Using Social Media Data for Online Television Recommendation Services at RTÉ Ireland

Andrea Barraza-Urbina, Hugo Hromic,
Ioana Hulpus, Benjamin Heitmann, Conor Hayes
Insight Centre for Data Analytics
National University of Ireland
Ireland, Galway
firstname.lastname@insight-centre.org

Neal Cantle
Raidió Teilifís Éireann (RTÉ)
Donnybrook
Dublin, Ireland
firstname.lastname@rte.ie

ABSTRACT
Raidió Teilifís Éireann (RTÉ) is the public service television and radio broadcaster in Ireland. Through on demand video services, RTÉ allows their users to catch up on television broadcasts via the RTÉ Player. The company interacts with their users by means of social media platforms such as Twitter, Facebook and YouTube. Aiming to improve user engagement and in order to deal with the potential information overload caused by a broad catalogue such as RTÉ's, a recommendation service would be a desirable feature for the RTÉ Player. However, the distinctive requirements of this use case, such as (*a*) lack of ratings, (*b*) lack of user sessions data, and (*c*) limited lifespan of items (*i.e.*, available videos), keep us from applying traditional recommendation approaches. Yet, we believe that immersed within social media data, there is valuable information about user viewing preferences, and that this information can be used as input for Recommendation Services. This paper examines the use of social media data for personalization in the RTÉ industrial use case, giving special focus to the use of Twitter data for Recommendation Systems.

Categories and Subject Descriptors
H.3.3 [**Information Search and Retrieval**]: Information Filtering

Keywords
Recommender Systems, Television, Microblogging.

1. INTRODUCTION
In this paper, we propose Recommender System (*RS*) solutions which draw inspiration from transfer learning, to solve an industrial use case put forward by Raidió Teilifís Éireann (RTÉ). What sets our approach apart, is that we rely solely on preferences inferred from social media for generating recommendations.

RTÉ is the national provider of public-service television (TV) and radio in Ireland. The company makes its TV content available online through the RTÉ Player, which is Ireland's most popular broadcaster video-on-demand service [11]. The Player allows users to access live video content as well as past TV shows which are made available online for a limited amount of time. To further relate with their audience, RTÉ also publishes content using social media channels such as Twitter, Facebook and YouTube. On average, there are more than 500 hours of programming available on the RTÉ Player at any given time. To better support users to discover interesting content, a *Recommender System* is a desirable feature, as it would help users overcome possible information overload and potentially boost user engagement.

Nevertheless, the distinctive requirements of this use case highly affect the suitability of traditional *RS* approaches. An important requirement is to minimize the reliance on personal user data, due to RTÉ's privacy concerns of sharing data with external organizations. Furthermore, the current version of the RTÉ Player does not support user ratings. As such, our aim is to provide recommendations to users, without accessing their RTÉ Player accounts. In this fashion, RTÉ could also offer recommendations to users who have decided not to register. Consequently, our use case is defined by three constraints: (*a*) lack of ratings, (*b*) lack of user sessions, and (*c*) limited lifespan of items.

For a solution, we draw inspiration from *transfer learning* [10] approaches that deal with the application of a model learnt in a source domain, to draw conclusions about a separate target domain. First, we observe RTÉ has a prominent presence in *social media*, which provides us a key opportunity in using data immersed within social media to better understand user show preferences. The analysis of social media can give us great insights into how user communities are formed around TV shows, the impact of the TV shows, the sentiment(s) expressed with respect to the programmes, as well as insights into their content, and ultimately provide users with valuable recommendations. Furthermore, we intend to make use of the vast amount of knowledge published as Linked Data in order to understand the content of TV programmes. Lastly, we also propose the integration with sentiment analysis solutions, that help us imply "ratings" of programmes, by assessing the sentiment of user posts.

In this paper, we propose *RS* solutions in which an important source of knowledge is the content produced in social media. We believe that solutions found in this project can be applied far beyond our particular use case. The project is at an early stage, and as a first step, we focus on data from Twitter's microblogging platform to offer programme recommendations. *The paper is structured as follows*: First, we describe the RTÉ use case. Next, we present the challenges faced when designing a *RS* for the RTÉ Player service, along with the research goal targeted by the project. Subsequently, we outline our approach, which places retrieved data in a graph data model to serve as the basis for the proposed *RS* solutions. Next, we describe preliminary experiments where we (*i*) capture relevant RTÉ-related Twitter data, and (*ii*) evaluate the quality of captured Twitter data for *RS* purposes. Subsequently, we present related work. Finally, conclusion and future work are presented.

