manuscript No.
(will be inserted by the editor)

A User-Centric Evaluation of Context-aware Recommendations for a Mobile News Service

Toon De Pessemier · Cédric Courtois · Kris Vanhecke · Kristin Van Damme · Luc Martens · Lieven De Marez

Received: date / Accepted: date

Abstract Traditional recommender systems provide personal suggestions based on the user's preferences, without taking into account any additional contextual information, such as time or device type. The added value of contextual information for the recommendation process is highly dependent on the application domain, the type of contextual information, and variations in users' usage behavior in different contextual situations. This paper investigates whether users utilize a mobile news service in different contextual situations and whether the context has an influence on their consumption behavior. Furthermore, the importance of context for the recommendation process is investigated by comparing the user satisfaction with recommendations based on an explicit static profile, content-based recommendations using the actual user behavior but ignoring the context, and context-aware content-based recommendations incorporating user behavior as well as context. Considering the recommendations based on the static profile as a reference condition, the results indicate a significant improvement for recommendations that are based

T. De Pessemier - K. Vanhecke - L. Martens
iMinds - WiCa - Ghent University, Dept. of Information Technology
G. Crommenlaan 8 box 201, 9050 Ghent, Belgium
Tel.: +32-09-33-14908
Fax: +32-09-33-14899
E-mail: toon.depessemier@ugent.be
E-mail: kris.vanhecke@ugent.be
E-mail: luc1.martens@ugent.be

C. Courtois - K. Van Damme - L. De Marez
iMinds - MICT - Ghent University, Dept. of Communication Sciences
Korte Meer 7-9-11, 9000 Ghent, Belgium
Tel.: +32-09-264 97 67
E-mail: cedric.courtois@ugent.be
E-mail: kristin.vandamme@ugent.be
E-mail: lieven.demarez@ugent.be

on the actual user behavior. This improvement is due to the discrepancy between explicitly stated preferences (initial profile) and the actual consumption behavior of the user. The context-aware content-based recommendations did not significantly outperform the content-based recommendations in our user study. Context-aware content-based recommendations may induce a higher user satisfaction after a longer period of service operation, enabling the recommender to overcome the cold-start problem and distinguish user preferences in various contextual situations.

1 Introduction

Recommender systems are software tools and techniques providing suggestions for items to be of interest to a user [30]. 'Item' is the general term used to denote what the system recommends to users such as videos [14], songs [22], or events [12]. The suggestions provided are aimed at supporting their users in various decision-making processes, such as what movies to watch, what music to listen, which events to attend, what products to buy, or what news to read. These suggestions are often offered as a ranked list of items. In performing this ranking, recommender systems try to predict what the most suitable items are, based on the user's preferences. These user preferences are acquired by collecting users' explicit ratings for items or by interpreting user actions such as clicks, selections, or purchases [33]. A rating or a user interaction with an item is commonly referred to as a 'consumption' of an item. Consumptions can be explicit or implicit. Explicit evaluations for content items, such as ratings, likes, or thumbs up/down options are denoted as explicit feedback. Implicit feedback are observational measures of the user's preference for an item, e.g., selecting an article to read, purchasing a product, watching a movie, or listening to a song.

Recommender systems have proven to be a powerful and popular tool for online users to cope with the information overload and assist in the decision-making process [24]. Over the years, several different approaches for generating recommendations have been proposed. These can be classified according to a taxonomy that was based on the work of Burke [10] and that has become a classical way of distinguishing between recommender systems and referring to them [33]. Six different classes of recommendation techniques can be distinguished based on the knowledge source: demographic recommendations, knowledge-based recommendations, community-based recommendations, content-based recommendations, collaborative recommendations, and hybrid recommendations. Demographic recommendations are based on the demographic profile of the user (e.g., age, gender) to infer personal recommendations. Knowledge-based systems use specific domain knowledge about how certain item features meet user preferences. Community-based recommendations rely on the user's connections or relations with other users or friends in a social network. The idea of content-based recommender systems is to suggest items that are similar to the items that the user liked in the past. These sys-

tems learn which type of items the user likes based on the user's consumption behavior and the attributes that describe an item such as the title, a description, keywords, categories, etc. Collaborative filtering is based on similarities in the consumption history of users to derive personal recommendations. Hybrid recommender systems combine two or more of the preceding recommender techniques. In a footnote, Burke argues that there is another knowledge source not yet included in the classification: context, which is becoming important, particularly for mobile applications.

The consumption of audiovisual media and the accessing of information always happen in a certain context [32], i.e. conditions or circumstances that significantly affect the decision behavior. This gave rise to the development of context-aware recommender systems (CARS), which take this contextual information into account when providing recommendations. To calculate recommendations, traditional recommender systems estimate preferences for all (User, Item) combinations for which no feedback is available, based on the known preferences of a few other (User, Item) combinations.

$$User \times Item \rightarrow Preference$$

To take into account the influence of context, CARS are the extension towards a multidimensional model, used to estimate preferences for all (User, Item, Context) combinations [1].

$$User \times Item \times Context \rightarrow Preference$$

This research aims to assess the usefulness of contextual information for algorithms that assists people in the decision-making process. More specifically, the main objective is to determine the added value of CARS with respect to traditional recommender systems for a functional online news service, called Stream Store. Since a user study can provide reliable explanations as to which recommendation method is the best, and why one method is better than the other [35], three alternative recommendation methods for the news service are compared through such a user study.

More specifically, this paper aims to address the following research questions (RQ):

- RQ1: "Is the dataset of a start-up news service dense enough for collaborative filtering (i.e. a commonly-used recommendation technique based on similarities in users' consumption pattern?)".
- RQ2: "Do users consume news content in several different contextual situations (e.g., at different times of the day, or using different devices?)"
- RQ3: "Do users of a news service have another consumption behavior (e.g., different preferred news categories) in different contextual situations?"
- RQ4: "How do recommendations based on explicit static preferences, content-based recommendations, and context-aware content-based recommendations compare for a news service?"
- RQ5: "Are explicitly-stated news preferences, the run time profile of a context-aware recommender, and the actual user behavior consistent?"

The remainder of this paper is organized as follows. Section 2 provides an overview of related work regarding CARS. Section 3 provides some details about Stream Store, i.e. the online news service, and discusses the setup of our experiment. The three alternative recommendation methods, the data they require, and their functioning are discussed in Section 4. An analysis of the usage of the news service (RQ1) is presented in Section 5. The usage of the news service in different contextual situations (RQ2) is investigated in Section 6. Section 7 elaborates on this and investigates if the users of the news service have another consumption behavior in different contextual situations (RQ3). In Section 8, the recommendations are evaluated by analyzing the feedback of the users (RQ4). Section 9 investigates the consistency of the user profile, on which the recommendations are based, and the actual user behavior, which reflects the user's interest (RQ5). Section 10 draws conclusions and points to future work.

2 Related Work

Driven by the increased availability of mobile Internet and the rise of a variety of mobile devices, context-aware applications experienced a growing interest during the last few decades. Context-aware systems use context to provide relevant information and/or services to the user, where relevancy depends on the user's task [16,28]. Multiple definitions of context exist in literature [7]. In the literature pertaining to context-aware systems, context was initially defined as the location of the user, the identity of people near the user, the objects around, and the changes in these elements [34,4]. In subsequent studies, the definition of context is broadened to elements such as the date, the season, and the temperature [9]. Afterwards, the context is completed with the user's emotional status and any information which can characterize - and is relevant to - the interaction between user and application [17].

Also in recommender systems for various application domains, the user context has gained an increased interest from researchers [4]. For context-aware music recommendations for example, the user's emotions can be used as input by using support vector machines as emotional state transition classifier [22]. In the application domain of tourism, various applications use the current location or activity of the user to personalize and adapt their content offer to the current user needs [31,15]. Personal recommendations for points of interest can be provided based on the user's proximity of the venue [26]. For context-aware media recommendations on smartphones, the situation of the user (location, activity, time) [20], as well as the device and network capabilities are considered as important contextual parameters [40]. In a news recommendation scenario, the importance of contextualization is demonstrated by the News@hand system [11].

Given the short-term life of news content, an alternative way to deliver news content is proactively pushing news articles to mobile users. These just-in-time personalized recommendations can be selected based on user's contextual

information as well as the news content [39]. Since users commonly consider the presentation of any information not explicitly requested as intrusive [25], this kind of push-recommendations is not used in our experiment. Pull-based systems, in which the delivery of recommended content is driven by queries, i.e., by user requests, are considered as less intrusive, since users maintain control on information delivery [21].

