ECAI 2014
T. Schaub et al. (Eds.)
© 2014 The Authors and IOS Press.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License.
doi:10.3233/978-1-61499-419-0-1227

Orwellian Eye: Video Recommendation with Microsoft Kinect

Tomáš Kliegr[1] and **Jaroslav Kuchař**[2]

Abstract. This paper demonstrates Interest Beat (InBeat.eu) as a recommender system for online videos, which determines user interest in the content based on gaze tracking with Microsoft Kinect in addition to explicit user feedback. Content of the videos is represented using a semantic wikifier. User profile is constructed from preference rules, which are discovered with an association rule learner.

1 Introduction

Eye tracking was recently proposed as an effective way of obtaining highly detailed user feedback [9]. Content-based recommendation has so far relied either on explicit information on user interest, or on those user actions (implicit feedback), which could be interpreted as a manifestation of user (dis)interest in a certain object [7]. While the latter does not require the user to perform any extra activity, the information obtainable in this way on a particular content item is often restricted to several discrete actions (e.g. user opening a web page) and the duration between them (the time spent on a web page). Gaze tracking, and physical behaviour tracking in general, helps to solve the knowledge-acquisition bottleneck by providing a rich-stream of interest data on a specific user.

InBeat.eu is a generic[3] web service for user tracking and preference learning. Its "SMART-TV use case" was first introduced at ACM RecSys'13 [3]. In this paper, we present an extended version of the system, which, in addition to explicit user actions such as clicking a button, performs physical behaviour tracking using Microsoft Kinect, a widely accessible commodity hardware.

2 System Walk-through

The user, Ivan, starts by opening the video player on a computer (see Fig. 1). The preloaded set of videos covers news from several domains. Ivan is a sportsperson and football fan. The news show opens with the politics topic. Ivan watches the program with mild interest, skipping most of the news. As the show gets to the sports, his attention increases. Ivan presses the bookmark button when his favourite footballer, Ronaldinho, appears on the screen. In a few moments, the commentator mentions a new football talent, Neymar. Since Ivan is curious about this player, he pauses the video and follows the auto-suggested link to Wikipedia. The sports news are followed by an advertisement spot for visiting Berlin. Since this topic is of no interest to Ivan, he picks his tablet and starts playing a game, which is recorded by the behaviour tracking module as "not looking at the screen".

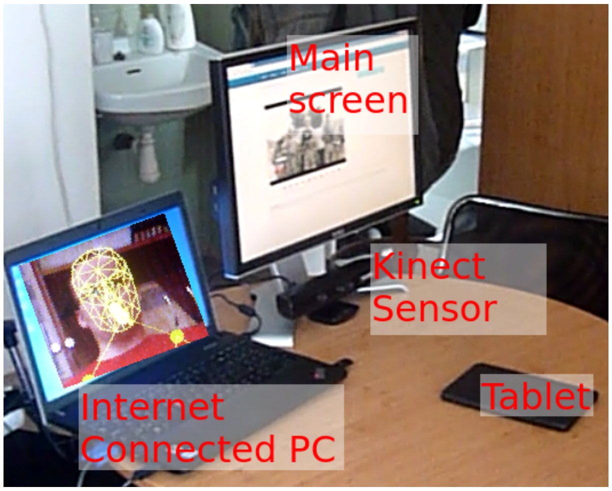

Figure 1. User-side setup

All explicit Ivan's actions as well as his gaze direction provide interest clues to the system. By the end of the news block, enough feedback has accumulated to learn rules such as "SoccerPlayer(yes) → interest(positive)", "Berlin(yes) ∧ Politics(yes) → interest(neutral)".[4] These rules are used to provide personalized recommendation of new videos from a shortlist.

3 Behind the Scenes

This section explains how the system arrived at the recommendations by showing the output of the InBeat components (ref. to Fig. 2).

Analyzing subtitles. InBeat retrieves subtitles from the videos and sends them via a web service for analysis to our entity recognition system entityclassifier.eu [1]. The results are returned within several seconds and contain a list of recognized entities for each subtitle. Entities are assigned a DBpedia URI and DBpedia Ontology Type (where available).

