Semantic User Modelling for Personal News Video Retrieval

Frank Hopfgartner and Joemon M. Jose

Dept. of Computing Science, University of Glasgow, Glasgow, G12 8RZ, UK
{hopfgarf, jj}@dcs.gla.ac.uk

Abstract. There is a need for personalised news video retrieval due to the explosion of news materials available through broadcast and other channels. In this work we introduce a semantic based user modeling technique to capture the users' evolving information needs. Our approach exploits the Linked Open Data Cloud to capture and organise users' interests. The organised interests are used to retrieve and recommend news stories to users. The system monitors user interaction with its interface and uses this information for capturing their evolving interests in the news. New relevant materials are fetched and presented to the user based on their interests. A user-centred evaluation was conducted and the results show the promise of our approach.

1 Introduction

A challenging problem in the user profiling domain is to create profiles of multimedia retrieval system users. Due to the Semantic Gap, it is not trivial to understand the content of multimedia documents and to find other documents that the users might be interested in. A promising approach to ease this problem is to set multimedia documents into their semantic contexts. For instance, a video about Barack Obama's speech in Ghana can be put into different contexts. First of all, it shows an event which happened in Accra, the capital of Ghana. Moreover, it is a visit by an American politician, the current president. Retrieving a video about Obama's visit to Ghana might indicate that someone is interested in either Barack Obama, Ghana, or in both.

Another challenge in user profiling research is the identification of users' interests in various events. Multiple interests lead to a sparse data representation and approaches need to be studied to tackle this sparsity.

In this paper, we introduce a semantic user profiling approach for news video retrieval, which exploits a generic ontology to put news stories into a context. In order to identify a user's interest in specific topics, we exploit his/her relevance feedback which is provided implicitly while interacting with the system. The remainder of this paper is structured as followed: In Section 2, we introduce various research domains which are relevant within our study and discuss research challenges in Section 3. In Section 4, we introduce our system from capturing daily news to presenting them to the users and evaluate it in Section 5. Section 6 provides a conclusion and a discussion of our findings.

2 Related Work

In this section, we introduce state-of-the-art methodologies to address research challenges that our work builds upon.

User profiling is the process of learning a user's interests over a long period of time. Several approaches have been studied to capture users' news interests in a profile. Chen and Sycara [5] analyse internet users during their information seeking task and explicitly ask them to judge the relevance of the webpages they visit. Exploiting the created user profile of interest, they generate a personalised newspaper containing daily news. However, providing explicit relevance feedback is a demanding task and users tend not to provide much feedback [8]. Bharat et al. [2] created a personalised online newspaper by unobtrusively observing the user's web-browsing behaviour. Although their system is a promising approach to release the user from providing feedback, their main research focus is on developing user interface aspects, ignoring the sophisticated retrieval issues. The web-based interface of their system provides a facility to retrieve news stories and recommends stories to the user based on his/her interest.

An interesting approach for news personalisation is to map relationships between concepts in the user profile by using ontologies. Fernández et al. [7] argue that ontologies can be exploited to structure news items and to annotate them with additional information. In the news video domain, Bürger et al. [3] have shown that such structured data can be used to assist the user in accessing a large news video corpus. Dudev et al. [6] propose the creation of user profiles by creating knowledge graphs that model the relationship between different concepts in the Linked Open Data Cloud. This collection of ontologies unites information about many freely available different concepts. The backbone of the cloud is DBpedia, an information extraction framework which interlinks Wikipedia content with other databases on the Web such as Geonames or WordNet. In this paper, we exploit this data cloud to link automatically segmented story videos. Challenges and open research questions are introduced in the next section.

3 Research Challenges

Various challenges arise when aiming at creating semantic user profiles in the multimedia domain.

State-of-the-art user profiling approaches exploit the textual content of relevant documents to identify user's interests. Considering the short length of news video stories, creating useful profiles is rather problematic. The development of Semantic Web technologies promise a solution for this problem. If a story contains various concepts, additional information about these concepts might help to model user interests more accurate. Järvelin et al. [10] already showed that a concept-based query expansion is helpful to improve retrieval perfomance. Adapting their approach, we hypothesise that ontologies can be exploited to organise user interests and also the pre-fetched relevant documents.