Most of the work on context-aware recommender systems has been conceptual [4], where a certain method has been developed, but testing is limited to an offline evaluation or a short-term user test with only a handful of people, often students or colleagues who are not representative for the population.

In contrast, this research investigates the role of context in users' consumption behavior and personalized recommendations for news, based on a large-scale user study. Users could utilize a real news service that offers content of four major Flemish news companies on their own mobile devices, in their everyday environment, where and when they wanted, i.e., in a living lab environment. Living lab experiments are an extension towards more natural and realistic research test environments [19]. Although less transparent and predefined, living lab experiments aim to provide more natural settings for studying users' behavior and their experience. Especially for context-aware applications, in which the user's environment has an influence on the way the application works and/or the offered content, a realistic setting is essential for a reliable evaluation. Therefore, this paper investigates the influence of context and the benefit of context-aware recommendations for a real news service by means of a large-scale user panel, in a realistic environment, over a longer period of time.

3 Experimental setup

Nowadays, users have access to multiple online news sources, also on the mobile platform. Various news services have a mobile website, and almost every newspaper has its own mobile app. This indicates the strong interest of users in news, but also the saturation of the market with many alternative services, besides the traditional printed press. Stream Store differentiates from existing news service on three fronts: aggregating news of different providers, bundling news articles into packages with news articles about a specific topic (i.e. 'streams'), and personalized context-aware recommendations.

Stream Store aggregates content of different premium content providers in Belgium: newspapers, magazines, but also content of television as short video clips. This aggregated content provides users a more complete and varied overview of the news than traditional services do.

To provide a clear overview of the content, news items can be grouped into streams about a certain topic such as typhoon Haiyan, Barack Obama, or the FIFA World Cup. Figure 1 shows a screenshot of the user interface of Stream Store, in which some streams are visible.

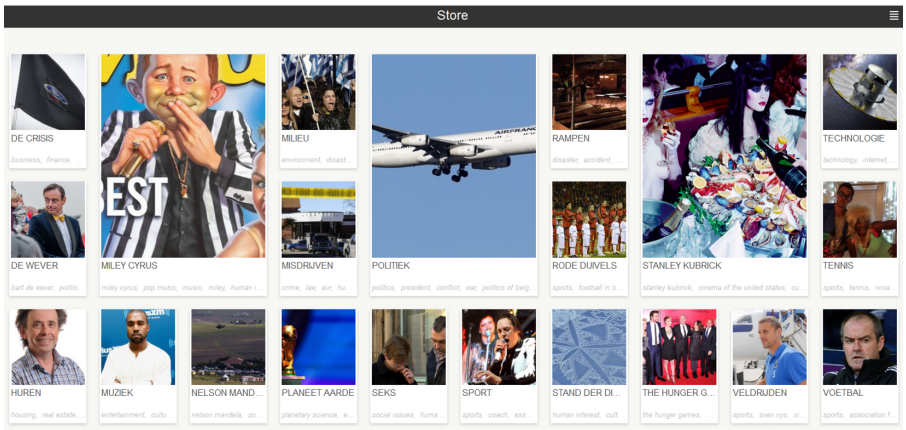

Fig. 1 A screenshot of the client application of the Stream Store news service

To anticipate the abundance of news content and the associated choices that people have to make, Stream Store offers the user personalized, context-aware recommendations. These are suggestions for news items, tailored to the user's personal interests, and taking into account the current contextual situation of the user. CARS can take into account various contextual aspects, which might have an influence on users' decision-making process. In this experiment, the time of the day and the device type are investigated in detail as contextual aspects.

Compared to the well-established application domains of recommender systems, such as movies or books, news items have a shorter lifespan and are frequently updated. Consequently, consulting the news on a daily basis, or even multiple times a day, can be interesting, which makes the time aspect an important contextual factor. The time is closely related to the location of the user, as was also witnessed during the analysis of users' consumption behavior (Section 6). A frequently recurring pattern was as follows: in the morning, users are at home; during daytime, they are at work; and during the evening, they are again at home.

The Stream Store news service is accessible through a mobile application, which is available on Android and iOS for tablets as well as smartphones. As a result, the type of device that is used for consuming the news content is also an interesting contextual factor.

To conclude, the time of the day (which is closely related to the location) and the device type (smartphone or tablet) are studied as contextual influences of the decision-making process in Stream Store. Because recommendation algorithms require sufficient time to learn variations in users' interests in different contextual situations, the number of contextual dimensions is limited to two in this experiment. Given the availability of online information services, e.g., for the weather or traffic, and the presences of various sensors in mobile devices,

such as GPS, accelerometers, gyroscopes, and illumination sensors, context can be extended with additional aspects in the future.

To evaluate the Stream Store news service and answer the research questions listed in Section 2, a user study was conducted. For this user study, 120 test users were recruited by an experienced panel manager from iMinds-iLab.o[1] (i.e. a research division with a strong expertise in living lab research and panel management). These test users, all owning a smartphone and/or tablet, belong to the target group of an online news service such as Stream Store.

The test users could install the Stream Store app on their smartphone and/or tablet and freely use the service during the evaluation period of around 5 weeks. These test users were divided into three groups, each receiving a different type of recommendations, as explained in Section 4.

To analyze the consumption behavior of the test users, their interactions with the service were logged from November 8th of 2013 (the start of the experiment) to December 13th of 2013. These logging data provided insight in the actual use of the application: 10 test users did not install the app, or did not use the Stream Store service during the evaluation period. They are excluded from the analysis, so that the number of actual participant was reduced to 110.

4 Recommendation Types

For this experiment, three types of recommendations were generated and evaluated in terms of user interaction. Table 1 lists the types of recommendations that are generated, the input data they use, and the number of test users who received that type of recommendations during the evaluation period. Each user received only one recommendation type during the whole evaluation period. Because some test users dropped out just before the evaluation period, the different recommender types are not evaluated by the same number of test users.

In user studies, it is beneficial to hide the true purpose of the experiment from the test users, to avoid people's natural subconscious tendency to meet the expectations of the study [35]. Therefore, test users were not aware of the existence of multiple types of recommendations.

4.1 Recommendations Based on Explicit Static Preferences

Before the actual experiment, candidate test users were asked about their preferences for different categories of news content through an online questionnaire (N = 464). In this questionnaire, users could specify their interests for the news categories (National, International, Culture, Economy, Lifestyle, Politics, Sports, and Interesting facts) on a 5-point rating scale. Users could

[1] http://www.openlivinglabs.eu/livinglab/iminds-ilabo

refine their preferences by specifying a score for each category for different times of the day (in the morning, at noon, during the daytime, and during the evening). The answers on this questionnaire are used as initial profile of the user in order to generate personal recommendations for news items. During the evaluation period, these preferences are considered static; user profiles are not updated based on the user's explicit or implicit feedback on the content, and the recommender is not learning from the user's behavior.

4.2 Content-based Recommendations

These recommendations are not based on a prior questionnaire but rather on the implicit and explicit feedback users provide during the evaluation period. Users can request news articles or videos to read or watch, which is considered as an implicit signal of interest in that news item. Besides, users can evaluate the news item that they consumed by means of a 'Thumbs Up' and a 'Thumbs Down' icon in the user interface, which is explicit feedback for the article or video.

Because feedback is gradually collected during the usage of the service, the profiles on which these recommendations are based, are dynamic and change constantly as users consume news content [13]. As a result, the recommender system is learning the user's preferences as the user is utilizing the news service. Because of the sparsity of the data set (see Section 5) and the availability of informative metadata about the content items, a content-based recommendation algorithm was used, namely the 'InterestLMS algorithm' of the Duine framework [36]. The InterestLMS algorithm builds a profile for every user based on the metadata describing the news items that are requested (implicit feedback) or evaluated (explicit feedback) by that user. This feedback is transformed into a rating score (e.g., thumbs up or like corresponds to the maximum rating, thumbs down is mapped on the minimum rating). For implicit feedback, the amount of time spent on a news item (reading time or watching time) is translated into a rating. These item ratings are then used to create or update the terms of interest in the user profile that correspond to the metadata fields of the content item (equation 1).

$$newTerm_i = currentTerm_i + updateModerator * (rating - predition) * factor \tag{1}$$

Here, the updateModerator is a constant that specifies the rate in which interests in subjects are updated in the user profile. The factor corrects for the number of terms ($\#terms$) that describe a content item: $factor = 1/\#terms$. The interest terms in the user profile are used to infer a prediction of the user's personal interest for a news item based on the terms describing the item (equation 2).

$$predition = \sum_i currentTerm_i * weight(Term) \tag{2}$$

Here, the weights can specify the relative importance of different terms (e.g., categories are considered as more important than keywords). By ordering and filtering the news items according to their prediction, the recommendations are generated based on the user's personal preferences for news content of different categories and characterized by different keywords.

