Modeling the Flow and Change of Information on the Web

Nataliia Pobiedina
Supervised by Hannes Werthner
Institute for Interactive and Software Systems
TU Vienna, Austria
pobiedina@ec.tuwien.ac.at

ABSTRACT

The proposed PhD work approaches the problem of information flow and change on the Web. To model temporal dynamics both of the Web structure and its content, the author proposes to apply the framework of stochastic graph transformation systems [13]. This framework is currently widely used in software engineering and model checking. A quantitative and qualitative evaluation of the framework will be performed during a case study of the short-term temporal behavior of economics news on selected English news websites and blogs over selected time period.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database applications-Data mining

General Terms

Algorithms, Design, Experimentation

Keywords

Web mining, Online media analytics, Stochastic modeling

1. INTRODUCTION

Since its creation, the Web has grown tremendously and has proven to be one of the major successes as a technology. Currently it is the most used application in the history of computing and even of human communication. The Web has changed the ways teaching, communication, publishing and research in academia are performed. It has created an entire sector of e-commerce in industry and has affected the work of governments. Other interesting aspects of the use of the Web include social networking, tagging, data integration, information retrieval and Web ontologies. All these changes in our daily life led to the creation of a new area - Web Science [14], and the questions to be studied among others include: (1) Can we forecast future directions of the Web's growth? (2) Can we predict the social impact of the Web? (3) Can we engineer applications on the Web with a desirable impact on the society?

On the one hand, the Web is an infrastructure of artificial languages and protocols, and on the other hand, it implies an interaction of many users with one another in often-unpredicted ways. This means that a successful research of the Web incorporates the study both of technical and social effects at the micro and macro scales. Thus, an efficient combination of the techniques from natural and social sciences is required.

Much research is already going on in this area, but like the Web itself this area is quite young. These facts inspired me to start my PhD and to pursue the goal of further unveiling a curtain on the Web's dynamics.

The main task of the proposed PhD project is to investigate and model the flow and change of information on the Web. The space of investigation has been limited to make the project feasible within limited time of the PhD studies. I focus on the short-term temporal variation of economics news on selected news websites and blogs in English over a selected time period. Only textual information is taken into account. However, in case of successful results a variety of future possible directions will open up.

I propose to use the framework of stochastic graph transformation systems [13] to model the dynamics both of content as well structure features of the Web. I intend to construct prototype tools to explore and predict news propagation and their change across websites over time. I also expect to obtain temporal patterns of news popularity and to identify those features of websites which are important to the news spread. The exploration tool is mainly addressed to meet the needs of news readers. While the prediction tool as well as obtained patterns will be useful for governments and businesses.

The rest of the paper is structured as follows. The next section provides arguments on the relevance of the problem and describes it in detail. Afterwards, the state of the art is discussed. Section 4 presents a novel approach to tackle the stated problem. In Section 5 the adopted methodology is outlined. Section 6 talks about expected results and progress so far. Finally, the conclusion is drawn and possible future work is presented.

2. PROBLEM DESCRIPTION

Relevance. An average human is overwhelmed with the abundance of information on the Web, and while reading news related to specific topics he/she may have the necessity to understand the prehistory of the topic described there.

On the other hand, media on the Web with its rapid ability to reach vast amounts of people is reshaping the modern society. For example, the Pentagon is already concerned about the large roles which social media play in nourishing unrest in countries like Egypt and Iran [21]. The Pentagon

WWW 2012 Companion, April 16–20, 2012, Lyon, France.
ACM 978-1-4503-1230-1/12/04.

shares the opinion that social media will change the nature of warfare. Therefore, it is crucial to understand the spread and change of information through the Web, and to be able to forecast its future flow.

Furthermore, businesses may benefit from the ability to predict future popularity of a piece of news. For example, advertising and marketing companies will be able to choose for a piece of their information such placement on the Web so that it would ensure with certain probability to provide desirable coverage of the target audience.

Problem Specification. The main task of the proposed PhD project is to develop a framework which will allow to investigate the flow and change of information on the Web. I will evaluate the developed framework during a case study for economics news from news websites and blogs. Though currently news propagate in different forms like text, video, pictures, I take into account only textual information. So, in my case study I have a limited amount of pre-defined websites with textual content in English and a pre-defined time period. The focus is on the short-term temporal behavior of economics news, i.e. one time point corresponds to one day.

