Tracking and analyzing TV content on the web through social and ontological knowledge

Alessio Antonini, Ruggero G. Pensa, Maria Luisa Sapino, Claudio Schifanella, Raffaele Teraoni Prioletti, Luca Vignaroli
Tracking and analyzing TV content on the web through social and ontological knowledge
Editor: ACM Press
2013
ISBN: 9781450319515

EuroITV '13 Proceedings of the 11th european conference on Interactive TV and video
13 - 22
11th European Conference on Interactive TV and Video, EuroITV '13
Como, Italy
June 24-26, 2013

Tracking and Analyzing TV Content on the Web through Social and Ontological Knowledge

Alessio Antonini
University of Torino
Torino, Italy
antonini@di.unito.it

Ruggero G. Pensa
University of Torino
Torino, Italy
pensa@di.unito.it

Maria Luisa Sapino
University of Torino
Torino, Italy
mlsapino@di.unito.it

Claudio Schifanella
University of Torino
Torino, Italy
schi@di.unito.it

Raffaele Teraoni Prioletti
RAI-CRIT
Torino, Italy
raffaele.teraoni@rai.it

Luca Vignaroli
RAI-CRIT
Torino, Italy
l.vignaroli@rai.it

ABSTRACT

People on the Web talk about television. TV users' social activities implicitly connect the concepts referred to by videos, news, comments, and posts. The strength of such connections may change as the perception of users on the Web changes over time. With the goal of leveraging users' social activities to better understand how TV programs are perceived by the TV public and how the users' interests evolve in time, we introduce a knowledge graph to model the integration of the heterogeneous and dynamic data coming from different information sources, including broadcasters' archives, online newspapers, blogs, web encyclopedias, social media platforms, and social networks, which play a role in what we call the "extended life" of TV content. We show how our graph model captures multiple aspects of the television domain, from the semantic characterization of the TV content, to the temporal evolution of its social characterization and of its social perception. Through a real use-case analysis, based on the instance of our knowledge graph extracted from (the analysis of) a set of episodes of an Italian TV talk show, we discuss the involvement of the public of the considered program.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*clustering, information filtering*; H.5.1 [**Information Interfaces and Presentation**]: Multimedia Information Systems

Keywords

interactive television, social television, social networks, information integration

EuroITV'13, June 24–26, 2013, Como, Italy.

1. INTRODUCTION

In recent years the way users watch television is radically changing. With the introduction of digital television and the growing number of generic and thematic channels the final user tends to use new forms of navigation in the television content space. To enable users' navigation the broadcasters provide new enriched metadata services such as EPGs (Electronic Program Guide) which describe the scheduled programs. Also the home ambient environment in changing, many smart users watch television while using a portable PC or a tablet as secondary screen related or not to the broadcasted programs [7, 5].

At the same time, social networks allow the final user to be immersed in a collaborative environment providing a powerful reflection of the structure and dynamics of the society. User-generated contents are revolutionizing all phases of the value chain of contents: people can very easily produce content, they can distribute their produced material, and they can experience multiple forms of interactions, such as leaving comments, sharing opinions, supporting other users' contributions, or posting fragments extracted from already online material. We observe in particular that a very large number of user-generated content which users share in their social networks include significant portions of content already broadcasted by the TV broadcaster.

In this direction a number of Social TV services are emerging, which provide the final user with tools to support the social interaction while watching the television, or some media content related to a particular TV program. If properly leveraged, these collaborative social environments can be seen as information-rich data sources, indirectly returning to the broadcasters and the content producers some form of implicit feedback from the final users. A number of services, including user behavior profiling, brand reputation, and recommendation systems for contents and advertisements could benefit from the analysis of the social network data flow. In this paper we address the challenge of exploiting the information gathered from the users' activity in their social networks.

1.1 The extended life of television content

Television content evolves in time: after a television content is produced and broadcasted, a copy of it, enriched with its description (in natural language) together with a collec-

tion of related metadata is statically stored in the TV archive to be reused if needed. The broadcaster fills the EPG with the description. Big content producers also make their TV contents available in their Internet site. After a TV product is broadcasted, the content producers usually estimate the associated users' satisfaction by means of quantifiable data, such as audience data (in the case of television) or impression/view count in the case of the internet site, thus ending the "broadcasting phase" of the content's life.

Interestingly, for a significant portion of television content the life cycle spans much beyond that point. In fact, TV programs which better capture the users' interest will probably be published online, either entirely or, more often, in part. During this new phase the television content potentially attracts new users in the network. It will be watched, tagged, liked, commented about and shared again and again. The TV content will act as a magnet attracting the users in the network; it will become a "Social Object".

It is a fact that YouTube is (at least in Italy) the first place where people come to search for a television content they recently missed, although this might violate some copyright requirements. Content producers can choose to contrast this form of piracy, or they can choose to exploit it to their benefit as well, being aware that those users who upload TV content in YouTube are in fact conducting (for free) effective dissemination, description and publishing activities. The upload in Internet of a fragment of a television content starts its "extended life". As an example, we observe that it is very easy to find in YouTube segments of TV programs that have been uploaded long time ago and are now very often watched, commented and liked. This is the case for a short video of Roberto Benigni acting the Dante's "Divina Commedia" published by an unknown YouTube user in 2007 and still watched and commented very frequently in 2013. This content would otherwise be just stored in the archives of the broadcaster and it would be inaccessible to users. As time passes and the users' social context changes, the way any specific television content is perceived also evolves. Capturing how can the TV content evolve and detecting which phenomena can emerge from the contents' evolution are among the objectives of our work. For example, a content can attract a new community of users interested in it, or it might change its own meaning because of a new fact happened in the world. If timely discovered these phenomena could be leveraged by the broadcaster: some of the contents already available in the archives could be considered for repurposing and retransmission.

Some programs also undergo a "short term" evolution. This is the case for news talk shows: some uploaded video burst in number of views and comments. This gives to the public the opportunity to express their opinions about the program and the guests of the show. Analyzing the feedback provided by the public is potentially very beneficial for a number of stakeholders including broadcasters, content producers, ads and media companies.

