Journal of
Print and Media Technology
Research

3-2013

September 2013

Thematic issue

Content technologies

Guest Editor

Caj Södergård

J. Print Media Technol. Res. 2(2013)3, 125-214

Contents

A word from the Guest Editor

Caj Södergård
VTT - Technical Research Centre of Finland, Espoo

E-mail: Caj.Sodergard@vtt.fi

Content technologies provide tools for processing content to be delivered via any media to the target audience. These tools are applied in numerous ways in media production. Research into content technologies is very active and opens new possibilities to improve production efficiency as well as to enhance the user experience and thereby the business value of media products and services.

This thematic issue focuses on several applications of content technologies. All papers address the user, and the ability to objectively measure and predict the responses various content causes in users is a much needed tool for the media professional. An emerging application proposed in this issue helps journalists find interesting topics for articles from the excessive information available on the internet. Another class of applications dealt with here is recommending content to the users. Relevant recommendations motivate the user to visit and spend time on a web service. Recommenders are therefore important in designing attractive - and monetizable - digital services. As a consequence, this technology is found in many services recommending media items such as music, books, television programmes and news articles. The papers on recommenders in this issue cover the three main methods in the field - content-based, knowledge-based and collaborative - and they bring new perspectives to all three. One such novel perspective which has been evaluated in user studies is that of a portable personal profile.

Most of the included papers are outcomes of the Finnish *Next Media* research program (www.nextmedia.fi) of Digile Oy. Next Media has run from 2010 through 2013 with the participation of 57 companies and eight research organisations. The volume of the program has been substantial; annually around 80 person years with half of the work done by companies and half by research partners. The program has three foci: e-reading, personal media day, and hyperlocal. The papers in this issue represent only a small part of the results of Next Media. As an example, during 2012 the program produced 101 reports, most of which are available on the web.

Even if this thematic issue is centred on work done within the Finnish Next Media program, content technologies are of course studied in many other places around the world. The paper by NTNU in Norway presented here is just one example. Computer and information technology departments at universities and research institutes often pursue content related topics ranging from multimedia "big data" analysis to multimodal user interfaces and user experience. In the upcoming EU Horizon 2020 program, "Content technologies and information management" is a major topic covering eight challenges. This will keep the theme for this thematic issue in the forefront of European research during the years to come.

Caj Södergård, guest editor of this issue of JMTR, holds a doctoral degree in Information Technologies from the Helsinki University of Technology. After some years in industry, he has held positions at VTT as researcher, senior researcher, team manager and technology manager. His work has resulted in several patents and products used in the media field. Currently Caj Södergård is Permanent Research Professor in Digital Media Technologies at VTT.

JPMTR 025 | 1312
UDC 004.78:004.58

Research paper
Received: 2013-07-08
Accepted: 2013-11-11

Learning user profiles in mobile news recommendation

Jon Atle Gulla[1], *Jon Espen Ingvaldsen*[1], *Arne Dag Fidjestøl*[2], *John Eirik Nilsen*[1], *Kent Robin Haugen*[1], *Xiaomeng Su*[2]

[1] Department of Computer and Information Science
Norwegian University of Science and Technology
Sem Sælands vei 7, Gløshaugen
N-7499 Trondheim, Norway

E-mail: jag@idi.ntnu.no

[2] Research and Future Studies
Telenor Group, Norway
N-1331 Fornebu, Norway

Abstract

Mobile news recommender systems help users retrieve relevant news stories from numerous news sources with minimal user interaction. The overall objective is to find ways of representing news stories, users and their relationships that allow the system to predict which news would be interesting to read for which users. Even though research shows that the quality of these recommendations depends on good user profiles, most systems have no or very simple profiles, because users are reluctant to giving explicit feedback on articles' desirability. In this paper we present a user profiling approach adopted in the SmartMedia news recommendation project. We are building a mobile news recommender app that sources news from all major Norwegian newspapers and uses a hybrid recommendation strategy to rank the news according to the users' context and interests. The user profiles in SmartMedia are built in real-time on the basis of implicit feedback from the users and contain information about the users' general interests in news categories and particular interests in events or entities. Experiments with content-based filtering show that the profiles lead to more targeted recommendations and provide an efficient way of monitoring and representing users' interests over time.

Keywords: recommender systems, personalization, Big Data, user click analysis, news apps, content-based filtering

1. Introduction and background

1.1 Rationale for this study

We have in the last few years witnessed the introduction of a number of commercial mobile news apps. These are applications that help users find relevant news from a number of news sources without going to each individual source or browsing through all the news available at each source. Due to the limitations of mobile devices in terms of screen sizes and input methods, these news apps are mostly gesture-based with news headlines or news stories compressed to fit small, buttonless displays. Some of these apps, such as Summly, Wavii and Circa, present summaries of news articles to make better use of the small screens, whereas others have introduced categories and sharing with friends to help users focus on the news interesting to them and filter out the irrelevant parts (Circa, 2013; Haugen, 2013; Yeung, 2013).

It seems tempting to introduce technology that can help the news apps recommend the news that are most likely to be of interest to their users. This requires some knowledge about the news content and the users' opinions on both particular news articles and news categories in general. Explicit signals about user desirability or interests are, however, weak and the recommender system would normally need to use implicit signals such as user click patterns to infer preferences and priorities. Moreover, since a mobile user is also situated in a particular context, located at a particular place at a particular time, we may also need to assume that her context is relevant to which articles she would deem interesting. Even though a user has a general preference for sports news, for example, she might want to know that there is a traffic accident just a few blocks ahead of her.

Recommender systems have been used extensively for music and movie recommendations as well as for product reviews in general (Schafer, Konstan and Riedl, 1999) and there are already major online stores such as Amazon that offer product recommendations as part of their services. News recommendation differs in several ways from these well-known types of recommender systems: (i) articles have short life-cycles, and freshness and location may often be as important to the user as the arti-

cle's content relevance, (ii) news articles are unstructured and more complex to analyze than objects with structured properties such as product reviews or networks of friends, (iii) the volatility and unlimited reach of news lead to rapid changes in both terminologies and topics over time, (iv) serendipities, or the need for variety and unexpected news, have to be addressed, and (v) cold-start problems linked to users that have no history and news that have not yet been discovered by enough users are notorious.

