Development of a test collection for studying long-term user modelling

Frank Hopfgartner
International Computer Science Institute
1947 Center St, Suite 600
Berkeley, CA 94704, USA
fh@icsi.berkeley.edu

Joemon M. Jose
University of Glasgow
18 Lilybank Gardens
Glasgow G128RZ, UK
Joemon.Jose@glasgow.ac.uk

Abstract

Evaluating user modelling systems is a complex process and yet there currently exists no standard evaluation methodology. The interactive user evaluation paradigm cannot easily be applied since it does not allow users to follow their personal information need over a long time period. In this paper, we suggest a novel evaluation methodology for long-term user modelling which is based on user simulation. We first create individual ground truth data which we then use to simulate users interacting with a retrieval system. We also provide an example study exploiting these simulated profiles, illustrating how our method can be used to further study long-term user modelling techniques.

1 Introduction

In recent years, research on adaptive information retrieval has achieved a massive boost, indicated by the growing amount of research papers focusing on either adapting retrieval results to the users' personal information need or on recommending related documents. In short, adaptation systems can be defined as systems that personalise their output based on user interaction. Even though promising approaches have been introduced to adapt results to satisfy users' short term interests, research on adapting content based on users' long term interests has hardly been studied. The main hindrance for such studies is the lack of an evaluation methodology. As [Belkin, 2008] pointed out in his keynote speech at ECIR 2008, a grand challenge in the evaluation of (adaptive) information retrieval approaches is to bring the user into the evaluation process.

Evaluation in information retrieval can broadly be categorised into two paradigms. The most dominant one is system-centred evaluation. TREC is an example of this. System-centred experiments are defined by a strict laboratory-based setting. Automatically generated retrieval results are compared with a list of assessed documents, referred to as the ground truth, and standard evaluation metrics such as precision and recall are computed. The metrics of both systems are then used to evaluate the effectiveness of the introduced method. Even though system-centred evaluation is suitable for experiments, it cannot easily be applied to study some research approaches which are focused around the user [Voorhees, 2008; Kelly *et al.*, 2009]. This is especially problematic in adaptive information retrieval which is based on adapting retrieval results to satisfy users' personal interests. In user-centred evaluation, user satisfaction is used as evaluation measure. User-centred evaluation schemes are very helpful in getting valuable data on the behaviour of interactive search systems. Various problems, however, arise when solely relying on this paradigm [Sparck Jones and Willett, 1997]. First of all, user satisfaction is highly subjective. Moreover, it is almost impossible to test all the variables involved in an interaction and hence compromises are needed on many aspects of testing. Furthermore, such a paradigm is inadequate in benchmarking various underlying retrieval algorithms. In order to evaluate the performance of long-term adaptation, user studies will be required where users interact with the system over several iterations. As argued before, both system-centred and user-centred evaluations are not suitable for this scenario. In this paper, we propose to combine both evaluation schemes to generate a test collection for long-term user modelling. The collection can be used to study questions such as how user interests can be captured over a longer period of time in a profile or how recommendation techniques can exploit such profiles. We first collect personal interests of volunteer subjects on broadcasted news over several weeks. Based on their feedback, we provide them with a number of news video stories related to their needs and ask them to assess their relevance to their defined interests. By this, we avoid the situation where users have to evaluate all broadcasted material. The assessment results in individual relevance lists containing users' interests in news topics covering several weeks. By introducing a user simulation scheme, we present an example of how this test collection can be used to evaluate research questions related to long-term user profiling. The paper is structured as follows. In Section 2, we outline a framework aiming at studying long-term user profiles and discuss why evaluating this framework is challenging. Section 3 reviews related work which is relevant in the context of this study. In Section 4, we introduce our approach to generating ground truth data. In Section 5, we illustrate how this ground truth data can be exploited to simulate long-term user profiles. Section 6 concludes this work.

2 Long-term User Profiling Scenario

In [Hopfgartner and Jose, 2009], we propose a novel news video recommender system. The aim of the study was to model user's evolving interests in multiple topics over multiple iterations. The user's interaction with the retrieval interface plays a key role in the underlying recommendation method. User interaction is exploited to create personalised user profiles. Based on these profiles, further documents are recommended that match the user's personal information need. In order to evaluate the quality of the recommendations over a longer time period, a long-term user experiment is required where users are able to use the system to satisfy their personal information need. The following scenario illustrates the setting of such a long-term experiment:

> "Imagine a user who is interested in multiple news to pics. He registered with our news recommender system with a unique identifier. For several months, he logs into our system, which provides him access to the latest news video stories of the day. On the systems graphical interface, he has a list of the latest news stories. He interacts with the presented results and logs off again. On each subsequent day, he logs in again and repeats the above process."