The next problem is how the user's evolving interests can be captured in a long-term user profile. What a user finds interesting on one day might be

completely irrelevant on the next day. In order to model this behaviour, we incorporate the Ostensive Model of developing Information Need [4]. In this model, providing feedback on a document is considered as ostensive evidence that this document is relevant for the user's current interest. As argued before, however, users tend not to provide constant feedback on what they are interested in. Thus, one condition we set is that a user profile should be automatically created by capturing users' *implicit* interactions with the retrieval interface. Our next hypothesis is hence that implicit relevance feedback techniques can efficiently be employed to create efficient long-term user profiles.

Another problem in the context of user profiling is the users' multiple interests in various topics. For example, users may be interested in Sports and Politics or in Business news. Further, they can even be interested in sub categories such as Football, Baseball or Hockey. A specification for a long-term user profile should therefore be to automatically identify these multiple aspects. We hypothesise that separating user profiles based on broader news categories can lead to a structured representation of the users' interests. This will lead to a more indicative presentation of materials. Moreover, we hypothesise that a hierarchical agglomerative clustering of the content of these category-based profiles can be used to effectively identify sub categories.

Summarising, we address the following hypotheses in this work:

1. Implicit relevance feedback techniques can be exploited to create efficient long-term user profiles.
2. Separating user profiles based on broader news categories can lead to a structured representation of the users' interests.
3. Hierarchical agglomerative clustering of the content of these category-based profiles can be used to effectively identify sub categories.
4. Ontologies can be exploited to organise user interests and also the pre-fetched documents.

In order to study these hypotheses, we introduce a novel news video retrieval system which automatically captures users' interests. The system and its components will be introduced in the next section.

4 System Description

The architecture of the introduced news video retrieval system can be segmented into three conceptual parts: A data processing phase, the graphical user interface and the profiling module. Since we want to provide an up-to-date news video collection, the data processing phase is called twice a day, starting with the actual capturing of the broadcast and the decoding of the teletext transmission. In this study, we focus on the daily BBC One O'Clock News and the ITV Evening News, the UK's largest news programmes. Each bulletin with a running time of thirty minutes is enriched with a teletext signal. Following Hopfgartner et al. [9], we segment these news videos into coherent news stories. In the remainder of this section, we introduce the steps from annotating these news stories using

external sources and indexing them. Moreover, we introduce the system interface and discuss our user profiling approach.

4.1 Semantic Annotation

Usually, news content providers classify their news in accordance to the IPTC standard, a news categorisation thesaurus developed by the International Press Telecommunications Council. We use OpenCalais[1], a Web Service provided by Thomson Reuters, to classify each story into one or more of the following IPTC categories: Business & Finance, Entertainment & Culture, Health, Medical & Pharma, Politics, Sports, Technology & Internet and Other.

In a next step, we aim to identify concepts that appear in the stories. Once these concepts have been positively identified, the Linked Open Data Cloud can be exploited to further annotate the stories with related concepts. Three problems arise when conducting this procedure.

First of all, how can we determine concepts in the story which are strong representatives of the story content? In the text retrieval domain, named entities are considered to be strong indicators of the story content, since they carry the highest content load among all terms in a document. Therefore, we extract persons, places and organisations from each story transcript using OpenCalais.

The second question is, how can these named entities be positively matched with a conceptual representation in the Linked Open Data Cloud. For resolving the identity of an entity instance, we again rely on the OpenCalais Web Service, which compares the actual entity string with an up-to-date database of entities and their spelling variations. Once entities have been disambiguated, OpenCalais maps these entities with a uniform resource identifier (URI) and their representation in DBpedia.

Since the link between the story and the Linked Open Data Cloud has been established, the next problem is how can the structured knowledge represented in the Linked Open Data Cloud be exploited to augment the story. A long-term user profile which is created using implicit evidence will contain many entries, which makes a weighted semantic network approach as suggested by Dudev et al. [6] infeasible. Therefore, we consider only direct links from the identified concept to other concepts in the Semantic Web. We therefore augment the stories with all URIs that are directly associated with these entities in the Cloud.

