Countering Contextual Bias in TV Watching Behavior: Introducing Social Trend as External Contextual Factor in TV Recommenders

Felix Lorenz[1], **Jing Yuan**[2], **Andreas Lommatzsch**[2], **Mu Mu**[3],
Nicholas Race[4], **Frank Hopfgartner**[5], **Sahin Albayrak**[2]
[1]Technische Universität Berlin, Berlin, Germany, f.lorenz@campus.tu-berlin.de
[2]TU Berlin – DAI-Lab, Berlin, Germany, {firstname.lastname}@dai-labor.de
[3]The University of Northampton, Northampton, United Kingdom, Mu.Mu@northampton.ac.uk
[4]School of Computing & Communications, Lancaster University, Lancaster, UK, n.race@lancaster.ac.uk
[5]University of Glasgow, Glasgow, UK, frank.hopfgartner@glasgow.ac.uk

ABSTRACT

Context-awareness has become a critical factor in improving the predictions of user interest in modern online TV recommendation systems. In addition to individual user preferences, existing context-aware approaches such as tensor factorization incorporate system-level contextual bias to increase predicting accuracy. We analyzed a user interaction dataset from a WebTV platform, and identified that such contextual bias creates a skewed selection of recommended programs which ultimately locks users in a filter bubble. To address this issue, we introduce the Twitter social stream as a source of external context to extend the choice with items related to social media events. We apply two trend indicators, Trend Momentum and SigniScore, to the Twitter histories of relevant programs. The evaluation reveals that Trend Momentum outperforms SigniScore and signalizes 96% of all peaks ahead of time regarding the selected candidate program titles.

ACM Classification Keywords

H.5.1 Information Interfaces and Presentation (e.g., HCI): Multimedia Information Systems; H.3.3 Information Storage and Retrieval: information filtering, relevance feedback; H.2.8 Database Applications: Data mining

Author Keywords

privacy reserving recommender; video on demand; user experience; context-aware applications; trend detection

INTRODUCTION

With the rapid growth of Internet connectivity, movies and videos are increasingly available online. Whilst streaming services continuously expand their market share, channel-based linear TV also remains very popular [17]. IP Television (IPTV) providers therefore typically offer their customers access to both Live and VoD content. Recommenders in such TV systems play a key role in unburdening users from the choice of whether to watch linear TV channels or browsing through a confusingly vast catalog of recorded TV programs [28]. TV program recommenders have their own unique characteristics when compared with other recommending scenarios. Usually, TV broadcasters try to incorporate their assumptions about viewers preferences and habits in their schedule to achieve high viewing figures. For example, news and weather reports are mostly arranged in the morning while dramas and sports-related content is generally scheduled in the evening or during the weekend. Thus when we analyze users' viewing behavior in a linear TV scenario, the patterns we observe reflect the arrangement of TV schedules to some degree. This characteristic often limits the benefit of a traditional Context-aware Recommender (CAR) for TV content since the observed user preferences are highly biased by the broadcasted TV programs.

A CAR usually defines a system's internally traceable auxiliary information (e.g. "time of day", "day of week", "location") as contextual factors [9]. Using logged clickstream data with detailed playback statistics, we analyze users' viewing habits in the "Vision" system, a production-level live and on-demand streaming service used by a large Living Lab user community at Lancaster University, UK. We find a significant portion of the users' consumption habits with respect to both live linear and on-demand TV content to be attributed to *contextual bias*. It is not surprising that recommendation algorithms that reliably predict items according to this *filter bubble* bias show high performance in laboratory tests. However, the resulting recommendations can be monotonous and repetitive to the end users. In order to escape from this self-fulfilling prophecy, additional external information is required. In this paper, we propose using TV domain relevant late-breaking events detected from social media as an external contextual factor to provide more diverse and precise recommendations.

Incorporating social media and crowd-sourced supplemental attributes has become an increasingly popular approach in recent recommender designs [41, 16, 27]. However, most social aware recommenders consider such factors only from the

TVX '17, June 14-16, 2017, Hilversum, Netherlands

ISBN 978-1-4503-4529-3/17/06. . . $15.00
DOI: http://dx.doi.org/10.1145/3077548.3077552

perspective of analyzing large-scale social graphs constructed from user-item relations. Though methods known from time series analysis such as "anomaly detection" and "trending topics analysis" are widely used for use cases like earthquake detection [35], breaking news event detection [29], and reputation monitoring [39], they havn't been attemped in social source analysis in recommender area to our best knowledge. There is a distinct lack of research investigating the utility of hot events or trends in social media as a recommendation factor. Our work make up for this vacancy by introducing trend analysis of domain related social media streams as an external contextual factor in a TV recommendation scenario. At the same time, since no internal information is needed prior to producing recommendations, the *cold start* problem with new users is addresses by our approach.

The introduction of social media trends as an external contextual factor for TV recommenders involves a number of challenges. For example, there is no concise statistical model of a trend, though some work summarizes frequently used features [33] and creates taxonomy of different trend types [4]. In addition, the relationship between the events in the social domain (e.g. Twitter stream) and the Online TV domain (e.g. users' content browsing and playback activities) is not well understood. To address these challenges, the present paper makes two main contributions: 1) we provide quantifications for the measurements of trend in both TV user watching data streams and Twitter social media text streams; and 2) we show that the trends detected in the Twitter social stream highly cover peaks observed in the user data (but not necessarily vice versa), proving it's not contrary to user behavior and applicable in a TV platform. Specifically, we use two indicators known from stock market analysis – Trend Momentum [29] and SigniScore [37] to capture the hotness of specific TV programs in the "Vision" user data via their related Twitter timelines, which we obtained by crawling the results from a keyword based Twitter search. Through investigation of the two parallel datasets, we identify the best hyper parameters for both trend measures through a grid search. The quantified evaluation of Omission rate (OR) and time difference between the trend indication and consumption peaks demonstrates that: 1) Trend Momentum (TM) can better predict points of high user demand in IPTV services; 2) OR can serve as a criterion for filtering out the programs which do not lend themselves well to the proposed approach.

The remainder of the paper is structured as follows. Sect. 2 summarizes related work on recommenders in TV services and social trend analysis. Contextual bias within online TV dataset is analyzed in Sect. 3, while the definition of external context is introduced in Sect. 4. Our approach towards TV domain trend analysis in Twitter streams combined with user activity data is presented in Sect. 5. Finally, Sect. 6 evaluates the trend measures and a conclusion and an outlook on future work is given in Sect. 7.

RELATED WORK

In this section, we review the literature with respect to three aspects: Recommender systems for IPTV services, the use of contextual factors in recommenders, and trend analysis in social networks.