The constrictions of laboratory-based interactive experiments with pre-defined search tasks do not allow the above scenario, since users will not be allowed to search for the content they are really interested in. Moreover, test collections such as TREC News collections or TRECVid News videos are outdated, which is a big drawback for potential user-based evaluation of profiling approaches. Users will behave differently when searching for old news instead of the latest news, hence biasing the outcome of such studies. [Sanderson, 2006] proposes to create individual, context-specific collections. Using up-to-date test collections can motivate the user to retrieve information they are personally interested in. They can hence act more naturally while accessing the data collection. In [Misra *et al.*, 2010], we introduce the creation of a video collection, consisting of the daily news bulletins of two large news programmes. Each bulletin has a running time of thirty minutes and is broadcast on week days. The channels enrich their broadcasts with a closed caption (teletext) signal that provides textual transcripts. We captured the broadcasts of both channels and segmented the bulletins into semantically related news stories, the unit of retrieval.

Various challenges, however, arise when user experiments are based on non-standard test collections. Considering that every participant will be allowed to search for topics of personal interest, no common assessment lists can be created. Participants are unlikely to show interest in the same documents. One possibility to achieve this is to ask the users to judge relevance for every document in the collection. Considering the size of modern data collections, this approach is not feasible. In order to reduce the manual assessment task, we propose to reduce the number of documents that users have to assess by providing them with subsets of the news collection which matches their reported interest. Another problem is that optimising recommendation approaches, e.g. by fine tuning various parameters, cannot be done using a user-centred evaluation since it would require many users to repeat the same steps a number of times. Considering the long time duration for accurate user profiling evaluations, this is not an option. We therefore propose to evaluate the long-term effect of our recommendation approach by exploiting the relevance assessments lists to create simulated long-term user profiles. After introducing related state-of-the-art research in Section 3, we introduce our approach of generating ground truth lists in Section 4. We then exploit these lists to create individual user profiles in Section 5.

3 Background

In classical user studies, participants are asked to find as many relevant documents to a given search topic as possible, usually within a pre-defined time period. Relevance assessment lists are then used to compute standard evaluation measures of the outcome of the user experiment. In this section, we first discuss how these assessment lists are generated, followed by an outline on simulation-based user experiments.

3.1 Ground Truth Generation

According to [Voorhees, 2001], assessment lists used in TREC are typically binary; a document is either relevant or not relevant to the given topic. A simple approach of creating assessment lists is to manually assess the documents of the test collection. Considering the large human effort involved, this approach is very expensive and therefore not suitable for large-scale collections. [Spärck-Jones and van Rijsbergen, 1965] argue for the creation of assessment lists using subsets of the actual collection. Assuming that the highest ranked documents of multiple independent retrieval runs will contain a large number of relevant documents, they propose to merge these results in a "pool" of documents. Assessors are then asked to judge relevance of these documents. This approach, referred to as pooling, is the primary assessment method within TREC. [Sanderson and Joho, 2004] evaluate various other approaches which can compete with the pooling approach. None of the introduced assessment approaches, however, result in complete lists containing all relevant documents of the collection.

3.2 Simulation-Based User Experiments

User studies are very helpful in getting valuable data on the behaviour of interactive search systems. However, employing this evaluation scheme, it is difficult to optimise retrieval or recommendation techniques, e.g. by comparing runs using different parameters. Hence, such a methodology is inadequate in benchmarking various underlying adaptive retrieval algorithms. An alternative, well-established way of evaluating such systems is the use of simulations. A survey on user simulation is given by [Ivory and Hearst, 2001]. Most simulation schemes rely on pre-defined interaction patterns, often backed by statistical click analyses. Stereotype users are mimicked, e.g. by analysing how often and under which conditions are performed by real users. Most simulations are rather generic and based on heuristic user interactions. Due to these limitations, user simulations can only be seen as a pre-implementation method which will give further opportunity to develop appropriate systems and subsequent user-centred evaluations. More recently, [White *et al.*, 2005] proposed a simulation-based approach to evaluate the performance of implicit indicators in textual retrieval. They simulated user actions such as viewing relevant documents, which were expected to improve the retrieval effectiveness. Further, [Hopfgartner and Jose, 2007] employed a simulated evaluation methodology which simulated users interacting with state-of-the-art video retrieval systems.