Example. Consider subtitle fragment: *Luiz Felipe Scolari wants to combine the experience of the former Barcelona, AC Milan and Paris Saint-Germain star with young talent like Neymar.* The underlined words are recognized as entities and disambiguated to DBpedia

[1] Faculty of Informatics and Statistics, University of Economics, Prague, Czech Republic, and Multimedia and Vision Research group, Queen Mary University, London, U.K.

[2] Faculty of Information Technology, Czech Technical University in Prague and Faculty of Informatics and Statistics, University of Economics, Prague, Czech Republic.

[3] InBeat was recently also used to provide links to related articles to website visitors at the News Recommender Challenge'13, where it obtained the runner-up award (https://sites.google.com/site/newsrec2013/challenge).

[4] It should be noted that thresholds for rule learning have been set to low values to allow for fast learning for demo purposes. A realistic setting would require more instances.

resources. Consequently, for the first entity (Luiz Felipe Scolari) the following types are retrieved: SoccerManager, Agent and Person.

Subtitles are also used to detect *pseudo-shots*, fragments of the video between the start and end offsets of a subtitle. Training instances are created on this granularity.

Capturing Interest Clues. A modified version of the YouTube player is used with button actions linked to the InBeat server. In this way, all user interactions with the player are recorded. Most information on user interest comes from the observation of physical user behaviour with Microsoft Kinect [4]. Kinect is connected directly to a PC, which sends the tracking data over the Internet to InBeat.eu.

Aggregating Content Descriptions. Additional structure over the result of entity classification is provided by the DBPedia ontology. Appearance of an entity in a subtitle activates one or more types in the ontology, the activation is spread up to the root class (see Table 1). Due to the small span of a pseudoshot, weights of concepts are binary: either the concept appears in the shot or not.

Aggregating Interest Clues. Multiple interest clues can be recorded for a specific pseudo-shot (duration of a subtitle). These are are all aggregated using a list of hand-coded rules to a single scalar value of interest. An example of such a rule is: *if user looks at the screen, the interest is increased by 0.3*.

The result of this process is a value for the *interest* target variable, which is appended to the pseudo shot's content vector, creating one complete training instance. The value of interest is discretized into three categories: *negative, neutral, positive*.

Exporting Data. The result is a matrix containing one instance (row) for each pseudo-shot. Columns correspond to classes (*dbo:*) and resources (*dbp:*) from the DBpedia ontology, with uninformative ones omitted, and names of recognized entities. The last column is the interest value (label). This output, exemplified on Table 1, is in a form which allows to directly most classifiers.

Identification			Description			
userId	sessionId	pseudo shotId	dbp: Neymar	dbo:Soc cerPlayer	...	Interest
user1	1124541	125	yes	yes	...	positive

Table 1. Example training instance

Learning Preference Rules. The matrix with labeled pseudo-shots is subject to association rule learning. This data mining technique was selected for its favourable scalability on large datasets [2].

For demo purposes, the rule learning process can be visualized in the EasyMiner (`easyminer.eu`) web interface.

Recommendation. The recommendation is performed on a shortlist of related objects (videos).[5] The videos are subject to the same type of content aggregation that is performed on pseudo-shots. For each unlabeled video, a rule engine identifies matching rules, and uses the strongest one to assign an interest value. Since the set of learnt association rules may not cover the entire instance space, some videos may not be ranked.

4 Contribution and Future work

We see the main contribution of the presented system in two areas:

- To the best of our knowledge, InBeat.eu is the first on-line recommender employing motion sensing with Microsoft Kinect.
- A wikifier is used to bridge the the difference between natural language descriptions of the labeled data and the items to recommend ("semantic gap") by representing the text with a feature vector containing related concepts retrieved from the Linked Data Cloud.

While the individual parts of the system have been subject to experimental evaluation, including the use of Microsoft Kinect to detect user attention changes in screen content [5], a comprehensive validation of our system remains for future work. Finally, it should be noted that this research hints at immediate commercial applications[6], the use of physical behaviour tracking of TV users by governments has also been foreseen [6]. A demo is available at `InBeat.eu`.