1.2 Contribution

In this paper, we define a model for the integration of the heterogeneous data coming from the knowledge sources (broadcasters archives, EPGs, collected audience data, social networks, etc.) which play a role in the "extended life" of TV content, starting from its production phase, going through the on-air phase, and continuing with the on-line phase. The model highlights the tight interactions between the Web world and the TV world. A key characteristic of our model is that it is designed to be generic, and it enables a uniform treatment for the different information sources. More specifically, the integrated domain is modelled as a knowledge graph (Section 3.2), in which nodes represent the concepts, while edges capture the relationships existing among them. The key idea that we convey in our model is that the meaning of each entity and relationship within the knowledge graph depends on the context in which they are considered. Thus, persons might be considered as, for instance, authors, reviewers, consumers and so on. Context dependent qualification of entities is not limited to people. For example, a video concerning a piece of news may be regarded in different contexts as part of a news broadcast, a political comment, or a comic sketch (because of some anchorman's gaffe). Network actors and interactions are gathered from existing information sources. Users interact with each other using common social network/media platforms, more or less oriented to TV broadcasting (e.g., YouTube, Dailymotion, Facebook, Twitter, Google+). In this paper, we are interested in extraction and analysis of interactions that are related to TV contents, like videos, TV shows produced by some commercial/public broadcaster (such as RAI). Information from these sources is collected using standard search API's, web crawling techniques or, when possible, by means of social applications. Our framework supports the extraction of both metadata associated to the media contents and related information like user comments. Videos are posted by users by both uploading new content on video-sharing websites, or sharing other people's video content. Usually the original source of these videos are personal home recording. In this case, video content recognition algorithms to map videos to the exact part of TV shows they have been extracted from can be applied [3]. Moreover, our framework is able to support more reliable sources of information produced by a dedicated web-tv platform (e.g., Rai.tv) which enables users to extract portions of TV streams and post them on their preferred social/video-sharing platform. Hence, the original source of posted videos is certified and not subjected to errors. Moreover, citations, posts and comments related to this videos may be directly tracked by the TV service provider. Through a case study, we show how our model captures multiple aspects of the considered domain, from the semantic characterization of the TV content, to the temporal dimension of the problem, to the social characterization and the social perception of a TV event. Last but not least, we provide a non trivial cross-domain analysis scenario on real data gathered from YouTube and Twitter, and related to an Italian TV talk show on politics, broadcasted by RAI.

The paper is organized as follows: in Section 2 we briefly explore some related research work. In Section 3 we introduce the general structure of our integration framework and present the social graph which models the social-driven knowledge in the television context. In Section 4 we apply the model to a real use-case on TV-Web integration. We conclude the paper (Section 5) with a discussion highlighting the potential impact of the model, and showing how it can be used in a number of innovative applications.

2. RELATED WORK

Nowadays we can find an increasing number of emerging

services that aim to enhance the TV experience by offering both extra contents and social platforms on second-screen devices like tablets, smartphones and PCs [7, 5]. Among the most widely used we can mention different enterprise-level products:

- Rai.tv[1] gives to registered (Rai.tv, Facebook or Twitter) users Social TV events linked to the broadcasted television main stream in which the user can watch the program also via IP streaming, comments the program, interacts with comments from other connected users, expresses the feeling about the program, sees the liking of the other users, knows the argument in real time using the associated tag cloud, watches extra contents published by the editorial staff during the social event and answers to real time questions linked to the argument of the program. All these features are available in the web site and also on the secondary screen using the dedicated Apps;
- Tok.tv[2] enables friends to interact with each other within a virtual living room while watching American football matches on TV;
- GetGlue[3] lets users check-in to television shows;
- Miso[4] enables the users to create side shows to support user-generated content;
- IntoNow[5] serves contextual stories from the Web based on real-time mentions;
- Zeebox[6] provides an electronic program guide where the media content is weighted based on social network and also enables social engagement during the viewing experience.

A second screen interactive TV experience conducted by Basapur *et al.* [4] evidenced that the application prototype allows users to better connect with their TV shows and have an enriched social life around live as well as time-shifted TV content. This type of service opens a new kind of television usage and creates a new channel of information from the final user back to broadcasters and content producers. In this case the big challenge is: how can we use this new source of information? The trend of the exploitation of this new type of user interaction has one main direction: the exploitation of user activities to help the broadcaster and the content provider for the acquisition of new audiences, to help with the conservation of the audience and to help maximize the revenue produced by the audience. For these reasons a number of analytics tools are emerging in order to analyze the crowd buzz to try to extract information about the user behavior [8]. The idea of building a framework, based on a knowledge graph, able to capture and track the evolution of television content in the network is our attempt to give a novel approach to efficiently and effectively exploit the huge flow of information coming from social media. Our work is also inspired by a number of projects such as the NoTube project[7] which provided an integration framework between TV, Web, and Semantic Web to build services based on the enrichment and the personalization of the TV content [25, 24]. WinaCS [25] (Web-based Information Network Analysis for Computer Science) is another project that incorporates many recent developments in data sciences to construct a Web-based computer science information network and to discover, retrieve, rank, cluster, and analyze such an information network. However, the scope of WinaCS is limited to scientific content and digital bibliographies. From a more television oriented point of view, in the context of TV and social Web integration and in particular in the social media analytics tools, the recently instituted company Bluefin Labs[8] released a suite of tools to explore the social content related to Social TV programs analyzing the data generated by this mapping between social media and TV media, referred to as the TV Genome. This software is largely based on researches on natural language processing, speech to- text and video-entity recognition carried out by the two co-founders [10, 6].

[1] is a short preliminary version of this paper. While [1] is mostly focused on the presentation of the social and ontological knowledge integration framework, in this paper we discuss the use of the integrated data source, through the detailed analysis of use cases. In particular, our framework allows the application of most network analysis algorithms and tools, such as clustering [19], tensor factorization [14], analysis of diffusion and influence in social networks [12], recommender systems [26], and other social network analysis measures and methods [17].

3. A FRAMEWORK FOR SOCIAL MEDIA DATA INTEGRATION AND ANALYSIS

In this section we introduce the framework which enables the integration of various social and non social information sources in a unique knowledge base. The knowledge base, modelled as a knowledge graph integrating domain and general purpose ontologies as well as social interactions among users and social media, can be queried and analyzed as a whole, enabling the discovery of new and interesting cross-domain patterns.

3.1 The integration framework

Figure 1(a) presents an overview of our integration framework. It consists of three main layers: a source processing layer, a knowledge graph layer and a knowledge query and analysis layer.