4 Generating Relevance Assessment Lists

In the long-term user profiling scenario, users are asked to use a recommender system to satisfy their personal information needs. The evaluation of such a scenario using standard measures requires relevance assessments for every search topic. In this scenario, no predefined search tasks exist. Further, relevance is relative, which makes pooling assessed documents not possible. Even if users are interested in the same topic, they will probably be interested in different aspects and will judge the relevance of documents differently [Cuadra, 1967]. Individual assessment lists are therefore needed. In this section, we propose how to generate individual assessment lists for long-term user profiling using the following approach: (1) Recruiting representative news consumers, (2) asking them to identify interesting news topics and (3) provide them a reduced number of video stories to assess.

4.1 Assessment User Group

In order to generate necessary ground truth data, we recruited 18 volunteers. The assessment task was split into two parts, each ended with a questionnaire where the participants could express their opinions. Before the actual assessment, the assessors were asked to fill in an entry questionnaire to provide demographic information. The group consisted of 12 male and 6 females with an average age of 26.4 years. A majority of them held either an undergraduate or postgraduate degree with a background in IT technologies. We were first interested to find out which sources they usually rely on to gather the latest news. The most named answers they selected from a predefined list were news media websites, followed by television news and radio reports. These replies indicate that the participants accept online news, but also rely on television broadcasts. Our assessment group corresponds to the most active group in online services [Choicestream, Inc., 2008]. We hence conclude that they are an appropriate group to base our study on.

4.2 Gathering of User Interests

In the first part of the assessment task, we aimed to identify the participants' specific interests for news events. The following assumption underlies this subtask: We assume that each day, national news media report the most important news events. More specifically, we assume that the BBC, the world's largest news gatherer, reports these events on their news website. This website is one of the most popular news websites in the UK and well-known for its detailed content. Further, we assume that events with the highest media attention are the most important news events. In order to identify those stories on the BBC News website which received the highest media attention on that day, we rely on Google News which clusters similar news stories from multiple sources and ranks them based on their popularity. For each day of our experiment, we retrieved the URL, the headline and a short snippet from the BBC News website as provided by the Google API. For the assessment task, we generated lists of all retrieved stories, separated by the date and split into blocks of two weeks each. Our participants were now asked to mark all stories in each list, seven lists in total, which they find interesting. In the second step, they had to categorise the selected articles into related groups and provide each group with a common label. Table 1 provides an overview of assessed news stories and identified news categories.

The questionnaire aimed at evaluating their assessment experience. Using Five-Point Likert scales, we first asked them to judge the difficulty of the assessment task. The majority claimed that they found the task very simple. The main difficulty they reported was that some news stories could be classified as belonging to more than one category, which our interface did not support. Since the assessment task took place a few months after the time period of the data corpus, we were interested if this time difference caused troubles for the participants. The assessors stated that before starting the task, they had a general idea of which news events happened in the given time period. Moreover, they claimed that they already knew which kind of stories they were interested in before looking at the collection. As we expected, they claimed that they discovered various news events which they were not aware of before. We assume that this might be partly due to the time difference, but also due to a less intensive following of the news events. The majority did not agree with the statement "I marked various news events as interesting even though I was not interested in them at the given time period". We conclude that the time difference did not influence the assessors judgment on what they find interesting. The selected categories should therefore be a realistic representation of the assessors interests in news within the time period.

4.3 News Video Assessment

Knowing the users' categories of interest, the second part of the assessment aimed at identifying news reports in the video corpus for each category of int erest. In an ideal case, the participants would be asked to assess the full data corpus in order to identify the video clips which are relevant to their identified interests. Due to the size of the data collection, however, this approach is not feasible. Hence, it is necessary to provide the participants with a subset of the corpus which they could assess accordingly. In order to identify a good subset for each category of interest, we exploit a simple observation: Studies (e.g. [Lioma and Ounis, 2006]) have shown that nouns and foreign names play a key role in the semantics conveyed by language. The news documents which have been marked and classified in the preceding subtask mainly consist of reports or interviews and hence contain many terms of this type. Assuming that the same news events which are broadcast have also been reported online, these terms should also be mentioned in the video report about the same event. Considering that both textual and video news are published by the same news content provider (BBC in our case), it is even more likely that the same terms are used. Moreover, since the textual reports usually contain more details than short video clips, there is a high probability that all terms which are mentioned by the reporter in the video also appear in the text report. The most important nouns from the textual documents should hence provide a good presentation of the content of each category. Further, retrieving news stories using these nouns as a search query should provide a significantly smaller subset of the data corpus which can then be assessed by the participant.

