, A hybrid user model for news story classification, in: Proc. of Int. Conf. on User Modeling (UM), 1999, pp. 99–108.
[8] D. Billsus, M.J. Pazzani, A hybrid user model for news story classification, in: Proc. of Int. Conf. on User Modeling (UM), 1999, pp. 99–108.
[9] R. Burke, Hybrid recommender systems: survey and experiments, User Model. User-adapted Interact. 12 (4) (2002) 331–370.
[10] H. Cao, D. Jiang, J. Pei, Q. He, Z. Liao, E. Chen, H. Li, Context-aware query suggestion by mining click-through and session data, in: Proc. of ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD), 2008, pp. 875–883.
[11] P.A. Chirita, C.S. Firan, W. Nejdl, Personalized query expansion for the web, in: Proc. of Int. ACM SIGIR Conf. on Research and Development in Information Retrieval, 2007, pp. 7–14.
[12] W. Chu, S.-T. Park, Personalized recommendation on dynamic content using predictive bilinear models, in: Proc. of Int. Conf. on World Wide Web (WWW), 2009, pp. 691–700.
[13] A.S. Das, M. Datar, A. Garg, S. Rajaram, Google news personalization: scalable online collaborative filtering, in: Proc. of Int. Conf. on World Wide Web (WWW), 2007, pp. 271–280.
[14] G. De Francisci M., A. Gionis, C. Lucchese, From chatter to headlines: harnessing the real-time web for personalized news recommendation, in: Proc. of Int. Conf. on Web Search and Data Mining (WSDM), 2012.
[15] J. Demšar, Statistical comparisons of classifiers over multiple data sets, J. Mach. Learn. Res. (JMLR) 7 (2006) 1–30.
[16] Z. Dou, R. Song, J.-R. Wen, A large-scale evaluation and analysis of personalized search strategies, in: Proc. of Int. Conf. on World Wide Web (WWW), 2007, pp. 581–590.
[17] B. Fetahu, K. Markert, A. Anand, Automated news suggestions for populating wikipedia entity pages, in: Proc. of ACM Int. Conf. on Information and Knowledge Management (CIKM), 2015, pp. 323–332.
[18] E. Gabrilovich, S. Dumais, E. Horvitz, Newsjunkie: providing personalized newsfeeds via analysis of information novelty, in: Proc. of Int. Conf. on World Wide Web (WWW), 2004, pp. 482–490.
[19] F. Garcin, C. Dimitrakakis, B. Faltings, Personalized news recommendation with context trees, in: Proc of Int. ACM Conf. on Recommender Systems (RecSys), 2013, pp. 105–112.
[20] M. Harvey, F. Crestani, M.J. Carman, Building user profiles from topic models for personalised search, in: Proc. of ACM Int. Conf. on Information and Knowledge Management (CIKM), 2013, pp. 2309–2314.
[21] C.-K. Hsieh, L. Yang, H. Wei, M. Naaman, D. Estrin, Immersive recommendation: news and event recommendations using personal digital traces, in: Proc. of Int. Conf. on World Wide Web (WWW), 2016, pp. 51–62.
[22] H. Husin, J. Thom, X. Zhang, News recommendation based on web usage and web content mining, in: ICDE Work., 2013, pp. 326–329.
[23] K. Järvelin, J. Kekäläinen, Cumulated gain-based evaluation of IR techniques, ACM Trans. Inf. Syst. (TOIS) 20 (4) (2002) 422–446.
[24] N. Jonnalagedda, S. Gauch, Personalized news recommendation using twitter, in: Proc. of IEEE/WIC/ACM Int. Joint Conf. on Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2013, pp. 21–25 vol. 3.
[25] G. Kazai, I. Yusof, D. Clarke, Personalised news and blog recommendations based on user location, facebook and twitter user profiling, in: Proc. of Int. Conf. on Research and Development in Information Retrieval (SIGIR), 2016, pp. 1129–1132.
[26] D. Lagun, M. Lalmas, Understanding user attention and engagement in online news reading, in: Proc. of Int. Conf. on Web Search and Data Mining (WSDM), 2016, pp. 113–122.
[27] K. Lang, NewsWeeder: learning to filter netnews, in: Proc. of Int. Conf. on Machine Learning (ICML), 1995, pp. 331–339.