The **source processing layer** has the role of collecting all the data which will be conveyed in the model. It accesses a number of predefined web/social/media sources (e.g., broadcasters official web sites, social networks, TV channels, etc) and processes them in order to extract those information units which will be represented as nodes in the knowledge graph, as well as those information that support the existence of relationships (modelled as edges in the graph) among them.

The **knowledge graph layer** manages the knowledge graph, which is the core of our proposal. The graph contains essentially three types of nodes: social objects, subjects and

[1] http://www.rai.tv/
[2] http://www.tok.tv/
[3] http://getglue.com/
[4] http://gomiso.com//
[5] http://www.intonow.com
[6] http://zeebox.com/tv/home
[7] http://www.notube.tv
[8] http://bluefinlabs.com/

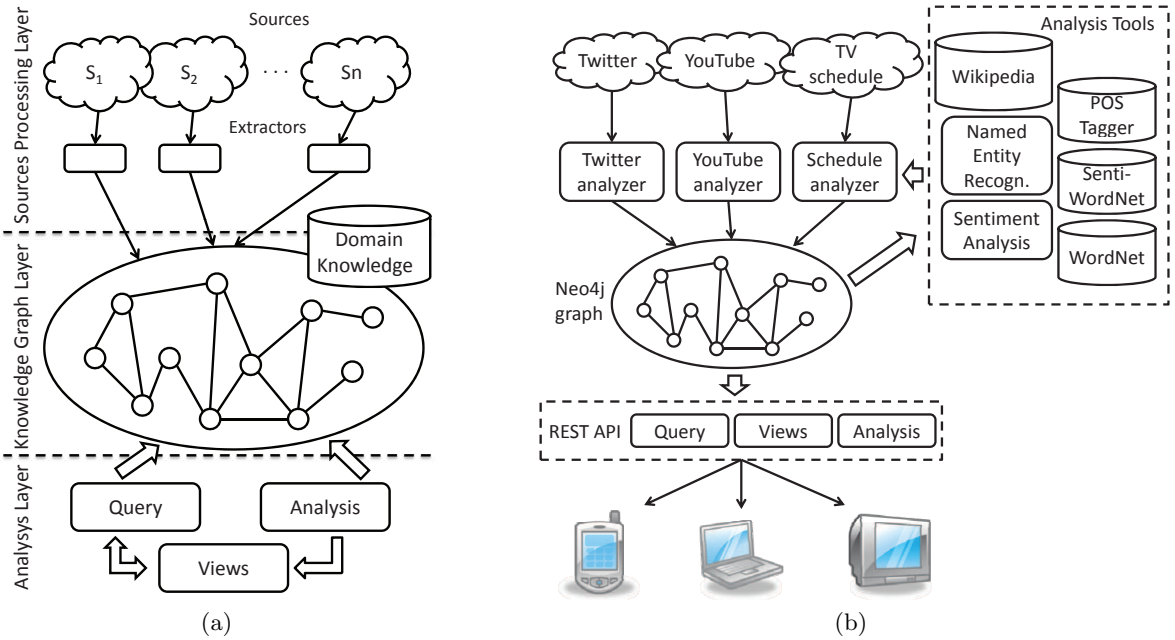


Figure 1: The integration framework and the related system architecture

concepts, and all social representations and structural interactions among them.

The **knowledge query and analysis layer** consists in a set of components for querying, browsing and analyzing the knowledge graph. A query module extracts subgraphs from the knowledge graph based on user's requirements and constraints. Each extracted subgraph can be seen as a "view" over the complete knowledge graph, only containing nodes and edges potentially relevant to the user query. An analysis module provides a set of analysis and data mining tools to obtain models and patterns from the knowledge graph. It can act directly on the knowledge graph, or it can handle the subgraphs extracted from the query module also in terms of matrices or tensors.

The core of our framework is the knowledge network. In particular, we are interested in capturing the dynamic evolution in time of the graph by using temporal nodes associated to social objects and describing their lifecycle.

Notice that in our integration framework a fundamental role is played by a *semantic engine* in two places. First, it is adopted in the source processing layer to provide an interpretation to web/social/media elements taken by the heterogeneous sources. Within this layer, the semantic engine helps understand whether the considered entities should be modelled as a node or an edge in the graph, and helps provide a congruent set of features based on their characteristics. Second, it plays an important role in the graph query and analysis layer, where it is employed to assign a semantic role to each selected node/edge.

Figure 1(b) depicts the actual architecture implementing our framework. In the following sections, we describe each layer in details.

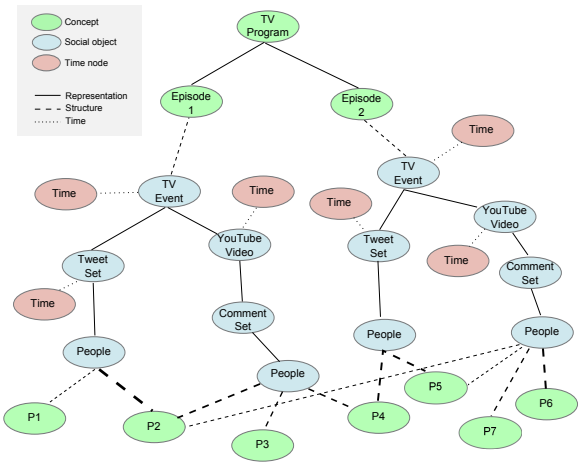
3.2 The knowledge graph

The core of our framework is the knowledge graph that represents the result of public actions of users in social environments, combining different theories from cognitive science [3, 16, 13, 22], language philosophy [20] and social ontology [9, 21]. In this domain we recognize three entities (corresponding to three types of nodes in the knowledge graph): *subjects*, users that act, *social objects*, the result of public acts, and *concepts*, physical and ideal objects referred to by subjects via their public actions. Any act (or set of acts) that can be identified by its trace, and has a recognized social value, is a social object. However, we do not represent single subjects' actions but a unique social object for each group of similar actions. A special subgraph is the one consisting of all concepts, i.e., the forest of the ontology of concepts, or the users' shared conventional knowledge.

