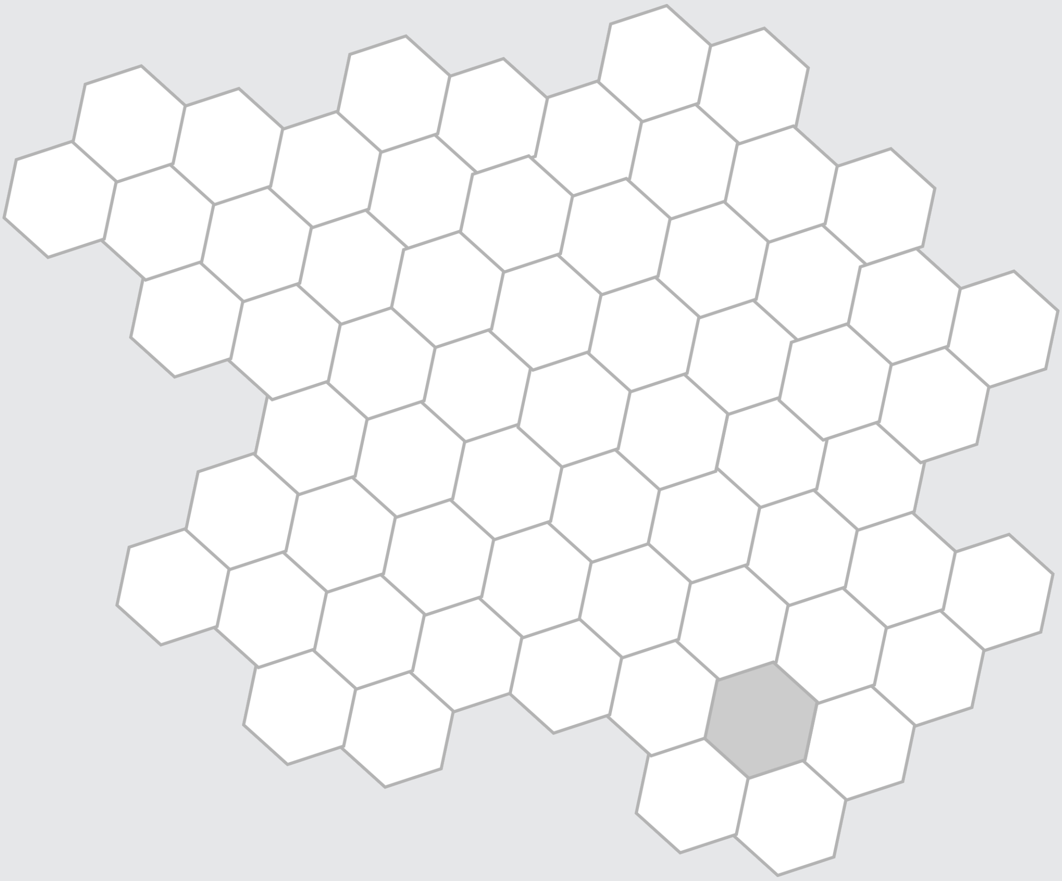
International Journal on

SOCIAL MEDIA

MMM: Monitoring, Measurement, and Mining



I, 2010, 1
ISSN 1804-5251

...OFFPRINT...

International Journal on SOCIAL MEDIA

MMM: Monitoring, Measurement, and Mining

I, 2010, 1

Editor-in-Chief: Jan Žižka

Publisher's website: www.konvoj.cz
E-mail: konvoj@konvoj.cz, SoNet.RC@gmail.com
ISSN 1804-5251

A Proposal for an Approach to Extracting Conceptual Descriptions of Hyper-linked Text Documents

Gabriel Lukáč
Department of Cybernetics and Artificial Intelligence, FEI, TU Košice, Letná 9, 042 00 Košice, Slovakia
e-mail: `Gabriel.Lukac@tuke.sk`

Abstract

This paper is a brief description of work that is planned to be realised during the PhD study of the author. It describes a proposal for a methodology for extracting conceptual descriptions of text documents published within a social network of authors. Additionally, the hyper-linked nature of particular documents (e.g. a citation network of blogs) provides a very good framework in which to take an advantage from the networked environment and it allows one to use algorithms for information diffusion to track influential concepts spreading down the network. The paper is divided into several sections. It gives an overview of related work that serves as a foundation for the presented method. At the end an outline of a particular method proposal is provided. Its main purpose is to foster further discussions about the topic and serve as a baseline for future work.