This paper discusses an approach for personalizing mobile news services by means of implicitly inferred user profiles. After a review of the current state of the art of recommender systems and user click analysis in section 1, we present our SmartMedia news recommendation project in section 2 and go through the steps from logging user behavior to constructing user profiles for news recommendation. Section 3 demonstrates the use of user profiling in the news recommender app and shows how news articles are ranked differently as a user's profile is updated over time. There are many complex dependencies in news recommender systems, and some of them are discussed in section 4. Conclusions and plans for further work are laid out in section 5.

1.2 Recommender systems

Recommender systems as a scientific discipline emerged in the early nineties as a particular branch of *information filtering* (Belkin and Croft, 1992). The general idea is to define techniques for predicting user responses to a given set of options on the basis of information about the options, the users and their interdependencies. The discipline draws on research from cognitive science, information retrieval, prediction theories and management science, as well as lately from semantic web and data mining (Borge and Lorena, 2010).

Formally, the problem in news recommendation is that of estimating and ranking the evaluations of articles unknown to the user. To compute these estimates, evaluations of other articles by the same users or evaluations by other users with similar interests may be used. Following Borges and Lorena (2010), we can define a set of users *U* and a set of news articles *A*; let *s* be a utility function (Equation 1) that defines the evaluation of an article *a* for a user *u*:

$$s: U \times A \rightarrow V \quad [1]$$

in which *V* is a completely ordered set formed by non-negative values within an internal, e.g., 0 to 1 or 0 to 100. The system is to recommend an article a' that maximizes the utility function (Equation 2) for the user:

$$a' = \arg\max_{a \in A} s(u, a) \quad [2]$$

An element in U can be defined by a number of characteristics that constitute the user profile of a particular user. Similarly, elements from A may be given different characteristics, depending on what information is available about the article. For example, a news article may have characteristics such as title, category, publication date, publisher, entities and locations.

It is important to note that the utility function *s* is not defined in the whole space $U \times A$. Estimating or extrapolating evaluations for the blanks in this space is the goal of the recommender system itself. In doing so, a whole battery of techniques may be applied, including decision trees, Bayesian classifiers, support vector machines, singular value decomposition, neural networks, clustering and information retrieval similarity scores (Adomavicius and Tuzhilin, 2005; Rajaraman and Ullman, 2011).

Recommender systems use a number of different technologies that can be classified into two broad groups: content-based filtering and collaborative filtering.

In *content-based filtering,* a new article is recommended to a user if it exhibits important similarities with her user profile. Since the user profile is constructed on the basis of her previously read articles, the recommended articles are those that are similar to articles in which she has shown interest in the past. In a vector-based system, both the user profiles and the news articles are represented as vectors in which term frequencies indicate the prominence of topics and entities, and a simple cosine similarity score may be computed to assess an article's relevance to a user. More sophisticated methods also take advantage of semantic reasoning and domain ontologies (e.g., Cantador, Bellogin and Castells, 2008). According to Borges and Lorena (2010), content-based filtering methods are effective at recommending unrated news articles, though the methods find it difficult to analyze the quality of articles or to recommend new or surprising stories that are not encoded in the user's reading history (serendipitous recommendations).

C*ollaborative filtering* can be seen as an automation of *word-of-mouth* recommendation. The idea is to recommend news articles to a user if they have been well evaluated in the past by people with similar preferences as the user. The approach can be further categorized into two types, memory-based and model-based, both of which are dependent on efficient techniques for grouping similar users together.

Compared to content-based filtering, collaborative filtering is able to make serendipitous recommendations, since similar users may still read articles that are not in the current user's own history. However, collaborative filtering needs a substantial amount of data in order to be effective. There are sparsity and cold-start problems that prevent the system from recommending new and relevant articles that have no historical ratings among the network of similar users (Borges and Lorena, 2010).

Recent *hybrid filtering* approaches try to combine the best features of content-based filtering and collaborative filtering. As demonstrated on a fraction of live traffic on Google News website by Liu, Dolan and Pedersen (2010), these combined approaches may both improve the quality of the recommendations and attract more frequent visits to the news site. More details about the various types of recommender systems, including knowledge-based filtering, are found in Jannach et al. (2010). Additional techniques that make use of contacts on social networks are presented in, for example, De Francisco Morales (2012), O'Banion, Birnbaum and Hammond (2012) and Shuai, Liu and Bollen (2012).

1.3 User click analysis

Successful news recommendation requires good and updated models of users' preferences. Unfortunately, users are often reluctant to give explicit feedback on news articles that can be inspected to construct and maintain user profiles (Thurman, 2011). This leaves us with the option of analyzing user click behavior to build user profiles that are consistent with their reading history and presumably useful in recommending interesting articles in the future. Most research on user click streams comes from the web search domain. Lee, Liu, and Cho (2005) build user models on the basis of click streams to enhance personalized web search. They infer search goals from analyzing how other users in the past have used the results of the same queries and their results suggest that the goals of up to 90 % of the search queries can be identified in this manner. Speretta and Gauch (2005) analyze click logs that consist of queries and documents clicked for every query. Like Kim and Chan (2003), they use the logs to learn user profiles that contain taxonomies of concepts, in which weights indicate the strengths of the relationships. In Nasraoui et al. (2008), clustering techniques are used to summarize user sessions into clusters that may serve as user profiles for the users in questions.

Billsus and Pazzani (2000) have developed a system for interpreting implicit user feedback on news articles presented in the Daily Learner. If a user clicks on the headline of an article, they assume that there is some basic interest in the article and set an initial score of 0.8. This score is gradually increased as the user requests more pages of the story, until a final score of 1.0 is reached if all pages have been viewed.