4.2 User Interface

Figure 1 shows a screenshot of the news video retrieval interface. It can be split into three main areas: Search queries can be entered in the search panel on top, results are listed on the right side and a navigation panel is placed on the left side of the interface. When logging in, the latest news will be listed in the results panel. Search results are listed based on their relevance to the query. Since we are using a news corpus, however, users can re-sort the results in chronological

[1] http://www.opencalais.com/

Fig. 1. Graphical User Interface of the System

order with latest news listed first. Each entry in the result list is visualised by an example keyframe and a text snippet of the story's transcript. Keywords from the search query are highlighted to ease the access to the results. Moving the mouse over one of the keyframes shows a tooltip providing additional information about the story. A user can get additional information about the result by clicking on either the text or the keyframe. This will expand the result and present additional information including the full text transcript, broadcasting date, time and channel and a list of extracted named entities. In the example screenshot, the third search result has been expanded. The shots forming the news story are represented by animated keyframes of each shot. Users can browse through these animations either by clicking on the keyframe or by using the mouse wheel. This action will center the selected keyframe and surround it by its neighboured keyframes. The keyframes are displayed in a cover-flow view, meaning that the size of the keyframe grows larger the closer it is to the focused keyframe. In the expanded display, a user can also select to play a video, which opens the story video in a new panel.

The user's interactions with the interface are expoited to identify multiple topics of interests (see Section 4.3). On the left hand side of the interface, these interests are presented by different categories. Clicking on any of these categories in the navigation panel will reveal up to four sub categories for the according category. The profiling approach will be introduced in the following section.

4.3 User Profiling

When a user interacts with a result, he leaves a "semantic fingerprint" that he is interested in the content of this item to a certain degree. In this work, we

employ a *weighted story vector* approach to capture this implicit fingerprint in a profile. The weighting of the story will be updated when the system submits a new weighted story to the profile starting a new iteration j. Hence, we represent the interaction I of a user i at iteration j as a vector of weights

$$\boldsymbol{I}_{ij} = \{W_{ij1}...W_{ijs}\}$$

where s indexes the story in the whole collection. The weighting W of each story expresses the evidence that the content of this story matches the user's interest. The higher the value of W, the closer this match is. In this work, we define a static value for each possible implicit feedback feature:

$$W = \begin{cases} 0.1, & \text{when a user browses through the keyframes} \\ 0.2, & \text{when a user uses the highlighting feature} \\ 0.3, & \text{when a user expands a result} \\ 0.5, & \text{when a user starts playing a video} \end{cases}$$

Note that some of these features are independent, while others depend on a previous action (e.g. a video cannot be played without being clicked on).

As explained before, each news story has been classified as belonging to one or more broad news categories C. Since we want to model the user's multiple interests, we use this classification as a splitting criteria. Thus, we represent user i's interest in C in a category profile vector $\boldsymbol{P}_i(C)$, containing the story weight $SW(C)$ of each story s of the collection:

$$\boldsymbol{P}_i(C) = \{SW(C)_{i1}...SW(C)_{is}\}$$

In the user interface, each category profile is represented by an item in the navigation panel.

In our category profile, the story weight for each user i is the combination of the weighted stories s over different iterations j: $SW(C)_{is} = \sum_j a_j W_{ijs}$. Following Campbell and van Rijsbergen [4], we include the ostensive evidence

$$a_j = \frac{1 - C^{-j+1}}{\sum_{k=2}^{j_{max}} 1 - C^{-k+1}} \tag{1}$$

to introduce an inverse exponential weighting which will give a higher weighting to stories which have been added more recently to the profile, compared to stories which were added in an earlier stage.