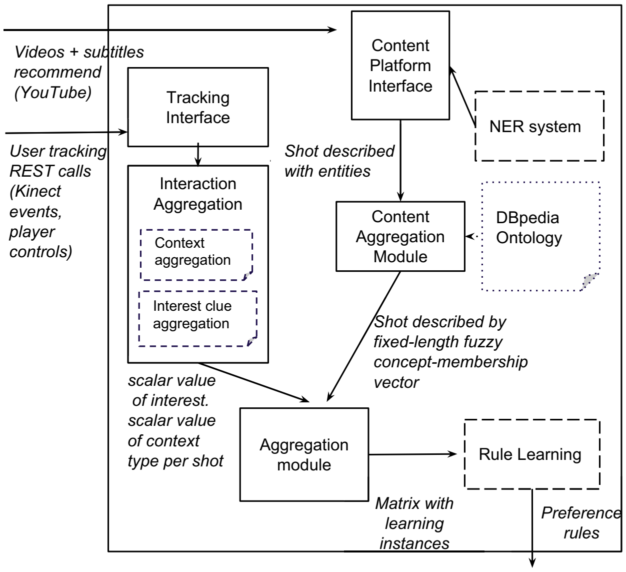
Videos + subtitles recommend (YouTube)
Content Platform Interface
Tracking Interface
NER system
User tracking REST calls (Kinect events, player controls)
Shot described with entities
Interaction Aggregation
Context aggregation
Interest clue aggregation
Content Aggregation Module
DBpedia Ontology
Shot described by fixed-length fuzzy concept-membership vector
scalar value of interest. scalar value of context type per shot
Aggregation module
Rule Learning
Matrix with learning instances
Preference rules

Figure 2. InBeat architecture

ACKNOWLEDGEMENTS

We would like to thank our project colleagues Julien Leroy, Matei Mancas and team from the University of Mons for providing the head pose estimation software. This work is supported by the European Union under grant no. FP7-ICT-2011-7 LinkedTV (Television Linked To The Web).

REFERENCES

[1] Milan Dojchinovski and Tomáš Kliegr, 'Entityclassifier.eu: real-time classification of entities in text with Wikipedia', in *ECML'13*, pp. 654–658, (2013). Springer.

[2] Trevor Hastie, Robert Tibshirani, and Jerome Friedman, *The Elements of Statistical Learning*, Springer Series in Statistics, Springer New York Inc., New York, NY, USA, 2001.

[3] Jaroslav Kuchař and Tomáš Kliegr, 'GAIN: Web service for user tracking and preference learning - a Smart TV use case', in *Proceedings of the 7th ACM Conference on Recommender Systems*, RecSys '13, pp. 467–468, New York, NY, USA, (2013). ACM.

[4] Julien Leroy, Francois Rocca, Matei Mancas, and Bernard Gosselin, '3d head pose estimation for TV setups', in *Intelligent Technologies for Interactive Entertainment*, volume 124, 55–64, Springer, (2013).

[5] Julien Leroy, Francois Rocca, Matei Mancas, Radhwan Ben Madhkour, Fabien c Grisard, Tomáš Kliegr, Jaroslav Kuchař, Jakub Vit, Ivan Pirner, and Petr Zimmermann, 'Innovative and creative developments in multimodal interaction systems', in *KINterestTV - Can we measure in a non-invasive way, the interest that a user has in front of his television displaying its content?*, eds., Yves Rybarczyk, Tiago Cardoso, and Joao Rosas. Springer, (2014).

[6] G. Orwell [Eric Arthur Blair], *1984*, Secker and Warburg, 1949.

[7] Michael J. Pazzani and Daniel Billsus, 'Content-based recommendation systems', in *The Adaptive Web*, LNCS, 325–341, Springer, (2007).

[8] A. Pradeep, R.T. Knight, and R. Gurumoorthy. Methods and apparatus for providing personalized media in video, June 11 2013. US Patent 8,464,288.

[9] Songhua Xu, Hao Jiang, and Francis C.M. Lau, 'Personalized online document, image and video recommendation via commodity eye-tracking', in *RecSys '08*, pp. 83–90, New York, NY, USA, (2008). ACM.

[5] Selection of these videos is out of the scope of the current demo, it can be performed for example using nearest neighbour algorithms.

[6] E.g. a recent patent [8] on personalized media during commercial breaks.