Similarly, a skipped article is assumed to be uninteresting and is given a negative score that is subtracted from the system's prediction score for the article. All these rated articles are afterwards combined to produce a user profile that lists weighted informative words associated with each user.

Liu, Dolan and Pedersen (2010) build user profile vectors that express users' evolving interests in specific news categories. For each user, they record the distribution of clicks and associate click rates with categories on a monthly basis. This allows them to analyze the proportion of time the user spends reading about each category as well as to reflect on the development of her interests from one month to another.

Interestingly, the design of the user interface heavily influences the user profiling techniques available to the system.

The Daily Learner can use a more fine-grained analysis than Google News because their users need to go through a series of clicks to confirm their interest and read the full news story. This may encourage the introduction of more complex user interfaces, though usability studies show that users are not very happy with news apps that require too much interaction.

2. Methods

2.1 SmartMedia news recommendation

The SmartMedia project at the Norwegian University of Science and Technology (NTNU) was initiated in 2011 in close collaboration with the regional media industry and the Norwegian telecom operator Telenor Group. Central in the project is the development of an iOS[1] news recommender system app for publishing and recommending news from a number of Norwegian newspapers. An architecture based on Big Data processing pipelines and search technologies is employed to deal with the constant flow of news that is added to the SmartMedia news index. The project focus is on recommendation technologies and semantic search, making use of NTNU's experience with large-scale advanced search platforms (see, for example, Gulla, Auran and Risvik (2002), Brasethvik and Gulla (2002) or Solskinnsbakk and Gulla (2010) for the technological background of SmartMedia).

A hybrid approach to news recommendation is adopted in SmartMedia. Freshness and locational information extracted from news events and users' mobiles are part of the recommendation strategies to make sure that new events in a user's neighborhood are given sufficient attention. Addressing users' particular interests and behavior, the system combines content-based and collaborative filtering to promote news that are consistent with her previously accessed articles or preferred by other users with similar interests.

[1] Apple's operating system for mobile devices

Due to the limitations of mobile devices, the system does not assume any direct user feedback on the articles presented to her. The user will not explicitly remove or promote any articles in news streams recommended to her, as opposed to what is common in most news reader apps today (Haugen, 2013). Also, the system does not access user data from other sources that may be used to construct user profiles in the system. The only information available to the system is the observed behavior of users retrieving and reading news in the SmartMedia iOS app. Since the explicit signals about the user's interests is so weak, it has been paramount to extend the analysis of user behavior to include a broad array of complementary implicit indicators of users' interests. We consider pre-read actions such as clicking on news articles, reading characteristics such as time spent in the article view, and post-read actions such as favoriting, sharing and e-mailing article links as indicators of the user's interests. This calls for a rather complex analysis, as there are dependencies between the user actions and not all actions should contribute in the same way and to the same extent in the resulting user profile.

Figure 1: Swiping movements are used to turn pages and navigate in news app

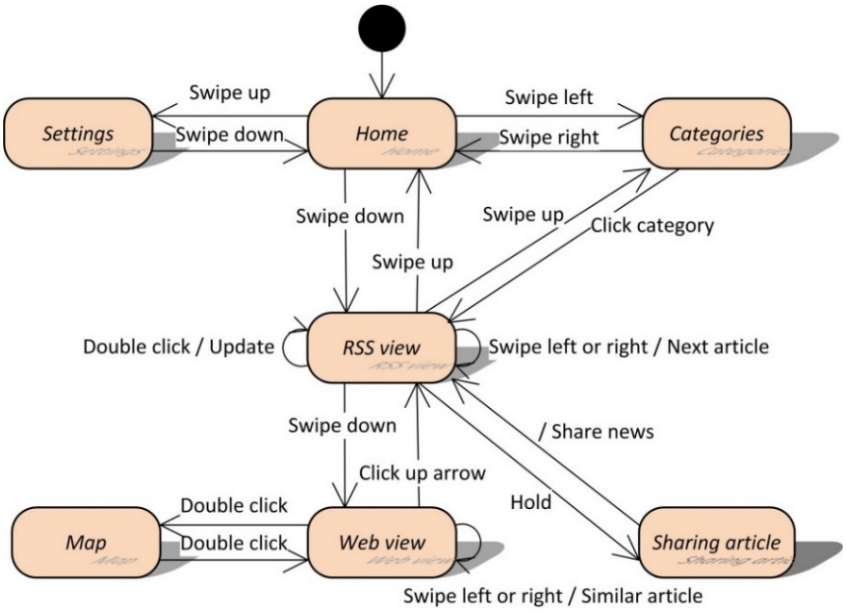

Figure 2: State transition diagram for app user interface

The SmartMedia news app has a pure gesture-based user interface as illustrated by the RSS (Rich Site Summary) and map screen shots in Figure 1. To accommodate the small screen and the lack of a proper keyboard, there are no buttons, and all navigation is done with swiping or clicking movements. The main page of the news app gives a small summary of the latest news and allows the user to log in if she would like to share her user profile across devices. Sliding side-menus are available from the main page to configure the app or select particular news categories. The user can swipe down to the RSS view, which lists - one story per page - the news recommended to this particular user. Horizontal swipes in the RSS view moves from one recommended article to the next, whereas vertical swipes take the user up to the main page or down to the full web view of the article. In the web view, the user may swipe horizontally to access related stories or double click to get a map showing where the news took place.

The state transition diagram in Figure 2 shows how the user is using gestures to navigate from page to page in the app. For more details about the implementation of the app, the reader is referred to Mozhgan et al. (2013).

2.2 Logging user behavior

The client-server architecture of the SmartMedia news app is a combination of a standard Solr[2] index for new articles, a Hadoop[3] cluster for generating user profiles, and a MongoDB[4] for event logs and generated user profiles. This Big Data approach ensures that the system can deal with the number of user actions that need to be recorded at the client side and analyzed and stored on the server side of the system.

