Social Media Analytics in Support of Documentary Production

Giorgos Mitsis, Nikos Kalatzis, Ioanna Roussaki,
Eirini Eleni Tsiropoulou, Symeon Papavassiliou
Institute of Communications and Computer Systems
Athens, Greece
e-mails: {gmitsis@netmode, nikosk@cn,
ioanna.roussaki@cn, etsirop@netmode,
papavass@mail}.ntua.gr

Simona Tonoli
Mediaset
Milan, Italy
e-mail: Simona.Tonoli@mediaset.it

***Abstract*—Recent market research has revealed a globally growing interest on documentaries that have now become one of the most populated content-wise genre in the movie titles catalog, surpassing traditionally popular genres such as comedy or adventure films. At the same time, modern audiences appear willing to immerse into more interactive and personalized viewing experiences. Documentaries, even in their linear version, involve high costs in all phases (pre-production, production, post-production) due to various inefficiencies, partly attributed to the lack of scientifically-proven cost-effective Information and Communications Technology (ICT) tools. To fill this gap, a set of innovative ICT tools is delivered that focus on supporting all stages of the documentary creation process, ranging from the documentary topic selection to its final delivery to the viewers. This paper provides an overview of the respective tools, elaborating on two specific tools that primarily focus on the interests and satisfaction of the targeted audience: the Integrated Trends Discovery tool and the Social Recommendation & Personalization tool, elaborating on their design, functionality and performance, and concludes with exposing the future plans and potential regarding these tools.**

Keywords-documentary production; social-media analytics; Integrated Trends Discovery tool; Social Recommendation & Personalization tool.

I. INTRODUCTION

From the earliest days of cinema, documentaries have provided a powerful way of engaging audiences with the world. They always had social and market impact, as they adapted to the available means of production and distribution. More than any other type of films, documentarians were avid adapters of new technologies, which periodically revitalized the classical documentary form. The documentary is a genre which lends itself straightforwardly to interaction. People have different knowledge backgrounds, different interests and points of view, different aesthetic tastes and different constraints while viewing a programme. Therefore, it becomes evident that some form of personalized interactive documentary creation will enhance the quality of experience for the viewers, facilitating them to choose different paths primarily with respect to the documentary format and playout system. The convergence between the documentary production field and of digital media enables the realization of this vision.

As the range of ICT platforms broadens, documentary makers need to understand and adopt emerging technologies in order to ensure audience engagement and creative satisfaction, via the use of personalization and interactive media. One of the major challenges for stakeholders in the arena of documentary creation is the development of processes and business models to exploit the advantages of those technical achievements, in order to reduce the overall cost of documentary end-to-end production, to save time and to deliver enhanced personalized interactive and thus more attractive documentaries to the viewers.

PRODUCER [1] is an H2020 EU project that aims to pave the path towards supporting the transformation of the well-established and successful traditional models of linear documentaries to interactive documentaries, by responding to the recent challenges of the convergence of interactive media and documentaries. This is achieved via the creation of a set of enhanced ICT tools that focus on supporting all documentary creation phases, ranging from the user engagement and audience building, to the final documentary delivery. In addition to directly reducing the overall production cost and time, PRODUCER aims to enhance viewers' experience and satisfaction by generating multi-layered documentaries and delivering more personalized services, e.g., regarding the documentary format and playout.

In order to provide the aforementioned functionality, the PRODUCER platform implements 9 tools, each focusing on a specific documentary production phase. These tools are: Integrated trends discovery tool, Audience building tool and Open content discovery tool (that support the documentary pre-production phase), Multimedia content storage, search & retrieval tool and Automatic annotation tool (that support the core production phase), Interactive-enriched video creation tool, 360° video playout tool, Second screen interaction tool and Social recommendation & personalization tool (all four focusing on the documentary post-production phase). The architecture of the PRODUCER platform is presented in more detail in [2]. This paper elaborates on two of the PRODUCER tools: the Integrated Trends Discovery tool and the Social Recommendation & Personalization tool.

In the rest of the paper, Section 2 elaborates on the design & functionality of the Integrated Trends Discovery tool, presenting initial evaluation results for one of its mechanisms. Section 3 focuses on the description of the Social Recommendation & Personalization tool, while it elaborates on specific performance evaluation/benchmarking results related to its functionality. Finally, in Section 4, conclusions are drawn and future plans are presented.

II. INTEGRATED TRENDS DISCOVERY TOOL

This section elaborates on the ITD tool, i.e., its innovations, architecture, user demographics inference mechanism and respective evaluation.