Recommender System in IPTV Services

Recommender systems have been broadly applied by IPTV providers to increase users' Quality of Experience (QoE) when they watch TV programs over the internet [5]. The classical recommendation strategy Collaborative Filtering (CF) essentially clusters users based on their choice of content in the past [40] and infers potentially interesting items using similarity between clusters [21]. Unfortunately neither item-based CF nor user-based CF addresses the *cold start* problem, which arises when a new user or item is added to the system without any prior information on usage. A similar issue surfaces for outlier content (i.e. *gray sheep* problem) and is highly undesirable in the TV domain [15]. Solutions include using demographic user information obtained from user profiles or a linked social media account to generate group-based recommendations for new users [7]. However, profile information creates new concerns about user privacy, many systems thus deploy hybrid algorithms that use Content-based Filtering (CB) to compensate CF drawbacks.

In the TV domain, Electronic Program Guide (EPG) data serves as a standardized source for CB models. The similarity between programs is often calculated through semantic analysis of their EPG descriptions [32, 6, 11]. Extensions via online databases like Internet Movie Database (IMDb) and DBpedia[1] [19] can help to enhance the descriptiveness of representations. Given the availability of meta-data, a wide range of well-known techniques from the field of information retrieval and extraction are then employed to compute similarity scores in textual feature space [2]. Nevertheless, CB recommenders tend to over-specialize and constrain their recommendations to a *filter bubble* of similar items [6]. As a result, recommenders often fail to adapt to new trends and changes in user preferences.

Contextual Factors in Recommenders

Abreu et al. conducted a survey of TV viewer behaviors, according to which more than half of the relevant determinants for program selection depend on the situational context of the user [1]. Furthermore, 60% of respondents state that the presence of company and the available time are important contextual factors to select a program. Context-aware recommender algorithms, e.g. based on tensor factorization [24, 23] incorporate time [26, 45, 9] or location [14, 28] as additional parameters to encode contextual information. In addition, some advocate to model the local social environment [43, 30] and propose strategies to improve recommendations in households with multiple users sharing a single device [38].

Although existing context-aware approaches take influential factors into consideration to improve diversity of recommendations, contextual bias still exhibits significant influence even in state of the art model encodings [34]. The problem is mainly attributed to the restriction of the training set to system-level internal user behavior data. In the TV domain such contextual bias from user behavior is more prevalent, because programs

[1] http://wiki.dbpedia.org last accessed Oct. 23, 2016

are intentionally arranged by TV stations to match temporal preferences and target audiences. To address this issue, we extend the notion of context to domain relevant trends detected in online social media streams like Twitter, such that the selection presented to a user is not restricted to content that is biased towards the usual choices at the specific moment in time.

Social Trend Analysis

Defining "hot" events as trends in the external context of social media creates the need for techniques such as trend detection and prediction, which have been successfully applied to social streams to identify trending topics [18, 31], political opinions [42, 39], and news stories [46, 20]. Directly modeling the time series with keyword co-occurrence [36] is a typical approach to detect and observe trending conditions. Others apply clustering methods [3] and incorporate user authority information to improve detection rate [10]. Qualitative examinations by Asur et al. [4] reveal typical emergence and decay patterns of twitter trends that can be exploited at model creation time, e.g. when selecting sensible ranges for hyper parameter search. Since 2015, Twitter provides an API endpoint for locally sensitive trending topics[2] [22]. To quantitatively define the trend indication, Lu and Yang introduce TM, a smoothed version of Moving Average Convergence-Divergence (MACD) and use a threshold crossing point as trend indication signal [29]. Meanwhile, Schubert et al. [37] propose a *z-score* based on Exponentially Weighted Moving Average (EWMA) and Exponentially Weighted Moving Average Variance (EWMAVar) as a quantitative measure of trendiness. However, to our knowledge, there is currently no research adopting mentioned methods for the purpose of TV program recommendation.

CONTEXTUAL BIAS

In this section, we introduce the concept of *contextual bias* to the evaluation of recommender systems. We intuitively explain its influence on user behavior using a concrete scenario, namely the IPTV service "Vision", operated by Lancaster University.

The computation of recommendations is usually seen as a data mining or machine learning task. Thus a specific evaluation protocol and adequate metrics (like *precision* and *recall*) are used to determine how well the predicted results fit with the ground truth of static datasets (c.f. [12, 25]). Normally, the more accurate the algorithm predicts missing values, the better it is considered to be. Recommender algorithms taking into account contextual factors in their models often outperform "normal" approaches. This has been observed in some recent studies where time is used to track the evolution of user preferences [44] and to identify periodicity in user behavior [8]. However, once bias exists in the dataset, its fingerprint is visible in the statistical patterns allowing a targeted improvement of recommendation quality. To the present date there has not been extensive research on the impact of this bias information. We analyze the phenomenon that users become increasingly trapped in a limited selection of items and receive very few new recommendations due to *contextual bias*.

In IPTV services, TV programs are often scheduled for a long duration of one to several hours. Therefore, temporal bias of users' choices is more prevalent than in other services such as YouTube. We use users' behaviors as captured by "Vision" to quantitatively analyze *contextual bias* in detail. In "Vision", TV programs are broadcasted as live TV and also recorded by a cloud-based service for on-demand retrieval beyond their original broadcast time. Users can decide themselves how and when they retrieve arbitrary content. The dataset from "Vision" contains fine details of user interactions with the service including clickstream data and playback statistics (e.g., play, pause, resume, etc). The dataset also encapsulates EPG data of 106,710 programs videos from a set of 12,809 unique titles over 23 genres and 62 channels. Over a period of 26 months, 2241 users made 204,920 requests to the site that are labeled "live" (142,011) or "vod" (62,909) depending on how the request was served. The cumulative playback duration over all requests was about 90,000 hours, about a third of which were streamed via Video-On-Demand (VOD). The majority of the programs are provided with genre specification via EPG and a third party TV catalog service. Exploration of this dataset reveals typical "time of day" *contextual bias* in the selection of channels, genres and programs in both Live and VOD mode.