To describe the content of the news items, eight different categories are used. The same category labels were used in the questionnaire to elicit the explicit static preferences of the users (Section 4.1). The keywords of the news articles are not predefined but are extracted from the text of the article using OpenCalais [37]. OpenCalais is a Web service that automatically creates rich semantic metadata for the content. It analyzes the news article and finds the entities within it, but it returns the facts and events hidden within the text as well. This way, the news article is tagged and analyzed with the aim of checking whether it contains information what the user cares about.

For the content-based recommendations, contextual aspects are not taken into account. In other words, contextual data, such as the device type and the time of the day, are ignored during the creation of the profile and the calculation of the recommendations.

4.3 Context-aware Content-based Recommendations

Just like the content-based recommendations, the context-aware content-based recommendations are not based on a prior questionnaire but self-learning based on the explicit and implicit feedback users provide during the evaluation period. For this type of recommendations, the 'InterestLMS algorithm' of the Duine framework is extended to take into account the context of the user. Before generating the recommendations, explicit and implicit feedback is processed by a contextual pre-filter [3]. The contextual information is used to select or filter the most relevant data for generating recommendations. Therefore, the day is partitioned into the four non-overlapping parts that are used in the filtering phase of the recommender: morning from 6:00 to 11:00, daytime from 11:00 to 12:00 and from 13:00 to 18:00, noon from 12:00 to 13:00 and evening/night from 18:00 to 6:00.

The relevance of implicit and explicit feedback for the context-aware news recommendations is determined by an *exact pre-filter* [4]. This means for example that if a user wants to read news during the evening, only feedback gathered during the evening is used to recommend news articles.

Different pre-filtering techniques have proven their efficacy in literature [6]. They all have to cope with the problem of context over-specification: focusing on the exact context is often a too narrow limitation. An overly specified context may not have enough training examples for accurately estimating the user's interests. For example, if a user rarely utilizes a tablet during noon to read news articles, the exact context (noon + tablet) may not provide enough data (feedback from the user) for accurately calculating the recommendations, which gives rise to the 'sparsity' problem. The sparsity problem in recom-

Table 1 Types of recommendations and the number of test users assigned to it

Recommender Type	Input Data	Number of Test Users
Explicit Static Preferences	Preliminary questionnaire	38
Content-based	Personal feedback	37
Context-aware Content-based	Personal feedback + context	35

mender systems literature refers to the situation that feedback or rating data are lacking or are insufficient for generating reliable recommendations [29] .

An appropriate solution for context over-specification is to use a more general context specification by applying context generalization [4]. Since certain aspects of the overly specific context may be less significant, the data filtering can be made more general in order to retain more data after the filtering for calculating recommendations.

For this experiment, in cases where not enough data was available, context generalization is applied in two phases. In a first phase, the time frame is broadened. E.g., if recommendations are requested for a user who is reading news on a tablet during noon, and the user has not yet provided enough feedback in that specific context in the past, then the data of that specific context is supplemented with the user's feedback coming from reading news on a tablet during other time periods of the day. If the amount of feedback is still insufficient after this first generalization, the context is further generalized. In a second phase, the device type is broadened. More specifically, the user's feedback provided on a specific type of device (e.g., a tablet) is supplemented with the user's feedback provided on other device types (e.g., a smartphone). We opted to apply the generalization first on the time aspect of the context, and in a second phase on the device type, since many users are utilizing the service during different time periods (Section 6.1) but on only one type of device (Section 6.2).

One major advantage of the contextual pre-filtering approach is that it allows deployment of any of the traditional recommendation techniques [2]. This makes it possible to use the same underlying algorithm for the context-aware content-based recommendations as for the content-based recommendations, which enables the comparison of both types of recommendations and investigate the influence of contextual information.

5 Analysis of the Usage of the News service

In view of generating personalized recommendations for news articles, the first question that is investigated is RQ1: "Is the dataset of a start-up news service dense enough for collaborative filtering?". Collaborative Filtering (CF) is a commonly-used recommendation technique that is based on the similarity in consumption pattern of different users or items. CF systems require some overlap in the consumptions of users or items in order to find neighboring users or items for the calculation of the recommendations [8]. As a result, a big challenge for CF systems is to generate suitable recommendations when there

is relatively few consumption data (cfr. the sparsity problem). In addition, new services have to cope with the cold start problem [23]. This is a special case of the sparsity problem which involves the difficulty to generate recommendations for new users that have not yet consumed any item (new-user problem) and the issue of dealing with items that have not been consumed yet (new-item problem).

For the Stream Store news service, the overlap in consumption behavior is analyzed by Figure 2, which shows how many times the news items are requested for viewing during the evaluation period. The graph shows that the majority of the news items (55.3%) is requested only once, i.e. by one user of the system. Other news items are requested 2 (22.8%), 3 (10.0%), or up to 6 times. Only a minority of the news items (2.6%) is requested by more than 6 users.

This limited overlap in the consumptions of users can be explained by the application domain and the characteristics of the service. The large content offer (more than 71,000 items) makes the news service very attractive from a user point of view, but induces a sparse consumption matrix for the recommender system, i.e. only a limited number of the available news items is viewed by each user. Because of this sparsity, CF solutions are unusable to generate recommendations for this specific case (rapidly-varying content, a large number of content items, and a relatively small number of users). E.g., in case of a user-based collaborative filtering algorithm, articles that are viewed only once, will not be considered to create neighborhoods of similar users, since these items cannot be part of the overlap of two user profiles. In case of an item-based collaborative filtering algorithm, articles that are viewed only once, will never be recommended, so that more than half of the content pool is excluded from the recommendations.

In contrast, content-based recommender systems have the big advantage that they do not require a large community of users and overlap in the consumptions to achieve a reasonable performance. As a result, content-based recommendations are less affected by the sparsity problem. In addition, newly added items can be immediately recommended once the item attributes are available. In other words, content-based recommendations in the domain of news do not suffer from the new-item problem.

To conclude, a content-based recommendation algorithm is the most suitable solution for a start-up news service that has to build a community of active users and has to cope with the sparsity and new-item problem.

6 Usage in Different Contexts

To verify the usefulness of context-aware recommendations for news content, the variability of the context is investigated through the second research question, RQ2: “Do users consume news content in several different contextual situations?”. Context-aware recommender systems offer users a content offer that is adapted to their current contextual situation. If it turns out that users

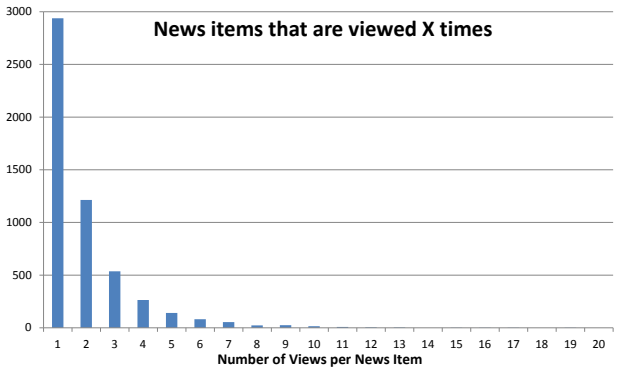

Fig. 2 Histogram showing how many times the news items are requested for viewing.

utilize the service in only one context, then it is impossible to detect variations in users' consumption behavior depending on the context and consequently, the context-aware algorithm falls back on the traditional, 'context-ignorant' recommendations.

6.1 During Different Times of the Day

Figure 3 shows how many test users utilized the news service on different times of the day. During the evaluation period, most users were active during multiple time periods. For the combined activity on both device types, 10 users (9.1%) utilized the service during only one time period (morning, noon, daytime or evening), 17 users (15.5%) have viewed news items during two different time periods, 36 users (32.7%) were active during three different time periods, and 47 users (42.7%) have requested news content during morning, noon, daytime, and evening (all four time periods).