A. Rationale and Innovations

In recent years, there is an increasing trend on utilizing social media analytics and Internet search engines analytics for studying and predicting behavior of people with regards various societal activities. The proper analysis of Web 2.0 services utilization, goes beyond the standard surveys or focus groups and has the potential to be a valuable source of information leveraging internet users as the largest panel of users in the world. Analysts from a wide area of research fields have the ability to reveal current and historic interests of individuals and to extract additional information about their demographics, behavior, preferences, etc. One of the c aspects of this approach is that the user base consists of people that the researchers have never considered.

Some of the research fields that demonstrate significant results through the utilization of such analytics include epidemiology (e.g., detect influenza [3][4] and malaria [5]) epidemics), economy (e.g., stock market analysis [6], private consumption prediction [7], financial market analysis and prediction [8], unemployment rate estimation [9]) politics (e.g., predicting elections outcomes [10]).

On the other hand, there are limitations on relying only on these information sources as certain groups of users might be over- or under-represented among internet search data. There is a significant variability of online access and internet search usage across different demographic, socioeconomic, and geographic subpopulations.

With regards content creation and marketing, the existing methodologies are under a major and rapid transformation given the proliferation of Social Media and search engines. The utilization of such services generates voluminous data that allows the extraction of new insights with regards the audiences' behavioral dynamics. In [11], authors propose a mechanism for predicting the popularity of online content by analyzing activity of self-organized groups of users in social networks. Authors in [12] attempt to predict IMDB (http://www.imdb.com/) movie ratings using Google search frequencies for movie related information. In a similar manner, authors in [13] are inferring, based on social media analytics, the potential box office revenues with regards Internet content generated about Bollywood movies.

The existing research approaches are mainly focusing in post-production phase of released content. Identifying the topics that are most likely to engage the audience is critical for content creation in the pre-production phase. The ultimate goal of content production houses is to deliver content that matches exactly what people are looking for. Deciding wisely on the main documentary topic, as well as the additional elements that will be elaborated upon, prior to engaging any resources in the documentary production process, has the potential to reduce the overall cost and duration of the production lifecycle, as well as to increase the population of the audiences interested, thus boosting the respective revenues. In addition, the existence of hard evidence with regards potential audience's volume and characteristics (e.g., geographical regions, gender, age) is an important parameter in order to decide the amount of effort and budget to be invested during production.

There are various social media analytics tools that are focusing on generic marketing analysis e.g., monitoring for a long time specific keyword(s) and websites for promoting a specific brand and engaging potential customers. These web marketing tools rely on user tracking, consideration of user journeys, detection of conversion blockers, user segmentation, etc. This kind of analysis requires access to specific websites analytics and connections with social media accounts (e.g., friends, followers) which is not the case when the aim is to extract the generic population trends. In addition, these services are available under subscription fee that typically ranges from 100 Euros/month to several thousand Euros/month, a cost that might be difficult to be handled by small documentary houses.

The ITD Tool aims to support the formulation, validation and (re)orientation of documentary production ideas and estimate how appealing these ideas will be to potential audiences based on data coming from global communication media with massive user numbers. The ITD tool integrates existing popular publicly available services for: monitoring search trends (e.g., Google Trends), researching keywords (e.g., Google Adwords Keyword Planner), analyzing social media trends (e.g., Twitter trending hashtags). In more details, the ITD tool innovations include the following:

- Identification and evaluation of audience's generic interest for specific topics and analysis/inference of audience's characteristics (e.g., demographics, location)
- Extraction of additional aspects of a topic though keyword analysis, quantitate correlation of keywords, and association with high level knowledge (e.g., audience sentiment analysis)
- Discovery and identification of specific real life events related with the investigated topic (e.g., various breakthroughs of google/twitter trending terms are associated with specific incidents)
- Utilisation of data sources that are mainly openly accessible through public APIs which minimises the cost and increases the user base.

B. Architecture

A functional view of ITD tool's architecture is provided in Figure 1. Its core modules are described hereafter.

RestAPI: This component exposes the backend's functionality through a REST endpoint. The API specifies a set of trend discovery queries where the service consumer provides as input various criteria such as keywords, topics, geographical regions, time periods, etc.

Trends Query Management: This component orchestrates the overall execution of the queries and the processing of the replies. It produces several queries formulated properly that are forwarded to the respective connectors/wrappers to dispatch the requests to several existing TD tools/services available online. Given that each external service will reply in different time frames (e.g., a call to Google Trends discovery replies within a few seconds while Twitter stream analysis might take longer time periods) the overall process is performed in an asynchronous manner, coordinated by the Message Broker. The Query Management enforces querying policies tailored to each service in order to optimize the utilization of the services and to avoid potential bans. To this end, results from calls are also stored in ITD tool's local database in order to avoid unnecessary calls to the external APIs that have recently performed.