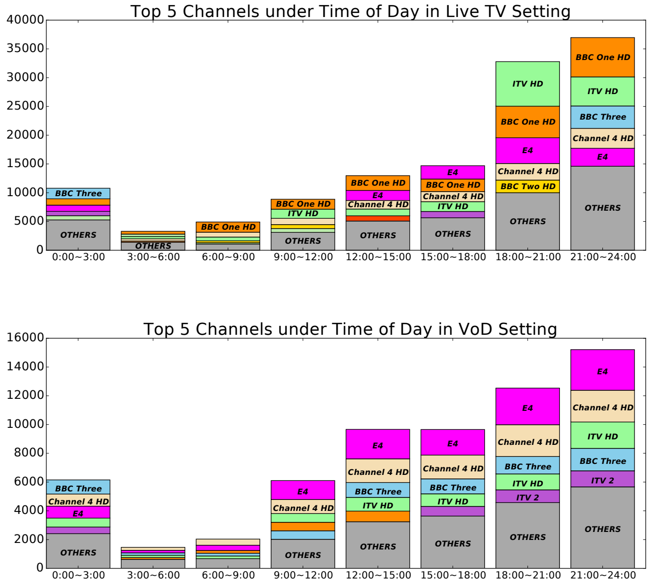

Figure 1: Frequency distribution on channels per "time of the day" for VOD and live.

For visualization, we split a day by grouping every three hours as a time segment, such that 8 time segments span a whole day. Figure 1 shows the ranked frequency statistics of channels (channels ranked after 5 are colored gray) in different time segments. From such ranked channels under both live (upper sub-figure) and VOD (bottom sub-figure) environment, we can see that within each time segment, the top 5 out of 62 distinct channels always hold more than half of the watching consumption. Though the channels consumed in a live environment are more diverse than under VOD, five most dominant channels in each categorical time bin are already sufficient to make a

[2] https://dev.twitter.com/rest/reference/get/trends/place Oct. 26, 2016

satisfactory suggestion for over half of the users of Live TV. Regarding the VOD environment, the pattern is more obvious that the rank of the top 5 channels is consistently *E4*, *Channel 4 HD*, *BBC Three*, *ITV HD*, and *ITV2* over all time segments. The observation is a typical example where relying on the statistical property of temporal bias can serve as a "good" recommendation strategy for both linear Live TV and a VOD yet at the expense of sacrificing diversity.

In Figure 2 we present a bar chart for the viewing frequency statistics on genres with respect to "time of day" in both Live and VOD consumption mode. Similar to our observations regarding the channel popularity distribution, the top 5 genres (*Sitcom*, *Drama*, *Entertainment*, *Soap*, *Comedy*) rank stably under VOD in each time segment, while such a distribution is slightly more diverse in the Live setting.

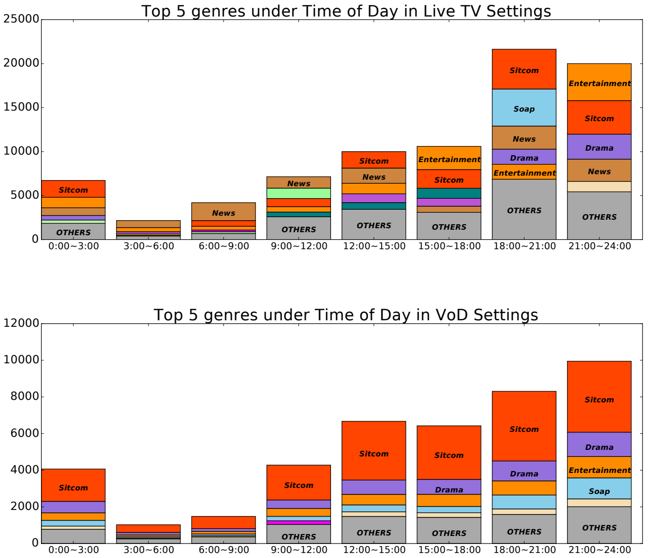

Figure 2: Frequency distribution on genres per "time of the day" for VOD and live.

In Figure 3, the frequency of viewing of the top 30 most retrieved programs is plotted (black dots) for each time segment in both Live and VOD settings. The most viewed program title and its proportion in this time category is also filled in a box in the place near its own solid circle. In addition, the percentage of the top 30 programs in each time segment is also written as text in middle area of every sub-figure. The frequency distribution of programs in each time segment depicts power law distribution. Out of 12,809 unique program titles, the proportion occupied by the top 30 programs in each temporal category ranges from 30% (21:00 - 24:00 in live settings) to 56.6% (3:00 - 6:00 in live settings), with an outlier of 70.18% (6:00 - 9:00 in live settings). It appears that a very small proportion (0.23%) of the programs attract around 30% to 50% of user playback requests within every temporal category. This proves that *contextual bias* exists not only in high-level user selections (such as channels and genres) but also at the individual program-level.

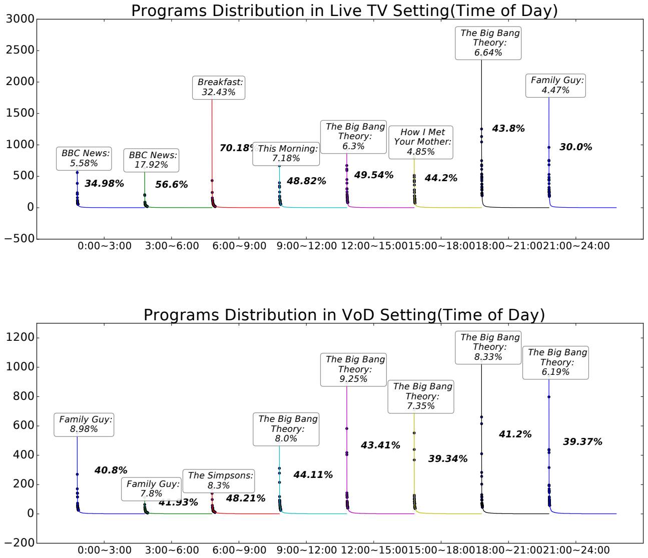

Figure 3: Frequency distribution on programs per "time of the day" for VOD and live.

Figure 4 in the next page displays frequency statistics concerning the time lag between airing and subsequent VOD request for four popular programs over the whole time-span in the dataset. We observe distinctive user behaviors on two types of programs. First, programs like *Gogglebox* and *Match of the Day* receive periodic attractions over a long period of time after their initial broadcast. This reflects the stable "time of day" regularity, which we identified in Figure 3. For this type of program, it makes sense to break its temporal bias and also recommend content at special moments determined by external factors. Conversely, some programs such as *This Morning* and *Coronation Street* can be grouped as the second type, which receive most of their on-demand playback requests within the first few hours after broadcast. Their popularity then suddenly drops to a minimal level, and no obvious periodic pattern can be caught. For these programs, only relying on system internal record information limits the possibility of their appearance in the recommendation list, thus introducing an external source thereby creates new opportunities for them to come back to users' sight.

Our exploration of the user interaction data indicates some strong contextual bias, which suppresses the diversity of recommendations in IPTV services. We thus introduce TV domain related trends in social media as an external contextual factor to help the users escape from the unfavorable filter bubble.