On the server side, real-time RSS news streams from Norwegian newspapers are continuously analyzed and indexed in Solr for later recommendations. As shown in Figure 3, the indexing process accesses the RSS news before it retrieves the corresponding HTML documents.

After extracting the body texts from these HTML documents, the system extracts named entities from the texts, identifies the locations of the news using Google Maps, and stores the information about every news article in a structured Solr index.

If the article is not already categorized by the publisher, a simple classifier is used to annotate the news with the appropriate news categories. Associated with every article in the index are meta-data such as publication time, geo locations, key phrases/entities, categories, and publisher.

On the client side, a middleware layer is used to retrieve ranked news articles from the Solr index and present these to the user. The user may also inspect her own user profile in case she would like to overrule or reset the profile. The iOS client logs every gesture and user click and sends these logs back to the server for storing in the MongoDB database.

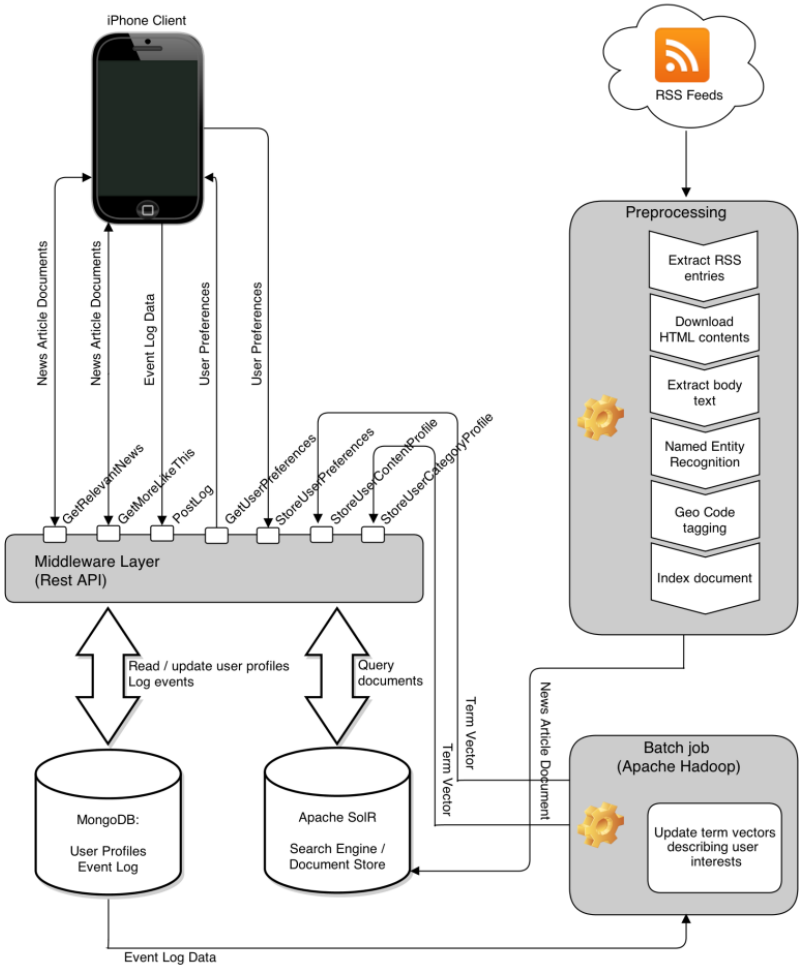

Figure 3: Architecture for user click behavior analysis

As seen in Table 1, the log information includes not only information about the user and the type of user action, but also properties of the news article itself that may be

[2] Solr is an open source enterprise search platform from the Apache Lucene projects.
[3] Hadoop is an open source software framework for processing of large data sets using clusters of hardware.
[4] MongDB is an open source document-oriented database system.

needed for the subsequent construction of user profiles. The fields 'Tags' and 'Categories' list the key-phrases and categories annotated with the article in the document index.

Table 1: Each user action is recorded as a separate record in a log database

User log data field	Description
ID	Unique identifier for the user action
User ID	ID of the user performing the action
Article ID	ID of the article subjected to the action
User Action Type	Type of user action
Timestamp	Time the action occurred
Geographical Location	Coordinates of the mobile user at the time of the action
Tags	Entities and keywords extracted from the article
Categories	Classification of article in news categories
Properties	Extra field for additional data

Since the user is not giving any explicit rating of news articles' attractiveness, the system needs to reason about her behavior to assess their value to her. The only information available, though, are the gestures and user clicks used to navigate around the user interface. On the basis of the user interface model in Figure 2, we have identified 10 user actions that may be taken as indicators of users' interests in the news article (see Table 2). For example, if the user is swiping down from the RSS view of an article to see the web view text, we assume that she has found the article interesting enough to read the full text. Similarly, we assume that she liked the article if she decides to share the article on Twitter or Facebook, store the article among her favorites, or send the article link by e-mail. If she is checking the map or requesting similar articles, it must also strengthen the view that she appreciated the article's content.

More complicated are the user actions linked to the time spent reading an RSS article or a full-text article. It seems reasonable to assume that the article is of interest to the user if she spends more time reading it than reading most other articles.

Of course, it varies from one user to another how much time is needed to read a news article. With regard to the implementation, this means that we need to monitor and store each individual user's reading habits and only consider these actions when the reading time exceeds the average reading time for this particular user.

Table 2: User actions that are used to construct user profiles

User action type	Description
Opened article view	User opened full text version of article
Time spent article view	Time the user spent viewing the article
Time spent preview	Time the user spent viewing the RSS version of article
Clicked category	User selected a news category
Shared twitter	User shared the article on Twitter
Shared facebook	User shared the article on Facebook
Shared mail	User shared the article on mail
Starred article	User added the article to favorites
Viewed map	User viewed the location of the article on a map
Viewed similar article	User accessed another article similar to current article

2.3 User profiles

A user profile is constructed for a chosen time interval. If there already exists a user profile for the current mobile device, the new profile is combined with the older one using a technique that renders the old profile less important than the new one.