We use the LingPipe toolkit[1], at default settings (trained on the Brown corpus) to extract all nouns and foreign names from every assessed document. In a next step, we combine the top ten percent most frequent terms of each category of interest using the "or" operator to form a search query. Each category is thus represented by a search query. Using the interface introduced in [Hopfgartner and Jose, 2010; Hopfgartner, 2011], the participants were now presented a result list of each category of interest. The label of the category was given on top of the list. Results were ranked using BM25. In addition, each retrieved story had an additional ranking bar where users were asked to assess how much this result is relevant to the given category. Search results were split into several pages containing 15

[1] `http://alias-i.com/lingpipe`

Table 1: Summary of the BBC Online News Assessment Task

	U1	U2	U3	U4	U5	U6	U7	U8	U9
# stories	188	340	117	33	90	178	183	84	157
# categories	19	21	28	10	21	29	17	13	43
	U10	U11	U12	U13	U14	U15	U16	U17	U18
# stories	83	40	157	191	97	38	166	118	127
# categories	68	22	32	18	29	17	46	27	15

Table 2: Summary of the News Video Assessment Task

	U1	U2	U3	U4	U5	U6	U7	U8	U9
# days with annotated results	70	76	65	39	50	59	86	88	59
# relevant assessed stories	234	297	217	101	112	155	302	99	203
	U10	U11	U12	U13	U14	U15	U16	U17	U18
# days with annotated results	44	52	69	58	36	51	69	71	32
# relevant assessed stories	156	137	200	187	69	124	187	160	95

results each and the participants were asked to assess at least the first three pages.

Table 2 shows the summary of the news video assessment task. As can be seen, the assessment task ended with diverse results, indicated by the different number of relevant assessed stories and different number of days with annotated results.

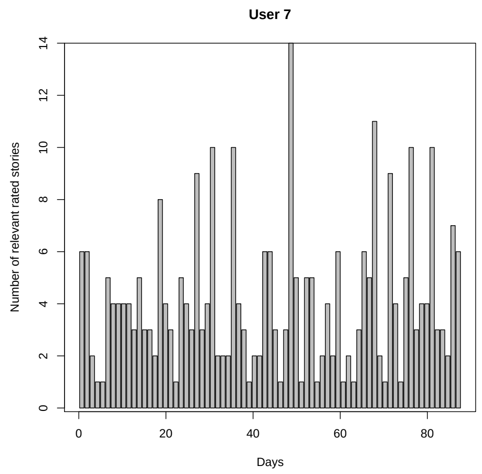


Figure 1: Number of relevant rated stories per day (User 7)

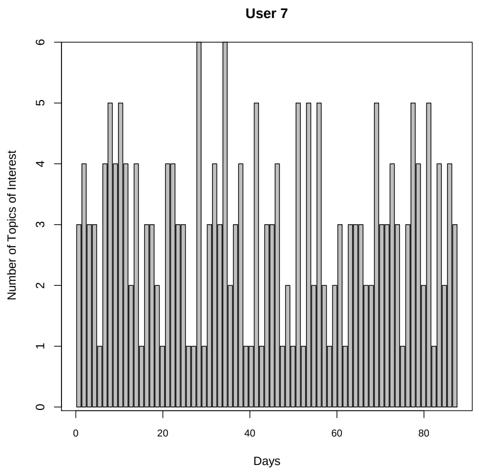


Figure 2: Number of topics of interest per day (User 7)

Figures 1 and 2 show the number of relevant rated stories and the frequency of topics of interest per day for User U7. Similar patterns can be observed for all participants. As these figures illustrate, the occurrence frequency of topics of user's interest is highly variable. Since users will show diverse interest in news stories on various days, we thus conclude that these assessment lists reflect realistic user interests. In the final questionnaire, we aimed at evaluating whether the presented subset of the data corpus was appropriate. Using Five-Point Likert scales, we asked the participants to judge whether the displayed news stories were rela ted to the corresponding news topic. Even though the majority had a neutral perception towards this statement, 43% slightly agreed to it. Moreover, they were asked to judge whether the news stories covered most subtopics of the according topic on a Five-Point Likert scale. Again, the participants tended to agree with the statement. We therefore conclude that using the news article assessments to identify good search queries resulted in sensible subsets of the actual video data corpus.