EXTERNAL CONTEXT

To address the problem of contextual bias, we opt to include Twitter as external context source for TV domain recommendation. Though Twitter is one of the most popular social media platforms for second screen usage, its user base is eventually heterogeneous and its applicability to the TV domain needs to be evaluated. We obtained our data by crawling only tweets related to relevant program names. To narrow the investiga-

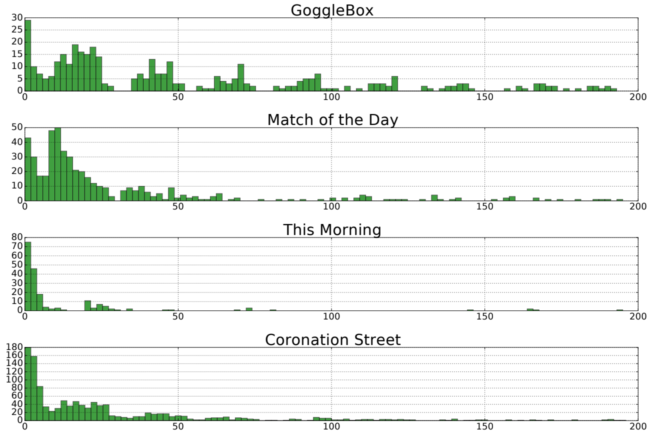

Figure 4: Delayed hours of watching from the first time aired for selected programs under VOD setting.

tion range, we select 12 titles from the 30 most requested TV programs on the "Vision" platform and perform a Twitter keyword search[3] to crawl Twitter histories. In total, the crawler gathered 4.7 million tweets for the 12 TV programs across a 26 month period parallel to user request dataset. Detailed statistics about the crawled tweets and user requests are provided in Sect. 6.

We group tweets according to the associated program title into one hour time bins. The histogram of tweets frequency in each time bin is first drawn to tell whether Twitter data really contains trend information regarding TV programs. Figure 5 depicts the frequency distribution of two typical kind of programs: *EastEnders* and *Gogglebox*. Programs like *EastEnders* show a generally stable frequency throughout the observed period, with rare bursts in the trending moments further called *peaks*, i.e. timeslots with a significantly higher number of tweets, that occur at uncertain points in time. In this case, the trend information is particularly valuable, because it reflects the uncertain external trending moment rather than other fixed bias regularity. On the other hand, *Gogglebox* represents another kind of trend phenomenon. The *peaks* of these TV programs in tweets even shows periodic regularities, which often correlate with their airing time.

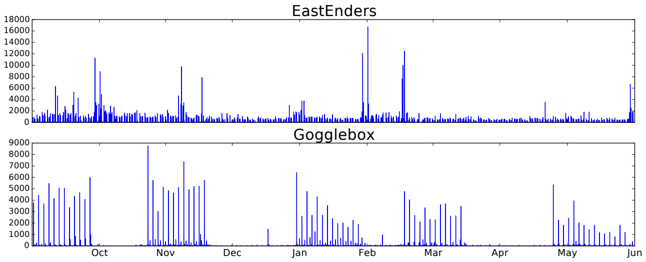

Figure 5: Tweet frequency distribution of two selected programs.

Knowing that *trends* exist in the external source, we investigate how trend information correlates with users' behavior in choosing programs to watch. In Figure 6, frequencies of user requests for the program *Coronation Street* are plotted in red, while the tweet frequency of the same program is plotted in green. This data from a short period (1 week) of time series demonstrates how both tweeting and playback request activities correspond to the same social events. A *peak* in tweet frequency often follows a *peak* in playback requests, while the rising flank of the tweet *peak* can actually happen ahead of the *peak* in requests. Crossing points of Trend Momentum and SigniScore w.r.t. the highest peak of program viewing frequency have been marked in green and blue scatters respectively, which show the forecasting effect of trend indicators in Twitter regarding frequency peaks in TV user requests data. Our ultimate aim is to use this earlier detection effect of Twitter *trends* on user request *peaks* to improve recommendations. To this end, the concept of both *trends* and *peaks* must be quantified.

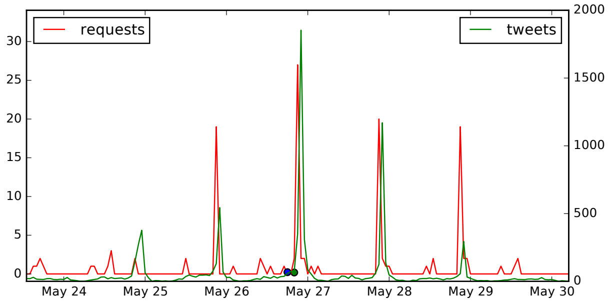

Figure 6: Parallel comparison of data distributions between tweets and user requests for program *Coronation Street*. The two dots mark the threshold crossings of both trend indicators closest to the center peak.

TREND MEASUREMENTS

To model the dynamics in popularity of a specific program in both internal user request data and external tweet data, we unify two concepts. First, a *peak* represents a point with significantly higher number of tweets/requests compared to the average number over the full period of discrete time-bins. Second, a *trend* is described by the rising phase of a *peak* and often appears shortly before the corresponding *peak*. That is to say *trends*, the early indications of upcoming *peaks*, are more valuable because they can help foretell an increase in popularity of a program in the near future, such that users receive recommendations for the program prior to its extensive discussion in social media. At the end of our literature review, we identified two promising trend measurements that can be used in our scenario.

$$MA(n,k) = \frac{\sum_{i=n-k+1}^{n} x(i)}{k} \tag{1}$$

$$MACD(n) = MA(n,k_{\text{fast}}) - MA(n,k_{\text{slow}}) \tag{2}$$

$$TM(n) = MA(n,k_{\text{fast}}) - MA(n,k_{\text{slow}})^{\alpha} \tag{3}$$

$$Momentum(n) = MA(TM(n),k_{\text{smooth}}) \tag{4}$$

[3] https://twitter.com/i/search/

The first one is the TM score listed in Eq. 3 proposed by Lu and Yang [29]. It is a smoothed version of MACD stock trend indicator and has been deployed to detect trending news in the Twitter stream. The definition stems from the concept of Moving Average (MA) (as Eq. 1 shows), which captures at the n_{th} time bin, the average frequency of k previous time bins. Considering that this average is not enough to represent a rising or decreasing trend, MACD (as shown in Eq. 2) utilizes the difference between the MA in k_{fast} (shorter) time windows and the MA in k_{slow} (longer) time windows to determine whether there is a trend appearing. In addition, with the discount parameter α assigned as exponential term to longer period MA in MACD, TM is defined as Eq. 3, and the sign change of its value from negative to positive or reversely indicates the appearance of a rising or declining trend. Furthermore, to avoid a volatile condition, MA is applied again with a third, even shorter time window k_{smooth} to further smooth the trend indicator as presented in Eq. 4. Throughout the remainder of this paper, we will simply use the name TM to refer to the final *Momentum* value (Eq. 4). By using this measure of momentum, a trend is said to be emerging when there is a turning point from negative to positive. Apart from the typical values recommended by the textbook, the four hyperparameters (k_{fast}, k_{slow}, k_{smooth} and α) can be tuned to improve accuracy of predicting *peaks*.