Besides, some differences are visible between users' consumption behavior on smartphone and tablet. Figure 3 reveals that on smartphone, users are typically reading news throughout the whole day: 47.8% of the smartphone users have requested news items during all four time periods, 32.7% during three time periods. In contrast, reading news on tablet is often limited to one or two periods during the day. 21.7% of the tablet users are active during only one time period, and 21.7% of the tablet users are reading news on two times of the day. This difference can be explained by the fact that people generally keep their phone more closely throughout the whole day than their tablet, which is often used only once during the day, typically at home.

As stated in Section 3, the time and the location of the user are closely related. Links between the device type and patterns in users' consumption behavior can be identified. E.g., some users have the habit of reading news with their tablet, at home, during the evening.

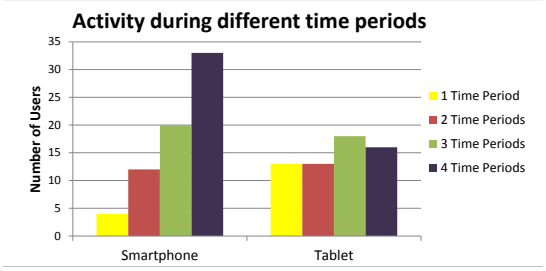

Fig. 3 Histogram showing how many test users are using the service on different times of the day.

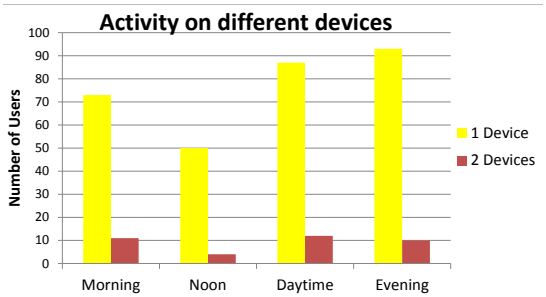

Fig. 4 Histogram showing how many test users are using the service using different devices.

6.2 On Different Devices

Figure 4 shows how many test users utilized different devices to access news content. The graph indicates that during a specific time period, most users utilize only one type of device to consult the news: either smartphone or tablet. The same conclusion applies to the entire evaluation period (all time periods together): 91 users (82.7%) utilized the service on either smartphone or tablet, and 19 users (17.3%) used both devices at least once for reading news. The most obvious reason for this result is that some users possess only one device. Users that have both, a smartphone and a tablet, may have the habit or preference to read news on only one of them.

To conclude, most users utilize the news service on different times of the day, but only a minority of them (17.3%) utilizes multiple types of devices to access the content. As a result, different contextual situations in which users consume news content can be identified, which offers potential for context-aware recommender systems.

7 Content Preferences in Different Contexts

Since users consume news content in different contextual situations, this context can have a significant influence on the user's selection and consumption behavior, which is crucial information for a recommender system. Therefore,

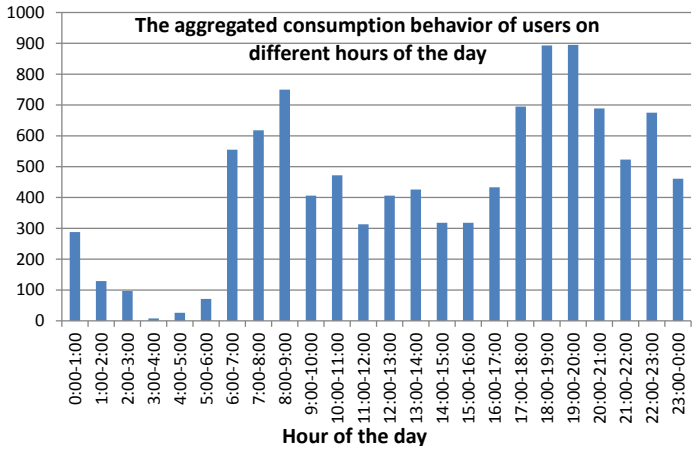

Fig. 5 The amount of user interaction with the news service, partitioned according to the hour of the day.

the follow-up question is RQ3: "Do users of a news service have another consumption behavior in different contextual situations?". To answer this question, two aspects are investigated: 1. The influence of context on the service usage intensity (i.e., the amount of content that is consumed) and 2. The influence of the context on users' content selection (i.e., which content that is consumed).

7.1 Service Usage Intensity

7.1.1 Influence of the Time

Figure 5 shows the amount of user interaction with the service (i.e., selecting a news item to view), aggregated over all users, for each hour of the day. A clear pattern in the consumption behavior is visible throughout the whole day. The close relation between time and location may have strengthened this pattern. As expected, the amount of activity with the service is limited during the night. In the morning, users are very interested in the news, which is reflected in a peak in the service usage. During the day, the amount of consumptions varies slightly per hour with a slight increase around noon. During the evening, users spend more time reading news, which is revealed in Figure 5 by the increased user activity.

The influence of the time on the service usage intensity is further investigated by analyzing users' news consumption per category. Figure 6 shows the amount of user interaction with the service for each news category, partitioned according to the time of the day. This figure uses the same partitioning of the day into four non-overlapping parts as the context-aware recommender does. The amount of user activity is clearly dependent on the time period (morning, noon, daytime, evening) and the duration of this period. Both Figure 5 and 6

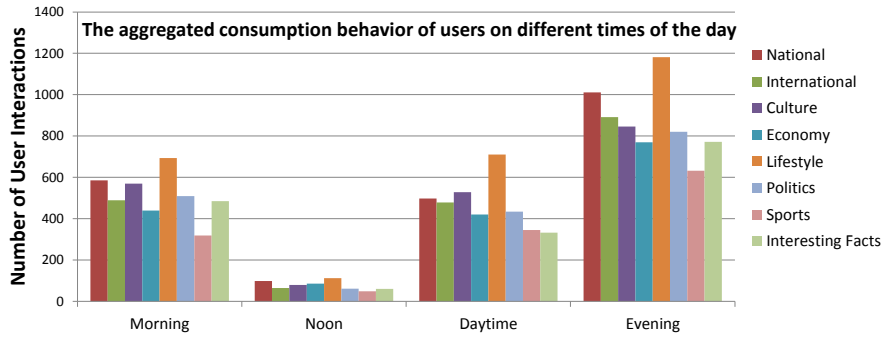

Fig. 6 The amount of user interaction with the news service for each news category, partitioned according to the time of the day.

demonstrate that users spend most time on reading news during the evening, followed by the morning. The increase in user activity during noon, compared to the activity per hour during daytime (Figure 5), showed to be limited. This analysis confirms the assumption that the usage of a news service and the amount of news content that is read, is strongly linked to the time of the day.

7.1.2 Influence of the Device

Figure 7 shows the amount of user interaction with the service (i.e., selecting a news item to view) for each news category, partitioned according to the type of device. The graph illustrates the varying number of consumptions, aggregated over all users, for each type of device and news category. Various trends are visible in this graph. E.g., the news category with the highest number of article selections (2696 in total) is 'Lifestyle'. The popularity of this category is observable for all time intervals (Figure 6) and devices (Figure 7). The category 'Sports' has the lowest number (1345 in total) of article selections during the evaluation period.

To investigate if a news category is selected more on one specific type of device than on another type of device, a Wilcoxon signed rank test is performed. The Wilcoxon signed rank test is a non-parametric statistical hypothesis test used to compare two related samples or repeated measurements on a single sample to assess whether their population mean ranks differ. Through a number of Wilcoxon signed rank tests (one for every news category), the number of consumptions of a category were compared for the different device types. More specifically, for each user who participated in the evaluation, the number of news items (i.e. the dependent variable) requested on smartphone or tablet (i.e. the grouping variable, consisting of two related samples) are compared for each news category. The null hypothesis of the test is that the median difference between the number of consumptions on smartphone and the number of consumptions on tablet (i.e. the pairs of observations) is zero.

Table 2 lists the results of the Wilcoxon signed rank test with the z-value and the p-value for each news category. For most categories, the Wilcoxon test

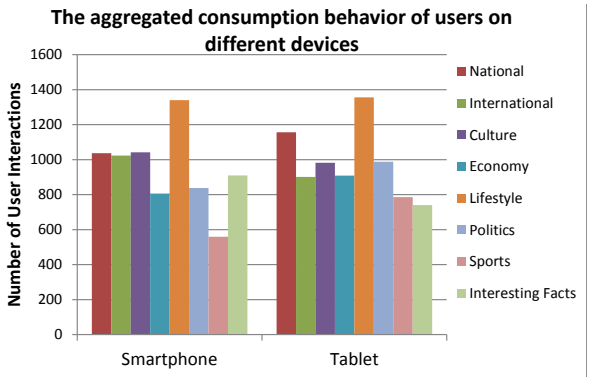

Fig. 7 The amount of user interaction with the news service for each news category, partitioned according to the type of device.