[28] L. Li, D.-D. Wang, S.-Z. Zhu, T. Li, Personalized news recommendation: a review and an experimental investigation, J. Comp. Sci. Tech. 26 (5) (2011) 754–766.
[29] L. Li, D. Wang, T. Li, D. Knox, B. Padmanabhan, Scene: a scalable two-stage personalized news recommendation system, in: Proc. of Int. Conf. on Research and Development in Information Retrieval (SIGIR), 2011, pp. 125–134.
[30] L. Li, W. Chu, J. Langford, R.E. Schapire, A contextual-bandit approach to personalized news article recommendation, in: Proc. of Int. Conf. on World Wide Web (WWW), 2010, pp. 661–670.
[31] J. Liu, P. Dolan, E.R. Pedersen, Personalized news recommendation based on click behavior, in: Proc. of Int. Conf. on Intelligent User Interfaces (IUI), 2010, pp. 31–40.
[32] A. Lommatzsch, Real-time news recommendation using context-aware ensembles, in: Proc. of European Conf. on Advances in Information Retrieval (ECIR), 2014, pp. 51–62.
[33] P. Lops, M. de Gemmis, G. Semeraro, Content-based recommender systems: state of the art and trends, in: Recommender Systems Handbook, Springer, 2011, pp. 73–105.
[34] H. Ma, X. Liu, Z. Shen, User fatigue in online news recommendation, in: Proc. of Int. Conf. on World Wide Web (WWW), 2016, pp. 1363–1372.
[35] A. Maksai, F. Garcin, B. Faltings, Predicting online performance of news recommender systems through richer evaluation metrics, in: Proc of Int. ACM Conf. on Recommender Systems (RecSys), 2015, pp. 179–186.
[36] C.D. Manning, P. Raghavan, H. Schütze, et al., Introduction to Information Retrieval, 1, Cambridge university press Cambridge, 2008.
[37] P. Pantel, T. Lin, M. Gamon, Mining entity types from query logs via user intent modeling, in: Proc. of Annual Meeting of the Association for Computational Linguistics (ACL), 2012, pp. 563–571.
[38] S.-T. Park, D. Pennock, O. Madani, N. Good, D. DeCoste, Naïve filterbots for robust cold-start recommendations, in: Proc. of ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD), 2006, pp. 699–705.
[39] O. Phelan, K. McCarthy, M. Bennett, B. Smyth, Terms of a feather: content-based news recommendation and discovery using twitter, in: Proc. of European Conf. on Information Retrieval (ECIR), 2011.
[40] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, J. Riedl, Grouplens: an open architecture for collaborative filtering of netnews, in: Proc of ACM Conf. on Computer-Supported Cooperative Work and Social Computing (CSCW), 1994, pp. 175–186.
[41] X. Su, T.M. Khoshgoftaar, A survey of collaborative filtering techniques, Adv. Artif. Intell. 2009 (2009) 4:2–4:2.
[42] A.-H. Tan, C. Teo, Learning user profiles for personalized information dissemination, in: Proc. of IEEE Int. Joint Conf. on Neural Networks (IJCNN), 1998, pp. 183–188 vol.1.
[43] B. Tan, Y. Lv, C. Zhai, Mining long-lasting exploratory user interests from search history, in: Proc. of ACM Int. Conf. on Information and Knowledge Management (CIKM), 2012, pp. 1477–1481.
[44] M. Trevisiol, L.M. Aiello, R. Schifanella, A. Jaimes, Cold-start news recommendation with domain-dependent browse graph, in: Proc of Int. ACM Conf. on Recommender Systems (RecSys), 2014, pp. 81–88.

[45] A. Umyarov, A. Tuzhilin, Leveraging aggregate ratings for better recommendations, in: Proc of Int. ACM Conf. on Recommender Systems (RecSys), 2007, pp. 161–164.
[46] A. Umyarov, A. Tuzhilin, Using external aggregate ratings for improving individual recommendations, ACM Trans. Web (TWEB) 5 (1) (2011) 3:1–3:40.
[47] H. Wen, L. Fang, L. Guan, A hybrid approach for personalized recommendation of news on the web, Expert Syst. Appl. 39 (5) (2012) 5806–5814.