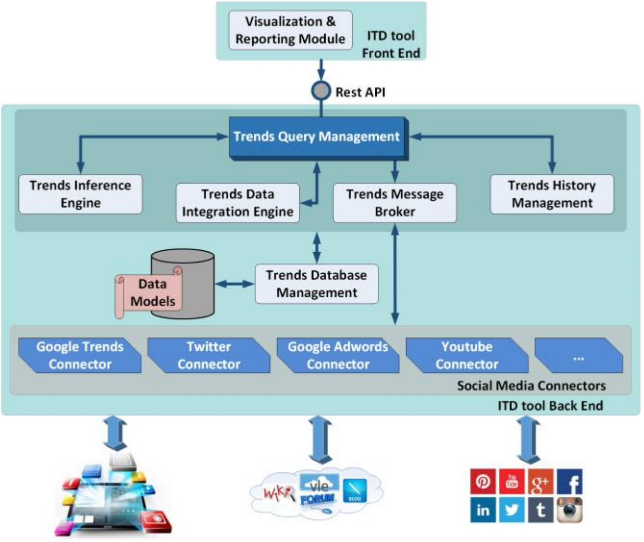


Figure 1. Architecture of the Integrated Trends Discovery Tool.

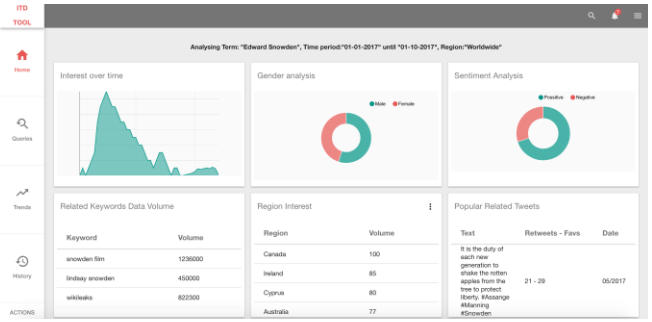
Trends Message Broker: This component realizes the asynchronous handling of requests. It is essentially a messaging server that forwards requests to the appropriate recipients via a job queue based on distributed message passing system.

Social Media Connectors: A set of software modules that support the connection and the execution of queries to external services through the provided available APIs. Connectors are embedding all the necessary security related credentials to the calls and automate the initiation of a session with the external services. Thus, the connectors automate and ease the actual formulation and execution of the queries issued by the Query Management component. Some example APIs that are utilized by the connectors are: Google Adword API, Twitter API, YouTube Data API v3.

Trends Data Integration Engine: This module collects the intermediate and final results from all modules, homogenize their different formats, and extracts the final report with regards the trends discovery process. The results are also modelled and stored in the local data base in order to be available for future utilization.

Trends Database Management & Data model: The ITD tool maintains a local database where the results of various calls to external services are stored. The Database Management module supports the creation, retrieval, update and deletion of data objects. This functionality is supported for both contemporary data but also for historic results (Trends History Management). Hence, it is feasible for the user to compare trend discovery reports performed in the past with more recent ones and have an intuitive view of the evolution of trend reports in time.

Front End: The Front-End visualizes the results providing the following output: (i) a graph of terms (each term is escorted by an audience popularity metric and is correlated with other terms, where a metric defines the correlation level), (ii) audience interest per location (country/city), (iii) interest per date(s) (significant dates, identification of seasonal habits), (iv) audiences sentiment analysis, (v) audiences gender analysis (vi) related questions with the topic. An ITD GUI snapshot is depicted in Figure 2.



Analysing Term: "Edward Snowden", Time period:"01-01-2017" until "01-10-2017", Region:"Worldwide"

Related Keywords Data Volume

Keyword	Volume
snowden film	1236000
lindsay snowden	450000
wikileaks	822300

Region Interest

Region	Volume
Canada	100
Ireland	85
Cyprus	80
Australia	77

Popular Related Tweets

Text	Retweets - Favs	Date
It is the duty of each new generation to shake the rotten apples from the tree to protect liberty. #Assange #Manning #Snowden	21 - 29	05/2017

Figure 2. Snapshot of the Integrated Trends Discovery Tool GUI.

Trends Inference Engine: In some cases, the external services are not directly providing all information aspects of the required discovery process. To this end, by applying the appropriate inference mechanisms on the available data allows the extraction of additional information escorted by a confidence level with regards the accuracy of the estimation. Details of this module are presented in the following section.

C. Inference of User Demographics

During the preproduction phase of a documentary, producers are highly interested in estimating trends in correlation with potential audiences' gender and age classification. This kind of information is not freely available from social media services due to user privacy protection data policies. There are various state of the art attempts that focus on inferring user demographics though probabilistic approaches based on user related data freely available on social media (e.g., tweets content, linguistic features, followers' profile) [14][15][16][17].