$$\Delta \leftarrow x - EWMA \tag{5}$$

$$EWMA \leftarrow EWMA + \alpha \cdot \Delta \tag{6}$$

$$EWMVar \leftarrow (1-\alpha) \cdot (EWMVar + \alpha \cdot \Delta^2) \tag{7}$$

$$\alpha = 1 - \exp\left(\frac{\log(0.5)}{t_{half}}\right) \tag{8}$$

$$\text{SigniScore}(x, \beta) = \frac{x - \max(EWMA, \beta)}{\sqrt{EWMVar} + \beta} \tag{9}$$

Another trend indicator as defined in Eq. 9 is called *SigniScore* and was introduced by Schubert et al. [37]. It is also a member of the MA family. With x being the frequency of occurrence within a time bin, the definition of Δ in Eq. 5 represents the deviation of this time bin from the EWMA calculated over preceding bins. After the current time bin has been observed, Δ is added to the EWMA in Eq. 6, where α is used as a weighting factor, similar to a learning rate. Corresponding to the accumulated mean, i.e. EWMA along the frequency stream, the formula for the accumulated variance is given in Eq. 7. As shown in Eq. 8, α can be derived from the half-life time t_{half} according to domain expert's knowledge. In our case, the critical parameter t_{half} is used as one of the hyperparameters to be optimized. On top of EWMA and EWMAVar, *SigniScore* is defined in Eq. 9 in the form of a *z-score*. Here β is the bias term that avoids division by 0 and at the same time filters noise. It constitutes the second hyperparameter which will be searched for in the case of *SigniScore*. Telling the (normalized) significance of a trend rather than solely relying on the sign change is an advantage of this measurement. It makes the comparison between different trending moments possible.

EVALUATION

Having introduced the two trend measures in the previous section, we proceed to apply them to the Twitter stream dataset over a timespan parallel to the user request data. Given the fact that "Vision" is a UK TV platform, for the 12 targeted TV titles, we choose 10 UK productions and 2 US productions. In addition, some program titles mostly consist of stop-word-like terms, e.g. *This Morning* and *My Wife and Kids*. For these program titles, the obtained tweets contain much more noise than others, and we include them to estimate the extend to which unrelated tweets influence the peak alignment. In Table 1, for both user request data and the crawled Twitter data, the number of data points, average number of points per non-empty bin μ, standard deviation σ, number of peaks occurring over the whole evaluation period #*peaks* and number of times where two concecutive peaks appear within 12 hours δ_{12} are displayed. We evaluate the precision of using *trends* in Twitter stream to predict *peaks* in user request data by two scores: OR and earliness of signal Δt.

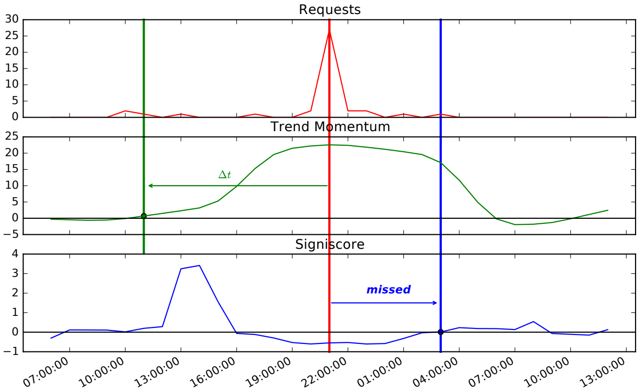

Figure 7: Visual comparison of a successful trend indication (TM) versus a *miss* (*SigniScore*).

The evaluation of our approach employs the concept of *peaks* in demand as time slots where the number of tweets/requests exceeds two standard deviations over the mean per (non-empty) time bin in the respective data stream. Following this definition, as Table 1 shows, the user dataset contains a total of 464 *peaks* for the targeted 12 TV programs, while in the comparably bigger Twitter dataset exist 3,455 such *peaks*. Even though there are relatively few peaks in the user request dataset, δ_{12} (the number of times where two consecutive peaks appear within a range of 12 hours) excludes the risk that they all belong to a few major *peak windows*. In accordance with the literature, we define the point when a trend measure crosses a threshold as the signal for the potential arrival of a *peak*. If the closest threshold crossing point occurs after the corresponding peak, we count the trend prediction as *missed* or as an *omission*. Otherwise, the crossing point can be seen as a successful indication of the incoming *peak*. A special example can be found in Fig. 7, where Trend Momentum successfully predicts the peak while Signiscore misses the chance because of the late capturing of this peak. For every successful trend indication, we compute the time delay between the threshold crossing and the highest *peak* point as Δt, which shows how

	User request data					Twitter data			
	N	μ	σ	#peaks	δ_{12}	N	μ	σ	#peaks
Made in Chelsea	1235	2.0	3.5	20	0	649523	36.2	312.4	111
EastEnders	1195	1.8	1.4	38	1	1276494	65.8	337.0	310
Hollyoaks	2070	1.7	1.1	91	13	1010904	52.7	223.8	375
Gogglebox	911	1.4	1.1	28	0	502954	31.5	257.0	128
Match of the Day	1122	1.7	1.7	23	2	189331	10.7	42.4	275
Emmerdale	1601	1.9	1.5	67	1	457419	24.5	149.7	175
Coronation Street	2624	2.0	2.1	43	7	295738	15.6	61.3	221
Britain's Got Talent	683	2.6	3.6	13	3	126100	7.1	34.8	102
Frasier	1413	1.4	0.7	72	23	116387	6.0	4.3	531
North West Tonight	1338	2.0	1.6	20	0	76512	1.8	1.3	430
This Morning	1046	1.5	0.8	21	0	17415	1.7	1.4	223
My Wife and Kids	756	1.8	1.1	28	6	10115	4.3	4.6	574

Table 1: Comparative statistics per evaluated program for the two datasets in use. N is the number of data points. For bin size of 1 hour, μ is the average number of points per non-empty bin and σ the respective standard deviation. Peaks are bins with more than $\mu + 2\sigma$ datapoints. δ_{12} counts the number of times where two consecutive peaks appear within a range of 12 hours.

