Towards an Intelligent Framework to Understand and Feed the Web

Anna Fensel[1], Julia Neidhardt[2], Nataliia Pobiedina[2], Dieter Fensel[1], and Hannes Werthner[2]

[1] Semantic Technology Institute (STI) Innsbruck, University of Innsbruck, Technikerstrasse 21a, 6020 Innsbruck, Austria,
{anna.fensel,dieter.fensel}@sti2.at
http://www.sti-innsbruck.at

[2] E-Commerce Group, Institute of Software Technology and Interactive Systems, Vienna University of Technology, Favoritenstrasse 9-11, 1040 Vienna, Austria,
{julia.neidhardt,pobiedina,hannes.werthner}@ec.tuwien.ac.at
http://www.ec.tuwien.ac.at

Abstract. The Web is becoming a mirror of the "real" physical world. More and more aspects of our life move to the Web, thus also transforming this world. And the diversity of ways to communicate over the Internet has enormously grown. In this context communicating the right thing at the right time in the right way to the right person has become a remarkable challenge. In this conceptual paper we propose a framework to apply semantic technologies in combination with statistical and learning methods on Web and social media data to build a decision support framework. This framework should help professionals as well as normal users to optimize the spread of their information and the potential impact of this information on the Web.

Key words: Social Media, Semantics, Web Mining, Online Marketing

1 Introduction

More and more aspects of our life move to the Web; the Web – as the underlying information infrastructure of today – is becoming a mirror of the "real" physical world. And it is obviously transforming this world, where it is hard to distinguish between the physical and the virtual. With developments such as the Semantic Web (moving from the syntax to the concept level) information can be automatically processed, and approaches such as Web analytics and network analysis facilitate statistical / logical inference of new knowledge, moving from data to knowledge. In this context the Web also emerged into a medium where users have become both the most important content consumers as well as the most important content producers – leading to so-called "prosumers".

This development has provided a multitude of interaction possibilities, leading to "the growth of the multichannel monster" [1]. Taking an organizational / business point of view one sees that organizations of all sizes, commercial and

not-for-profit, face the challenge of communicating with their clients and partners using a multiplicity of channels, e.g., Web sites, videos, PR activities, events, email, forums, online presentations, social media, mobile application as well as structured data. The social media revolution has made this job much more challenging. The number of channels has grown and communication has changed from a mostly unilateral "push" mode to an increasingly bilateral communication, where individual stakeholders expect one-to-one communication. And the content of communication becomes more granular and increasingly dependent on the identity of the receiver and the context of the communication.

Having such a huge amount of different communication channels as well as this overabundance of information raises crucial questions: Which communication channel should I use? How should I use them? What do I communicate? How does it scale? Current tools only support the simple feedback analysis of social media data. And a straight-forward automation of the information publishing to numerous channels, as currently state of the art, does not cover very complex aspects in the decision making on where to publish the information, particularly, with respect to the optimization of brand and reputation management.

The central issue is to manage and reduce complexity in this Web x.0 world, which we intend to achieve with our framework SCAN (Social Channel Analysis and Networking). The proposed framework is currently under construction and throughout the paper we describe the generic approach and outline the challenges.

SCAN is a smart combination of semantic technologies as well as statistical and learning methods, and achieves the following: It 1) models and analyses the communication in different online channels, 2) forecasts and measures impact of the performed communication and, 3) based on the model and measurements, delivers suggestions which communication channel a user should use in order to increase the reach and spread of information. The evaluation of the framework is based upon the metrics which are the outcome of the impact analysis.

The SCAN framework will decrease the costs of driving different social media marketing campaigns. Furthermore, these activities may lead to higher conversions in online bookings and higher revenues. Since the developed system core is generic and extensible to any type of SME, our approach can be applied to different economic sectors.

The structure of the paper is as follows. In Section 2, we discuss the related work and position our approach within its landscape. In Section 3, we describe the general framework set up and its architecture. In Section 4, we present our approach to "feed" the Web, and in Section 5 our approach to "understand" the Web. Section 6 concludes the paper.

2 State of the art

Here, we discuss the state of the art approaches from the two major activities of SCAN, namely, understanding and feeding the Web.

Modeling the flow and change of information for analysis and forecast. This is a multidisciplinary research, therefore the literature required to follow current state of the art covers Web Science [2], evolutionary modeling, graph transformation systems and text mining.