I model textual information on a Web document as a set of *memes* extracted from the document. In principal, a meme is anything that can be copied from one mind to another. In my setting, a meme is a phrase, a combination of words, which does not occur only once. However, considering the social aspect of memes, I take also into account how many different authors used certain meme. Thus, I filter out the phrases extracted from the whole set of Web documents according to their amount of occurrence and amount of different authors who used them. I measure the *popularity* of a meme by the amount of its occurrence in the set of Web documents.

Afterwards, I will identify *topics* for each meme. Thus, a topic will be a collection of memes. However, the name of the topic will be a meme itself. For example, if we have memes like "my mother", "my father", "my family", then "my family" will be the name for the topic which combines all three memes. The change of information is modeled through the changes in topics over time.

Tasks to be accomplished during my PhD work are as follows: (a) investigate flow of memes and topics in the set of Web documents over given time: identify first appearance and track further spread; (b) investigate change of topics in the set of Web documents over given time period; (c) forecast flow of memes and topics in the set of Web documents for the desired time period; (d) forecast further change of topics in the set of Web documents for the desired time period; (e) explore patterns of popularity of memes and topics, i.e. patterns of increasing or decreasing presence of memes and topics over time; (f) explore websites which provide high popularity.

3. STATE OF THE ART

Literature overview discussed here is by no means complete, but to a meaningful level it represents current state of the art in the area. Throughout the PhD work relevant literature will be reviewed, and the state of the art will be correspondingly updated.

The proposed project is a multidisciplinary research; hence the literature required to follow current state of the art covers such areas as: Web Science; text mining; evolutionary modeling of social networks; dynamic topic modeling; and graph transformation systems.

The book [8] provides a very good survey of application of social network analysis in the areas of dynamic topic modeling and text mining as well as in evolutionary modeling.

There is sizable literature on the Web structure and on significant results to model information diffusion on the Web. In [17] 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 [19] 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 effective 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 [18] and [24] the 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.

Topic identification in the set of static documents is a well-known problem in text mining, and there are lots of algorithms already developed. The classical algorithms are feature-based, i.e. every document is represented as a feature vector. However, graph-based algorithms for topic identification are being successfully developed recently [7, 20]. There are also algorithms which take into account hyperlinks between pages like in [10].

In the area of social network analysis there are several frameworks proposed to model the evolution of the network, e.g. event-based frameworks from [2], [23] and a generic framework from [9] which provides balance between two optimization criteria: history cost and snapshot quality.

In the area of dynamic topic modeling scientists solve problems of topic identification and tracking over time. In [1] on-line LDA represents documents as mixture of topics and topics as mixture of words, and iTopicModel [22] leverages both content and linkage information in the clustering process. However, these works are oriented at well-structured text corpora like DBLP [22], or Nips and Reuters [1].

In her survey [3] Berendt outlines the major challenges of text mining for online news and blogs analysis. She also provides a 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.

4. PROPOSED APPROACH

To model temporal dynamics both of the content on the Web and of its structure, I propose to apply the framework of stochastic graph transformation systems [13] - a formal approach which puts two major mathematical structures, topology and algebra, together. Currently graph transformation systems [11] are successfully applied in the areas of software engineering and model checking.

The idea is that at some point in time we have a graph snapshot which corresponds to the state of the modeled system. This graph is transformed into the graph at the next time point by applying corresponding transformation rules.

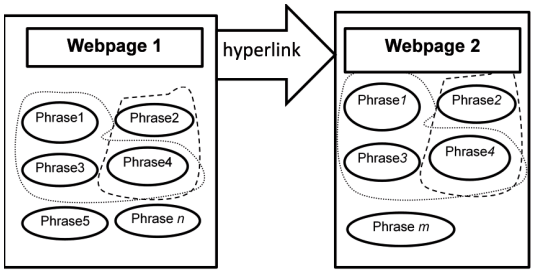

Figure 1: Snapshot of a webpage-based model at a certain timepoint.

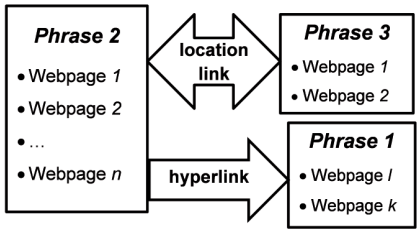

Figure 2: Snapshot of a meme-based model at a certain timepoint.