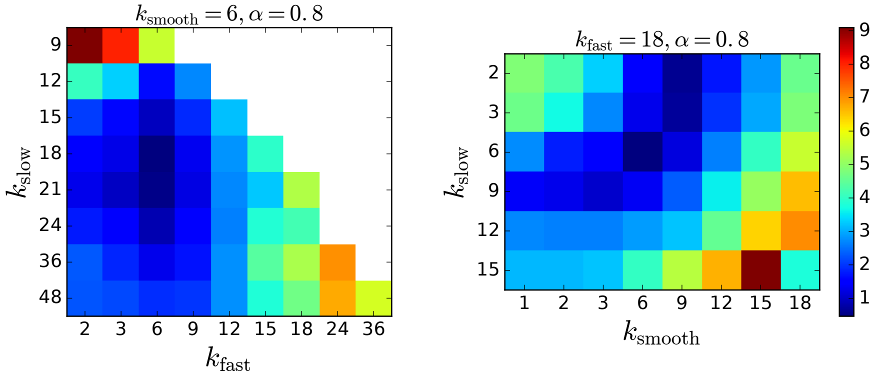

Figure 8: Omission rate under combinations of k_{fast}, k_{slow}, and k_{smooth}. Heat map shows the omission rate (in %)

early Twitter stream *trend* can predict user choice *peaks* (just as Δt shown in Fig. 7). The threshold can later be used to trade off between precision and recall but during the evaluation we use a fixed threshold $\theta = 0$ for finding crossings. In addition to the two trend measurements, we add a baseline trend indication in the evaluation. The baseline indication directly use the *peaks* in tweets frequency as trend indication signal, where the crossing threshold is a two standard deviations increase over the mean. The calculation of the baseline is straightforward and no parameters need to be tuned.

The performance of a trend indicator in a particular domain depends on the values assigned to its hyperparameters. We deploy grid search to find the best parameter settings for our target application. The ranges of the grids are derived from examination of contextual bias in Sect. 3. The performance criterion to be minimized is the average number of missed peaks (OR), over the 8 target programs. Since all preliminary trials reported the best value of α to be 0.8 in TM, we keep it constant throughout optimization. Then we conduct a 3-dimensional grid search with ranges $k_{slow} \in [2,\ldots,48], k_{fast} \in [2,\ldots,36], k_{smooth} \in [1,\ldots,18]$ and skip trials where $k_{fast} > k_{slow}$. For the 2-d heatmaps as display in Fig. 8 we fix k_{smooth} at 6 and k_{fast} at 18 respectively as they achieved lowest omission rate throughout the full search. The parameters settings falling in the dark blue area turned out to be better in reducing the omission rate. The results indicate that the best parameter settings for TM are $[k_{slow} = 18, k_{fast} = 6, k_{smooth} = 6, \alpha = 0.8]$.

For *SigniScore*, as described in Sect. 5, the bias term β and half-life setting t_{half} are the parameters to be tuned. Since there are only two parameters to be searched, one grid is directly applied in the range as depicted in Figure 9. The results show that $[\beta = 9, t_{half} = 9]$ is the optimal parameter combination for the problem at hand.

The best hyper parameters determined by the grid search are used to examine the per-program performance. In addition to the OR, for each program, we compute the average on Δt, i.e. time difference between threshold crossing and *peak*, to denote how early the trend indicator takes effect. Aside from that, the deviation σ on Δt is also provided. Contrary to the usual case where trend detection is favored to be as early as possible, in our setup, lower delays represents higher correlation between trend signal in tweets and *peaks* in user data. Thus we consider a small Δt (and small $\sigma(\Delta t)$) to be better.

Among the 12 programs under consideration, the evaluation result for 8 of them turned out to be better (Table 2). The

	Trend Momentum			SigniScore			Baseline		
	OR	Δt	σ	OR	Δt	σ	OR	Δt	σ
Made in Chelsea (20)	**5.0%**	**11.0**	1.5	**5.0%**	14.3	4.8	15.0%	3.9	1.6
EastEnders (38)	**2.6%**	**8.7**	2.1	7.9%	10.7	9.5	26.3%	11.3	15.1
Hollyoaks (91)	**3.3%**	8.0	2.6	20.9%	**6.6**	3.2	31.9%	17.6	23.4
Gogglebox (28)	**3.6%**	7.6	1.8	7.1%	**7.0**	2.0	35.7%	69.1	15.9
Match of the Day (23)	**0.0%**	**4.4**	2.3	39.1%	20.0	14.6	39.13%	33.9	14.2
Emmerdale (67)	**1.5%**	8.2	2.2	9.0%	**6.6**	2.8	49.3%	112.5	122.5
Coronation Street (43)	**2.3%**	9.8	2.5	25.6 %	**6.2**	4.0	51.2%	31.5	25.9
Britain's Got Talent (13)	30.77%	**9.4**	2.0	**7.69%**	10.7	9.5	100.0%	-	-

Table 2: Evaluation of trend indicators and baseline. The number of peaks per program is displayed in braces.

	Trend Momentum			SigniScore			Baseline		
	OR	Δt	σ	OR	Δt	σ	OR	Δt	σ
My Wife and Kids (28)	21.4%	7.8	6.9	25.0%	12.5	10.4	25.0%	106.8	220.6
Frasier (72)	47.2%	13.3	12.4	40.3%	1.6	1.8	45.8%	19.4	20.9
This Morning (21)	95.2%	17.0	0.0	61.9%	153.1	80.5	57.1%	11.6	8.5
North West Tonight (20)	70.0%	10.0	2.7	50.0%	476.5	335.8	60.0%	87.4	75.7

Table 3: Evaluation of trend indicators and baseline over the excluded programs. The number of peaks per program is displayed in braces.

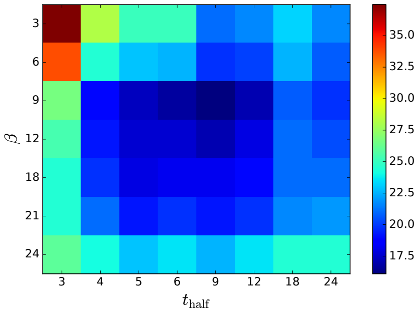

Figure 9: Omission rate for various combinations of t_{half} and β. Heat map shows the omission rate (in %)

comparison reveals that TM outperforms *SigniScore* and *baseline* for most of the target programs in terms of OR, Δt and $\sigma(\Delta t)$. The threshold crossing has a relatively consistent Δt before *peaks* of about 8 hours. This early increase in communication is observed for all soaps and reality shows (like *EastEnders* and *Gogglebox*) in the dataset. To a lesser extent, sports programs such as *Match of the Day* very well reflect a general characteristic of Twitter second screen usage. The reality TV show *Britain's Got Talent* constitutes the only case where *SigniScore* achieves lower OR, but its data is particularly sparse with only 13 peaks in total. Another special case is marked by *Made in Chelsea*, for which *baseline* already shows reasonable performance (the value of Δt is 3.9), while both trend measurements cross the threshold much earlier.