In summary, in this section, we introduced an approach to the generation of personalised ground truth lists. In order to reduce the amount of manual labour, we aimed to adapt the assessable documents to the assessors personal interests. Both quality and quantity of the resulting lists varies from user to user. While some users provide a large amount of assessments, other users assess a small amount of stories. Consequently, not all relevant documents are assessed to be relevant by the users. Nevertheless, since this is a well known problem that also influences other well- established relevance assessment approaches, we accept this to be an inevitable problem.

5 Simulating long-term user profiles

The relevance assessment lists which have been introduced in the previous section express the interests of 18 individual assessors. Now, we will exploit the above collection to illustrate how these assessments can be used to benchmark long-term user profiling approaches. Focusing on the usage scenario introduced in Section 2, we simulate a user profiling approach which captures users' implicit relevance feedback in a user profile. In this simulation, we model users interacting with a news video recommender interface as introduced in [Hopfgartner and Jose, 2010] over a longer period of time.

5.1 Training a User Interaction Model

The first step towards creating long-term user profiles is to simulate users interacting with the system. [Dix *et al.*, 1993] argue that user interactions in interactive systems can be represented as a series of low-level events, e.g. key presses or mouse clicks. User actions can be seen as a sequence of one or more of these events. In [Hopfgartner and Jose, 2007], the authors illustrate some possible user actions in a video retrieval interfaces. They argue that some events are independent, while other events depend on preceding events. Two events in the interface shown in [Hopfgartner and Jose, 2010] can be triggered independently from others: Users can always move the mouse over a result to get more information (tooltip event) and can always expand a search result (clicking event). Once a story is expanded, the user can browse through the shots (browsing event) or start playing the video (viewing event). The latter events are hence dependent on the clicking event. We describe possible event sequences as a Markov Chain since this allows us to model possible combinations of event sequences. Markov Chains consist of states and transitions. A state change is triggered by a certain event with a certain probability. Figure 3 illustrates the possible user interactions of users using the example interface. The probabilities of the above introduced events trigger the transitions between the different states. Note that for simplicity, we assume that users will do every event only once.
Following [Vallet *et al.*, 2008], the transitions can be defined as follows:

$$P(R|\text{Click}) = \frac{\text{\# relevant clicks}}{\text{\# total clicks}} \quad (1)$$

$$P(\neg R|\text{Click}) = \frac{\text{\# non-relevant clicks}}{\text{\# total clicks}} = 1 - P(R|\text{Click}) \quad (2)$$

$$P(\text{Clicking}|R) = \frac{\text{\# clicks on relevant stories in result set}}{\text{\# relevant rated stories}} \quad (3)$$

$$P(\text{Clicking}|\neg R) = \frac{\text{\# clicks on non-relevant stories in result set}}{\text{\# non-relevant rated stories}} \quad (4)$$

$$P(\text{Previewing}|R) = \frac{\text{\# tooltip highlighting on relevant stories}}{\text{\# relevant rated stories}} \quad (5)$$

$$P(\text{Previewing}|\neg R) = \frac{\text{\# tooltip highlighting on non-relevant stories}}{\text{\# non-relevant rated stories}} \quad (6)$$

$$P(\text{Viewing}|R) = \frac{\text{\# playing of relevant stories in result set}}{\text{\# relevant rated stories}} \quad (7)$$

$$P(\text{Viewing}|\neg R) = \frac{\text{\# playing of non-relevant stories in result set}}{\text{\# non-relevant rated stories}} \quad (8)$$

$$P(\text{Browsing}|R) = \frac{\text{\# browses of relevant stories in result set}}{\text{\# relevant rated stories}} \quad (9)$$

$$P(\text{Browsing}|\neg R) = \frac{\text{\# browses of non-relevant stories in result set}}{\text{\# non-relevant rated stories}} \quad (10)$$

Having defined a Markov Chain to simulate user interactions, the next step is now to determine realistic probabilities for each transition in the chain. The best way to simulate realistic user interaction patterns is to analyse how real users interact with the video retrieval system. In order to obtain a set of characterisation parameters, we use statistical information by exploiting the log files of a user study [Hopfgartner and Jose, 2010] to calculate the probability of certain types of actions. According to the log files, the average probability of clicking on a document and rating this document $P(R|Click)$ is 0.55. In other words, approximately every second story that the users interacted with was labelled as relevant by the user. Table 3 shows the averaged probabilities of an implicit action being performed on relevant and non-relevant documents.