With regards to the documentary preproduction phase, the task of age and gender estimation is tackled by the ITD tool via the utilization of classification algorithms trained with ground-truth data sets of a number of tweeter users. Twitter service proved to be the most proper for extracting user profile information as Twitter account data and content are openly available. The trained network is then utilized in order to generalize the training process and estimate missing information from wider networks of twitter users.

The inference process is coordinated by the Trends Inference Engine. The engine uses the TwitterAPI to retrieve tweets where the keywords connected with certain topics are mentioned. Based on the respective Twitter Account ids, profile information is collected for each account. Based on profile attributes (e.g., "name", "screen_name", "profile photo", "short description", "profile_color") each user is classified to age & gender category and each classification is escorted by a confidence level.

The actual classification process is based on a statistical model where recurring patterns of users' profile attributes are accompanying a certain age and/or gender class. Learning is performed based on a ground truth dataset containing records of real Twitter profile information and the respective gender/ age. The ITD tool is capable to utilise various classification algorithms but as a first proof of concept the Naive Bayes is evaluated. Naive Bayes (NB) is an algorithm that fulfills the requirements set by similar problems and has performed well in many complex real world situations [18]. NB follows a supervised learning approach for estimating parameters of the classifier, such as means and variances of the variables. The algorithm provides quantifiable probability distributions for each possible class and requires a small amount of training data. In addition, NB can handle both categorical and numerical attributes. Compared with Bayesian Networks, there is no need for domain expert interference in designing dependencies between input attributes. On the other hand, it assumes that attributes are independent from each other with respect to the classification outcome, something that it is not always the case, while the computing resource consumption can get significantly high.

A user's profile is modelled as s $= \{c_1, c_2, \ldots, c_n\}$, where c_i is the value of user profile information of type $i, (i = 1,2, \ldots, n)$. Gender classes are modelled as g_j (j =1,2,3) corresponding to: "Female", "Male" and "Unknown". Age classes are modelled as a_i (i=1,…,7) corresponding to the following 7 age states: 18-24, 25-34, 35-44, 45-54, 55-64, 65 or more, and Unknown.

Based on the ground truth dataset age and gender classes can be associated with specific user profiles in the form of tuples such as (gender, profile) => (g_j, s) and (age, profile) => (a_i, s). Bayes rule for calculating prediction probabilities according to the defined problem becomes:

$$P[g_j|s] = P[g_j] \times \frac{\prod_{j=1}^{n} P[c_j|g_j]}{P(s)}$$

where g_j is the expected gender classification outcome and $s = \{c_j\},\ j = 1, \ldots, n$ is the current evidence input.

Similarly, Bayes rule for estimating the user's age is:

$$P[a_j|s] = P[a_j] \times \frac{\prod_{j=1}^{n} P[c_j|a_j]}{P(s)}$$

Based on these rules the actual estimation is realised through the maximisation of these probabilities: $a = \arg\max\{P[a_j|s]\}$ and $g = \arg\max\{P[g_j|s]\}$.

D. *Evaluation*

The presented architecture is under implementation by the authors of this article and a first release is already available at: http://itd.lab.netmode.ece.ntua.gr/. The ITD backend tool is developed in Django-Python framework, the front-end is based on Angular-Material while the following services have been integrated through the respective APIs: Google Trends, Google Adwords, Twitter, Youtube. The first evaluation processes with regards the overall utilization of the tool are encouraging and allow to discover early in the development phase potential shortcomings.

Such an issue is related with the volume of calls to external services. For example, Twitter API limits the allowed calls to 15 every 15 minutes per service consumer. As this issue was expected, a caching mechanism is utilized where results from each call to the Twitter API are also stored in the local database. Hence the ITD builds each own information store in order to avoid unnecessary calls. To this end, as the tool is utilized from various user the local information store is getting more complete.

With regards the ITD inference engine a first evaluation attempt realized for the gender estimation mechanism. The evaluation has been based on a public data set (https://www.kaggle.com/crowdflower/twitter-user-gender-classification) of ground truth data containing information of 10021 twitter users' profiles. The dataset contains the gender of distinct twitter users escorted by profile information. As a first step on the evaluation process and given that stylistic factors are often associated with user gender, the Twitter profile colour has been initially utilized.

Each colour's RGB value (red, green, blue) is fed to the Bayesian classifier as a distinct numerical feature. Thus, each class (male, female, unknown) is associated with three numerical features. The aim is to handle colour features not as independent enumerated attributes, but as continuous numerical values, as shades of the same colour are expressed via close RGB values. The Bayesian classifier has been developed using the "scikit-learn" library (http://scikit-learn.org/), and given the fact that the colour attributes are expressed as continuous values, the Gaussian Naive Bayes algorithm has been adapted to the needs of the described problem.