Take the user action shown in Figure 4, which states that a user spent about 1.4 seconds reading the full text of a news article on June 2. The action, the article and the user are all given internal identifiers by the system. The event described in the article has been associated with a pair of geographical coordinates, is of the NEWS category, and seems to be dealing with an Indian man in the southwestern part of Norway that got into trouble and was imprisoned by the police.

If 1.4 seconds is longer than this user's normal reading time, the action should be taken into account when building the user profile.

The key phrases in Figure 4 point to a serious challenge in analyzing news stories. Since the terminology changes rapidly in the news domain and sentences may be both ambiguous and sometimes directly ungrammatical for literary effects, it is often difficult to extract proper key phrases and entities from stories. In this case, for ex-

ample, both *'havnet'* (ended up) and *'satt'* (sat) are unsuitable as key phrases, implying that our Named Entity Recognition (NER) component suffers from a substantial amount of noise. Experiments on some articles show that almost 50 % of the phrases suggested by the NER component may in fact be noise.


```
{
"_id" : "241B50BE-DFF5-4AAB-A12D-98D4A4606028" ,
"articleId" : "318218311" ,
"userId" : "bf4d2b7adec01da0ddc8c3317ø88bcdc6" ,
"eventType" : "TIME_SPENT_ARTICLE_VIEW" ,
"timestamp" : { "$date" : "2013-06-02T16:41:15.511Z"} ,
"geoLocation" :
        { "name" : " " ,
        "type" : " " ,
        "longitude" : 8.00354 ,
        "latitude" : 58.138821} ,
"properties" : { "duration" : "1.427272"} ,
"tags" :
        [ "agder politidistrikt" ,
        "havnet" ,
        "satt" ,
        "rebelsk mannen" ,
        "operasjonsleder" ,
        "kristiansund slo" ,
        "politiet" ,
        "kristiansand skallet" ,
        "Maharashtra" ,
        "India" ,
        "Egersund" ,
        "Rogaland" ,
        "Norway" ] ,
"categories" : [ "NEWS]}
```


Figure 4: A user action recorded by the iOS client

We assume that a user's general interests can be analyzed at two levels. At the top level she might have certain preferences for particular news categories, such as sports or lifestyle news. These categories correspond roughly to the standard categories used by newspapers to structure their own content. At a lower level, a user may prefer stories about particular events, persons, companies, products, etc. These are typically found as key phrases or entities in news articles, and a particular article may refer to many of these topics with various degrees of prominence. An article may for example be mostly about the Barcelona football club but there may also be parts of it referring to Real Madrid or other Spanish or non-Spanish football clubs, as well as to persons playing for these clubs.

Our user profile is comprised of two vectors:

$$P = < \vec{C}, \vec{K} >$$

- a **category vector** $\vec{C}$ in which each news category is given a weight that indicates the importance of this category to the user,
- a **content vector** $\vec{K}$ that lists all prominent key phrases and entities that the user may find interesting to read about. The weights indicate the user's relative interest in each key phrase and entity.

Both vectors are normalized so that the maximum weight of any category or key phrase is 100. The category vector below belongs to a user that has been reading mostly news and entertainment stories.

$$\vec{C} = < (\text{"NEWS"}, 100.0), (\text{"SPORTS"}, 13.1), (\text{"TRAVELING"}, 7.5), (\text{"LIFESTYLE"}, 80.3), (\text{"ECONOMY"}, 3.14) >$$

The interpretation of $\vec{C}$ is that this user prefers straight news stories and to some extent lifestyle news, and she is not very likely to read about economic affairs. The content vector below is more difficult to interpret, as it seems that the user's interests cover a wide spectrum of news categories. Again, the terms 'i', '2', '3' and 'ST 13' should probably not have been part of the content vector.

$\vec{K}$ = < ("46354", 1.866), ("Akatsi", 1.866), ("Forbruker", 1.866), ("mortenthomassen", 0.933), ("jeg", 0.9444), ("Lodzkie", 19.633), ("Strømstad", 1.8666, ("PaysdelaLoire", 1.866), ("150", 2.833) ("RueBenjaminFranklin", 1.866), ("Nordrhein-Westfalen", 1.866), ("rachelnordtømme", 1.866), ("Nordland", 7.555), ("¡", 0.933), ("3", 2.811), ("ST 13", 19.633), ("2", 81.844), ("BestWestern Anker Hotel", 1.888),... >

The content vectors grow as users read more stories and will ultimately contain thousands of entities that the user may have found interesting at some point. However, only the higher weighted terms will be important at the recommendation stage, and the less important terms can be readily ignored or even removed from the vectors.

2.4 Automatic construction of user profiles

User profiles are constructed in two steps: (1) build a time-constrained profile that covers the time from when the last profile was generate up until the current time, and (2) merge the old profile with the time-constrained profile.

The following steps show how user *u*'s time-constrained user profile is built for the time period from t_0 to t_1:

1. *Define a user action set S that contains all user actions from t_0 to t_1 for user u.*
2. *Assume a weight of 1 for all user actions in S (all actions are equally important).*
3. *Remove user actions from S about timed reading events if the time spent is less than the average time for u.*
4. *Form a category vector, in which the weight of each category represents the total number of occurrences of the category in the actions in S.*
5. *Form a content vector, in which the weight of each key phrase/entity represents the total number of occurrences of this phrase/entity in the actions in S.*
6. *Normalize category and content vectors.*

Merging the old user profile with the new time-constrained profile can be done in different ways. Intuitively, we would want to modify the old profile so that your later interests become more important than your old ones. This will make sure that irregular and important events that are unfolding right now receive enough attention, even if these topics are not necessarily found in the old user profile.

On the other hand, we need to be careful about deleting parts of the old user profile. If there are elements in the old profile that are not present in the time-constrained profile, the reason is not necessarily that the user's interests or preferences have changed. It might simply be that there has not been any news lately about those particular topics, and the user would be happy to have the topics recommended when there is news about them again.