In [3] the topological structure of the Web has been presented as a directed graph with Web pages as nodes and hypertext links between Web pages as edges between nodes. In [4] and [5] the evolution of the Web graph over time is studied, and some patterns on the emergence of new nodes and edges are presented. Though the "Web graph" model has proven to be efficient to investigate certain topological properties of the Web, e.g., in-degree and out-degree of the nodes follow the power law, it does not take into account the content placed on the Web pages. On the other hand, in [6] and [7] authors focus on the investigation of temporal dynamics of the content. However, the content under analysis is limited to the phrases extracted from quotes, and the Web structure is not taken into account.

In her survey [8], Berendt outlines the major challenges of text mining for online media. She also provides taxonomy of problems in this area. Despite the considerable successes in the areas mentioned above, the major challenge in dynamic modeling of Web documents still remains the problem how to efficiently account both for the content and structure information of Web documents over time. Currently graph transformation systems [9] are successfully applied in the areas of software engineering and model checking. We perceive stochastic graph transformation systems [10], which provide a generic framework to consider both content and structure information of the system, can be applied to the dynamic modeling of Web documents.

Content Dissemination and Impact Analysis. The field of semantics-based or enhanced Content Management Systems (CMSs) has already been quite thoroughly explored. One of the earlier approaches to ontology-based Web site management was the OntoWebber system described in [11]. OntoWebber introduced an integration layer which adapts to different data sources. This is related to our approach, but, in contrast, our approach adapts to different channels rather to different information sources. A year later, in [12], Sheth et al. introduced the SCORE system, which defines four key features: semantic organization and use of metadata, semantic normalization, semantic search, and semantic association. Although being written in the early days of the Semantic Web, the paper covers topics such as metadata extraction from unstructured text and automatic classification that may also become relevant to our approach.

The British national broadcaster BBC started to integrate semantic technologies (i.e., Linked Data) in 2009 in order to integrate various data and content sources distributed throughout the enterprise [13]. As a result of this, reported in [14], BBC's World Cup 2010 site is based on semantic repositories that enable the publishing of metadata about content rather than publishing the content itself. While the data input is fixed, different schemas for the output are defined. However, as only one channel for output is considered, the mapping is performed in a quite straightforward manner. In contrast, our system accounts for different

information needs of various and heterogeneous channels and therefore enables the distribution of content throughout different portals.

Marketing plays an important role in the prosperity of a company, and social media is recognized as one of the efficient ways to lower costs for marketing campaigns [15]. However, social media based promotions are less controllable by managers compared to traditional marketing strategies [16]. Therefore, managers are in need of tools and methods to organize promotional activities in social media to reach the performance goals. Initiatives like HootSuite, HubSpot and others (see Table 1) provide toolkits to manage communication via Web 1.0 channels (Email, Blog) and via social Web 2.0 channels, e.g. Facebook or Twitter. Many of them provide capabilities to post in many streams via one clique, using simple mechanisms to adapt the content with regard to channel specifics. Additionally, most of the toolkits allow the user to see activity statistics and retrieve feedback. However, these toolkits neither support the user in selecting the communication channel and the target audience in it, nor provide rules to adapt the content to optimize the efficiency of the campaign.

Moreover, managers require evidence of return on investment in such activities [17]. These questions have been raised in business environment and gained much interest from the side of behavioral scientists, but they still lack thorough quantitative research. In [18] scientists argue that the design and implementation of social media campaigns is specific to national source markets. They provide guidelines for marketers from Russia and FSU countries concerning social media. However, to our best knowledge, there are no up-to-date studies of influence of social media on the Austrian market.

3 Overall framework and architecture

Our framework SCAN supports effective social media marketing campaigns by combining analysis of distributed information on the one side and a decision-making process on the other side. In other words, the framework helps to choose "how" (specific phrases/keywords/images/mode), "where" (specific channels) and "when" (specific time) to deliver the information – assuming that the knowledge of "what" (information on the conceptual level) to deliver and "to whom" (target category of users according to certain criteria, for example according to gender, age, location, etc.) is available. To solve this problem, we focus on the following technical subproblems:

1. Semantic modeling of the content and its integration for information on multiple channels as well as the related dissemination mechanisms;
2. Monitoring, analyzing and forecasting information flow in social media, as well as impact analysis;
3. Development of a decision making system ("how", "where" and "when" to publish the specified content to reach the target audience and to increase the effectiveness of social media marketing campaign) and a publishing toolkit.