Table 2 Results of the Wilcoxon signed rank test for the different news categories

News Category	Z-value	p-value
National	1.1337	0.2569
International	1.5423	0.123
Culture	1.8324	0.0669
Economy	1.155	0.2481
Lifestyle	1.972	0.0486
Politics	1.0573	0.2904
Sports	1.3475	0.1778
Interesting Facts	1.7691	0.0769

concludes that the null hypothesis cannot be rejected. This means there is not enough evidence to believe that any difference between the groups is for any reason other than chance (e.g., sampling). Or in other words, there is not a significant difference in the median of the number of consumptions on smartphone and tablet. Except for news of the category "Lifestyle", the Wilcoxon test points to a significantly higher median of the number of consumptions on smartphone compared to tablet, on an individual user basis. (This difference is not visible in Figure 7, which shows an aggregation of the number of consumptions over all users per category.) This analysis in terms of the amount of news content that is read, points to differences in consumption behavior (for some categories) on different types of devices.

7.2 Users' Content Selection

7.2.1 Influence of the Time

In Figure 6, variations in the popularity of news categories are visible for the different time periods. These variations in the content selection behavior depending on the time, are further analyzed by a cross tabulation. Table 3 shows the number of news consumptions (count) aggregated over all users, per

news category and for different times of the day. Besides the observed number of consumptions, the table indicates the expected number of consumptions (expected count), in case there is no correlation between the selected news category and time. In each cell, additional information is provided about that particular combination of category and time period: the share of consumptions of that particular category within the specific time period (%within time period), the share of consumptions during that particular time period within the specific category (%within category), and the share of consumptions of that particular combination of time and category over all possible combinations (% of total). For each combination of time and category, the deviation between the observed and expected number of consumptions can be calculated.

To investigate whether these deviations are significant, or in other words, whether the time of the day has a significant influence on the category of the news that is read, a Pearson χ^2 test has been performed. The χ^2 test is a statistical hypothesis test in which the sampling distribution of the test statistic is a chi-squared distribution when the null hypothesis is true [38]. The null hypothesis states that the frequency distribution of certain events observed in a sample is consistent with a particular theoretical distribution [5]. The events that are considered in the test must be mutually exclusive and the sum of their probability must be 1.

In this analysis, the null hypothesis is that the category of the selected news item and the time period are independent. This means that there is no difference in users' content selection behavior during different time periods, or in other words, the distribution of consumptions over the different news categories is the same for every time period. The Pearson χ^2 test has a resulting value of 56.68, with 21 degrees of freedom, and a p-value < 0.01. The low p-value implies that the null hypothesis should be rejected; the data provides enough evidence to conclude that there is a significant correlation between the time period and the news category.

The following meaningful differences are visible (Table 3) by comparing the number of observed consumptions during a specific time period and the expected number of consumptions, calculated as if there would be no influence of the time period. In the morning, more consumptions are registered for categories such as culture (+5.9%), politics (+4.8%), and interesting facts (+10.5%) compared to the expected number of consumptions if no time dependency exists. Sports are less consumed in the morning (-10.8%). At noon, more news of the category national (+13.3%) and economy (+25.7%) is consumed than expected in an unconditional case without a temporal context. Politics (-14.8%) and international news (-15.3%) are less read at noon. During daytime, an increased interest in culture (+7.0%) and lifestyle (+8.1%) is witnessed, whereas less news items with interesting facts are read (-17.2%). During the evening, the number of read news items of the category international (+2.8%), sports (+4.3%), and interesting facts (+3.9%) is higher than expected without the knowledge of the temporal context. Cultural news items (-7.2%) are less consumed during the evening. Some differences are easy to

Table 3 The crosstabulation showing the number of consumptions per news category for different times of the day

		news category								
		national	international	culture	economy	lifestyle	politics	sports	interesting facts	total
morning	count	585	489	570	439	693	509	319	485	4089
	expected count	583.3	511.8	538.4	456.2	717.1	485.7	357.8	438.9	4089.0
	% within time period	14.3%	12.0%	13.9%	10.7%	16.9%	12.4%	7.8%	11.9%	100.0%
	% within category	26.7%	25.4%	28.2%	25.6%	25.7%	27.9%	23.7%	29.4%	26.6%
	% of total	3.8%	3.2%	3.7%	2.9%	4.5%	3.3%	2.1%	3.2%	26.6%
noon	count	99	65	80	86	112	62	49	60	613
	expected count	87.4	76.7	80.7	68.4	107.5	72.8	53.6	65.8	613.0
	% within time period	16.2%	10.6%	13.1%	14.0%	18.3%	10.1%	8.0%	9.8%	100.0%
	% within category	4.5%	3.4%	4.0%	5.0%	4.2%	3.4%	3.6%	3.6%	4.0%
	% of total	0.6%	0.4%	0.5%	0.6%	0.7%	0.4%	0.3%	0.4%	4.0%
daytime	count	498	479	528	420	710	434	345	333	3747
	expected count	534.5	469.0	493.3	418.0	657.1	445.1	327.8	402.2	3747.0
	% within time period	13.3%	12.8%	14.1%	11.2%	18.9%	11.6%	9.2%	8.9%	100.0%
	% within category	22.7%	24.9%	26.1%	24.5%	26.3%	23.8%	25.7%	20.2%	24.4%
	% of total	3.2%	3.1%	3.4%	2.7%	4.6%	2.8%	2.2%	2.2%	24.4%
evening	count	1011	891	846	770	1181	821	632	772	6924
	expected count	987.7	866.6	911.6	772.4	1214.3	822.4	605.8	743.2	6924.0
	% within time period	14.6%	12.9%	12.2%	11.1%	17.1%	11.9%	9.1%	11.1%	100.0%
	% within category	46.1%	46.3%	41.8%	44.9%	43.8%	45.0%	47.0%	46.8%	45.0%
	% of total	6.6%	5.8%	5.5%	5.0%	7.7%	5.3%	4.1%	5.0%	45.0%
total	count	2193	1924	2024	1715	2696	1826	1345	1650	15373
	% of total	14.3%	12.5%	13.2%	11.2%	17.5%	11.9%	8.7%	10.7%	100.0%

understand such as the increased interest for sports during the evening, since many sports events are scheduled in that time period.

Differences in users' content selection for different times periods, can also be illustrated by tag clouds, i.e. a visual representation for text data used to depict keyword metadata (tags) in which the importance of each tag is indicated with the font size.

Figure 8 shows the tag clouds composed of the keywords of the news articles that are read by the users during the evaluation period. For each time period, a tag cloud is providing an aggregated overview of users' content selections by showing the 50 most popular tags. Since these tag clouds are created based on the consumption behavior of the same users, but for another time period, many tags appear in all four tag clouds, such as 'Belgium', 'United States', and 'politics'. Still, interesting differences can be noticed. By comparing the category popularity within the time period, Table 3 shows that during the morning, the share of political news is larger (12.4%) than during other time periods (noon: 10.1%, daytime: 11.6%, evening: 11.9%), thereby introducing keywords such as 'president', 'New York', and 'energy' into the tag cloud. During noon, news of the category 'national' is chosen more frequently (16.2%) than during other time periods (morning: 14.3%, daytime: 13.3%, evening: 14.6%). Also economic news is more popular during noon (14.0%) (compared to morning: 10.7%, daytime: 11.2%, evening: 11.1%). This is reflected in the tag cloud by words such as 'ceo', 'company', 'economy', 'electronic', and 'web'. During daytime, lifestyle is the most selected category (18.9%) (compared to morning: 16.9%, noon: 18.3%, evening: 17.1%), giving rise to tags such as 'hospitality'. During the evening, sport (9.1%) and international news (12.9%) are more often read, compared to other time periods, which leads to words such as 'court' and 'London' in the tag cloud.

(a) Morning (b) Noon

(c) Daytime (d) Evening

Fig. 8 Tag clouds visualizing the consumption behavior per time period.

This analysis in terms of the type of news content that is read, demonstrates the differences in users' content selection behavior on different times of the day.