2. RTÉ Use Case
In this section, we describe the RTÉ use case. On the RTÉ Player service, RTÉ offers TV shows from both national and international productions. In addition, live TV broadcasts can be found for RTÉ managed channels: RTÉ One, RTÉ Two,

RTÉ News Now, and RTÉ Jr (for children). The Player offers different types of *programmes*, those with: (*a*) *dependent episodes*: require previous knowledge of the plot, or (*b*) *independent episodes*: do not follow a story line; *e.g.*, sport events. Each programme belongs to one of ten categories defined by RTÉ (*e.g.*, Entertainment, Drama). Other data includes: if the programme contains *mature content* and the *associated website* where programme-specific material is published. We refer to each published video as a specific *episode*, which belongs to a programme. Each episode is only available online for a limited time period after it is aired on live TV: from seven to thirty days. Lastly, programmes have to comply with geographical copyright restrictions, meaning some shows are only available for Ireland.

Within the RTÉ Player, users can browse episode content by name, category or time (latest shows). Furthermore, RTÉ offers a list of "*Most Popular*" shows, generated according to user viewing statistics within the service. When viewing a specific episode, users are offered a list of "*More from the same Category*", as a very basic form of episode recommendation.

A user may register to the RTÉ Player, either with an email address or a Facebook account. Each registered user has a "*Recently Watched*" list of episodes and a "*Your Favourites*" list. If linked with the Facebook account, users can easily express their views on *RTÉ* content using buttons such as "*Watching it"*, "*Like*" and "*Share*". Also, the Player has a "*Tweet*" button that allows users to publish in Twitter about a given episode. As can be seen, RTÉ highly encourages users to discuss and share in social media their opinions and preferences concerning programmes.

Finally, we highlight that RTÉ also offers content outside the RTÉ Player page. Besides possibly having their own website, programmes could also have: a Facebook page (*e.g.*, to announce related news), Twitter user accounts (*e.g.*, to promote events and gossip) and YouTube video channels (*e.g.*, posting trailers).

In summary, the RTÉ Player offers a representative case study opportunity, not only for the development of research related to online TV services, but specifically for *RS* that can use social media as a valuable knowledge source. It is important to emphasize that this project does not have access to private user account information due to user confidentiality purposes. However, we do retrieve publicly available data from the RTÉ Player site with a *web crawler*, which runs recurrently to extract updated programme and episode data.

In the following section, we describe the challenges faced when designing a *RS* for RTÉ and formalize our research goal.

3. Challenges and Research Goal

The RTÉ use case presents unique challenges that must be carefully considered when designing a *RS* solution, as these do not allow for the majority of current traditional *RS* approaches to be directly applied. These challenges are: *(a) Lack of personal preference data such as ratings:* Due to intellectual property and privacy issues, we do not have access to potential preference data found in user accounts (*e.g.*, user's favourites list). Furthermore, most users do not register to the RTÉ Player, resulting in little availability of user data anyway. Consequently, we have decided that for our first approach, users will be considered anonymous and only the user's current session is known. *(b) Lack of historical user session information:* Because users are considered anonymous, past viewing history is not known. This information would be useful in order to make accurate "*dependent episode*" recommendations. As a result, for our approach, the *RS* will offer **programme recommendation** by offering to the user the last episode available for the programme. (*c*) *Dynamic inventory and limited life span of recommendable items:* Programmes have a limited availability, new programmes and episodes are constantly added, while older are removed after expiration (view *section 2*). As a result, our system must use information about past episodes, to recommend programmes that are still available online.

Even though we do not have explicit user-programme preferences, we believe users express their opinions on programmes using social media. Thus, we define the following ***Research Goal***:

Design a RS that can offer ***programme suggestions*** *to an* ***anonymous user****, based solely on information about the user's* ***current session*** *and* ***inferred preferences of other RTÉ users*** *extracted* ***from social media****.*

4. Our Approach

In our work, we focus on translating the activities of users in social media with respect to RTÉ content, into knowledge about their tastes and preferences. Our most important assumptions are (*i*) that users' activities in social media with respect TV programmes are correlated with their appreciation of the programmes, and (*ii*) that social media provides sufficient content about our targeted TV programmes for us to be able to distill the main trends. The two assumptions are crucial for successfully applying the user-programme models learnt from social media, to the context of the RTÉ Player *RS*.