Table 1. Social Media Management Tools

Toolkits	**Number of Supported Channels**	**Multi Channel Posting**	**Content Adaption Strategy**	**Feedback & Statistics**	**Decision Support**
HootSuite http://hootsuite.com	9	Yes	Truncation of long text	Yes	No
HubSpot http://www.hubspot.com	6	Yes	No	Yes	No
Media Funnel http://sti.mediafunnel.com	6	No	Length restriction	Yes	No
Moderation Market http://moderationmarketplace.com	6	No	Length restriction	Yes	No
Ping.FM http://ping.fm	32	Yes	Truncation of long text	No	No
Seesmic https://seesmic.com	24	Yes	Common denominator	Yes	No
Sendible http://sendible.com	20	Yes	No	Yes	No
SproutSocial http://sproutsocial.com	6	Yes	Common denominator	Yes	No

In subproblem 1), central to the conceptualization is the abstraction of content domain from the communication channels using semantic technologies. Furthermore, interweaving, i.e., content to channel mapping, is combined with an intelligent mechanism for channel selection.

Subproblem 2) could be addressed by a model and a collection of techniques to analyze and forecast distributed information spread over the Web. That would help to estimate the impact of information spread on customer behavior and company performance on the one side, and to forecast the impact over time on the other side, providing means to define return on investment from social media spending.

The analytical part described above is integrated within an information dissemination platform (subproblem 3)). The architecture of the SCAN platform is based on the semantic data model between user and communication channels. It is depicted in Figure 1 and will be discussed in more detail in Section 4. Furthermore, a decision support framework facilitates targeted marketing based on analysis of information spread and the impact of this spread over various Web channels like Facebook, Twitter, etc.

Our approach is based on an iterative cycle, depicted in Figure 2 and to be discussed in Section 5. Austrian eTourism with its huge set of different players and Web sites [19] would represent a concrete case to evaluate our approach.

In the following two sections we discuss the issues of feeding and understanding the Web, following the same structure: “Aim and Outcome”, “Methods” and “Challenges”.

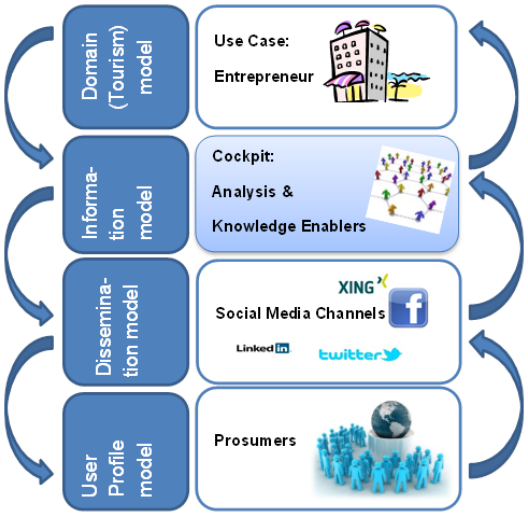

Fig. 1. SCAN Semantic Modeling Framework

4 Feeding the Web

Aim and Outcome. Employing well-developed semantic technology practices on ontology engineering and information integration, we address an under-investigated problem of dissemination and communication in the conditions of the modern multi-channeled social media.

We would use semantics to manually define the information items, the channels, the user groups, and the relationships between them. The overall SCAN semantic modeling framework is depicted in Figure 1 and is comprised of the following types of models (seen at the left hand side of the figure):

- *Domain model (Tourism)* refers to the tourism domain and can be adapted to other use cases; it defines the concepts in the domain and their relationships;
- *Information model* contains information items to be disseminated, such as news items, hotel room prices, pictures, information about services (when applicable, in terms of ontologies such as schema.org, and Dublin Core, GoodRelations) and includes mappings of information items on channels through weavers. The Information model is refined on the basis of the Web analysis, which is described in Section 5.
- *Dissemination model* describes the various channels (e.g., multiple ways to post information on Facebook) and their target groups;
- *User Profile model* contains and represents information about the user – it can be in e.g., FOAF or proprietary social network formats.