The varied evaluation results over different programs suggest that the applicability of the trend indicator is program dependent. For the other 4 programs out of 12 candidates, the performance as displayed in Table 3 sets a negative example of the trend indicators adaptability. Considering programs like *This Morning* and *North West Tonight*, their terms collocations are not unique, thus the tweet crawler tends to retrieve a significant number of unrelated tweets. This might be one important explanation for the poor performance of trend indicators. Similarly expected are the failures over the US programs like *My Wife and Kids*, for which the matching between Twitter trends and program interest in Vision, a system based in the UK, is skewed due to different time zones. In summary, we benefit from our evaluation in terms of both performance comparison and adaptable program grouping.

CONCLUSIONS AND FUTURE WORK

This paper introduces a novel approach to enhancing TV program recommendation based on external social context. We analyzed user interaction with a hybrid VOD/linear TV platform and identified a prevalent *contextual bias*. TV consumers' choices tend to fall in a strong temporal regularity, in which a few dominant programs or channels account for the majority in consumption. To overcome the issue of *contextual bias* and improve the diversity in recommendation, we harvest Twitter conversations as a source of social context. Using trend scores to detect early signs of increasing interest in program-related Twitter streams we explain *peaks* in the user request data. After hyperparameter optimization, we find that the MACD-based Trend Momentum indicator can very well achieve that goal, successfully forecasting about 96% of all peaks in TV programs' consumption dataset.

Our method comes with numerous advantages over previous approaches to context-sensitive recommendation. First and foremost, it alleviates the issue of *contextual bias* by extending the notion of context to a societal level, thus increasing the diversity of recommendations. Secondly, exploiting social context helps address *cold start* problem, because a significant portion of programs is new in TV recommendation scenario. Since Twitter is an external source, its utilization deprecates the need for any user- or program-related data prior to model deployment. A further strength of our method is that it particularly lends itself to the detection of individually popular episodes of repetitive programs without the scaling issues that come with CB alternatives. Finally, recommendation based on Twitter trends obtains the potential to create a feedback loop when it leads to more participants joining social conversation and reinforcing the trend. Appropriate access points can help the user directly engage the relevant discussion thereby facilitating VOD collaborative watching and the increasingly popular second screen usage.

Recommendation based on Twitter trends is not a silver bullet. Its applicability depends on the means by which program related conversations are collected. Filtering for programs with distinctive names is trivial as demonstrated by our analysis. In other cases, if we intend to discern TV program through the whole context in a Tweet, the approach proposed by Cremonesi et al. in [13] can be a good work to refer to. In addition, Tweets about programs with an international audience may show reduced correlation to *peaks* in consumption on platforms with a specific national target audience. Geolocation information as an additional filter can be used to reduce noise and augment tweets relevance in the future [41]. Admittedly, the analysis in the current paper can only illustrate the applicability of social trend in TV recommendation tasks. Yet to which degree it can enhance the user experience or how much they can overcome the contextual bias, is to be determined through future work.

ACKNOWLEDGEMENTS

This work was supported in part by the German Federal Ministry of Education and Research (BMBF) under the grant number 01IS16046. It was also partially supported by the European Commission through the Horizon 2020 programme under grant agreement No. 687921, project MPAT (Multi-Platform Application Toolkit). The second author of the paper has been supported by China Scholarship Council (CSC). We are also grateful to Jamie Jellicoe for assistance with data extraction from the “Vision” platform.

REFERENCES

1. J. Abreu, P. Almeida, B. Teles, and M. Reis. Viewer behaviors and practices in the (new) television environment. In *Procs. of the 11th european conf. on Interactive TV and video*, pages 5–12. ACM, 2013.
2. G. Adomavicius and A. Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *Knowledge and Data Engineering, IEEE Transactions on*, 17(6):734–749, 2005.
3. C. C. Aggarwal and K. Subbian. Event detection in social streams. In *SDM*, volume 12, pages 624–635. SIAM, 2012.
4. S. Asur, B. A. Huberman, G. Szabo, and C. Wang. Trends in social media: Persistence and decay. *Available at SSRN 1755748*, 2011.
5. R. Bambini, P. Cremonesi, and R. Turrin. A recommender system for an iptv service provider: a real large-scale production environment. In F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor, editors, *Recommender Systems Handbook*, pages 299–331. Springer US, 2011.
6. A. B. Barragáns-Martínez, E. Costa-Montenegro, J. C. Burguillo, M. Rey-López, F. A. Mikic-Fonte, and A. Peleteiro. A hybrid content-based and item-based collaborative filtering approach to recommend tv programs enhanced with singular value decomposition. *Information Sciences*, 180(22):4290–4311, 2010.
7. P. Bellekens, G.-J. Houben, L. Aroyo, K. Schaap, and A. Kaptein. User model elicitation and enrichment for context-sensitive personalization in a multiplatform tv environment. In *Proceedings of the seventh european conference on European interactive television conference*, pages 119–128. ACM, 2009.
8. P. G. Campos, A. Bellogín, F. Díez, and J. E. Chavarriaga. Simple time-biased knn-based recommendations. In *Proceedings of the Workshop on Context-Aware Movie Recommendation*, pages 20–23. ACM, 2010.
9. P. G. Campos, F. Díez, and I. Cantador. Time-aware recommender systems: a comprehensive survey and analysis of existing evaluation protocols. *User Modeling and User-Adapted Interaction*, 24(1-2):67–119, 2014.
10. M. Cataldi, L. Di Caro, and C. Schifanella. Emerging topic detection on twitter based on temporal and social terms evaluation. In *Proceedings of the Tenth International Workshop on Multimedia Data Mining*, page 4. ACM, 2010.
11. N. Chang, M. Irvan, and T. Terano. A tv program recommender framework. *Procedia Computer Science*, 22:561–570, 2013.
12. P. Cremonesi, Y. Koren, and R. Turrin. Performance of recommender algorithms on top-n recommendation tasks. In *Proceedings of the fourth ACM conference on Recommender systems*, pages 39–46. ACM, 2010.
13. P. Cremonesi, R. Pagano, S. Pasquali, and R. Turrin. Tv program detection in tweets. In *Proceedings of the 11th european conference on Interactive TV and video*, pages 45–54. ACM, 2013.
14. F. S. da Silva, L. G. P. Alves, and G. Bressan. Personaltvware: A proposal of architecture to support the context-aware personalized recommendation of tv programs. In *European Interactive TV Conference (EuroITV 2009), Leuven, Belgium*, 2009.
15. D. Das and H. ter Horst. Recommender systems for tv. In *Recommender Systems, Papers from the 1998 Workshop, Technical Report WS-98-08*, pages 35–36, 1998.
16. Y. Dong, J. Tang, S. Wu, J. Tian, N. V. Chawla, J. Rao, and H. Cao. Link prediction and recommendation across heterogeneous social networks. In *Proceedings - IEEE Intl. Conf. on Data Mining, ICDM*, pages 181–190, 2012.
17. Ericsson ConsumerLab. Tv and media 2015. Technical report, Ericsson ConsumerLab, 2015.