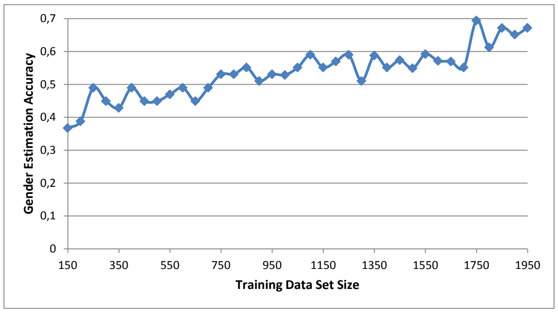


Figure 3. Evaluation of the gender estimation mechanism performance.

In order to evaluate the gender inference algorithm, the initial dataset (~10000 records) has been divided into 40 parts each containing about 250 records. Each dataset part was gradually incorporated to the classifier, while the last 250 records were used for evaluation. The initial evaluation attempts didn't provide high performance results. A data cleansing process was subsequently performed removing records that had the default predefined Twitter profile colors that resulted in a dataset of ~2000 records. The same evaluation process was then conducted where each of the 40 parts contained 50 records. The respective evaluation results are presented in Figure 3 and are rather encouraging, demonstrating about 70% of accurate classification when the entire training data set is incorporated.

The evaluation process is planned to proceed with further testing of the proposed approach based on more datasets, originating from additional social media (not only Twitter), to compare with similar existing approaches and to incorporate additional user profile attributes, including text analysis of provided profile description and Tweets text.

III. Social Recommendation & Personalization Tool

This section elaborates on the SRP tool, i.e., its functionality, architecture, recommendation extraction algorithm and respective evaluation/benchmarking.

A. Functionality & Design

Personalization & Social Recommendation are dominant mechanisms in today's social networks, online retails and multimedia content applications due to the increase in profit to the platforms as well as the improvement of the Quality of Experience (QoE) for its users and almost every online company has invested in creating personalized recommendation systems. Major examples include YouTube that recommends relevant videos and advertisements, Amazon that recommends products, Facebook that recommends advertisements and stories, Google Scholar that recommends scientific papers, while other online services provide APIs such as Facebook Open Graph API and Google's Social Graph API for companies to consume and provide their own recommendations [19].

The Social Recommendation & Personalization (SRP) tool of PRODUCER holistically addresses personalization, relevance feedback and recommendation, offering enriched multimedia content tailored to users' preferences. The tool's functionalities can be used in any type of content that can be represented in a meaningful way, as explained later. The application is thus not restricted to documentaries.

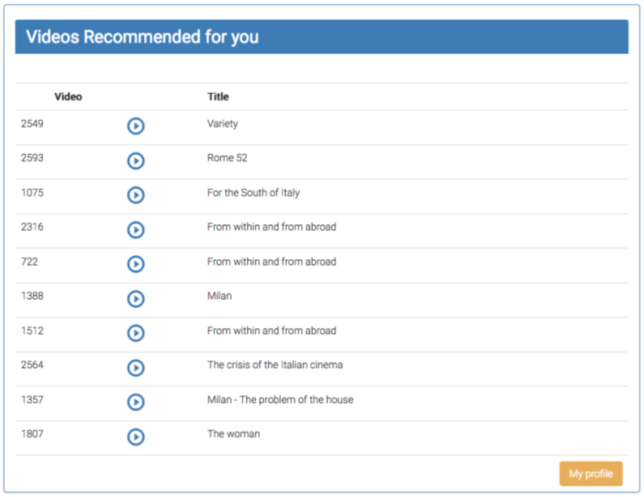
The recommendation system we built is not restricted to the video itself, but applies also to the set of enrichments accompanying the video. Interaction with both video and enrichments is taken into consideration into updating the user's profile, thus holistically quantifying the user's behaviour. Its goal is to facilitate the creation of the documentary and allow the reach of the documentary to a wider audience. To do so, the SRP tool is responsible for proposing appropriate content for specific target groups to the producer of the film via a personalization mechanism.

The first process the SRP tool has to perform is to index the content in a meaningful way, an important step as also indicated in [20][21]. Each video/enrichment is mapped to a vector, the elements of which are the scores appointed to the video/enrichment expressing the relevance it has to each category we have defined. The categories used come from the upper layer of DMOZ (http://dmoztools.net/), an attempt to create a hierarchical ontology scheme for organizing sites, that fits the generic nature of the PRODUCER videos.