Methods. For feeding the Web we take the following steps:

- *Semantic models for content and channels:* Abstracting from the channel of communication and focusing only on the actual content is a prerequisite for

scalable online communication. Here, we use semantic technologies to create models for domain content and channels. These semantic models are then used to formalize social media content in order to maximize the effect of the information spread and forecast algorithms. A semantic model is used to capture the content in domain (and not channel specific) terms allowing direct multi-channel communication for non-communication experts. These models and their interweaving with communication channels can also be reused in a vertical domain for various enterprises and other organizations.

- *Web channels identification and collection of relevant data:* For an efficient and high impact communication identifying the right audience, the right time, and especially the right channels on the Web, is crucial. In this task we will specify Web channels in social media which will be used in further analysis.
- *Content and channel interweaving:* Here, our new approach is applied, which is based on distinguishing and explicitly interweaving content and communication as a central means for achieving reusability and thereby scalability. We develop techniques for linking the Domain model with the Dissemination model through a weaver and implement them as a component of the SCAN tool. Formally, a weaver is an ordered list of tuples (see [20] for more technical details):
  1. An information item is defined as an information category that should be disseminated through various channels.
  2. An editor defines the agent that is responsible for providing the content of the information item.
  3. An editor interaction protocol is defined as the interaction protocol defining how an editor collects the content.

To allow the user to abstract from the channel level to the content level, we design the Information model and corresponding methods in our analytical part (see Section 5).

Challenges. Introducing a semantic layer on top of communication channels is required to enable common value management. However, such combination opens a broad variety of new challenges yet to be solved, in particular, as follows:

- *Modeling and interweaving feedback:* Feedback is an important part of all effective communication. Without feedback, the sender - the one who intends to convey information - has no means to validate whether or not the recipient received or understood the message. It is also often preferable to have a fully-fledged two-way conversation instead of simple one-way broadcasting. Therefore, it will be necessary to model feedback and interweave it with content items that we previously published.
- *Modeling target groups:* Companies that pursue common value management usually have a very restricted target group of people they want to address. So far in our channel model we do not distinguish between different target groups in different channels. However, different target groups reside on different communication platforms, even though there is some overlap. For example, you will find more young and hip people on Facebook, and more professional users

on Xing or LinkedIn, but there are quite a few users that have a profile on both platforms. Nonetheless, they expect a different way of being engaged in different platforms.
- *Adapting content:* This problem encompasses transformations of the given information item into different formats, such as extracting images, videos or extracting and shortening Web links from piece of content. However, adapting content in a way that requires creativity and human intelligence is still a strenuous problem that reaches the borders of computability. Examples of such adaptations are shortening or translating an essay, or rewriting a text in a way that matches the target group it addresses.
- *Quality management:* An important part of targeted communication is assessing and improving the quality of conveyed content. Whereas trust, reputation and brand management are influenced by how information is perceived, quality assurance is an inbound process. The business processes for quality management and what this actually mean have yet to be defined for common value management. The bigger the campaign is, the more visible the effect of proper quality management.

5 From data to knowledge

Aim and Outcome. The goal of the analytical part is to develop appropriate statistical and learning techniques to analyze, forecast and measure the impact of information spread across various Web channels over time. This starts with the specification of Web channels and data to be collected from the Web following our Dissemination and User Profile models. Based on the collected data and Domain model (in our use case Tourism model), we would develop a data model as well as corresponding methods to accomplish the above stated goal.

Methods. The main criteria for the data model is to efficiently account both for structure and content present on the social media. That is why it will be represented as a directed attributed graph, and will encompass the following components: 1) content in social media; 2) users; 3) user attributes, e.g., gender, age, location, etc.; 4) relationships between users. However, the social structure differs for each social media channel, especially the type and structure of relationships between users, e.g., consider Twitter and Facebook. Thus, a separate data model is needed for each social media channel. These models are required for the analysis and prediction mechanisms, and complement the Information model, one of our semantic models.