We assume that the new time-constrained vector should be given more weight than the old user profile. There are different ways of implementing this "forget me" function, but we have adopted the following formula in our project:

$$\overrightarrow{V_u} = \overrightarrow{V_n} + c\overrightarrow{V_o} \qquad [3]$$

The new user profile $\overrightarrow{V_u}$ is given by the sum of the new time-constrained profile $\overrightarrow{V_n}$ and the multiplication of the old profile $\overrightarrow{V_o}$ with a constant with a value between 0 and 1. If *c* is set to 1, old and new user behavior count as equally important in the updated user profile. Only the latest user behavior is considered in the new user profile if *c* is set to 0.

3. Results

3.1 User profiles from interaction with news app

The SmartMedia news recommender app is already in operation and contains close to 150 000 news articles. Each day, around 1 500 articles are added from a total of 89 newspapers in Norway. The average newspaper article is 220 words long, if we exclude finance news that are usually substantially longer than news from other categories. Statistically, each article contains 1.6 location names, 2.3 person names, 2.3 organization names and 0.8 role names.

On the left-hand side of Figure 5 we show the user profile of a particular user at time t_0. The profile is automatically built from logging and analyzing all actions by the user up until time t_0. To simplify the presentation, only the top 20 terms of the content vector are shown, ordered according to their weights.

The category vector reveals that the user is using the app to read standard news stories. On the content level, we notice that there are numerous local geographical terms as well as some general terms from the latest news. Before any user profile is established, news articles are presented to the user on the basis of freshness and geographical proximity.

As the user profile is constructed, new articles are recommended and ranked according to how they match this user profile. In the current implementation, a con-

tent-based match is computed as the cosine similarity between the user profile and the vector representations of the news articles. The left-hand side of Table 3 lists the top 10 news recommended to the user at t_0. Whereas seven of the stories fall into the news category, there are also two articles about sports and one about lifestyle.

Most of the news stories concern events that take place in the vicinity of the mobile user. One explanation for this may be that the user has a strong preference for local news and has already in the past preferred such news articles. However, since geographical proximity is also used to recommend articles, her exposure to local news articles may have been so high from the outset that it was difficult not to view mostly local news.

3.2 Learning user profiles over time

Figure 5 also demonstrates the learning effect of the user profiling approach. The user profile is updated at regular intervals by combining the old profile with an analysis of what the user has read after the old profile was constructed. Whereas the user profile at time t_0 reflects her behavior up to time t_0, the new profile in t_1 for the same user is constructed from the profile in t_0 and a time-constrained profile that covers the time from t_0 to t_1. We can see that the user has gradually moved more into sports, either because she is genuinely more interested in sports or because there are relatively few proper news stories at this point. Her interests in finance news have fallen, though she seems more inclined to like travelling news now than in the past.

Time t_0			*Time t_1*	
Category C_0			Category C_1	
NEWS	*100.0*		*NEWS*	*100.0*
SPORTS	*1.47*		*SPORTS*	*27.2*
LIFESTYLE	*4.41*	→	*LIFESTYLE*	*2.91*
TRAVELING	*1.47*		*TRAVELING*	*18.4*
ECONOMY	*8.82*		*ECONOMY*	*5.83*
Key phrases K_0			Key phrases K_1	
Norway	*100.0*		*Norway*	*100.0*
Trondheim	*87.5*		*Trondheim*	*79.6*
Sør-Trøndelag	*85.2*		*Sør-Trøndelag*	*78.4*
for	*39.8*		*politiet*	*24.1*
Nidelva	*38.6*		*for*	*22.2*
Espen Sandmo	*38.6*		*Nidelva*	*21.0*
Morten Karlsen	*38.6*		*Espen Sandmo*	*21.0*
politiet	*38.6*		*Morten Karlsen*	*21.0*
NTNU	*12.5*		*Borussia Dortmund*	*13.0*
Rogaland	*10.2*	→	*lionel messi*	*13.0*
nettportalen	*9.09*		*vif*	*13.0*
Stavanger	*9.09*		*robin van persie*	*13.0*
Internasjonalt hus	*7.95*		*bayern munchen*	*13.0*
Internasjonalt kontor	*7.95*		*lambert*	*13.0*
University of Stavanger	*7.95*		*mourinho*	*13.0*
S O Bragstads Plass	*7.95*		*ole gunnar solskjær*	*13.0*
United States	*6.82*		*arve lote*	*13.0*
mms	*6.82*		*manchester united*	*13.0*
sms	*6.82*		*dortmundhelten*	*12.3*
john stene	*6.82*		*eva hongshagen*	*12.3*
...	...		...	...
...	....		...	...

Figure 5: Old and new user profile for a particular user

The news articles recommended at time t_1 reflect the changes of the user profile and are now totally dominated by sports news. As shown on the right hand side of Table 3, the top 10 stories recommended are in fact sports news, of which three are about local affairs, three concern national sports and four relate to sports events outside Norway.

Because the user profiles are updated on the basis of the users' current behavior, they tend to improve over time as the system learns more about the users' interests and preferences. This learning effect is important, as it means that the news recommendations will also improve if the user spends more time using the app. The dynamic nature of news streams, however, poses some particular problems to the news recommender systems. The selection of news stories changes continuously, since old stories grow outdated and new stories are added as events unfold. This means that the set of recommended stories will change from one point in time to another, even if the user profile is not changed, simply because the set of available stories is not the same. Consequen-

tly, the success of a particular user profile is not only decided by the content of the profile, but also by the availability of articles consistent with the profile. If there are no desirable articles available, the user risks watering down her profile when quickly inspecting the not so interesting articles presented by the app.