Each multimedia content item is therefore described as follows: $X_P = [X_{P_1}, X_{P_2}, \dots, X_{P_N}]$, where P_i are the specified categories and X_{P_i} are appointed using the Doc2Vec algorithm [22]; the metadata of each item are passed through a neural network which represents the item with a multidimensional vector. The same procedure is done with the defined categories, and the vector X_P is constructed by finding the similarity of the multi-dimensional vectors of the item with each of the categories.

In order to be able to identify content relevant to target audiences, the tool needs to collect information and preferences of viewers since user profiles constitute another integral part of a recommendation system. Within the platform the SRP tool operates, the viewer registers and provides some important demographics (i.e., gender, age, country, occupation and education), as well as some of his/her preferences on specified topics, that will be used to identify the audience group that the viewer is part of. Alternatively, instead of providing this information explicitly, the viewer can choose to login with his/her social network account (e.g., Facebook, Twitter) and this information could be extracted automatically.

The user profile created via this process is static and is not effective for accurate recommendation of content since: a) the user is not able to accurately express his/her interests and b) his/her interests change dynamically. Thus, in addition to the above process the SRP tool implicitly collects information for the user's behavior and content choices. Using information about the video he/she watched or the enrichments that caught his/her attention, the SRP tool updates the viewer's profile to reflect more accurately his/her current preferences.



Videos Recommended for you

Video		Title
2549	⊙	Variety
2593	⊙	Rome 52
1075	⊙	For the South of Italy
2316	⊙	From within and from abroad
722	⊙	From within and from abroad
1388	⊙	Milan
1512	⊙	From within and from abroad
2564	⊙	The crisis of the Italian cinema
1357	⊙	Milan - The problem of the house
1807	⊙	The woman

Figure 4. Recommendation of content to PRODUCER viewers based on their user profile.

The created user profile, allows the tool to suggest content to the viewer to consume (Figure 4), as well as a personalized experience when viewing the content by showing only the most relevant enrichments for his/her taste. Through a content-based approach, the user's profile is matched with the content's vector by applying the cosine similarity measure as:

$$sim_{up}^{cf}(i,j) = \frac{U_i \cdot X_P^j}{\|U_i\| \left\|X_P^j\right\|} \quad (1)$$

where U_i is the user's profile vector and X_P^j is the content's vector.

The collaborative approach is complementary with the content-based recommendation using information from other viewers with similar taste, to increase diversity. The idea is to use already obtained knowledge from other users in order make meaningful predictions for the user in question. To do so, the similarity between users is computed as follows:

$$sim_{uu}(i,j) = \frac{U_i \cdot U_j}{\|U_i\|\|U_j\|} \qquad (2)$$

where the H more similar users are denoted as neighbors. We then compute the similarity of the neighbors to the item:

$$sim_{up}^{cbf}(i,j) = \sum_{s=1}^{H} sim_{up}^{cf}(i,s) \cdot sim_{uu}(s,j) \qquad (3)$$

and the final similarity between the user and the item is calculated via a hybrid scheme by using the convex combination of the above similarities:

$$sim_{up}^{h}(i,j) = (1-\theta) sim_{up}^{cbf}(i,j) + \theta sim_{up}^{cf}(i,j) \qquad (4)$$

where $\theta : 0 \leq \theta \leq 1$ is a tunable parameter denoting the importance of the content-based and the collaborative approach on the hybrid scheme. A value of $\theta = 0.5$ has been shown to produce better results than both approaches used individually [23].

Based on the collected data above and constructed viewers' profiles, the producer of the documentary can filter the available content based on the preferences of the targeted audience. For this purpose, the k-means algorithm [24] is used to create social clusters of users. Based on the generated clusters, a representative user profile is extracted and is used to perform the similarity matching of the group with the content in question. The SRP tool assigns a score to each item and ranks the items based on that score.

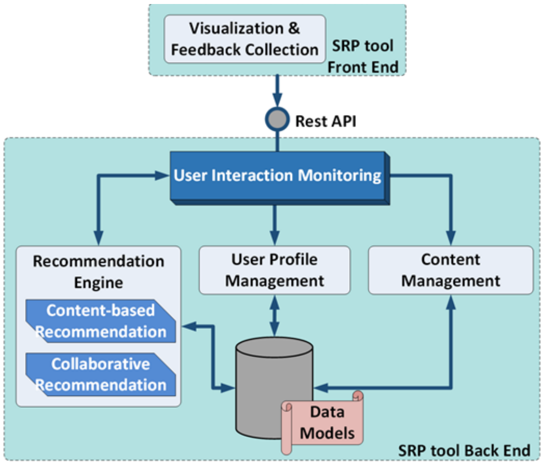

Visualization & Feedback Collection
SRP tool Front End
Rest API
User Interaction Monitoring
Recommendation Engine
Content-based Recommendation
Collaborative Recommendation
User Profile Management
Content Management
Data Models
SRP tool Back End

Figure 5. Architecture of the Social Recommendation & Personalization Tool.