Table 3: Probability values of possible action types

Action Type	Probability
$P(\text{Clicking}|R)$	0.34
$P(\text{Clicking}|\neg R)$	0.04
$P(\text{Previewing}|R)$	0.21
$P(\text{Previewing}|\neg R)$	0.02
$P(\text{Viewing}|R)$	0.42
$P(\text{Viewing}|\neg R)$	0.043
$P(\text{Browsing}|R)$	0.97
$P(\text{Browsing}|\neg R)$	0.01

5.2 Simulated Evaluation

Exploiting the possible user actions and the determined probability values, we aim to simulate the scenario introduced in Section 2, resulting in simulated long-term user profiles. Starting with the first day contained in the individual users assessment list, we simulate a user interacting with the news stories of the day according to the introduced user patterns. This interaction is then captured in the user profile. This procedure is repeated for every day with assessed news documents. The outcome of this simulation is 18 user profiles which contain news documents that the simulated user interacted with. Each simulated user profile has been created iteratively. For every day which is covered in the ground truth data, new documents have been added, resulting in a daily update of the user profile. In order to evaluate the personalisation technique, standard evaluation measures can be computed for every day represented in the user profile.

Figure 4 plots the performance of two simulation runs R(1) and R(2) for every evaluated day with respect to MAP using User U7's assessments list. Note that these runs serve as an example to illustrate the use of the data collection for evaluation. A description of the approaches is therefore outside the scope of this paper. Various observations can be made from this graph. First of all, in both cases, the recommendation performance fluctuates. The peaks, however, appear synchronously in both runs. We assume that

Figure 3: Markov Chain of User Actions

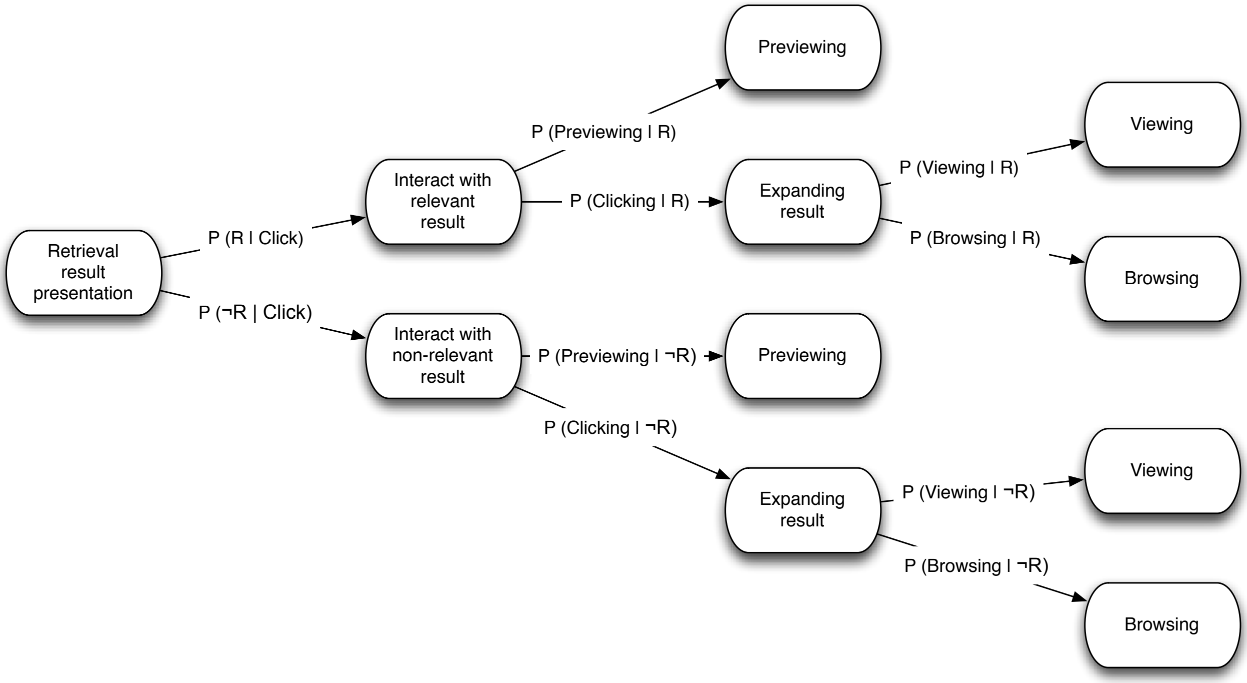
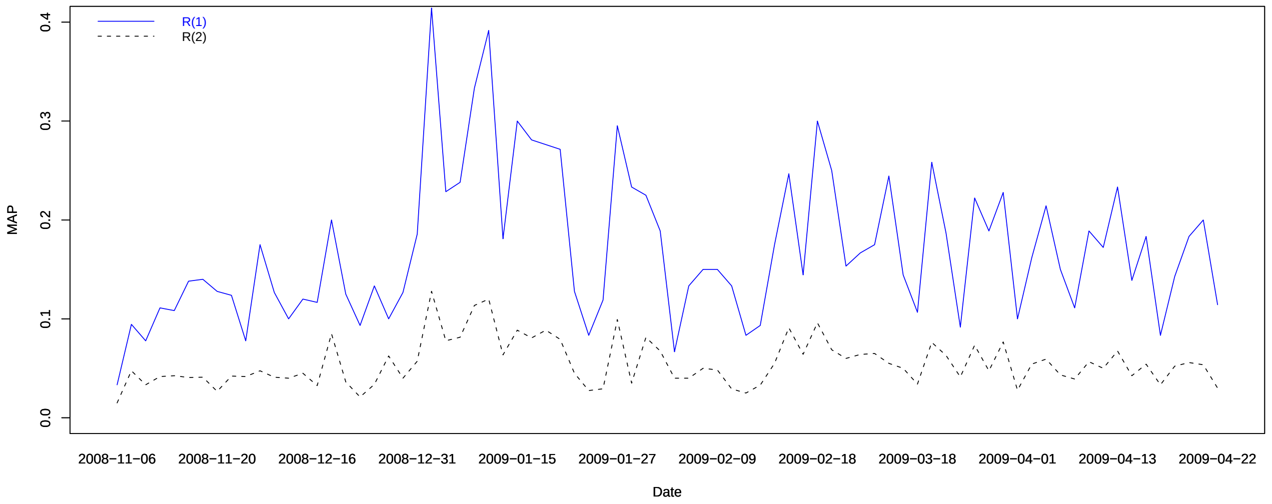