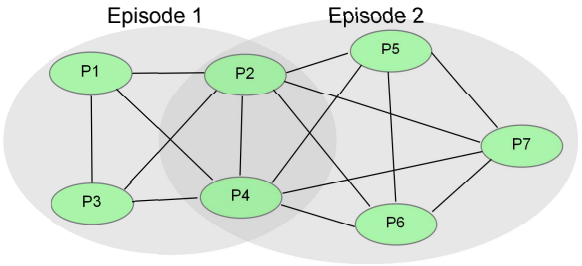
We introduce relationships between subjects and social objects and between social objects and concepts as follows: a group of subjects that recognize a social value of an act *supports* the resulting social object (e.g. the contractors *support* the contract); a social object *represents* a social instance of some concepts on a precise context (e.g. a video may represent a volleyball match). Other relationships involve entities of the same type. We call these relationships *structural dependencies*. A social object o_1 is *structural* of another object o_2 if o_1 is part of o_2 (e.g. a comment is part of a video). A subject is *structural* of a group of subjects (e.g. a subscriber is part of playlist subscribers) that performed the same kind of actions on the same social object. A concept may be *structural* of a more general concept (e.g. hilarity is a specialization of joy).

Finally, social objects evolve in time. Hence, as a special case of representation relationship, we consider the *temporal representation* of a social object against a special type of concept called *time objects* (e.g. a video has been posted in a specific time instant, and has been viewed during a specific time period).

The implementation of our knowledge graph is realized



(a)



(b)

Figure 2: A portion of the knowledge graph (a) and the resulting social network (b).

and stored in Neo4j[9], the well known NoSQL graph database: it offers a comprehensive REST interface, an object-oriented API, and it scales up to billions of nodes and relationships with properties. To populate the knowledge graph, our framework may interact with different and heterogeneous information sources. Each source is first analyzed, then relevant items and relationships are extracted and added to the graph. In the following, we explain how the sources of interest are analyzed.

3.3 Updating the knowledge graph

Our knowledge graph can be fed from any information source. However we distinguish between two kinds of sources: *social sources* and *non social sources*. The first ones consist essentially in social networking platforms, social media platforms and blogs. The second group of sources consists in general purpose or domain ontologies, online newspapers, news feeds, broadcasting websites that are needed to provide a human view on the results of social interactions. In our framework external sources are analyzed in order to extract resources that can be added to the knowledge graph following a set of specific rules.

For each source, we must set an extractor agent that should map each resource into a valid set of social objects, subjects, concepts and relationships among them. To correctly identify each entity, the extractor relies on a set of ontologies. To map each identified entity into a congruent set of vertices and edges in the graph, the extractor leverages a set of rules whose complexity depends on the specific source to be analyzed. In particular, as we mentioned earlier, we use two basic types of extractors: one for social sources, and one for non social sources.

Source Extractors.

Each source (both social and non social) is associated to an analyzer module (the boxes with solid line borders in Figure 1(b)), whose task is to collect the data from the sources and extract concepts, subjects, social objects and their relations through the combined use of different shared modules (the boxes with dotted line borders). The knowledge base extracted by each analyzer will be used to properly update the graph. More in detail, for each TV program that a Schedule Analyzer inserted in the knowledge graph, the Twitter module collects in real-time all related tweets, grouping them into time dependent slices, called tweet sets, where each slice contains the tweets published from time t to $t + \Delta$. Each tweet set is then processed in order to detect the named entities (people, places and events) trough the use of a NER (Named-Entity Recognition) module, while a Sentiment Analysis module allows to extract the opinions contained in a tweet set. Similarly, at each time slice, the YouTube analyzer looks for new videos or new user comments that belong to previously analyzed media and performs the same type of analysis described for Twitter.

Named-Entity Recognition.

Within the Named-Entity Recognition (NER) module, we can detect two different phases: entity detection and entity disambiguation [15]. Entity detection is performed by a combined use of the Freeling POS Tagger [18] and Wikipedia articles[10] as reference knowledge base. In particular, through the use of the Wikipedia search API, the NER module is able to detect the presence of entities starting from hashtags: for example, the hashtag *#barackobama* will be recognized by Wikipedia as the string "Barack Obama". Nevertheless, the most challenging task in Named Entity Recognition is represented by the entity disambiguation (or resolution) [15]. Since our scenario is characterized by the presence of short and sparse texts (both for Twitter and YouTube comments), many of the existing approaches based on the Bag of Words model will fail: for this reason our NER module tries to leverage additional information provided by the context defined by the TV program in which the resolution process is involved, in order to establish which entity is the best among the set of the candidate real-world entities. In details, the context of a TV program is defined by using the Wikipedia categories it belongs to and the set of all entities contained in the knowledge graph previously associated with the program. In this manner, for each detected entity, the NER module tries to establish an order among all real-world candidates extracted from Wikipedia. For example, if the text "Michael Jordan" is contained in a tweet set related to a TV sports program, it is very likely that the tweeter is referring to the famous basketball player rather than the Berkeley's professor, and this is computed by a comparison between the Wikipedia categories of the candidates and the corresponding categories of the TV program. Moreover, if, for example, Michael Jordan is present within the knowledge graph as a real-world entity recognized and associated

[9] `http://www.neo4j.org`

[10] `http://www.wikipedia.org`

with the considered TV program (i.e. because he is the presenter or a frequent guest), the NER module will choose it among all the possible real-world entity candidates. Finally, our module supports the integration of external knowledge generated by a supervised scenario and it allows for user feedback, using an active learning process. In our application, we filter out infrequent recognized entities with the energy cutoff method.

Sentiment Analysis.

The Sentiment Analysis module is used to extract polarity values and emotions from tweet sets. Concerning the former, a first phase of lemmatization is performed by the Freeling POS tagger, while SentiwordNet [2] is used to extract the polarity values: hence, an aggregation function allows us to enrich each tweet set in the knowledge graph with a degree of positivity, negativity and neutrality. With the same approach, WordNet-Affect[23] is used to extract emotions. Where necessary, MultiwordNet[11] is used for cross-language purposes.

Once the extractor agent has analyzed the source, it provides a set of concepts, subjects and social objects that should now be translated into new or updated vertices and edges in the graph. Thanks to this structure, it becomes possible to extract new cross-domain patterns.

3.4 Knowledge Graph in TV domains

In real applications, the graph will not be instantiated with all possible resources extracted from any social or non social source. The reasons are essentially twofold: on the one hand, the huge amount of information could be untractable in practice; on the other hand, many social sources set a limit to the number of resources that can be retrieved in a time slice. For this reason, the way the knowledge graph is populated is somehow constrained by the specific application. We come back now to our case study.