After the creation of the documentary, the SRP tool can provide a filtering on the enrichments that are paired with the video, so that they do not overwhelm the viewer, filtering out less interesting enrichments. After specifying the target audience, the SRP tool can provide the list of suggested enrichments that the producer can either accept automatically or select manually based on his/her preferences, enabling the delivery of personalized documentary versions, tailored to audience interests.

SRP tool's architecture is presented in Figure 5.

B. *Evaluation & Benchmarking*

In order to perform an initial evaluation of the SRP tool, actual user studies were performed during the MECANEX project [25]. The targeted group of users requested to participate in the study where approximately 150 students from the National Technical University of Athens, because their technical, informatics and/or marketing background would be useful in evaluating the tool. Eventually, 40 subjects participated and successfully completed the provided questionnaire, mainly students at the Techno-economics Master's program, an interdisciplinary graduate program designed for professionals.

During the study, each user had to register to the system by providing a username and a password, as well as some demographic information (e.g., name, age, education). He/she could then explicitly choose some initial topics of interest, resulting in a diversified set of preferences that were used by the algorithm to perform some initial recommendations. Based on this initial profile, ten videos from a set of available 2500 videos were shown to the user, who could then choose which one to watch and interact with. Using the information regarding the user interactions with the content, the SRP tool updated the respective user's profile, and a new set of videos was provided to the user. The users were asked to stop using the system as soon as they believed they were ready to rate its quality. The overall results of the study are presented in Figure 6.

More specifically, in Figure 6.a we can see that for the majority of the users, the SRP tool succeeded in predicting their expected profile after the use of the system, with 55% rating the matching of their profile with 4 or 5 stars. The above results come as verification to the simulations of the effectiveness of the algorithm performed in [26]. The overall experience of the tool was also rated highly by the subjects (Figure 6.b) with more than 50% giving 4 or 5 stars rating once more, which indicates that the proposed SRP tool is a well performing recommendation system.

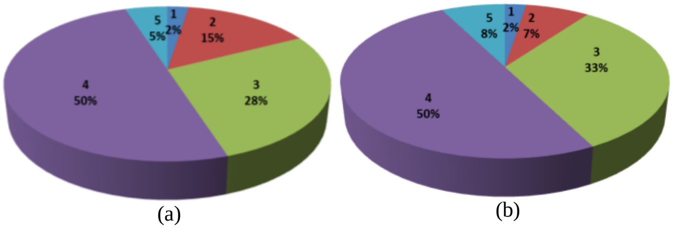


Figure 6. (a) Matching of final users' profile with their likes/preferences (1: Not really, 5: Matched exactly), (b) Rating of overall experience of the tool (1: Very Bad, 5: Very good).

More results concerning the impact and the effectiveness of the SRP tool in the MECANEX platform can be found in the public deliverable of the project [26]. Further, more thorough evaluation of the tool will be performed within the PRODUCER project's timeline.

IV. Conclusions

This paper introduced the PRODUCER platform for personalized documentary creation based on trend discovery. It briefly presented the set of tools offered by this platform, as well as its high level architecture. It then elaborated on two of its tools focusing on the targeted audience interests, identification and satisfaction. On the one hand, the ITD tool allows the identification of the most engaging topics to specified target audiences in order to facilitate professional users in the documentary preproduction phase. On the other hand, the SRP tool significantly improves the viewers' perceived experience via the provision of tailored enriched documentaries that address their personal interests, requirements and preferences. Initial prototype implementations of these tools are already available, while final prototypes will be delivered by spring 2018.

Both tools will be demonstrated and evaluated for a period of 3 months (March–May 2018) in an operational environment from an Italian broadcaster and a Belgium documentary production SME. This evaluation process will provide valuable feedback for further improving the overall functionality of the tools. Future plans also include the tools' integration with proprietary documentary production support services/infrastructures, as well as their extended evaluation and benchmarking against the various user requirements identified and against the Key Performance Indicators targeted (such as: cost reduction, time saving, increase of revenue in the entire documentary creation process).

Acknowledgment

This work has been partially supported by the European Commission, Horizon 2020 Framework Programme for research and innovation under grant agreement no 731893.

References

[1] The PRODUCER project. http://www.producer-project.eu, 2017. [*Retrieved January 2018*]

[2] G. Mitsis et. al, "Emerging ICT tools in Support of Documentary Production", 14th European Conference on Visual Media Production, 2017.

[3] J. Ginsberg, et. al, "Detecting influenza epidemics using search engine query data", Nature 457, pp. 1012-1014, 2009.