Figure 4: Recommendation performance for every evaluated day (measured in MAP)



the recommendations are directly influenced by the quality and quantity of the assessment lists, as shown in Figure 1. If a user shows less interest in news on a specific day, recommendations will reflect this fact. Another interesting observation is, however, that in all iterations, run R(1) outperforms R(2). Moreover, the figure does not show lower recommendation performance after numerous iterations. The results hence indicate that the recommendation approaches can be used over a longer term. Both results could not have been achieved using a classical evaluation scheme, since they do not support the evaluation of different runs over several iterations.

6 Conclusion and Future Work

In this paper, we suggest the development of a new test collection used for studying long-term user modelling techniques in video retrieval. We first introduced an approach of generating independent ground truth lists. In order to reduce the amount of manual labour, we aimed at adapting the documents to assess to the assessors' personal interests. Therefore, volunteers were asked to assess a textual news corpus and to identify news stories they are interested in. Further, they were asked to categorise these news stories into specific news topics. This first assessment step enables us to identify the assessors' interests in news topics. We further exploit this knowledge and identify potential relevant videos in a news video corpus. The assessors were then asked to assess the relevance of this subset. In order to study long-term profiling, we propose a simulation based evaluation scheme. We defined unique interaction patterns and identified usage patterns by exploiting a preceding user study. Moreover, we employ both patterns and ground truth lists to generate long-term user profiles. Finally, we illustrate how these user profiles can be used to evaluate long-term personalisation approaches. The developed test collection enables us to evaluate the performance of different adaptation approaches over multiple iterations. Using a classical evaluation scheme, such an evaluation would have been challenging. The main conclusion which can therefore be drawn is that the introduced data collection can be used for the benchmarking of long term recommendation approaches. Since all results are achieved by employing a simulation, further runs can be performed to fine tu ne recommendation parameters. Nevertheless, even though simulations can be used for benchmarking, it cannot replace real user studies. Future work includes therefore a thorough analysis of long-term adaptation approaches, followed by a real user study.

Acknowledgment

The first author was supported by a fellowship within the Postdoc-Programme of the German Academic Exchange Service (DAAD).

References

[Belkin, 2008] Nicholas J. Belkin. Some(what) grand challenges for information retrieval. *SIGIR Forum*, 42(1):47–54, 2008.