The instantiation of the framework to the TV domain involves a decision process in which we have to choose and define the social and non-social sources, define the resource prototypes and the policies for the source analyzers and decide the detail level of the representation. This last decision depends on what we can extract from social sources, what we want to know about the domain and what we can know about the users' actions. In the next section, we present our solution for the social TV domain. In our view, a user can create and enrich new social uses of the TV media with new metadata, comments, tags, sharing actions and rates. The objects that we will detect and capture are the new correlations introduced between a canonical description of a television event and other new, possibly surprising, concepts. In a nutshell, we want to audit the evolution of the social perception of TV events. The choice of the non-social sources is also critical for the domain definition because it contributes to form the core of the monitored topics on the social sources.

4. A CASE STUDY ON ITALIAN POLITICS

In this section, we describe a real use-case of our framework on an Italian TV show (Ballarò) dealing with politics and broadcasted by RAI. We focused our analysis on the episodes scheduled from October 2, 2012 to November 27, 2012 (nine episodes). This period is interestingly full of political events for many reasons: the past or future elections in many big Italian regions (Sicily, Lazio and Lombardy); the upcoming Italian general elections; the recession; the rise of the populist extra-parliamentarian group M5S (Movimento

[11] http://multiwordnet.fbk.eu

Table 1: Top betweenness centrality scores of nodes from Twitter social network in Fig. 3(a)

Rank	Person	Betweenness centrality
1	Maurizio Crozza	0.2410
2	Mario Monti	0.1783
3	Giovanni Floris	0.0901
4	Matteo Renzi	0.0597
5	Silvio Berlusconi	0.0235
6	Gianfranco Polillo	0.0207
7	Leoluca Orlando	0.0192
8	Bruno Tabacci	0.0172
9	Pier Luigi Bersani	0.0161
10	Luigi Angeletti	0.0135
11	Beppe Grillo	0.0104
12	Guido Crosetto	0.0102
13	Massimo Giannini	0.0051
14	Roberto Formigoni	0.0029
15	Concita De Gregorio	0.0016

Table 2: Top betweenness centrality scores of nodes from YouTube social network in Fig. 3(b)

Rank	Person	Betweenness centrality
1	Beppe Grillo	0.3413
2	Pier Luigi Bersani	0.1907
3	Matteo Renzi	0.1307
4	Giovanni Floris	0.1271
5	Mario Monti	0.1107
6	Gianfranco Fini	0.0512
7	Pier Ferdinando Casini	0.0484

Table 3: Top betweenness centrality scores of nodes from the combined social network in Fig. 3(c)

Rank	Person	Betweenness centrality
1	Maurizio Crozza	0.1710
2	Beppe Grillo	0.1710
3	Mario Monti	0.1305
4	Giovanni Floris	0.0954
5	Pier Luigi Bersani	0.0868
6	Matteo Renzi	0.0723
7	Gianfranco Polillo	0.0259
8	Silvio Berlusconi	0.0236
9	Bruno Tabacci	0.0234
10	Gianfranco Fini	0.0217
11	Luigi Angeletti	0.0213
12	Leoluca Orlando	0.0202
13	Pier Ferdinando Casini	0.0190
14	Guido Crosetto	0.0188
15	Roberto Formigoni	0.0176
16	Concita De Gregorio	0.0165
17	Massimo Giannini	0.0162
18	Rosario Crocetta	0.0162
19	Alessandro Sallusti	0.0162
20	Gianni Alemanno	0.0162