7.2.2 Influence of the Device

The influence of context on users' selection behavior is further investigated by analyzing the variations in the popularity of news categories for the different device types, as visible in Figure 7. The cross tabulation of Table 4 shows the number of news consumptions (count) aggregated over all users, per news category and for different types of devices (smartphone and tablet). The table also indicates the number of consumptions as expected in case there is no correlation between the selected news category and device type.

Table 4 The crosstabulation showing the number of consumptions per news category for different devices

		news category								
		national	international	culture	economy	lifestyle	politics	sports	interesting facts	total
phone	count	1037	1023	1042	806	1340	838	559	910	7555
	expected count	1077.7	945.5	994.7	842.8	1324.9	897.4	661.0	810.9	7555.0
	% within device	13.7%	13.5%	13.8%	10.7%	17.7%	11.1%	7.4%	12.0%	100.0%
	% within category	47.3%	53.2%	51.5%	47.0%	49.7%	45.9%	41.6%	55.2%	49.1%
	% of total	6.7%	6.7%	6.8%	5.2%	8.7%	5.5%	3.6%	5.9%	49.1%
tablet	count	1156	901	982	909	1356	988	786	740	7818
	expected count	1115.3	978.5	1029.3	872.2	1371.1	928.6	684.0	839.1	7818.0
	% within device	14.8%	11.5%	12.6%	11.6%	17.3%	12.6%	10.1%	9.5%	100.0%
	% within category	52.7%	46.8%	48.5%	53.0%	50.3%	54.1%	58.4%	44.8%	50.9%
	% of total	7.5%	5.9%	6.4%	5.9%	8.8%	6.4%	5.1%	4.8%	50.9%
total	count	2193	1924	2024	1715	2696	1826	1345	1650	15373
	% of total	14.3%	12.5%	13.2%	11.2%	17.5%	11.9%	8.7%	10.7%	100.0%

For each particular combination of category and device type, additional data is provided in Table 4: the share of news articles of that particular category, if only articles consumed on the specific device type are considered (%within device), the share of consumptions on that particular device type within the specific category (%within category), and the share of consumptions of that particular combination of device type and category over all possible combinations (% of total). For each combination of device type and category, there is a deviation between the observed and expected number of consumptions.

The significance of these deviations is investigated by a Pearson χ^2 test, which tests whether the type of device has a significant influence on the category of the news that is read. The null hypothesis is that the category of the selected news item and the type of device are independent. If this null hypothesis is true, there is no difference in users' content selection behavior on different device types, or in other words, the distribution of consumptions over the different news categories is the same for every device type.

The Pearson χ^2 test has a resulting value of 85.93, with 7 degrees of freedom, and a p-value < 0.01, which means that the null hypothesis should be rejected. This means the data provides enough evidence to conclude that there is a significant correlation between the device and the news category.

The following meaningful differences are visible (Table 4) by comparing the number of observed consumptions on a specific device and the expected number of consumptions, calculated as if there would be no influence of the device. On smartphones, more news of the categories international (+8.2%) and interesting facts (+12.2%) is consumed compared to the general statistics that make no distinction on device type. News of the category sports is less frequently consumed on smartphones (-15.4%) than expected in a case with no device context. On tablet, the opposite is witnessed: more news consumptions of the category sports (+14.9%) and less international (-7.9%) and interesting facts (-11.8%). Again, these findings can be explained: interesting facts are mostly small articles, perfect for reading on smartphones. Sports are often coupled with video material which is more suitable for tablets.

Differences in users' content selection for different device types, can also be illustrated by tag clouds. Figure 9 shows two tags clouds. One is composed of

(a) Smartphone (b) Tablet

Fig. 9 Tag clouds visualizing the consumption behavior per device type.

the keywords of the news articles that are read on smartphones. The other one is an aggregated overview (based on the 50 most popular tags) of the consumed new articles on tablets. Although both tag clouds show a lot of resemblance, subtle differences in the tag clouds can be detected due to differences in users' content selection on both types of devices.

On smartphones, more articles (12.0%) of the category 'interesting facts' are read compared to tablets (9.5%). This can be noted by keywords such as 'games', 'software', and 'web' in the tag cloud. Conversely, on tablets, articles regarding sports are chosen more frequently (10.1%) than on smartphones (7.4%). These sports articles introduce terms such as 'manager' and emphasize 'football' and 'sports' in the tag cloud for tablets.

As a result, in terms of the type of news content that is read, differences in users' content selection behavior are demonstrated on different types of devices.

8 Evaluating the Recommendations by a Feedback Analysis

To quantify the added value of a dynamic profile and contextual information, the next question is RQ4: "How do recommendations based on explicit static preferences, content-based recommendations, and context-aware content-based recommendations compare for a news service?"

During the evaluation period, explicit user feedback was gathered by asking users to rate every fourth news article that they read through a pop-up with a 'Thumbs Up' or 'Thumbs Down' rating. Per user, a rating ratio is computed, summing all positive (+1) and negative ratings (-1). Next, a regression model is built with the rating ratio as a dependent variable. The null hypothesis is that there is no significant linear correlation between the dependent variable and the

independent variables. The regression coefficients of the independent variables with an associate p-value < 0.05 are considered as significant, meaning that there is a significant linear correlation between the independent variable and the dependent variable.

Both learning-conditions (i.e. the content-based and context-aware content-based recommender which learn from actual user behavior) are set as independent variables (content-based and context-based in Table 5), while controlling for each user's number of article consumptions (N Consumptions in Table 5). This means that the type of recommendations and the number of article consumptions are used to predict the rating ratio. The different recommendation types are represented as boolean variables (content-based:yes or no, context-based: yes or no). Moreover, interaction terms are added (Content-based x N Consumptions and Context-based x N Consumptions in Table 5) to investigate the effect of conditions on the relation between increased consumption and rating ratios. As such, the factors explaining user satisfaction with the recommendations are statistically inferred.

The model yields good fit as the likelihood ratio-χ^2 has a resulting value of 49.20, with 5 degrees of freedom, and a p-value < 0.01. The results, enumerated in Table 5, reveal that as the intercept significantly differs from 0, all conditions display substantial positive rating ratios. (The intercept, often labeled the constant, is the expected mean value of the dependent variable Y, when all independent variables are zero, X=0). The unstandardized regression coefficients, B, are the maximum likelihood estimate of the parameters of the regression model. The standard error, SE, gives an indication of the variability on these coefficients. The Wald test is a parametric statistical test used to investigating the relationship (e.g., a dependency) between the variables.

However, using the static profile condition as a reference category, the analysis shows that the content-based recommendation condition leads to a significantly higher rating ratio. Or in other words: for all conditions (explicit static preferences, content-based, and context-based) the rating ratio is positive; but for content-based, and context-based a higher rating ratio is obtained. This is regardless of the number of consumptions, which has an evident positive effect, as this variable is also included in the equation. Still, the significant interaction terms indicate that for the learning recommendation conditions, there is a significant increase in the rating ratio that is related to increased consumption. This reflects and confirms the self-learning properties of the recommendation algorithm.

In sum, the analysis confirms the added value of recommendations, albeit for the content-based version and not the context-aware content-based version. Because the results are only based on the evaluation period of approximately 5 weeks, and CARS require sufficient time to learn user preferences in different contextual situations, the context-aware content-based recommender induces only a limited difference in terms of received feedback. We believe that because of the cold start problem, the potential of the contextual information has not fully exploited in this experiment. Additional data can provide the system the opportunity to learn patterns in the users' behavior regarding the

Table 5 Regression model of the different recommendation types' evaluations. B = unstandardized regression coefficient, SE = standard error, Wald χ^2 = test for independence of the variables, p-value = significance whether the given coefficient is different from zero

Parameter	B	SE	Wald χ^2	p-value
Intercept	5.42	1.96	7.68	0.01
Content-based	3.67	1.55	5.59	0.02
Context-based	1.54	1.59	0.93	0.34
N Consumptions	0.07	0.01	29.84	0.00
Content-based x N Consumptions	0.04	0.01	30.53	0.00
Context-based x N Consumptions	0.03	0.01	6.20	0.01

consumption of news, thereby further improving the accuracy of context-aware content-based recommendations. These results can be used by developers to decide on the usefulness of a context-aware recommender system. In the short-term, content-based recommendations outperform context-aware recommendations because of the cold start problem. In the long-term, a context-aware recommender system can help to distinguish the user preferences in different contextual situations as observed in Section 7.