To retrieve information from social media, we consider open source tools such as Terrier (`http://terrier.org`) and Web mining packages from Python (like BeautifulSoup). Facebook and Twitter APIs are to be used to obtain data from the relevant social networks. Google Refine (`http://code.google.com/p/google-refine`) and Talend (`http://www.talend.com/index.php`) can be applied for data cleaning and transformation. The possible tools to analyze rich semantic text are Open Calais service (`http://www.opencalais.com`), NLTK

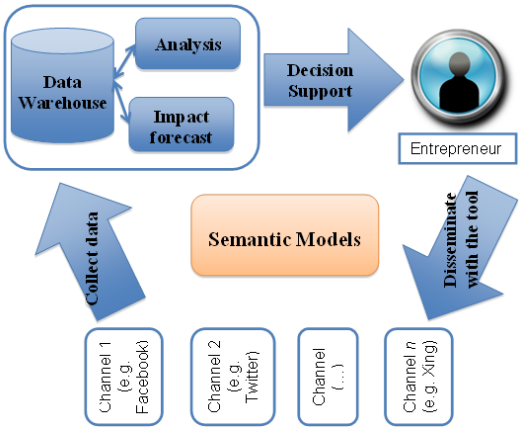

Fig. 2. SCAN Process Cycle

(natural language toolkit `http://www.nltk.org`), as well as GATE for sentiment analysis and opinion mining (`http://gate.ac.uk`).

Based on the social media channel data model, we would analyze the spread of information in the extracted data by combining text mining and network analysis methods. Afterwards, methods to forecast flow and change of information over the desired time in the given collection of Web channels could be developed by using Markov chains and graph transformation rules together with graph mining.

Ongoing, we specify and measure the impact of information spread on user behavior and company performance. The crucial question in this context is whether there is a correlation between 1) customer based indicators: amount of messages; mentions; comments; polarization of messages (negative or positive), etc.; and 2) company based indicators: amount of bookings; visitors to the company Web site; views of company profile, etc.

Based on the impact analysis and domain knowledge, an objective function would be constructed out of customer and company based indicators to optimize the feeding of social media channels with respect to the company goals and user interests. In the end, the companies will get support in the decision on the design of marketing campaigns. Specifically, by providing "what" (information on the conceptual level, e.g., special offer for Valentine's day) and "to whom" (the targeted audience in terms of age, gender, location, interests, etc.) to our decision support framework, the company will get advice on the aspects: "where" (set of specified social media channels),"how" (set of specified users and keywords) and "when" (specified time period for the marketing campaign) to publish.

The solution approach of SCAN is depicted in Figure 2. Our process is iterative which allows to refine the prototypes of our decision support framework and dissemination tool in each iteration. Semantic models are used throughout the whole process cycle, both contributing to each step and learning from its results.

Challenges. While analyzing collected data, we encounter a set of problems. We classify them and outline possible solutions below.

- *Missing / unavailable data:* Many social networks provide the users possibilities to restrict information visible to individuals outside of their network as well as amount of information coming from the outside. We consider opportunities to upgrade accounts on social networks to gain access to profile information of a bigger amount of users. Additionally, data mining techniques to deal with missing values are applied.
- *Low quality of data:* Data and text pre-processing techniques (e.g., stop words elimination, stemming) are applied to ensure coherent data.
- *Inappropriate choice of time periods:* To evaluate a marketing campaign it is essential to select a proper time period to monitor user behavior and company performance before and after the campaign since campaigns may have either immediate or remote effect. As an option to overcome this challenge, experiments with different time periods are conducted.
- *Scalability in forecast calculation:* Since the number of social media data that is available is rather high and decision support that is provided with a very high latency is useless, the amount of time needed to calculate the predictions needs to be within a feasible time frame.

6 Conclusions

We propose a conceptual framework, SCAN, that combines semantic as well as statistical and learning techniques for efficient and high impact communication, targeting the right audience at the right time through the right channels on the Web. Central to SCAN is the process of interweaving content and communication channels. That would enable channel selection based on the semantics of the information item to be distributed. Moreover, it would provide means for content transformation with regard to the specifics of the communication channels for the distribution of this information item. Thus, non-communication experts would not need to worry about the technical aspects of how to interact with various channels; instead, they would just focus on the content being communicated. Furthermore, by combining methods from Web Science, network analysis and text mining, SCAN provides intelligent analysis of the communication in different social media channels.