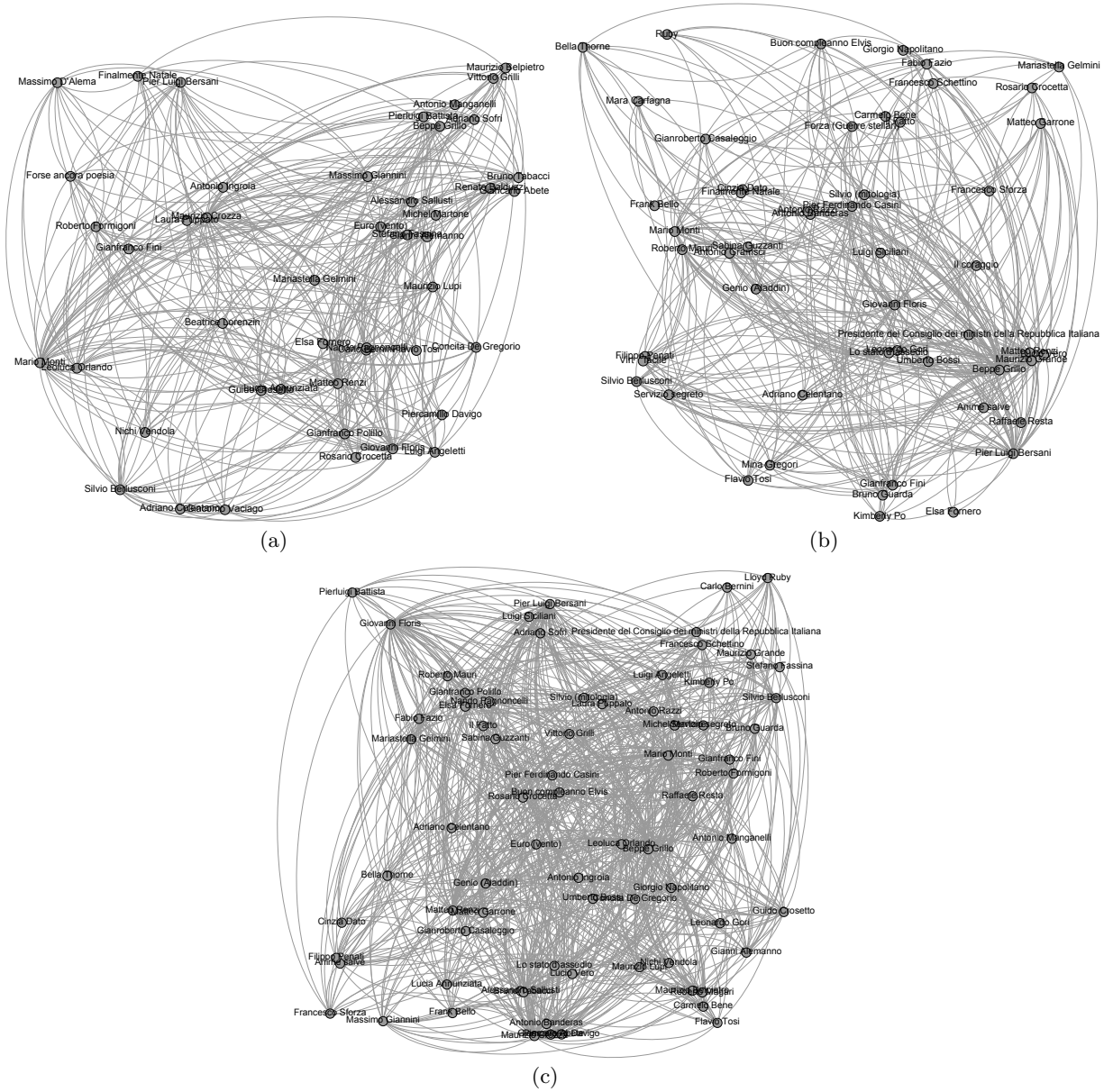


Figure 3: Ballarò social networks extracted from Twitter (a), Youtube (b) and both sources (c)

5 Stelle) that many polling institutes were considering as one of the favorite parties for the next elections in Italy.

We considered two social sources: Twitter and YouTube. For each episode, we collected all tweets containing *#Ballarò* (the official program hashtag) or *@RaiBallaro* (the official program username). YouTube videos were extracted at once by including in the search fields the keyword related to the TV program title ("Ballarò") and the date each episode was broadcasted (e.g., "2-10-2012" or "2 ottobre 2012").

4.1 Social Centrality Study

The first example we consider here concerns the study of the importance (in terms of centrality) of persons (politicians, television people, presenters, hosts), during the observation period. To perform this analysis, we consider all the persons referred by the tweets and videos associated to the nine episodes of the TV show and build the underlying social network. Figure 2 shows how we extracted the social network involving TV people. We add an edge between two person nodes if there exists a path between these two persons, traversing at most one *People* node or at most one *TV Event* node. For instance, following our decision, in Figure 2(a), there is a path between $P4$ and $P6$, but no path between $P1$ and $P7$ exists. Consequently, in Figure 2(b), $P4$ and $P6$ are connected, while $P1$ and $P7$ are not. Notice also that these paths may involve cross-source nodes, i.e., the analysis of an individual source, without our knowledge integration framework, would have led to a different, less precise, social network. Since more than one tweet set and YouTube video may exist during the week associated to each episode, for each episode, all the tweet sets and YouTube videos have been merged to obtain an aggregated episode representation. Each of them is then associated to the set of the most mentioned persons during the considered week.

On our Twitter data, the above described analysis produced the social network presented in Figure 3(a). By computing the betweenness centrality [11, 17] of each node (i.e., the number of shortest paths from all vertices to all others that pass through that node), we obtain the results in Table 1. These results show that Maurizio Crozza is very central for this TV program. He is a satirist that leads a 10 minutes' intervention during each episode of Ballarò TV programs. As such, he usually performs imitations of politicians (like Pierluigi Bersani and Matteo Renzi). Mario Monti, the Italian Prime Minister when these episodes were broadcasted, has been ranked second even if he never participated to the show during the observation period. Among less known politicians, Crosetto (ranked twelfth), had a certain popularity during that period, since he was creating a new political party, in disagreement with Silvio Berlusconi. Among the other top-ranked people, Giovanni Floris is the presenter of Ballarò, while Pier Luigi Bersani and Matteo Renzi were the two main competitors for the leadership of the center-left party, during the observation period.

The same analysis conducted on YouTube data, produced the social network in Fig. 3(b). The betweenness centrality computed for different TV people belonging to this network is reported in Table 2. Interestingly, this analysis shows that the best ranked person is Beppe Grillo. This is probably due to the fact that Grillo's supporters are particularly active in this social media platform. Thus, in this social network, the position of Grillo is more central than in the previous one.

By combining the two information sources (see the social network in Figure 3(c)), we may notice that all the relevant information for both sources are preserved, as shown by the ranked betweenness scores in Table 3. In particular, Grillo and Crozza are equally central, Prime Minister Mario Monti is still in a privileged position, while almost all the most important Italian politics actors are in the first positions of the ranking.

4.2 Popularity Study

The second experiment consists in computing the "episode popularity" of each person. The popularity of a given node is related to the percentage of citations of the associated persons' names in tweets and YouTube comments. Notice that this information is stored in the knowledge graph as the weight of the edge connecting each person to the *People* node (see Figure 2(a)), by the resource extractors. Hence, to conduct this analysis, we only need to aggregate the weights of the out-edges of each person node. Within a single source the aggregation is performed by merging all social objects (tweet set or YouTube video) related to a given episode. Then, each edge weight is multiplied by the total number of occurrences of the concept node *People*. Finally, the cut-off method based on energy is employed to filter out the less important entries. To consider the popularity in both Twitter and YouTube as a whole, we merged the YouTube video nodes and Tweetset nodes associated to each episode. The resulting weight for each person node i is then computed as $w(i)_{all} = \alpha \cdot w(i)_t + (1 - \alpha) \cdot w(i)_y$ where $w(i)_t$, $w(i)_y$ and $w(i)_{all}$ are, respectively, the node weights of the edge connecting i to the Tweetset node, the node weights of the edge connecting i to the YouTube video node, and the resulting weight associated to the edge connecting i to the aggregated social object node. In this experiment, we considered all sources with the same weight, i.e., $\alpha = 0.5$.

Figure 4 shows the results for the top-ranked personalities, as computed before. While most popularity values are quite stable during the observation period, the popularity of Matteo Renzi has two peaks, corresponding to the two episodes in which he was hosted in the show. We may also observe that Renzi is more popular on Twitter while Grillo appears to be mentioned more often on YouTube. Berlusconi is almost never mentioned: during the observation period, in fact, he was still not expected to be a key candidate of the center-right party campaign. Monti is mentioned regularly every week, except during the October 30 episode, when Renzi reports the first peak. In those days, in fact, the primary election of the center-left party took place, and Renzi was one of the most observed candidates because he is young and dynamic, and he effectively uses social media and the Web. Bersani, another primary elections candidate, is mentioned regularly but he does not warm the hearts of Web users. The third important candidate is Vendola who seems quite ignored by the Web audience. This may have two explanations: first, his party does not attract many votes and, most interestingly, he is mostly active on Facebook (which we didn't analyze in the discussed use case).