The above introduced methodology results in a category-based representation of the user's interests. Each category profile consists of a list of weighted stories, with the most important stories having the highest weighting. A challenge is here to identify different contextual aspects in each profile. We approach this problem by performing a hierarchical agglomerative clustering of stories with the highest story weight at the current iteration. Following Bagga and Baldwin [1], we treat the transcripts extracted from these stories as term vectors and compare them by cosine. Unlike their approach, however, we use the whole transcript rather

than sentences linked by coreferences and use the square root of raw counts as our term frequencies rather than the raw counts. We use complete-link clustering since this approach results in more compact clusters. Moreover, we do not use inverse-document frequency normalisation since this value can be important for discremination. For tokenisation, we use standard filters (conversion to lower case, stop word removal and stemming). The numbers of clusters k is a parameter. Since each cluster should contain stories associated with an aspect of the user's interest, k should be equal to the number of different interests that a user has. In this study, we have set $k = 4$. In the interface, the clusters represent the four sub categories under each category in the navigation panel. The two most frequent named entities in each cluster are used as a label for each sub category.

The content of the users' profiles is displayed on the navigation panel of the left hand side of the interface. Since the idea of such navigation panel is to assist the users in finding other stories that match their interests, the next challenge is to identify more stories in the data corpus that might be of the users' interests. Assuming that each of the sub categories contains stories that cover one or more (similar) aspects of a user's interest, the content of each sub category can be exploited to recommend more documents belonging to that cluster. The simplest method is to create a search query based on the content of each cluster and to retrieve stories using this query. A promising source to create such queries is the use of most frequent named entities within each cluster. Due to the rather short length of the story transcript, we identify additional named entities by performing an additional pseudo relevance feedback step.

For the initial search, we first extract all URIs from the cluster and retrieve stories containing these URIs. Then, we extract all named entities from the stories in the result list and finally use the most frequent entities as a search query.

5 Evaluation

In order to evaluate the hypotheses which have been introduced in Section 3, we performed a user study which will be described in the remainder of this section.

5.1 Experimental Design

Since the proposed profiling approach includes the capturing of long-term user interests, we had to study the effectiveness of our system over several days. We therefore captured six month of news video broadcasts and paid participants to use the system as additional source of information in their daily news consumption routine. Their interactions with the system were logged to evaluate the approach. They were asked to use the system for up to ten minutes each working day for up to seven days to search for any topic that they were interested in. In addition, we also created a simulated search task situation. Our expectation was twofold: First of all, we wanted to guarantee that every user had at least one topic to search for. Moreover, we wanted the participants to actually explore the

data corpus. Therefore, we chose a scenario which had been a major news story over the last few months:

> "Dazzled by high profit expectations, you invested a large share of your savings in rather dodgy securities, stocks and bonds. Unfortunately, due to the credit crunch, you lost about 20 percent of your investment. Wondering how to react next and what else there is to come, you follow every report about the financial crisis, including reports about the decline of the house's market, bailout strategies and worldwide protests."

Each participant started with an individual introductory session, where they were asked to fill in an entry questionnaire and could familiarise themselves with the interface. Every day, they were asked to fill in an online report where they were encouraged to comment on the system as they used it. At the end of the experiment, everyone was asked to fill in an exit questionnaire to provide feedback on their experience during the study.

5.2 Participants

16 users with an average age of 30.4 years participated in our experiment. Their favourite sources for gathering information on the latest news stories are news media web portals, word-of-mouth and the television. The typical news consumption habit they described was to check the latest news online in the morning and late at night after dinner. We hence conclude that the participants represent the main target group for the introduced retrieval system.

5.3 Results

By asking for daily reports, our goal was to evaluate the users' opinion about the system at various stages of the experiment. The first question was to find out what the participants actually used the system for. The majority of participants used it to retrieve the latest news, followed by identifying news stories they were not aware of before.

One of our main research interests was to determine whether the system provides satisfactory access to the data collection. Therefore, we asked the participants to judge various statements on a Five-Point-Likert scale from 1 (Agree) to 5 (Disagree). The order of the agreements varied over the questionnaire to reduce bias. Figure 2 shows the average judgement of all users over all seven days for two statements that were aimed at determing the general usability of the system. The first statement posed was "The interface structure helped me to explore the news collection", denoted "explore collection" in the figure. The second statement was "The interface helped me to explore various topics of interest", denoted "explore topics". As can be seen, the average user found the interface useful and was satisfied with its usability.