[Choicestream, Inc., 2008] Choicestream, Inc. 2008 Choicestream Personalization Survey. Technical Report, Choicestream, Inc., 210 Broadway, Fourth Floor, Cambridge, MA, USA, 2008.

[Cuadra, 1967] Carlos A. Cuadra. Opening the Black Box of Relevance. *Journal of Documentation*, 42(1):291–303, 1967.

[Dix *et al.*, 1993] Alan Dix, Janet Finlay, and Russell Beale. Analysis of user behaviour as time series. In *HCI'92: Proceedings of the Conference on People and computers VII*, pages 429–444, New York, NY, USA, 1993. Cambridge University Press.

[Hopfgartner and Jose, 2007] Frank Hopfgartner and Joemon M. Jose. Evaluating the implicit feedback models for adaptive video retrieval. In *Multimedia Information Retrieval*, pages 323–331, 2007.

[Hopfgartner and Jose, 2009] Frank Hopfgartner and Joemon M. Jose. On user modelling for personalised news video recommendation. In *UMAP'09: Proceedings of the 17th International Conference on User Modeling, Adaptation, and Personalization*, pages 403–408, 2009.

[Hopfgartner and Jose, 2010] Frank Hopfgartner and Joemon Jose. Semantic user modelling for personal news video retrieval. In *MMM'10: Proceedings of the Conference on Multimedia Modeling*, pages 336–346. Springer Berlin / Heidelberg, 2010.

[Hopfgartner, 2011] Frank Hopfgartner. Adaptive interactive news video recommendation: An example system. In *SEMAIS'11 - Second International Workshop on Semantic Models for Adaptive Interactive Systems, Palo Alto, CA, USA*, pages 21–25, 2 2011.

[Ivory and Hearst, 2001] Melody Y. Ivory and Marti A. Hearst. The state of the art in automating usability evaluation of user interfaces. *ACM Comput. Surv.*, 33(4):470–516, 2001.

[Kelly *et al.*, 2009] Diane Kelly, Susan T. Dumais, and Jan O. Pedersen. Evaluation Challenges and Directions for Information-Seeking Support Systems. *IEEE Computer*, 42(3):60–66, 2009.

[Lioma and Ounis, 2006] Christina Lioma and Iadh Ounis. Examining the Content Load of Part of Speech Blocks for Information Retrieval. In *ACL'06: Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Sydney, Australia*, 2006.

[Misra *et al.*, 2010] Hemant Misra, Frank Hopfgartner, Anuj Goyal, P. Punitha, and Joemon M. Jose. Tv news story segmentation based on semantic coherence and content similarity. In *MMM'10*, pages 347–357, 2010.

[Sanderson and Joho, 2004] Mark Sanderson and Hideo Joho. Forming test collections with no system pooling. In *SIGIR'04: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 33–40, 2004.

[Sanderson, 2006] Mark Sanderson. Test Collections for all (Position Paper). In *AIR'06: Proceedings of the First International Workshop on Adaptive Information Retrieval, Glasgow, United Kingdom*, page 5, 10 2006.

[Spärck-Jones and van Rijsbergen, 1965] Karen Spärck-Jones and C. J. van Rijsbergen. Report on the need for and provision of an ideal information retrieval test collection. Technical report, Computing Laboratory, University of Cambridge, UK, 1965.

[Sparck Jones and Willett, 1997] Karen Sparck Jones and Peter Willett. *Evaluation*, pages 167–174. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1997.

[Vallet *et al.*, 2008] David Vallet, Frank Hopfgartner, and Joemon M. Jose. Use of implicit graph for recommending relevant videos: A simulated evaluation. In *ECIR*, pages 199–210, 2008.

[Voorhees, 2001] Ellen M. Voorhees. The philosophy of information retrieval evaluation. In *CLEF'01: Revised Papers of the Second Workshop of the Cross-Language Evaluation Forum, Evaluation of Cross-Language Information Retrieval Systems*, pages 355–370, 2001.

[Voorhees, 2008] Ellen M. Voorhees. On test collections for adaptive information retrieval. *Information Processing & Management: An International Journal*, 44(6):1879–1885, 2008.

[White *et al.*, 2005] Ryen W. White, Ian Ruthven, Joemon M. Jose, and C. J. van Rijsbergen. Evaluating implicit feedback models using searcher simulations. *ACM Transactions on Information Systems*, 23(3):325–361, 2005.