5. DISCUSSION AND CONCLUSIONS

In this paper we have proposed a model for the integration of the heterogeneous data coming from many different knowledge sources, including broadcasters archives, EPGs, ontologies, and social networks. The model highlights the tight interactions between the Web world and the TV world. We have also provided a concrete example of the potential applications of our framework on real data.

We expect the model will have a significant impact on the television production environment. In particular, the ability to track and monitor the second life of Television content will be useful to a number of stakeholders.

- **Broadcasters**: The framework allows the broadcaster to add new references to the static big legacy archive, thus enabling archivists to have a new vision of the evolution of contents that are now frozen inside a huge data base. This new feature makes it easier the personalization of already broadcasted services in which content is customized, adapted to preferences and characteristics of single users or groups of users and provided again to them, exploiting at best the "long tail" phenomena relevant to its own television content. From this viewpoint the broadcaster has the opportunity to reuse materials already exploited by services formerly provided to users, maximizing business logics for these contents that otherwise would be exploitable just in the short term. Furthermore the framework could be useful to provide an alternative approach to the audience analysis of television programs giving more punctual suggestions to optimize the schedule of programs.

- **Service providers** A generic service provider, that only rearranges contents owned by other subjects, will be able to provide new pay services starting from already broadcasted content enforced with a big variety of related content also coming from other media.

- **Final users**: The final user enjoys indirect benefits coming from the use of the model by the broadcaster

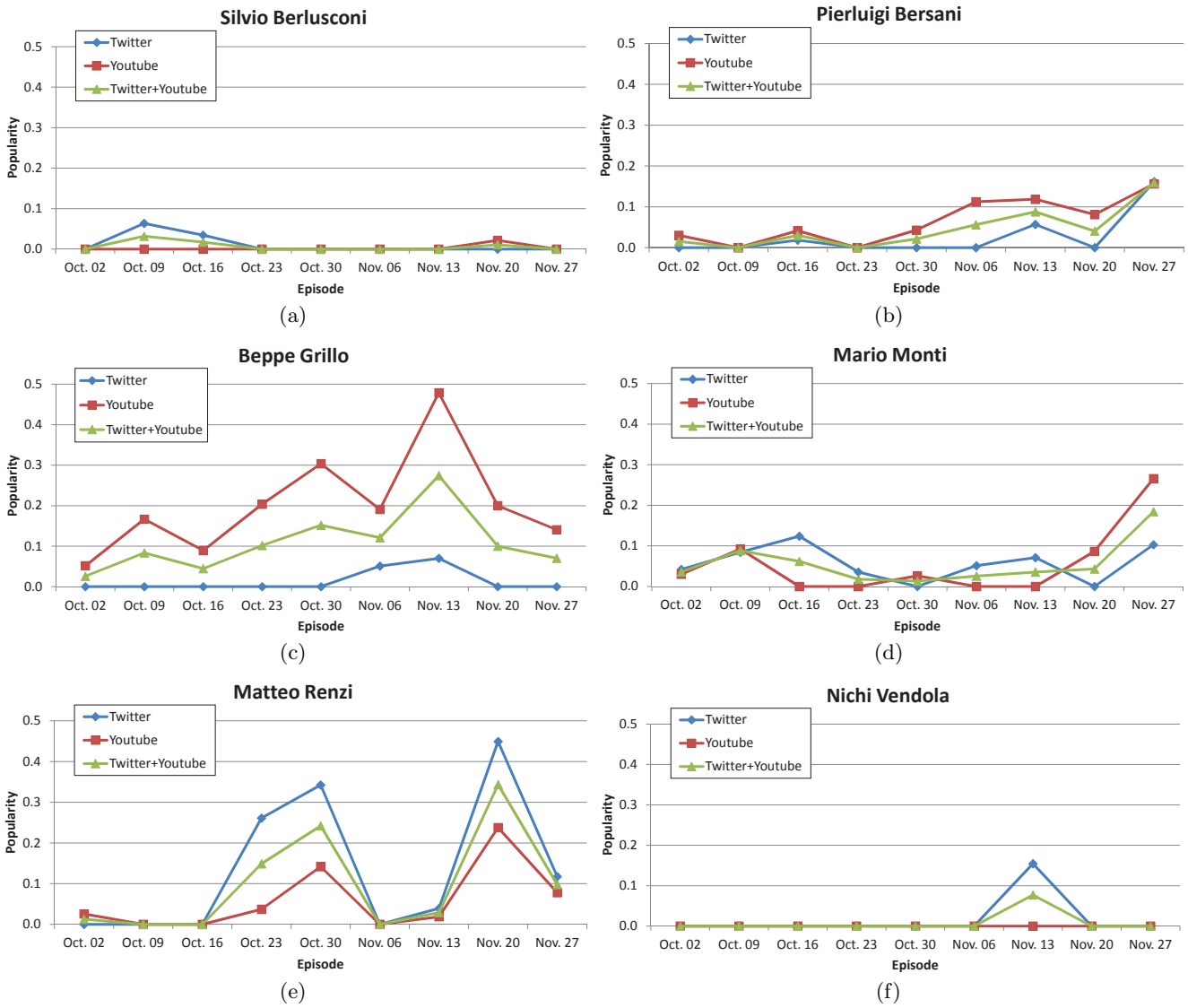


Figure 4: Episode popularity of some cited persons from our knowledge graph

or the service provider. In fact, he/she is able to interact with multimedia material of his own interest in a user suitable format and in time-independent and context-aware modality. Furthermore, the framework could be employed to enhance the interactive TV and "second screen" experiences, by enabling cross-source recommendation techniques, e.g., recommendation of YouTube videos triggered by the usage of particular Twitter hastags [1].

- **Research and industry**: From a research perspective, our framework provides a standard process for data gathering and analysis from different social media sources, thus enabling novel ways to approach sociological studies on the behavior of TV audiences. Moreover, it will relieve data scientists of the ungrateful task of designing ad-hoc data gathering techniques for testing their algorithms and proving their hypotheses. Additionally, from an industrial perspective, the framework could be employed as the underlying architecture for the development of new dedicated services and applications.

Future works will address some limitations of the current architecture. In particular, the correct identification of concepts in the knowledge graph lies in the accuracy of the named-entity recognition module. However, resolution of ambiguities is still an open problem involving information extraction, data mining, natural language processing and other related techniques. We believe that our knowledge graph may be employed to guide the correct identification of persons, places, emotions, events, and other relevant concepts. Hence, we will investigate new active learning techniques for the resolution of ambiguities leveraging the content of the graph, and thus minimizing the intervention of human experts in the named-entity recognition process.