9 Discrepancy between Profile and User Behavior

So although the recommendations based on explicit static preferences do not suffer from the cold start problem (because of the initial profile), the content-based and context-aware content-based recommender systems, which have to learn the preferences of the users over time, achieved a higher user satisfaction during the evaluation period (Section 8). In order to better understand the results of Section 8, the last research question is "Are explicitly-stated news preferences, the run time profile of a context-aware recommender, and the actual user behavior consistent?".

One might expect that the initial profile, as explicitly specified by the user, perfectly matches the user's preferences, and as a result, reflects the user's actual behavior with the system. In addition, radical changes in the user's preferences over time are unlikely given the time frame of the evaluation period. If this hypothesis is true, recommendations based on the explicit static preferences should perfectly match the user's actual preferences, thereby inducing a high user satisfaction. However, the evaluation of the recommendations of Section 8 revealed that recommendations based on explicit static preferences are the least appreciated by the users, and there is still much room for improvement by making the profile dynamic and taking into account the user interactions and context.

To investigate this, Figure 10 compares the initial explicit profile, the run time profile, and the actual user behavior based on the preference values for all news categories, aggregated over all users. The preference values of the initial explicit profile are the values as specified by the users (on a 5-point scale) through the questionnaire prior to the evaluation period. The preference values of the run time profile are the estimated user preferences as calculated by the

context-aware content-based recommender system. The preference values of the actual user behavior reflect the interaction behavior of the users, i.e. the amount of news items that are requested by the users to read. These preference values are plotted for the different time periods: morning, noon, daytime, and evening. Since most users utilized only one device during the evaluation period (Section 6.2), the device type is not included in this graph as a context variable.

To enable the comparison of the user profiles and the actual user behavior, which use a different scale (5-point vs. absolute values), all these preference values are converted to their Z-score. The Z-score is a dimensionless quantity obtained by subtracting the population mean from the individual score and then dividing the difference by the population standard deviation [27]. Thus, a positive Z-score represents a value above the mean, while a negative Z-score represents a value below the mean.

Figure 10 indicates some general trends, which were also witnessed in Figure 6: users tend to read the news especially during the evening, which is reflected in the high preference values for this time of the day. In contrast, during lunch break, the consumption of news is limited (partially because of the short time period), which explains the preference values below average.

Comparison between the initial explicit profile and the actual user behavior demonstrates that the stated preferences of the questionnaire do not always correspond to the users' actual preferences as can be derived from their interaction with the news content. The correlation between the preferences of the initial explicit profile and the actual user behavior is 0.19. This significant positive correlation indicates that the initial profile provides some valuable insights into the users' preferences; but as a model reflecting the actual user behavior, there is still room for improvement. This discrepancy between initial explicit profile and the actual user behavior can have different causes, which are worth to further investigate in future research. Firstly, users might have difficulties to state their preferences regarding news content on a 5-point rating scale, making them misjudgments. Secondly, the explicit profile might suffer from the social desirability bias (i.e., the tendency of respondents to answer questions in a manner that will be viewed favorably by others) [18]. The news categories hint to the fact that users might be influenced by the importance or the impact of the news. E.g., categories like 'international' and 'national news' are systematically graded too high in the initial profile, whereas categories such as 'lifestyle' received an evaluation that is too low in comparison with the users' actual demand for news.

Also the run time profile and the actual user behavior can be compared using Figure 10. The graph shows a better correspondence between the run time profile and the actual user behavior compared to the correspondence of the initial profile and the user behavior. The correlation between the preference values of the run time profile and the actual user behavior is 0.83. This strong positive correlation demonstrates that the user profile, as calculated by the recommender system, is an accurate model for the actual user preferences as reflected in the user behavior.

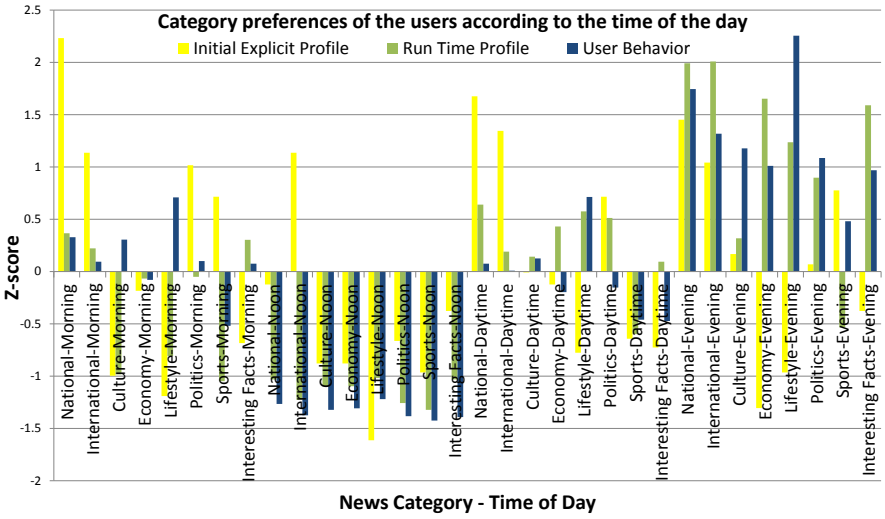

Fig. 10 Comparison of the initial explicit profile, the run time profile, and the actual user behavior.

Still, the correlation between user profile and user behavior is not perfect; differences can be witnessed for several news categories and at different times of the day. Additional data, gathered by monitoring the user behavior over a longer period of time, can contribute to the improvement of the correlation between user profile and user behavior. However a user model that perfectly represents the user's preferences and predicts the actual behavior remains a big challenge; after all, the user behavior and decision-making process is a very complex process.

10 Conclusions

In this paper, a start-up news service offering personal recommendations is evaluated from a user point of view. The typical characteristics of news content, such as the short-term life, and the limitations of a start-up service, such as the necessity for a community of active users, make collaborative filtering techniques unusable. Therefore, various types of content-based recommendations are evaluated through a user study: recommendations based on an explicit static profile, content-based recommendations using the actual user behavior but ignoring the context, and context-aware content-based recommendations incorporating user behavior as well as context. The study aimed to assess the importance of context in the news recommendation process by focusing on two contextual aspects: device type and time. Before considering contextual information as an added value for the recommendation process, a number of conditions have to be fulfilled.

Firstly, users have to utilize the service in different contextual situations. Logged user behavior showed that users typically consult the news service with only one device (either smartphone or tablet) at different times of the day, but especially in the evening.

The second condition is that users have other preferences, and as a result another consumption behavior, in these different contextual situations. Analysis of the user behavior showed a significant difference in service usage intensity (the amount of news content that is consumed) as well as in users' content selection behavior (what type of news content that is consumed) for the different contextual situations.

Thirdly, sufficient learning data has to be available to learn user preferences and variations in these preferences over the different contextual situations. Whereas traditional recommender systems have to overcome the cold start problem once, context-aware recommender systems have to overcome this problem for every contextual situation. Context generalization can be a partial solution, but choosing the right contextual aspect to optimally broaden the context is often difficult.

In this experiment, recommendations based on the actual user behavior outperformed recommendations based on an explicit static profile. However, including contextual information in the content-based algorithm did not significantly improve the user satisfaction with the recommendations. The cold start problem, strengthened by the fragmentation of the consumption data over different contextual situations, appears to be a serious obstacle in order to fully exploit the potential of contextual information in the recommendation process of a start-up service.

Acknowledgements This research was performed in the context of the iMinds-MIX Stream Store project. Stream Store is a project cofunded by iMinds (Interdisciplinary institute for Technology) a research institute founded by the Flemish Government. Companies and organizations involved in the project are De Persgroep, Roularta Media Group nv, iMinds-iLab.o, VMMA, and Limecraft with project support of IWT.

The authors would also like to thank the researchers of MIX for the development of the Stream Store client application and the team of the iMinds-MMLab research group for the processing of the metadata. For the setup of the user experiment and performing the evaluation, the authors would like to express their gratitude to the students of the iMinds-MICT research group.