[4] A. J. Ocampo, R. Chunara, and J. S. Brownstein, "Using search queries for malaria surveillance, Thailand", Malaria Journal, Vol. 12, pp. 390-396, 2013.

[5] S. Yang, et. al, "Using electronic health records and Internet search information for accurate influenza forecasting", BMC Infectious Diseases BMC series, inclusive and trusted, Vol. 17, pp. 332-341, 2017.

[6] F. Ahmed, R. Asif, S. Hina, and M. Muzammil, "Financial Market Prediction using Google Trends", International Journal of Advanced Computer Science and Applications,Vol. 8, No.7, pp. 388-391, 2017.

[7] N. Askitas and K. F. Zimmermann, "Google econometrics and unemployment forecasting", Applied Economics Quarterly, Vol. 55, pp. 107-120, 2009.

[8] S. Vosen and T. Schmidt, "Forecasting private consumption: survey-based indicators vs. Google trends", Journal of Forecasting, Vol. 30, No. 6, pp. 565–578, 2011.

[9] S. Goel, J. M. Hofman, S.Lahaie, D. M. Pennock, and D. J. Watts, "Predicting consumer behavior with Web search", Natl Acad Sci USA, Vol. 107, No. 41, pp. 17486–17490, 2010.

[10] B. O'Connor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith, "From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series", International AAAI Conference on Weblogs and Social Media, pp. 122–129, 2010.

[11] M. X. Hoang , X. Dang , X. Wu , Z. Yan , and A. K. Singh, "GPOP: Scalable Group-level Popularity Prediction for Online Content in Social Networks", 26th International Conference on World Wide Web, pp. 725-733, 2017.

[12] A. Oghina, M. Breuss, M. Tsagkias, and M. de Rijke, "Predicting IMDB movie ratings using social media", 34th European conference on Advances in Information Retrieval Springer-Verlag, pp. 503-507, 2012.

[13] B. Bhattacharjee, A. Sridhar, and A. Dutta, "Identifying the causal relationship between social media content of a Bollywood movie and its box-office success-a text mining approach", International Journal of Business Information Systems, Vol. 24, No. 3, pp. 344-368, 2017.

[14] J.D. Burger, J. Henderson, G. Kim, and G. Zarrella. "Discriminating gender on Twitter", Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, pp. 1301–1309, 2011.

[15] A. Culotta, N. R. Kumar, and J. Cutler, "Predicting the Demographics of Twitter Users from Website Traffic Data", AAAI, pp. 72–78, 2015.

[16] Q. Fang, J. Sang, C. Xu, and M. S. Hossain, "Relational user attribute inference in social media", IEEE Transactions on Multimedia, Vol. 17, No. 7, pp. 1031–1044, 2015.

[17] Y. Fu, G. Guo, and T. S. Huang, "Age synthesis and estimation via faces: A survey", IEEE transactions on pattern analysis and machine intelligence, Vol. 32, No. 11, pp. 1955–1976, 2010.

[18] I. H. Witten, E. Frank, and M. A. Hall, "Data Mining: Practical Machine Learning Tools and Techniques" book (3rd Edition), Morgan Kaufmann Series in Data Management Systems, Burlington, MA, USA, 2011.

[19] J. Osofsky. After f8: Personalized Social Plugins Now on 100, 000+ Sites. https://developers.facebook.com/blog/post/382, 2010. [*Retrieved January 2018*]

[20] A. Micarelli and F. Sciarrone, "Anatomy and empirical evaluation of an adaptive web-based information filtering system", User Modeling and User-Adapted Interaction, Vol. 14, No. 2-3 (2004), 159–200, 2004.

[21] G. Gentili, A. Micarelli, and F. Sciarrone. Infoweb: An adaptive information filtering system for the cultural heritage domain. Applied Artificial Intelligence, Vol. 17, No. 8-9, pp. 715–744, 2003.

[22] Q. Le and T. Mikolov, "Distributed representations of sentences and documents", 31st International Conference on Machine Learning, pp. 1188–1196, 2014.

[23] E. Stai, S. Kafetzoglou, E. E. Tsiropoulou, and S. Papavassiliou, "A holistic approach for personalization, relevance feedback & recommendation in enriched multimedia content", Multimedia Tools and Applications, 1–44. 2016.

[24] J. MacQueen, "Some methods for classification and analysis of multivariate observations", 5th Berkeley symposium on mathematical statistics and probability, Vol. 1. Oakland, CA, USA., pp. 281–297, 1967.

[25] The MECANEX project. http://mecanex.eu/, 2016. [*Retrieved January 2018*]

[26] MECANEX Deliverable D2.2: Multimedia Content Annotations for Rapid Exploitation in Multi-Screen Environments, 2016.