Another weakness of our framework is due to the fact that source analysis is guided by experts that define the correct queries and may possibly adapt them to the new

trends and/or needs. We will study self-adaptive strategies to automatically identify emerging keywords and add them to source analyzer queries, while removing obsolete ones.

Finally, since our knowledge graph is able to capture relationships among different types of information, we will investigate new data analysis and mining techniques that take into account the complexity and heterogeneity of the networks.

6. ACKNOWLEDGMENTS

We are grateful to Roberto Del Pero and Fulvio Negro from RAI CRIT for their constructive discussions during the formalization of the integration framework.

7. REFERENCES

[1] A. Antonini, L. Vignaroli, C. Schifanella, R. G. Pensa, and M. L. Sapino. MeSoOnTV: A media and social-driven ontology-based tv knowledge management system. In *Proc. of 24th ACM Conference on Hypertext and Social Media, HT'13, 1–3 May 2013, Paris, France*. ACM, 2013.

[2] S. Baccianella, A. Esuli, and F. Sebastiani. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proc. of LREC 2010, 17-23 May 2010, Valletta, Malta*, 2010.

[3] B. Bara. *Cognitive Pragmatics : The Mental Processes of Communication*. MIT Press, 2010.

[4] S. Basapur, H. M. Mandalia, S. Chaysinh, Y. S. Lee, N. Venkitaraman, and C. J. Metcalf. Fanfeeds: evaluation of socially generated information feed on second screen as a tv show companion. In *Proc. of 10th European Conference on Interactive TV and Video, EuroITV'12, Berlin, Germany, July 4-6, 2012*, pages 87–96. ACM, 2012.

[5] P. César, D. C. A. Bulterman, and A. J. Jansen. Usages of the secondary screen in an interactive television environment: Control, enrich, share, and transfer television content. In *Proc. of 6th European Conference, EuroITV 2008, Salzburg, Austria, July 3-4, 2008*, volume 5066 of *LNCS*, pages 168–177. Springer, 2008.

[6] P. DeCamp and D. Roy. A human-machine collaborative approach to tracking human movement in multi-camera video. In *Proc. of 8th ACM International Conference on Image and Video Retrieval, CIVR 2009, Santorini Island, Greece, July 8-10, 2009*. ACM, 2009.

[7] M. Doughty, D. Rowland, and S. Lawson. Co-viewing live tv with digital backchannel streams. In *Proc. of 9th European Conference on Interactive TV and Video, EuroITV'11, Lisbon, Portugal, June 29-July 1, 2011*, pages 141–144. ACM, 2011.

[8] M. Doughty, D. Rowland, and S. Lawson. Who is on your sofa?: Tv audience communities and second screening social networks. In *Proc. of 10th European Conference on Interactive TV and Video, EuroITV'12, Berlin, Germany, July 4-6, 2012*, pages 79–86. ACM, 2012.

[9] M. Ferraris. Documentality or why nothing social exists beyond the text. In *Cultures. Conflict - Analysis - Dialogue, Proc. of 29th International Ludwig Wittgenstein-Symposium, Kirchberg, Austria, August 6-12, 2006*, pages 385–401. Austrian Ludwig Wittgenstein Society, 2006.

[10] M. Fleischman and D. Roy. Grounded language modeling for automatic speech recognition of sports video. In *Proc. of 46th Annual Meeting of the Association for Computational Linguistics, ACL 2008, June 15-20, 2008, Columbus, Ohio, USA*, pages 121–129, 2008.

[11] L. C. Freeman. A set of measures of centrality based on betweenness. *Sociometry*, 40(1):pp. 35–41, 1977.

[12] M. Gomez-Rodriguez, J. Leskovec, and A. Krause. Inferring networks of diffusion and influence. *TKDD*, 5(4):21, 2012.

[13] P. N. Johnson-Laird. *Mental Models*. Cambridge University Press, 1983.

[14] T. G. Kolda and B. W. Bader. Tensor decompositions and applications. *SIAM Review*, 51(3):455–500, September 2009.

[15] H. Kopcke and E. Rahm. Frameworks for entity matching: A comparison. *Data and Knowledge Engineering*, 69(2):197 – 210, 2010.

[16] G. Lakoff. *Women, Fire, and Dangerous Things: What Categories Reveal About the Mind*. University of Chicago Press, 1987.

[17] M. Newman. *Networks: An Introduction*. Oxford University Press, 2010.

[18] L. Padró and E. Stanilovsky. Freeling 3.0: Towards wider multilinguality. In *Proc. of LREC 2012, Istanbul, Turkey, May 23-25, 2012*, pages 2473–2479, 2012.

[19] S. E. Schaeffer. Graph clustering. *Computer Science Review*, 1(1):27–64, 2007.

[20] J. R. Searle. *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press, 1970.

[21] J. R. Searle. *The Construction of Social Reality*. Free Press, 1997.

[22] R. E. Shaw and J. E. Bransford. *Perceiving, acting, and knowing: Toward an ecological psychology*. Lawrence Erlbaum, 1977.

[23] C. Strapparava and A. Valitutti. Wordnet affect: an affective extension of wordnet. In *Proc. of LREC 2004, May 26-28, 2004, Lisbon, Portugal*, 2004.

[24] L. Vignaroli, R. D. Pero, and F. Negro. Personalized newscasts and social networks: a prototype built over a flexible integration model. In *Proc. of WWW 2012, Lyon, France, April 16-20, 2012 (Companion Volume)*, pages 433–436. ACM, 2012.

[25] T. Weninger, M. Danilevsky, F. Fumarola, J. M. Hailpern, J. Han, T. J. Johnston, S. Kallumadi, H. Kim, Z. Li, D. McCloskey, Y. Sun, N. E. TeGrotenhuis, C. Wang, and X. Yu. Winacs: construction and analysis of web-based computer science information networks. In *Proc. of SIGMOD 2011, Athens, Greece, June 12-16, 2011*, pages 1255–1258. ACM, 2011.

[26] X. Yang, H. Steck, and Y. Liu. Circle-based recommendation in online social networks. In *Proc. of KDD '12, Beijing, China, August 12-16, 2012*, pages 1267–1275. ACM, 2012.