Table 3: Top recommended articles change as user's profile is developing

	Top 10 news at t_0			**Top 10 news at t_1**	
1	*"Avslutter søk etter mann i Nidelva"*	*NEWS*		*"Dette blir ingen vanlig kamp for Rekdal"*	*SPORTS*
2	*"Søker etter mann i Nidelva"*	*NEWS*		*"Mancini sparket ett år etter ligagullet med City"*	*SPORTS*
3	*"Liverpool-legende i Namsos"*	*SPORTS*		*"Hvert mål koster eliteserieklubbene over en million kroner"*	*SPORTS*
4	*"Her er kong Haakons flukthil"*	*NEWS*		*"Wigan rykker ned - Arsenal nærmer seg mesterligaen"*	*SPORTS*
5	*"Lettere skadd etter utforkjøring"*	*NEWS*		*"Dekket over tabber med å skjelle ut journalister"*	*SPORTS*
6	*"Posten må leie inn lastebiler"*	*NEWS*		*"Rooney buet ut av Uniteds parademarsj"*	*SPORTS*
7	*"Flere 5-åringer har hull i tennene"*	*NEWS*		*"Elisabeths drømmemål går verden rundt"*	*SPORTS*
8	*"Hittil har vi vært heldige"*	*NEWS*		*"Nå håper Aalesunds superspiss på mer"* TS)	*SPORTS*
9	*"Trondheim ble årets kulturkommune"*	*LIFESTYLE*		*"Stegavik til Kostad - om dama blir i Byåsen"*	*SPORTS*
10	*"Nå lever tjuvfiskerne farlig"*	*SPORTS*		*"Høgmo blir trener i Djurgården"*	*SPORTS*

4. Discussion

The experiments so far suggest that user profiles built from the content of user's already read news articles produce recommendations that are consistent with her general preferences. There are, however, a number of issues that affect the quality of the profiles as well as the quality of the recommendations in the next round.

Since the whole profiling process starts with the categories and key phrases associated with each news article, the quality of these categories and key phrases is of paramount importance. A simple k-nearest approach is used to classify the documents, and the tests so far show that annotated categories are very reliable. For the key phrases the situation is more challenging, as it is notoriously difficult to extract named entities and prominent phrases from domains that are characterized by rapidly changing terminologies and collapsed grammatical constructions. The evaluation of the initial Named Entity Recognition component shows an accuracy that is clearly not satisfactory, and we are now in the process of implementing an improved NER component.

The user profile construction process itself is complex with several strategies that need to be carefully calibrated and may also possibly interfere with each other:

- *Weighting scheme for user actions.* So far we have assumed that all user actions carry the same weight. Sensitivity tests indicate, however, that the *Opened article view* action dominates all other actions, and the other actions mostly serve to amplify the contribution from the *Opened article view* action. The one notable exception to this is the preview time action which added information about articles that were somewhat interesting to the user, but not interesting enough for her to access the full text. In practice, this additional information turned out to be substantial and amounts to 92% of the size of the final user profile.
- *Update frequency.* In the current system we do not have a clear policy for when a user profile is updated. We upload the existing profile when the user enters the app, and we update the profile when her session is finished. Since this implies that her latest articles will not normally be reflected in her profile, we are implementing a feature that allows the user to manually enforce an immediate update of her profile based on her ongoing session.
- *Merging old and new user profile.* Our formula for merging old and new profile vectors is rather coarse-grained using a simple multiplication constant between 0 and 1 for degrading the content of the old user profile. An alternative method would be to update each individual element of the old vectors on the basis of the time that has passed since its value was last set. This seems conceptually more correct, though it implies a computationally more expensive solution.
- *Short-term vs long-term user profiles.* Earlier research by Billsus and Pazzani (1999) and Liu, Dolan and Pedersen (2010) argues that user profiles need to be divided into short-term profiles and long-term profiles. Short-term profiles relate to popular topics such as big or surprising events and need to be updated rapidly as the events take place. Long-term interests account for the user's general preferences and are assumed to be fairly stable from one session to another. We have in our work only defined one user

profile that addresses mostly long-term interests, assuming that separate recommendation strategies based on freshness and collaborative filtering will bring in news that cater for the user's short-term preferences.

It is still early to conclude about the quality of the user profiles, as their influence on the final recommendations is difficult to separate from other aspects of the recommendation engine itself. In a hybrid recommenddation system there are several recommendation strategies that need to be weighted and combined to produce the final results (Borges and Lorena, 2010). In our case, the weights of freshness, geographical proximity, collaborative filtering and content-based filtering will severely affect the impact of the learned user profiles, and any changes to these weights may necessitate adjustments to how these profiles are generated.

A separate issue is the relative weighting of categories and content in the user profile, which also affects the profile's effect on the recommendations made.

Currently these weights are assumed to be equal, though the weights may need to be further refined as part of an analysis of the recommendation system's total weighting scheme.

Finally, there are two parties involved in news recommendation, the news *provider* and the news *reader*, that evaluate the quality of recommendations from two very different perspectives. Whereas the news reader is mostly interested in getting only news consistent with her current interests, the news provider wants to present news that extend her interests and thereby hold on to them as active news app users. Serendipity is important for this reason, but also the recommendation of not so relevant news in cases where no new relevant stories have come in.

Ultimately, news providers measure the success in terms of click-through rates, even though these may only be moderately correlated with the readers' perception of news relevance.

5. Conclusion

We have in this paper presented an approach for learning user profiles from observing the user's own actions on a mobile news app. No explicit information about user preferences is given or retrieved from other sources in this process. The user profiles are afterwards used by a hybrid news recommender engine to produce a personalized mobile news service.

Current research on recommendation technologies has had an emphasis on recommendation strategies and there is only limited research on the construction of user profiles from mobile user behavior. Our approach analyzes the content of news articles and associates the users with user profile vectors that aggregate the contents of previously read articles. The two profile vectors, one for representing category preferences and another for representing content preferences, are both extracted using standard text mining techniques for classification and entity extraction.