While our approach is theoretically applicable to any social media, in the following we are focusing on Twitter. RTÉ has a rich presence in the platform, with over sixty official registered user accounts. Furthermore, Twitter is used extensively in Ireland, with an average of one million daily tweets in April 2013 [12].

We connected Twitter data to the programme data collected by the RTÉ site crawler, all in a heterogeneous graph data model. This data model served as the basis for *RS* solutions proposed in *section 4.2*. In the following section, we describe the initial design of the graph data model for this project.

4.1 RTÉ Heterogeneous Graph Data Model

Our approach relies on the integration between three types of data: social media content, RTÉ content and Linked Data. For seamless integration, flexibility and scalability we propose a graph-based data model.

The initial design of the graph model can be viewed in *Figure 1*. In the graph, the User nodes represent users that have posted content in social media (*i.e.*, Twitter) related to RTÉ programmes. Each User is connected to the Tweets they have posted, and each Tweet can have attached Links or Media nodes. As a property of the tweet node, we add its associated sentiment from an external detection algorithm [13]. In reference to programme data, we have Episode nodes connected to Programme nodes. Most importantly, Tweet nodes are connected to a Programme or Episode node.

In *Figure 1*, the dashed arrows represent connections that associate the three types of data. These connections are automatically created. Currently, for the annotation of the Tweets with Programmes, we use a manually curated set of hashtags as described in *section 5.1*. Episodes can further be inferred by analysing the date and time of the tweet as well as its content. Named entity recognition, linking and disambiguation can be used for linking Tweet content to mentioned DBpedia concepts. Furthermore, most RTÉ programmes have their corresponding

DBpedia URI. The obtained knowledge graph can be further enriched with links to the IMDb pages of programmes, or other knowledge bases that are more specific to our use case, for instance Irish politics and government data.

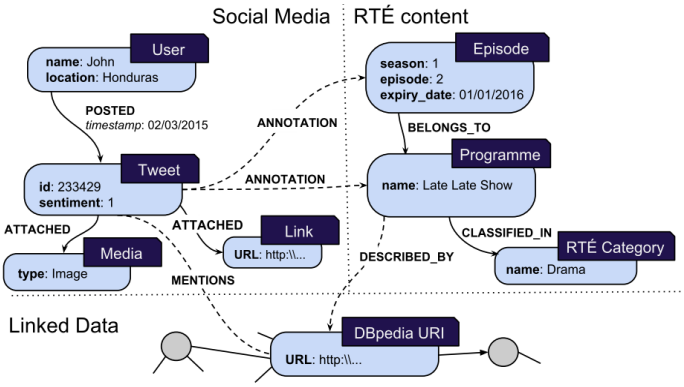

Figure 1. RTÉ Heterogeneous Graph Data Model

All connections ensure that we obtain a rich knowledge graph that can be used to provide recommendations as we describe in the following section.

4.2 Generating Recommendations

In this section, we outline *RS* approaches based on the proposed graph data model. The approaches receive as input the episode/programme the user is currently watching.

4.2.1 Collaborative Filtering Recommendation

In order to generate recommendations, we can obtain a User-Item matrix by using the information found in our knowledge graph. In this matrix, there would be a score or rating to determine how much a user likes/prefers a given programme. We are in the process of designing a strategy that can determine such a score, given the number of tweets the user posts about a programme and the sentiments associated to these tweets.

Next, if we assume that the target user likes the programmes he/she has been watching in the current session, we could potentially generate recommendations using a conventional Collaborative Filtering (CF) technique. Depending on the programme the user is watching and characteristics of the underlying data, we can choose to use a user-based CF or item-based CF approach. *Figure 2*, illustrates the main intuition behind our CF approach.

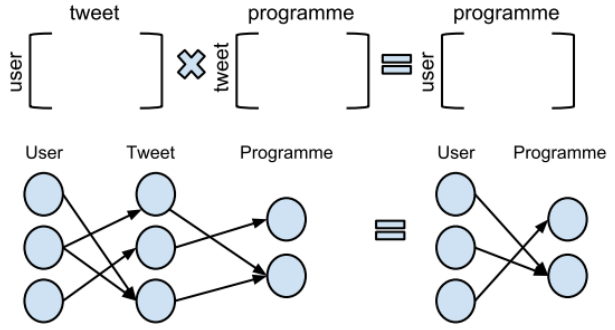

Figure 2. Using tweets to obtain a User-Programme matrix for CF

In *Figure 2*, a bi-partite graph of users and programmes is also shown. This graph could be used by several *RS* approaches described in [4], such as random walk similarity using ItemRank.