With the aim of evaluating our first hypothesis that implicit relevance feedback can be used to create long-term user profiles, we asked the participants

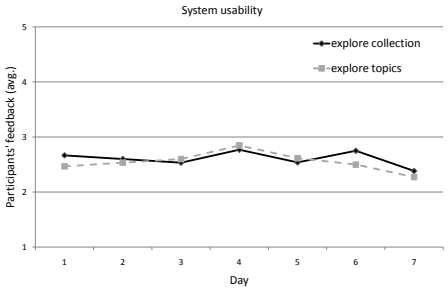

Fig. 2. Feedback on system usability (Lower is better)

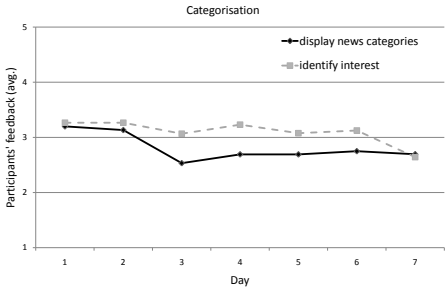

Fig. 3. Feedback on interest capturing (Lower is better)

to judge if the system was effective in automatically identifying their interests. Another statement was "the system sucessfully identified and displayed news categories I was interested in". As Figure 3 illustrates, the participants did neither agree or disagree to these statements, which corralates with the observation that the use of implicit features for short-term user modelling provides weaker evidence of relevance than explicit relevance feedback [11]. Nevertheless, a tendency towards a positive rating for the use of implicit indicators is visible, in particular towards the end of the user study. This suggests that implicit relevance feedback can be used to create long-term user profiling. However, further research is necessary to differentiate positive and negative indicators of relevance, which is beyond the scope of this work.

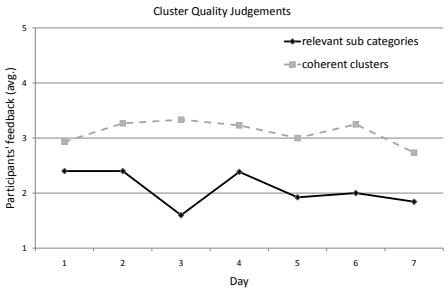

Fig. 4. Feedback on sub categories (Lower is better)

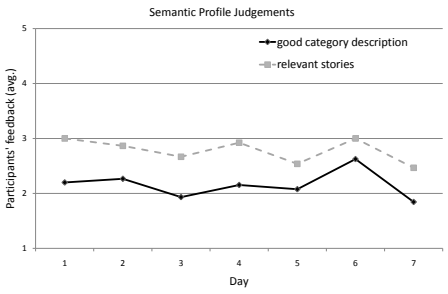

Fig. 5. Feedback on ontology-based recommendations (Lower is better)

With the goal of evaluating the news categorisation, we asked the users to judge the following two statements: "The displayed sub categories represent my interests in various topics" and "the displayed results for each sub category were

related to each other". Figure 4 shows the average answer over the whole time of the experiment. The first question, denoted "relevant sub categories", aimed to evaluate whether separating user profiles based on broader categories leads to a structured representation of the users' interests, our second hypothesis. The second question aimed to evaluate the coherence of each category-based profile, targeting the third hypothesis. As can be seen, the participants had a positive perception of the relevance of the sub categories. This could indicate that our clustering approach was successful in identifying diverse aspects of the same news category, supporting our hypothesis that categorising the user interests into broader categories provides a structured representation to the users' interests. Concerning the coherence of each cluster, the participants tended towards a neutral perception.

The last set of statements aimed to evaluate the fourth hypothesis that ontologies can be exploited to organise user interests and also the pre-fetched relevant documents. Thus, participants judged the following differentials: "The displayed results for each category matched with the category description" and "the displayed results for each category contained relevant stories I did not retrieve otherwise". Figure 5 shows the relevant responses. Again, a tendency towards a positive perception of the results which are determined using semantics is visible. In order to explore this hypothesis further, we analysed user transaction patterns which were captured in the log files. Our analysis revealed that the participants used the provided subcategories extensively. Roughly 40% of all search queries were triggered by clicking on one of these categories. This high percentage suggests that the participants found the results given by these categories to be useful, which would suggest that our semantics based user profiling is effective in recommending relevant results.