Key words

social networks; textual documents; diffusion of innovation, concepts, keyword extraction

Biographical note

Gabriel Lukáč is a PhD student in the Department of Cybernetics and Artificial Intelligence at the Faculty of Electrical Engineering and Informatics, Technical University of Košice. He is interested in research into the dynamics of complex networks with a focus on the spreading of information through networked structures. In 2007 he graduated (MSc degree) at the same faculty.

Introduction

The process of generating conceptual descriptions of textual documents automatically is an important challenge in the semantic web initiative (W3C Consortium, 2010). Without the possibility to make such automatic extractions, the vision of the semantic web will still remain a vision and will never be deployed in practice. There is another important use of such techniques: Keyword extraction for search engine optimisation (SEO). Many SEO companies and practitioners use it as a daily tool to extract keywords. In general, this process is mostly realised for documents, for which their hyper-linked structure is not considered. On the web there is an invaluable source of large-scale document data generated by the social network of authors and data itself has a very clean network structure. We believe that this advantage of networked structure can provide an opportunity to use principles of diffusion of information (Rogers, 2003) to track important concepts spreading through the network. Additionally this framework can give us the possibility to come-up with a measure that will produce a ranking of the concept importance for a given document.

Related work

In this section work related to our paper is provided. It is divided into two subsections, and each of them represents a topic that will be necessary to revisit in future research.

Spreading of information cascades

Models for catching the spreading of information through information channels are very often based on epidemic models (Satorras and Vespignani, 2001) describing the spread of viral diseases through social networks of people. An analogy to such epidemic models is the spread of so-called *information cascades*. Information cascades are phenomena in which an action or idea becomes widely adopted due to the influence of others, typically neighbors in some network (Leskovec et al., 2007). Every information cascade (or sometimes in literature it is denoted as a *conversation tree*) has one starting node called cascade initiator. In the case of emails or discussion forum posts the *cascade initiator* is the contribution starting an email or discussion thread. In the case of blogs, the cascade initiator is a blog article that comes-up with a certain unique idea for the first time. If the article is interesting enough, it will start an information cascade that will be successively built up by the progressive adding of new articles to the cascade – articles that make cite former blog contributions.

The idea of cascades spreading through networks has been studied from the point of view of various branches of science. Rogers (2003) has studied them as a sociological phenomenon called *diffusion of innovation*. More suitable for the purpose of this paper is the work of Kumar et al. (2004), where it was used to explain actual trends in the blogosphere.

To model the process of adoption of some idea that spreads through an information cascade, two groups of models are usually used: *Threshold models* (Granovetter, 1978), where adoption of an idea by some node (actor) is conditioned by the overall sum of weights of incident edges above a certain threshold t. The second class of cascade models are *independent cascade* models (Goldenberg et al., 2001), in which the chance that node i will adopt the behaviour of node j is given by probability p_{ij}.

Identification of spreading concepts

A critical problem when analysing information cascades is the identification of words, concepts, phrases (or sometimes the concept of a *meme* is used) that are being spread. It is a non-trivial assignment with a great complexity. In the literature there are a number of approaches to solving this problem. In this chapter we will show one approach that is directly related to networks and two others, that have their foundations in the field of information retrieval and the social web.

The first methodology we want to mention here is the *MemeTracker* of authors Leskovec et al. (2009) for extracting short textual phrases. It is an algorithm creating clusters by aggregating similar textual phrases and it declares this cluster as a standalone cascade. Particular phrases are perceived as an analogy of the 'genetic code' of various memes. This means that the similarity of several memes remains with increasing time still recognisable, but as in genetic structures, so also memes are subject to considerable mutations.

The work of Matsuo and Ishizuka (2003) is also related to the topic of identification of spreading concepts even if it is not related directly through network terminology as a common denominator. Their work has a basis in the theory of *information retrieval*. They based their algorithm for keyword extraction on the idea of frequent co-occurrence of some term with a certain subset of other terms. If one can find such a term in a text, then it is probably a key term of a given text.

Another important (and nowadays very popular) approach for obtaining conceptual descriptions of documents are *folksonomies* – online services for shared categorisation of web resources (Al-Khalifa and Davis, 2006). Their biggest advantage is the relevance of concepts that were contributed by respective users as annotations of their bookmarks. Moreover they provide public API that enables its users to access and download keyword descriptions for every URL that has been marked as public.