Thus, the goal is to determine the architecture of the graph and transformation rules. Then in order to predict the future state of the system each transformation rule is assigned application rates which indicate the probability that the current graph will be transformed according to this rule in the next time point.

I consider two approaches for the architecture of the snapshot graph: *webpage-based* and *meme-based*. In both cases the graph will be a so called E-graph [11] which can be regarded as an extension of a classical graph. Each node and each edge are additionally assigned some attributes - data labels from a data algebra. The construction of a data algebra is required to ensure structural and data preservation of the graph snapshots at different timestamps.

In the first approach (Figure 1) the Web structure is accounted the same way like in the Web graph model, i.e. nodes represent webpages, and directed edges between nodes represent hyperlinks between webpages [17]. Onwards, each node is assigned data labels representing memes extracted from the corresponding webpage. Each edge can be assigned a certain data label depending on the type of the hyperlink between webpages, e.g. it can be a navigational or transactional link, or this link can represent certain type of connection between people in blogs or social networks like friends or colleagues.

In the second approach (Figure 2), the graph nodes represent memes extracted from the webpages. The nodes are assigned data labels which represent the webpages containing the corresponding memes. As a variation of this approach, it is possible to assign weights to the nodes depending on the amount of webpages which contain the corresponding memes.

The edges are also assigned data labels depending on the type of connection between memes, so there may be a *link*-edge and a *location*-edge from one meme to another. A link-edge is drawn if there is a hyperlink from one of the webpages where the first meme is situated to one of the webpages where the second meme is situated. The edges can be also assigned weights which are proportional to the amount of hyperlinks from the set of websites where the first meme is located to the set of webpages where the second meme is located. A location-edge is drawn in case both memes are located on the same webpage. This model can be regarded as a social network where memes act as individuals, and topic identification can be considered as a community identification problem.

The rules of graph transformations over time can be learnt from historical data. Possible techniques to achieve a good level of learning at this point are discussed in [6]. However, it is also possible to construct transformation rules manually. Both approaches are to be considered during the PhD work.

Since my goal is to study the flow and change of information on the Web, the transformation rules are to be classified into the rules of flow and change. For example, the authors in [23] outline such rules like "form", "dissolve", "shrink", "continue", "merge", "split", "reform" to reason about community evolution in a social network. A similar approach is used in [2]. By regarding topics as communities and memes as individuals in the network, I will identify rules to track the changes in topics in the news cycle.

Finally, each transformation rule is assigned an application rate, i.e. a probability with which the corresponding event happens at a certain point of time given an initial graph snapshot. As a possible technique to account for the stochastic properties of the constructed system, authors in [13] propose to use Markov chains. In my setting, states of Markov chains will represent graph snapshots at different time points.

Introduction of Stochastic Logic [13] over the constructed Markov chain will provide machines with a formal framework to reason about future states of the system.

Since the framework considers the Web structure as well, it will allow identifying structural properties of efficient websites, i.e. websites which lead to the highest popularity of the desired content. Such features may include in-degree and out-degree of websites, centrality measures and connections to some specific websites.

Novelty. To my best knowledge, the closest techniques to above mentioned graph transformation systems are graph evolution rules proposed in [5] and an event-based diffusion model from [2], both developed to model evolution in social networks. None of these works consider the rich content found in online media.

5. METHODOLOGY

As a research methodology, I apply the design science research framework [16] to ensure a rigorous and relevant contribution to the Computer Science community. I have already presented the relevance of my problem in the Section 2. The rigor of my research is ensured by the knowledge base which encompasses a variety of techniques for text analysis and dynamic modeling.

The dynamic modeling is discussed in detail in the previous section. At this point I intend to make a major contribution to the knowledge base by adopting stochastic graph transformation systems to the study of Web evolution.