Furthermore, we propose a community-based *RS* approach. Given the Twitter data, communities can be found by analysing the interactions among users (*i.e.*, user-retweet-user, user-mention-user and user-reply-user) [8]. By identifying communities formed around RTÉ Twitter content, we can find potential patterns of programmes co-occurring across such communities. Co-occurring programmes can be used to generate recommendations.

A possible limitation of a CF approach is data sparsity. Specifically in our scenario, this would be the case if any user does not post about more than one programme in Twitter. This problem can be tackled by using a content-based or a hybrid approach. Given our data model, a straightforward content-based solution can be devised by modelling programmes based on their RTÉ description, category, or DBpedia Link. However, such an approach would be agnostic of the social media content. Therefore we prefer to integrate content-based aspects with aspects from CF, into a hybrid approach as follows.

4.2.2 Hybrid Recommendation

In this section, we propose a novel hybrid *RS* approach that can better exploit the graph-based nature of the data presented in s*ection 4.1*. The main intuition is to use graph algorithms over the Linked Data in order to assess semantic relatedness between tweets and ultimately between programmes. The idea is illustrated in *Figure 3*.

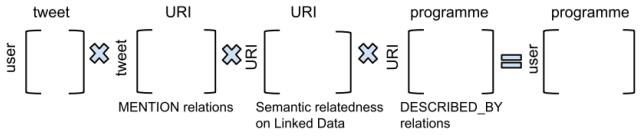

Figure 3. Using semantic relatedness over Linked Data for RS

In *Figure 3*, we show the matrices involved in the computation of the associations between users and programmes, in our proposed hybrid *RS*. The MENTION and DESCRIBED_BY relations are the ones illustrated in *Figure 1*. The key to this approach is the URI-URI matrix that captures semantic relatedness between URI's. For instance, given a tweet about *Ryan Tubridy*, the computation of semantic relatedness identifies films transmitted by RTÉ in which Ryan Tubridy plays. A novel method for computing the URI-URI associations is Heitmann *et al.*'s approach [8], a form of spreading activation to provide content-based recommendations, where recommended items are represented by URI's.

5. Preliminary Experiments

Currently, the approaches proposed in the previous section are work-in-progress. In this section, we present our approach on capturing Twitter data and a preliminary experiment to evaluate the quality of Twitter data for a CF solution approach.

5.1 Capturing Twitter Data

We use both the APIs provided by Twitter for accessing its data: (*a*) the REST API: retrieves data matching search criteria, and (*b*) the Streaming API: intended for real-time capturing of tweets. In both APIs, Twitter imposes request restrictions with the purpose of mitigating potential abuse of their platform. On the one hand, for the Streaming API, we use the approach proposed in [8]. This approach consists on defining a set of *seed* terms (hashtags, users to follow and keywords) to initially listen to, and then *extending* seed terms dynamically. On the other hand, for the

REST API, we decided to capture tweets using a curated list of terms in the form of Twitter Search Queries.

With respect to linking Tweets to the programmes/episodes they refer to, an important challenge is the low quality in the text of Twitter posts that strongly affects the linking performance. While we are researching for accurate approaches to this problem, we are listening and searching Twitter using a curated list of programme-specific hashtags, users and keywords (*i.e.*, seed terms). This way, we can directly annotate tweets with their associated programmes.

5.2 Analysis of Twitter Data Quality for RS

In this section, we offer an overview of preliminary tests carried out to assess the potential use of Twitter data for a CF approach. A requirement entailed by our CF approach (*section 4.2.1*) is that most users tweet about more than one program, and that we could potentially build a user-programme matrix for a CF approach that does not have significant data sparsity problems.

For this, we manually constructed a curated list of terms for 64 RTÉ programmes (39 national and 25 international). We queried the REST API search endpoint and collected their corresponding Tweets in the previous 10 days. The 10 days limit is imposed by Twitter to retrospective search. We captured 89,100 tweets posted by 38,901 users.

After accumulating these tweets, we computed the number of common users between all pairs of programmes, *i.e.*, users that had at least one tweet about both programmes. We found that only 1,373 users tweeted about more than one programme (3.5% from retrieved users). We found 46 programmes had at least one shared user with another programme. However, most of the programme pairs had less than ten overlapping users. The maximum user overlap found was of 145 users between the programmes: RTÉ News: Six One and The Sunday Game. These users generated 888 tweets commenting on these two programmes.