Concepts extraction method proposal

The purpose of this method is to provide a list of concepts (keywords) for any given document from the data set. That gives us a very clean list of inputs and outputs of the future algorithm. At the input there is a database of text documents along with an adjacency matrix that will cover citation bindings between particular texts. At the output one should have an opportunity to make a query for any document and obtain a list of the most representative keywords that characterise the document as accurately as it is possible.

We plan to divide our method into two parts. The first one will be the cascades reconstruction part, in which we will try to reconstruct as many cascades from the data as possible. This phase will be based on research on threshold and independent cascade models, trying to fit the data and extract conversation trees. For the second part we will use the MemeTracker methodology (Leskovec et al., 2009) combined with public APIs of online bookmarking services to identify concepts that spread through conversation trees.

Having a list of concepts available for every text document together with the nesting depth saying how deep the concept has spread within the conversation tree, it will be feasible to design a measure that will rank the representativeness of the concept to a particular document.

Conclusion

In this paper we give an overview of the future work that will be realised within the next year of the author's PhD study. Unfortunately it does not give any recent results, but its aim is to provide the foundation to raise discussions about the topic of generating representative concepts of texts and hopefully it will foster new ideas.

References

Al-Khalifa, H. S., Davis, H. C. (2006): Folksonomies versus automatic keyword extraction: An empirical study [on-line]. [cit. 2010-08-05]. Available at: `http://www.iadis.net/dl/Search_list_open.asp?code=2728`.

Goldenberg, J., Libai, B., Muller, E. (2001): Talk of the Network: A Complex Systems Look at the Underlying Process of Word-of-Mouth. *Marketing Letters*, *, pp. 211–223. (ISSN 0923-0645.)

Granovetter, M. (1978): Threshold Models of Collective Behavior. *American Journal of Sociology*, 83.6, pp. 1420.

Kumar, R. et al. (2004): Structure and evolution of blogspace. *Commun.ACM*, Vol. 47, No. 12, pp. 35–39. (ISSN 0001-0782.)

Leskovec, J., Backstrom, L., Kleinberg, J. (2009): Meme-tracking and the dynamics of the news cycle. In: *KDD '09: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining.* Paris (France) : ACM, pp. 497–506. (ISBN 978-1-60558-495-9.)

Leskovec, J., Faloutsos, Ch. (2007): Scalable modeling of real graphs using Kronecker multiplication. In: *ICML '07: Proceedings of the 24th international conference on Machine learning.* Corvalis (Oregon) : ACM, pp. 497–504. (ISBN 978-1-59593-793-3.)

Matsuo, Y., Ishizuka, M. (2003): Keyword Extraction from a Single Document using Word Cooccurrence Statistical Information. In: *Proceedings of the 16th International FLAIRS Conference.* St. Augustine (Florida) : AAAI Press, pp. 392–396. (ISBN 1-57735-177-0.)

Rogers, Everett M. (2003): *Diffusion of Innovations.* 5th ed. New York : Free Press, . ISBN 0-7432-2209-1.

Satorras, R. P., Vespignani, A. (2001): Epidemic spreading in scale-free networks. *Physical Review Letters*, vol. 86, no. 14, pp. 3200–3203.

W3C Consortium (2010): Semantic Web Activity [on-line]. [cit. 2010-08-05]. Available at: `http://www.w3.org/2001/sw/`.

The work presented in this paper was supported by the Slovak Grant Agency of the Ministry of Education and Academy of Sciences of the Slovak Republic within the 1/0042/10 project 'Methods for identification, annotation, search, access and composition of services using semantic metadata in support of selected process types'.

Call for Papers

Papers to be included in the next issue should be preferably focused on topics related to social-networks in one or more of the following subjects (the list is indicative rather than exhaustive):

Sentiment/Opinion Analysis in Natural-Language Text Documents

Algorithms, Methods, and Technologies for Building and Analysing Social Networks

Applications in the Area of Social Activities

Knowledge Mining and Discovery in Natural Languages Used in Social Networks

Medical, Economic, and Environmental Applications in Social Networks

Submitted papers should not have been previously published nor be currently under consideration for publication elsewhere. Each of the submitted research papers should not exceed 26 pages. All papers are refereed through a peer review process.

Submissions should be send in the PDF form via email to the following address: `SoNet.RC@gmail.com`

Accepted papers are to be prepared according to the instructions available at `http://www.konvoj.cz/journals/mmm/`*.*