18. Y. Fang, H. Zhang, Y. Ye, and X. Li. Detecting hot topics from twitter: A multiview approach. *Journal of Information Science*, page 0165551514541614, 2014.

19. Y. B. Fernández, J. J. P. Arias, M. L. Nores, A. G. Solla, and M. R. Cabrer. Avatar: an improved solution for personalized tv based on semantic inference. *Consumer Electronics, IEEE Transactions on*, 52(1):223–231, 2006.

20. J. Freitas and H. Ji. Identifying news from tweets. *NLP+ CSS 2016*, page 11, 2016.

21. D. Goldberg, D. Nichols, B. M. Oki, and D. Terry. Using collaborative filtering to weave an information tapestry. *Communications of the ACM*, 35(12):61–70, 1992.

22. S. Hendrickson, J. Kolb, B. Lehman, and J. Montague. Trend detection in social data. Technical report, Twitter, 2015.

23. B. Hidasi and D. Tikk. General factorization framework for context-aware recommendations. *Data Mining and Knowledge Discovery*, 30(2):342–371, 2016.

24. A. Karatzoglou, X. Amatriain, L. Baltrunas, and N. Oliver. Multiverse recommendation: n-dimensional tensor factorization for context-aware collaborative filtering. In *Proceedings of the fourth ACM conference on Recommender systems*, pages 79–86. ACM, 2010.

25. Y. Koren. Collaborative filtering with temporal dynamics. *Communications of the ACM*, 53(4):89–97, 2010.

26. M. Krstic and M. Bjelica. Context-aware personalized program guide based on neural network. *IEEE Trans. on Consumer Electronics*, 58(4):1301–1306, 2012.

27. J. Kunegis, G. Gröner, and T. Gottron. Online dating recommender systems: the split-complex number approach. In *ACM RecSys workshop on Recommender systems and the social web*, pages 37–44, 2012.

28. W.-P. Lee, C. Kaoli, and J.-Y. Huang. A smart tv system with body-gesture control, tag-based rating and context-aware recommendation. *Knowledge-Based Systems*, 56:167–178, 2014.

29. R. Lu and Q. Yang. Trend analysis of news topics on twitter. *International Journal of Machine Learning and Computing*, 2(3):327, 2012.

30. A. Q. Macedo, L. B. Marinho, and R. L. Santos. Context-aware event recommendation in event-based social networks. In *Procs. of the 9th ACM Conf. on Recommender Systems*, pages 123–130. ACM, 2015.

31. A. Madani, O. Boussaid, and D. E. Zegour. Real-time trending topics detection and description from twitter content. *Social Network Analysis and Mining*, 5(1):59, 2015.

32. A. B. B. Martínez, J. J. P. Arias, A. F. Vilas, J. G. Duque, and M. L. Nores. What's on tv tonight? an efficient and effective personalized recommender system of tv programs. *IEEE Transactions on Consumer Electronics*, 55(1):286–294, 2009.

33. M. Naaman, H. Becker, and L. Gravano. Hip and trendy: Characterizing emerging trends on twitter. *Journal of the American Society for Information Science and Technology*, 62(5):902–918, 2011.

34. U. Panniello, A. Tuzhilin, and M. Gorgoglione. Comparing context-aware recommender systems in terms of accuracy and diversity. *User Modeling and User-Adapted Interaction*, 24(1-2):35–65, 2014.

35. T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, pages 851–860. ACM, 2010.

36. H. Sayyadi, M. Hurst, and A. Maykov. Event detection and tracking in social streams. In *Icwsm*, 2009.

37. E. Schubert, M. Weiler, and H.-P. Kriegel. Signitrend: scalable detection of emerging topics in textual streams by hashed significance thresholds. In *Proc.s of the 20th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining*, pages 871–880. ACM, 2014.

38. C. Shin and W. Woo. Socially aware tv program recommender for multiple viewers. *IEEE Transactions on Consumer Electronics*, 55(2):927–932, 2009.

39. D. Spina. *Entity-based filtering and topic detection For online reputation monitoring in Twitter*. PhD thesis, Universidad Nacional de Educación a Distancia, 2014.

40. D. Véras, T. Prota, A. Bispo, R. Prudêncio, and C. Ferraz. A literature review of recommender systems in the television domain. *Expert Systems with Applications*, 42(22):9046–9076, 2015.

41. S. Wakamiya, R. Lee, and K. Sumiya. Towards better tv viewing rates: exploiting crowd's media life logs over twitter for tv rating. In *Proceedings of the 5th International Conference on Ubiquitous Information Management and Communication*, page 39. ACM, 2011.

42. J. Weng and B.-S. Lee. Event detection in twitter. *ICWSM*, 11:401–408, 2011.

43. C.-C. Wu and M.-J. Shih. A context-aware recommender system based on social media. In *International Conference on Computer Science, Data Mining & Mechanical Engineering*, 2015.

44. L. Xiang, Q. Yuan, S. Zhao, L. Chen, X. Zhang, Q. Yang, and J. Sun. Temporal recommendation on graphs via long-and short-term preference fusion. In *Procs. of the 16th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining*, pages 723–732. ACM, 2010.

45. H. Yin, B. Cui, L. Chen, Z. Hu, and Z. Huang. A temporal context-aware model for user behavior modeling in social media systems. In *Proceedings of the 2014 ACM SIGMOD international conference on Management of data*, pages 1543–1554. ACM, 2014.

46. A. Zubiaga, D. Spina, R. Martinez, and V. Fresno. Real-time classification of twitter trends. *Journal of the Association for Information Science and Technology*, 66(3):462–473, 2015.