These experiments show that although there are definitely some programmes that heavily co-occur, these pairs of programmes might be quite rare. Nevertheless, the scale of our experiment is rather small, and more similar experiments, ran over tweets collected over many months are needed before we can compute statistically significant results.

6. Related Work

Most approaches that use Twitter data in order to provide content recommendations [1][3] require a rich user profile. These approaches are therefore not applicable to our problem where the target users have neither a Twitter profile [1], nor a long-term interest profile in the target domain [3]. Another class of related approaches explore the use of Twitter simultaneously with TV watching [9][6] in order to understand Twitter's role and usage by TV fans. A similar work performs social media analysis in relation to TV watching, [7] captures real-time behavior of users while they watch their favourite shows. While the purpose of these works is not focused on *RS*, many of their analysis and experiments are valuable for understanding the correlations between Twitter and other social media usage and TV viewing.

TV programme recommendations have been also tackled, but most approaches do not consider any input from social media. A representative such approach is proposed in [2].

There is broad interest from the research community on the TV, social media and RS contexts; however, the studied solutions are still not sufficient for our described use case within RTÉ.

7. Conclusion and Future Work

The RTÉ project presents a representative use case to explore the potential use of microblogging data to enhance Recommendation Systems. Due to the requirements of this use case we have to treat all users of the RTÉ Player as anonymous.

In this paper, we proposed novel recommendation approaches that learn the user-item models from Twitter, and apply them to the context of an online TV Player. To this end, we designed collaborative filtering as well as hybrid approaches, and showed how Linked Data can potentially be used to further enrich the knowledge extracted from social media.

We plan to evaluate our proposed *RS* solutions with user testing, both in a controlled environment and on-line. The purpose of the *RS* service is to increase user retention and increase user engagement with content on the RTÉ Player. Promising metrics to measure user engagement are the length of user sessions, the number of recurrent users, the number of videos watched by a user and the number of tweets sent via the RTÉ Player. Evaluations must be first carried out in a controlled environment with a sample of user volunteers. Later, A/B testing could be used to temporarily test a *RS* approach on the live service.

Acknowledgements: This publication has emanated from research sup- ported in part by a research grant from Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289 (Insight).

8. REFERENCES

[1] Abel, F., Gao, Q., Houben, G.-J., & Tao, K. (2011). Analyzing User Modeling on Twitter for Personalized News Recommendations. Proc. of the 19th Int. Conf. on UMAP.

[2] Bambini, R., Cremonesi, P., & Turrin, R. (2011). A recommender system for an IPTV service provider: a real large-scale production environment. In Recommender systems handbook (pp. 299-331). Springer US.

[3] Deng, Z., Yan, M., Sang, J., & Xu, C. (2015). Twitter is Faster: Personalized Time-Aware Video Recommendation from Twitter to YouTube. ACM Trans. Multimedia Comput. Commun. Appl.

[4] Desrosiers, C., Karypis G. (2011). A Comprehensive Survey of Neighborhood-based Recommendation Methods. Recommender Systems Handbook. Springer US.

[5] Heitmann, B. (2014). An Open Framework for Multi-source, Cross-domain Personalisation with Semantic Interest Graphs. PhD dissertation. NUI Galway, Ireland.

[6] Highfield, T., Harrington, S., & Bruns, A. (2013). Twitter as a technology for audiencing and fandom: The# Eurovision phenomenon. Information, Communication & Society.

[7] Holanda, P., Guilherme, B., da Silva, A. P. C., & Goussevskaia, O. (2015). TV Goes Social: Characterizing User Interaction in an Online Social Network for TV Fans. In Engineering the Web in the Big Data Era. pp. 182-199.

[8] Hromic, H., Karnstedt, M., *et al.* (2012). Event panning in a stream of big data. In LWA Workshop on KDML.

[9] Narasimhan, N., & Vasudevan, V. (2012). Descrambling the social TV echo chamber. In Proc. of the 1st ACM workshop on Mobile systems for computational social science.

[10] Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. Knowledge and Data Engineering, IEEE Transactions.

[11] RTÉ Player Service. Retrieved June 2015, from: http://www.rte.ie/player/ie/help/

[12] Social Media Statistics Ireland. Retrieved June 2015, from: http://tinyurl.com/njz5ybh

[13] Thelwall, M., Buckley, K. (2013). Topic-based Sentiment Analysis for the Social Web: The role of mood and issue-related words. Journal of the American Society for Information Science and Technology.