The results so far suggest that the user profile captures important aspects of the user and leads to recommendations more consistent with her general preferences. These profiles, however, typically support the content-based recommendation part of the news app, and there are other strategies that should deal with short-term issues and news relevant to the geographical neighborhood. The weighting of recommendation strategies in such a hybrid recommender system is challenging and a topic for further research.

In our further work we plan to refine the construction of user profile vectors and evaluate different strategies for user action weighting, update frequencies, and merging of profile vectors. The same user profiles will also gradually be extended to support collaborative filtering, which means that they may need to incorporate aspects that have so far not been needed for content-based filtering.

Acknowledgements

This research was supported by Telenor Group as part of their collaboration with the Department of Computer and Information Science at the Norwegian University of Science and Technology in Trondheim, Norway.

References

Adomavicius, G. and Tuzhilin, A., 2005. Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions. *IEEE Transactions on Knowledge and Data Engineering*, 17(6), pp. 734-749

Belkin, N. J. and Croft, W. B., 1992. Information filtering and information retrieval: two sides of the same coin? *Communications of the ACM,* 35(12), pp. 29-38

Billsus, D. and Pazzani, M. J., 1999. A hybrid user model for news story classification. In *Proceedings of the Seventh International Conference on User Modeling* (UM'99). Berlin-Heidelberg: Springer. pp. 99-108

Billsus, D. and Pazzani, M. J., 2000. User Modeling for Adaptive News Access. *User Modeling and User-Adapted Interaction*, 10, pp. 147-180

Borges, H.L. and Lorena, A. C., 2010. A Survey of Recommender Systems for News Data. In: Szczerbicki. E., ed. *Smart Information and Knowledge Management*, SCI 260. Berlin-Heidelberg: Springer. pp. 129-151

Brasethvik, T. and Gulla, J. A., 2002. A conceptual modeling approach to semantic document retrieval. In: *Proceedings of the 14th International Conference on Advanced Information Systems Engineering (CAiSE'02).* Berlin-Heidelberg: Springer. pp. 167-182

Cantador, I., Bellogin, A. and Castells, P., 2008. Ontology-Based Personalised and Context-Aware Recommendations of News Items. In: *Proceedings of the 7th International Conference on Web Intelligence.* IEEE. pp. 562-565

Circa, 2013. *Catch up quick.* [online] Available at: http://cir.ca/. [Accessed 18 April 2013]

Das, A. S., Datar, M., Garg, A. and Rajaram, S., 2007. Google news personalization: scalable online collaborative filtering. In: *Proceedings of the 16th international conference on World Wide Web.* ACM. pp. 271-280

De Francisci Morales, G., Gionis, A. and Lucchese, C., 2012. From chatter to headlines: harnessing the real-time web for personalized news recommendations. In: *Proceedings of the Fifth ACM international conference on Web search and data mining.* ACM. pp. 153-162

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

Haugen, K. R., 2013. *Mobile News: Design, User Experience and Recommendation.* MSc thesis. Department of Computer and Information Science, Norwegian University of Science and Technology, Trondheim

Jannach, D., Zanker, M., Felfernig, A. and Friedrich, G., 2010. *Recommender Systems: An Introduction.* Cambridge University Press

Kim, H. R. and Chan, P. K., 2003. Learning implicit user interest hierarchy for context in personalization. In: *Proceedings of the 8th international conference on Intelligent user interfaces.* ACM. pp. 101-108

Lee, U., Liu, Z. and Cho, J., 2005. Automatic identification of user goals in web search. In: *Proceedings of the 14th international conference on World Wide Web.* ACM. pp. 391-400

Liu, J., Dolan, P. and Pedersen, E.R., 2010. Personalized news recommendation based on click behavior. In: *Proceedings of the 15th international conference on intelligent user interfaces.* ACM. pp. 31-40

Nasraoui, O., Soliman, M., Saka, E., Badia, A. and Germain, R., 2008. A web usage mining framework for mining evolving user profiles in dynamic web sites. *IEEE Transactions on Knowledge and Data Engineering.* 20(2), pp. 202-215

Nilsen, J. E. B., 2013. *Large-Scale User Click Analysis in News Recommendation.* MSc. Department of Computer and Information Science, Norwegian University of Science and Technology, Trondheim

O'Banion, S., Birnbaum, L. and Hammond, K., 2012. Social media-driven news personalization. In: *Proceedings of the 4th ACM RecSys workshop on Recommender systems and the social web.* ACM. pp. 45-52

Rajaraman, A. and Ullman, J. D., 2011. *Mining of Massive Datasets.* Cambridge University Press

Schafer, J. B., Konstand, J. and Riedl, J., 1999. Recommender Systems in E-Commerce. In: *Proceedings of the 1st ACM conference on Electronic Commerce (EC'99).* New York: ACM. pp. 158-166

Shuai, X., Liu, X. and Bollen, J., 2012. Improving news ranking by community tweets. In: *Proceedings of the 21st international conference companion on World Wide Web*. ACM. pp. 1227-1232

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

Speretta, M. and Gauch, S., 2005. Personalized search based on user search histories. In: *Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence.* IEEE. pp. 622-628

Tavakolifard, M., Gulla, J. A., Almeroth, K. C., Ingvaldsen, J. E., Nygreen, G. and Berg, E., 2012. Tailored News in the Palm of your HAND: A Multi-Perspective Transparent Approach to News Recommendation. In: *Proceedings of 22nd International World Wide Web Conference (WWW'13), Companion Volume.* Rio de Janeiro. pp. 305-308

Thurman, N., 2011. Making 'The Daily Me': Technology, Economics and Habit in the Mainstream Assimilation of Personalized News. *Journalism: Theory, Practice & Criticism*, 12(4), pp. 395-415

Yeung, K., 2013. *News curator Summly launches to help simplify the way we consume news on mobile devices.* [online] Available at: http://thenextweb.com/apps/2012/11/01/news-curator-summly-launches-to-help-simplify-the-way-we-consume-news-on-mobile-devices/#!p1eeM. [Accessed 17 April 2013]