6 Conclusion

In this paper, we introduced a semantic based user modeling technique which automatically captures the users' evolving information needs and represents this interest in dynamic user profiles. Therefore, we introduce a novel news video retrieval system which automatically captures daily broadcasting news and segments the bulletins into coherent news stories. The Linked Open Data Cloud is exploited to set these stories into context. This semantic augmentation of the news stories is used as the backbone of our user profiling methodology. The profiles can be used to identify the users' multiple interests in diverse aspects of news over a longer period of time. The semantic augmentations of the stories in the user profile are used to fetch new relevant materials.

Our preliminary study is based on four hypotheses, which are evaluated using a user-centred evaluation scheme. 16 participants were asked to include the news retrieval system into their daily news gathering routine and to judge the performance of the system on a daily basis. Differing from standard interactive information retrieval experiments, the evaluation was split into multiple sessions and performed under an uncontrolled environment, two necessary conditions for

a realistic evaluation of a long-term user profiling. This novel approach cannot rely on system-centred evaluation measures as common in information retrieval experiments. Thus, standardised evaluation measures need yet to be developed. The hypotheses were evaluated by analysing users' feedback which was provided during various stages of the experiment. The analysis of their feedback forms seem to support all hypotheses, suggesting that the introduced system can be effectively used to provide a personalised access to video news data. In fact, a majority of all participants claimed in the exit questionnaire that they would use a commercialised system with the presented features for their daily news gathering.

Future work includes a more thorough selection of concepts in the Linked Open Data cloud to be used for augmenting each story. Currently, every concept that is directly linked with the story's concept is used, resulting in many concepts of less importance. A better selection scheme can lead to stronger links between related stories, hence increasing the effectiveness of the introduced approach.

Acknowledgments

This research was supported by the EC under contract FP6-027122-SALERO.

References

1. A. Bagga and B. Baldwin. Entity-based cross-document coreferencing using the vector space model. In *Proc. Comput. Ling.*, pages 79–85, 1998.
2. K. Bharat, T. Kamba, and M. Albers. Personalized, interactive news on the web. *Mult. Syst.*, 6(5):349–358, 1998.
3. T. Bürger, E. Gams, and G. Güntner. Smart content factory: assisting search for digital objects by generic linking concepts to multimedia content. In *Proc. HT*, pages 286–287. ACM, 2005.
4. I. Campbell and C. J. van Rijsbergen. The ostensive model of developing information needs. In *Proc. Library Science*, pages 251–268, 1996.
5. L. Chen and K. Sycara. WebMate: A personal agent for browsing and searching. In K. P. Sycara and M. Wooldridge, editors, *Proc. Agents'98*, pages 132–139, New York, 9–13, 1998. ACM Press.
6. M. Dudev, S. Elbassuoni, J. Luxenburger, M. Ramanath, and G. Weikum. Personalizing the Search for Knowledge. In *Proc. PersDB*, 08 2008.
7. N. Fernández, J. M. Blázquez, J. A. Fisteus, L. Sánchez, M. Sintek, A. Bernardi, M. Fuentes, A. Marrara, and Z. Ben-Asher. NEWS: Bringing Semantic Web Technologies into News Agencies. In *Proc. ISWC*, pages 778–791, 2006.
8. M. Hancock-Beaulieu and S. Walker. An evaluation of automatic query expansion in an online library catalogue. *J. Doc.*, 48(4):406–421, 1992.
9. H. Misra, F. Hopfgartner, A. Goyal, P. Punitha, and J. M. Jose. TV News Story Segmentation based on Semantic Coherence and Content Similarity. In *MMM'10*, 2010. to appear.
10. K. Järvelin, J. Kekäläinen, and T. Niemi. ExpansionTool: Concept-Based Query Expansion and Construction. *Information Retrieval*, 4(3):231–255, 2001.
11. D. Kelly and J. Teevan. Implicit feedback for inferring user preference: a bibliography. *SIGIR Forum*, 37(2):18–28, 2003.