On the other hand, I use already available methods for text analysis like text preprocessing (elimination of stop words, stemming, lemmatization, WordNet for synonyms disambiguation); meme extraction (= phrase extraction); clustering (topic identification, community identification);

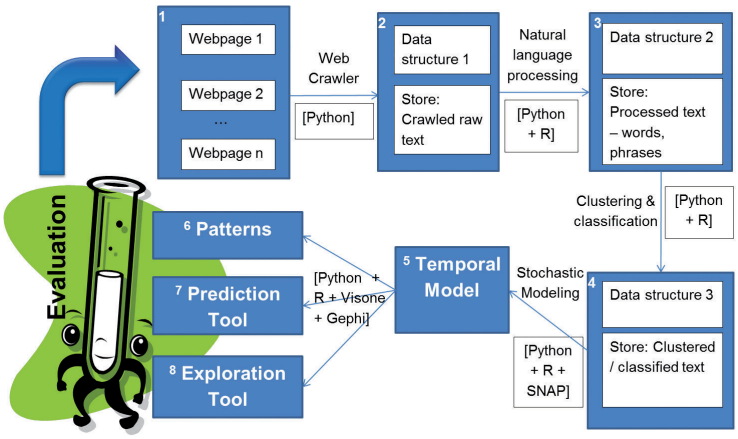

Figure 3: Iteration of the design cycle.

and classification (supervised topic identification; classification based on source type). Taking into account great amount of data on the Web and the need to construct scalable algorithms, I'm using stream algorithms [12], i.e. store summary data and use it to process future data.

To ensure the most efficient combination of techniques mentioned above, I have constructed a design cycle (Figure 3) which consists of nine blocks. Each block in this scheme represents the input to the next step in the cycle which is denoted by the arrow and defines the techniques and tools to be used. Since there are various methods available for text analysis, I will have several iterations of the design cycle where I will use different methods for natural language processing, clustering and classification which will lead to different results in the blocks 3-5, and as a consequence, in the blocks 6-8.

Evaluation. The last block in the design cycle (Figure 3) is evaluation. I intend to perform quantitative and qualitative evaluation of the constructed framework.

Quantitative evaluation constitutes of verifying prediction errors on different datasets, for example, changing amount of crawled webpages, time period, or geographic location of crawled webpages (like USA or UK news).

I will perform a qualitative evaluation with professional journalists by organizing a panel to discuss the validity of my results. Additionally, a use-case will focus on evaluating the tools to explore and visualize news spread.

Moreover, qualitative evaluation will include comparison of obtained results with the results from journalistic studies. For example, Boczkowski [4] outlines the phenomenon of "news-at-work" which arose due to the Web and observes that smaller news organizations mimic the bigger and more important ones. The authors in [15] argue that professional journalists tend to ignore user comments and other user involvement. Since these studies are qualitative, it is very interesting to verify whether their results can be checked quantitatively and whether the differences in my results can be explained.

6. RESULTS AND PROGRESS

I have classified the main contributions of my PhD into the following categories: (a) *sociological-economic*: the tools to forecast and explore the spread of economics news on the Web; (b) *representational*: the most optimal model to represent Web snapshots at different timestamps; (c) *algorithmic*: the most efficient combination of methods to model information spread and change on the Web; (d) *visual*: the suitable visualizations for the expected results.

I have recently started my work on the PhD thesis. So far I have set up the initial versions of the Web crawler, natural language processing and data storing packages, and I am looking into the data I get to develop the temporal model. I do not consider the flow and change of topics in the first iteration of my design cycle. My focus is solely on the problem of modeling the flow of memes in the form of extracted words.

7. CONCLUSION AND OUTLOOK

During my PhD studies I plan to develop a systematic approach to modeling the flow and change of textual information on the Web. By tailoring stochastic graph transformation systems to this problem I will design a framework with solid mathematical background to account both for the content and structure information on the Web. Considering the enormous amount of information on the Web and the limited time for the PhD work, I will evaluate the framework on the economics news from English news websites and blogs.

In case of successful results, the framework will take into account news from different categories (e.g., politics, social life and news in general). It will be interesting to extend the framework with sentiment analysis and opinion mining. Such extension may give an opportunity to better measure the change of information.

Social media is becoming more and more popular and is actively used to spread the information (for example, Twitter and Facebook), so this is a very desirable extension to the framework. Another future direction is the study of multilingual news sources. This is a big challenge since natural language processing techniques are much better developed for English in comparison to, e.g., German or Ukrainian.

The adaptation of the framework to reason about long-term temporal behavior of online news is planned. The comparison of long-term and short-term patterns in news variation may give a deeper understanding of the news cy-

cle. Finally, the framework could be extended to predict the impact of the news spread on the society.