The methods and techniques developed in SCAN would be applied to distributed data coming from Austrian eTourism. The two major challenges to exploit the benefits of social media are overabundance of information, out of which 75% is estimated to be redundant, and constant growth of structural complexity. Particularly, Austrian tourism operators are now left with the problem of how to use social media to its full extent. This is exactly the problem which our framework helps them to solve. The effectiveness of SCAN and its social media campaigns would be measured and evaluated through its impact on user behavior, for example, number of comments, and company performance, for example,

number of room bookings via the website of a hotel. Such evaluation will need a mixture of tools and approaches from on/off-line interviews with users up to economic analysis of company performance.

References

1. Mulpuru, S., Harteveldt, H.H., Roberge, D.: Five retail ecommerce trends to watch in 2011. Forrester Research Report (2011)
2. Hendler, J., Shadbolt, N., Hall, W., Berners-Lee, T., Weitzner, D.: Web science: an interdisciplinary approach to understanding the web. Commun. ACM **51**(7) (2008) 60–69
3. Kleinberg, J.M., Kumar, R., Raghavan, P., Rajagopalan, S., Tomkins, A.S.: The web as a graph: Measurements, models, and methods. In: Proc. COCOON'99. (1999) 1–17
4. Leskovec, J., Kleinberg, J., Faloutsos, C.: Graph evolution: Densification and shrinking diameters. ACM TKDD **1**(1) (2007)
5. Bringmann, B., Berlingerio, M., Bonchi, F., Gionis, A.: Learning and predicting the evolution of social networks. IEEE Intelligent Systems **25** (2010) 26–35
6. Leskovec, J., Backstrom, L., Kleinberg, J.: Meme-tracking and the dynamics of the news cycle. In: Proc. KDD'09. (2009) 497–506
7. Yang, J., Leskovec, J.: Patterns of temporal variation in online media. In: Proc. WSDM'11. (2011) 177–186
8. Berendt, B.: Text mining for news and blogs analysis. In: Encyclopedia of Machine Learning. Springer (2010) 968–972
9. Ehrig, H., Ehrig, K., Prange, U., Taentzer, G.: Fundamental theory for typed attributed graphs and graph transformation based on adhesive HLR categories. Fundam. Inf. **74** (2006) 31–61
10. Heckel, R., Lajios, G., Menge, S.: Stochastic graph transformation systems. Fundam. Inf. **74** (2006) 63–84
11. Jin, Y., Decker, S., Wiederhold, G.: Ontowebber: Model-driven ontology-based web site management. In: In Semantic Web Working Symposium (SWWS. (2001)
12. Sheth, A., Bertram, C., Avant, D., Hammond, B., Kochut, K., Warke, Y.: Managing semantic content for the web. IEEE Internet Computing **6**(4) (July 2002) 80–87
13. Kobilarov, G., Scott, T., Raimond, Y., Oliver, S., Sizemore, C., Smethurst, M., Bizer, C., Lee, R.: Media meets semantic web - how the bbc uses dbpedia and linked data to make connections. In: ESWC. (2009) 723–737
14. Bishop, B., Kiryakov, A., Ognyanoff, D., Peikov, I., Tashev, Z., Velkov, R.: Owlim: A family of scalable semantic repositories. Education **2**(1) (2011) 33–42
15. Kirtis, A.K., Karahan, F.: To be or not to be in social media arena as the most cost-efficient marketing strategy after the global recession. Procedia - Social and Behavioral Sciences **24**(0) (2011) 260 – 268 The Proceedings of 7th International Strategic Management Conference.
16. Mangold, W.G., Faulds, D.J.: Social media: The new hybrid element of the promotion mix. Business Horizons **52**(4) (2009) 357 – 365
17. Weinberg, B.D., Pehlivan, E.: Social spending: Managing the social media mix. Business Horizons **54**(3) (2011) 275–282

18. Fotis, J., Buhalis, D., Rossides, N.: Social media impact on holiday travel planning: The case of the russian and the fsu markets. International Journal of Online Marketing (IJOM) **1** (2011) 1–19
19. Piazzi, R., Baggio, R., Neidhardt, J., Werthner, H.: Destinations and the web: a network analysis view. Information Technology & Tourism **13**(3) (2011)
20. Bauereiss, T., Leiter, B., Fensel, D.: Effective and efficient online communication. Technical report, STI Innsbruck, Innsbruck (2011)