References

1. Adomavicius, G., Sankaranarayanan, R., Sen, S., Tuzhilin, A.: Incorporating contextual information in recommender systems using a multidimensional approach. ACM Transactions on Information Systems **23**(1), 103–145 (2005). DOI 10.1145/1055709.1055714
2. Adomavicius, G., Tuzhilin, A.: 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), 734–749 (2005). DOI 10.1109/TKDE.2005.99
3. Adomavicius, G., Tuzhilin, A.: Tutorial on context-aware recommender systems. In: Proceedings of the second ACM conference on Recommender Systems (RecSys '08) (2008)

4. Adomavicius, G., Tuzhilin, A.: Context-aware recommender systems. In: F. Ricci, L. Rokach, B. Shapira, P.B. Kantor (eds.) Recommender Systems Handbook, pp. 217–253. Springer US (2011). DOI 10.1007/978-0-387-85820-3_7
5. Bagdonavicius, V., Julius, K., Nikulin, M.S.: Chi-Squared Tests, pp. 17–75. John Wiley & Sons, Inc (2013). DOI 10.1002/9781118557716.ch2
6. Baltrunas, L., Ricci, F.: Context-based splitting of item ratings in collaborative filtering. In: Proceedings of the Third ACM Conference on Recommender Systems, RecSys '09, pp. 245–248. ACM, New York, NY, USA (2009). DOI 10.1145/1639714.1639759
7. Bazire, M., Brézillon, P.: Understanding context before using it. In: A. Dey, B. Kokinov, D. Leake, R. Turner (eds.) Modeling and Using Context, *Lecture Notes in Computer Science*, vol. 3554, pp. 29–40. Springer Berlin Heidelberg (2005). DOI 10.1007/11508373_3
8. Breese, J.S., Heckerman, D., Kadie, C.: Empirical analysis of predictive algorithms for collaborative filtering. In: Proceedings of the fourteenth conference on Uncertainty in artificial intelligence, UAI'98, pp. 43–52. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1998). URL `http://dl.acm.org/citation.cfm?id=2074094.2074100`
9. Brown, P.J., Bovey, J.D., Chen, X.: Context-aware applications: from the laboratory to the marketplace. IEEE Personal Communications **4**(5), 58–64 (1997). DOI 10.1109/98.626984
10. Burke, R.: Hybrid recommender systems: Survey and experiments. User Modeling and User-Adapted Interaction **12**(4), 331–370 (2002)
11. Cantador, I., Bellogín, A., Castells, P.: News@hand: A semantic web approach to recommending news. In: W. Nejdl, J. Kay, P. Pu, E. Herder (eds.) Adaptive Hypermedia and Adaptive Web-Based Systems, *Lecture Notes in Computer Science*, vol. 5149, pp. 279–283. Springer Berlin Heidelberg (2008). DOI 10.1007/978-3-540-70987-9_34
12. De Pessemier, T., Coppens, S., Geebelen, K., Vleugels, C., Bannier, S., Mannens, E., Vanhecke, K., Martens, L.: Collaborative recommendations with content-based filters for cultural activities via a scalable event distribution platform. Multimedia Tools and Applications **58**(1), 167–213 (2012). DOI 10.1007/s11042-010-0715-8
13. De Pessemier, T., Deryckere, T., Vanhecke, K., Martens, L.: Proposed architecture and algorithm for personalized advertising on idtv and mobile devices. IEEE Transactions on Consumer Electronics **54**(2), 709–713 (2008). DOI 10.1109/TCE.2008.4560151
14. De Pessemier, T., Dooms, S., Martens, L.: Comparison of group recommendation algorithms. Multimedia Tools and Applications pp. 1–45 (2013). DOI 10.1007/s11042-013-1563-0
15. De Pessemier, T., Dooms, S., Martens, L.: Context-aware recommendations through context and activity recognition in a mobile environment. Multimedia Tools and Applications pp. 1–24 (2013). DOI 10.1007/s11042-013-1582-x
16. Dey, A.K.: Understanding and using context. Personal and Ubiquitous Computing **5**(1), 4–7 (2001)
17. Dey, A.K., Abowd, G.D., Salber, D.: A conceptual framework and a toolkit for supporting the rapid prototyping of context-aware applications. Human-Computer Interaction **16**(2), 97–166 (2001). DOI 10.1207/S15327051HCI16234_02
18. Fisher, R.J.: Social desirability bias and the validity of indirect questioning. Journal of Consumer Research **20**(2), pp. 303–315 (1993). URL `http://www.jstor.org/stable/2489277`
19. Følstad, A.: Living labs for innovation and development of information and communication technology: A literature review. Electronic Journal of Organizational Virtualness **10**, 99–131 (2008)
20. Gavalas, D., Konstantopoulos, C., Mastakas, K., Pantziou, G.: Mobile recommender systems in tourism. Journal of Network and Computer Applications **39**(0), 319 – 333 (2014). DOI http://dx.doi.org/10.1016/j.jnca.2013.04.006
21. Gavalas, D., Konstantopoulos, C., Mastakas, K., Pantziou, G.: Mobile recommender systems in tourism. Journal of Network and Computer Applications **39**(0), 319 – 333 (2014). DOI http://dx.doi.org/10.1016/j.jnca.2013.04.006
22. Han, B.J., Rho, S., Jun, S., Hwang, E.: Music emotion classification and context-based music recommendation. Multimedia Tools and Applications **47**(3), 433–460 (2010)

23. Huang, Z., Chen, H., Zeng, D.: Applying associative retrieval techniques to alleviate the sparsity problem in collaborative filtering. ACM Transactions on Information Systems **22**(1), 116–142 (2004). DOI 10.1145/963770.963775
24. Jannach, D., Zanker, M., Felfernig, A., Friedrich, G.: Recommender Systems: An Introduction, 1st edn. Cambridge University Press, New York, NY, USA (2010)
25. Kabassi, K.: Personalizing recommendations for tourists. Telematics and Informatics **27**(1), 51 – 66 (2010). DOI http://dx.doi.org/10.1016/j.tele.2009.05.003
26. Kenteris, M., Gavalas, D., Mpitziopoulos, A.: A mobile tourism recommender system. In: Proceedings of the The IEEE symposium on Computers and Communications, ISCC '10, pp. 840–845. IEEE Computer Society, Washington, DC, USA (2010)
27. Kutner, M.H., Nachtsheim, C.J., Neter, J., Li, W.: Applied Linear Statistical Models, fifth edn. McGraw-Hill (2005)
28. Panniello, U., Tuzhilin, A., Gorgoglione, M.: Comparing context-aware recommender systems in terms of accuracy and diversity. User Modeling and User-Adapted Interaction **24**(1-2), 35–65 (2014). DOI 10.1007/s11257-012-9135-y
29. Papagelis, M., Plexousakis, D., Kutsuras, T.: Alleviating the sparsity problem of collaborative filtering using trust inferences. In: P. Herrmann, V. Issarny, S. Shiu (eds.) Trust Management, *Lecture Notes in Computer Science*, vol. 3477, pp. 224–239. Springer Berlin Heidelberg (2005). DOI 10.1007/11429760_16
30. Resnick, P., Varian, H.R.: Recommender systems. Communications of the ACM **40**(3), 56–58 (1997). DOI 10.1145/245108.245121
31. Ricci, F.: Mobile recommender systems. Information Technology & Tourism **12**(3), 205–231 (2010). DOI doi:10.3727/109830511X12978702284390
32. Ricci, F.: Contextualizing recommendations. In: ACM RecSys Workshop on Context-Aware Recommender Systems (CARS '12), In conjunction with the 6th ACM Conference on Recommender Systems (RecSys '12). ACM (2012)
33. Ricci, F., Rokach, L., Shapira, B., Kantor, P.B.: Recommender Systems Handbook, 1st edn. Springer-Verlag New York, Inc., New York, NY, USA (2010)
34. Schilit, B.N., Theimer, M.M.: Disseminating active map information to mobile hosts. IEEE Network **8**(5), 22–32 (1994). DOI 10.1109/65.313011
35. Shani, G., Gunawardana, A.: Tutorial on application-oriented evaluation of recommendation systems. AI Communications **26**(2), 225–236 (2013)
36. Telematica Instituut / Novay: Duine Framework (2009). Available at `http://duineframework.org/`
37. Thomson Reuters: Open Calais (2008-2013). Available at `http://www.opencalais.com/`
38. Weisstein, E.W.: Chi-squared test (1999)
39. Yeung, K.F., Yang, Y.: A proactive personalized mobile news recommendation system. In: Developments in E-systems Engineering (DESE), 2010, pp. 207–212 (2010). DOI 10.1109/DeSE.2010.40
40. Yu, Z., Zhou, X., Zhang, D., Chin, C.Y., Wang, X., men, J.: Supporting context-aware media recommendations for smart phones. Pervasive Computing, IEEE **5**(3), 68–75 (2006). DOI 10.1109/MPRV.2006.61