

IDENTIFICATION AND TRACKING OF NEWS STORIES USING HONEY BEE INSPIRED AGENTS

Štefan SABO

*Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 2, 842 16 Bratislava, Slovakia
sabo@fiit.stuba.sk*

Abstract. We propose method for identification and subsequent tracking of news stories. Our method utilizes a set of independent agents based on honey bees in order to acquire and analyse news articles from the Web. Each agent individually tracks a certain aspect of the news story, that we call a *story word*. The resulting landscape of stories is obtained as a result of social interaction between the agents. Main advantage of this approach is its dynamic nature, as stories are extracted on-the-fly, as new articles are issued, without the need of previous training, or supervision.

1. Introduction

In our work we are focusing on detection and tracking of the current news stories based on the analysis of news articles from the Web. Our goal is to be able to offer a fresh set of the most recent news stories, by taking a glance on the news articles currently published on the Web. No training set or previous knowledge of the domain is required. Furthermore, in addition to story extraction, it is necessary to provide a means of dynamic tracking of identified stories. As new articles are added, current news stories may undergo new developments, or new stories may unfold. All this needs to be taken into account, when proposing a dynamic method for news stories tracking.

2. Method outline

In order to identify and track news stories, we propose a dynamic method that consists of three main tasks, *data acquisition*, *story word extraction*, and *story identification*. The goal of data acquisition is to constantly provide a stream of up to date news articles, while keeping track of the changes in articles that have already been visited. This is achieved using the set of agents that act as web crawlers, visiting and retrieving news articles. In order to

coordinate individual agents, swarm insect inspired principles are utilized. The whole data acquisition process is designed for continuous operation, as our system is envisioned for long term operation over the Web, retrieving and updating articles, as new events unfold.

Story word extraction mechanism, that was the focus of our previous work [1] aims to model the connections between individual news articles, based on their mutual relations to story relevant terms, named *story words*. If two news articles discuss the same news stories, there should be some story related terms, shared by both of the articles. We look for these story words and try to map out the connections between articles. The advantage of social insect inspired approach to this task lies in the ability of agents to focus on exploration of most promising story word candidates. If a term relevant to multiple articles is found, more emphasis is given to exploration of articles relevant to this term.

Story identification is the last task and was outlined in our most recent work [2, 3]. In order to identify news stories discussed in the articles, we construct a graph of relationships between articles. We then identify communities within this graph, using Louvain algorithm for community detection. All subsequent articles may be classified on-the-fly, without the need to recalculate the structure of communities from the start.

3. Conclusion

This article gives an outline of our proposed method for dynamic detection and tracking of current news stories. In order to acquire news article and identify individual *story words*, we utilize a set of agents inspired by honey bees foraging for food. By comparing visited articles one to another, our agents identify the most prominent story related terms and map out the relevance of these terms to news articles. The information about the connection of news stories to story words and mutual connections between articles is projected onto a graph. In this graph we look for communities, which represent individual news stories. The main advantages of this approach are its flexibility, as no supervision or training is required and its ability to track news stories dynamically, as they unfold. Our future work will focus on visualization and presentation of the resulting stories.

Acknowledgement: This work was partially supported by the Slovak Research and Development Agency under the contract No. APVV-0208-10 and the Scientific Grant Agency of Slovak Republic, grant No. VG1/0971/11.

References

to other papers publishing the results that are summarized here

- [1] Navrat, P., and Sabo, S.: What's going on out there right now? A beehive based machine to give snapshot of the ongoing stories on the Web. In: *Proc. of Nature and Biologically Inspired Computing (NaBIC)*, 2012, IEEE, (2012), pp. 168 - 174.
- [2] Sabo, Š., Návrát, P.: Social insect inspired approach for identification and dynamic tracking of news stories on the Web. In: *Proc. of Nature and Biologically Inspired Computing (NaBIC)*, 2013, IEEE, (2013), pp. 226 - 231.
- [3] Navrat, P., Sabo, S.: Beehive Based Machine to Give Snapshot of the Ongoing Stories on the Web. *Trans. on Computational Science XXI*, Springer, LNCS 8160, (2013), pp. 296-314