8. REFERENCES

[1] L. AlSumait, D. Barbara, and C. Domeniconi. On-line LDA: Adaptive topic models for mining text streams with applications to topic detection and tracking. In *Proc. ICDM'08*, pages 3–12, 2008.

[2] S. Asur, S. Parthasarathy, and D. Ucar. An event-based framework for characterizing the evolutionary behavior of interaction graphs. *TKDD*, 3(4), 2009.

[3] B. Berendt. Text mining for news and blogs analysis. In *Encyclopedia of Machine Learning*, pages 968–972. Springer, 2010.

[4] P. Boczkowski. *News at Work: Imitation in an age of information abundance*. University of Chicago Press, 2010.

[5] B. Bringmann, M. Berlingerio, F. Bonchi, and A. Gionis. Learning and predicting the evolution of social networks. *IEEE Intelligent Systems*, 25:26–35, 2010.

[6] H. Bunke. Graph matching: Theoretical foundations, algorithms, and applications. In *Proc. Vision Interface 2000*, pages 82–88, 2000.

[7] H. Bunke and K. Riesen. Recent advances in graph-based pattern recognition with applications in document analysis. *Pattern Recogn.*, 44:1057–1067, 2011.

[8] C. C. Aggarwal, editor. *Social Network Data Analytics*. Springer, 2011.

[9] D. Chakrabarti, R. Kumar, and A. Tomkins. Evolutionary clustering. In *Proc. KDD'06*, pages 554–560, 2006.

[10] S. Chakrabarti, B. Dom, and P. Indyk. Enhanced hypertext categorization using hyperlinks. *SIGMOD*, 27:307–318, 1998.

[11] H. Ehrig, K. Ehrig, U. Prange, and G. Taentzer. Fundamental theory for typed attributed graphs and graph transformation based on adhesive HLR categories. *Fundam. Inf.*, 74:31–61, 2006.

[12] S. Guha, A. Meyerson, N. Mishra, R. Motwani, and L. O'Callaghan. Clustering data streams: Theory and practice. *IEEE Trans. Knowl. Data Eng.*, 15(3):515–528, 2003.

[13] R. Heckel, G. Lajios, and S. Menge. Stochastic graph transformation systems. *Fundam. Inf.*, 74:63–84, 2006.

[14] J. Hendler, N. Shadbolt, W. Hall, T. Berners-Lee, and D. Weitzner. Web science: an interdisciplinary approach to understanding the web. *Commun. ACM*, 51(7):60–69, 2008.

[15] A. Hermida and N. J. Thurman. A clash of cultures: The integration of user-generated content within professional journalistic frameworks at British newspaper websites. *Journalism Practice*, 2:343–356, 2008.

[16] A. R. Hevner, S. T. March, J. Park, and S. Ram. Design science in information systems research. *MIS Quarterly*, 28(1):75–106, 2004.

[17] J. M. Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan, and A. S. Tomkins. The web as a graph: Measurements, models, and methods. In *Proc. COCOON'99*, pages 1–17, 1999.

[18] J. Leskovec, L. Backstrom, and J. Kleinberg. Meme-tracking and the dynamics of the news cycle. In *Proc. KDD'09*, pages 497–506, 2009.

[19] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graph evolution: Densification and shrinking diameters. *ACM TKDD*, 1(1), 2007.

[20] A. Schenker, M. Last, H. Bunke, and A. Kandel. Classification of web documents using a graph model. In *Proc. ICDAR'03*, pages 240–244, 2003.

[21] D. Streitfeld. Pentagon seeks a few good social networkers. 2011. `http://bits.blogs.nytimes.com/2011/08/02/pentagon-seeks-social-networking-experts/`.

[22] Y. Sun, J. Han, J. Gao, and Y. Yu. iTopicModel: Information network-integrated topic modeling. In *Proc. ICDM'09*, pages 493–502, 2009.

[23] M. Takaffoli, F. Sangi, J. Fagnan, and O. R. ZaÃŕane. A framework for analyzing dynamic social networks. In *Proc. ASNA*, 2010.

[24] J. Yang and J. Leskovec. Patterns of temporal variation in online media. In *Proc. WSDM'11*, pages 177–186, 2011.